

SAPIENZA
UNIVERSITÀ DI ROMA

High-Resolution Synthesis Across Domains: A Wavelet-Driven Approach to Generative Modeling

Department of Information Engineering, Electronics and Telecommunications (DIET)
PhD in Information Communication and Technologies (ICT) (XXXVII cycle)

Luigi Sigillo
ID number 1761017

Advisor
Prof. Danilo Comminiello

Academic Year 2024/2025

Thesis defended on 19 September 2025
in front of a Board of Examiners composed by:

Prof. Marco Ricci (chairman)

Prof.ssa Stefania Bartoletti

Prof. Piero Tognolatti

High-Resolution Synthesis Across Domains: A Wavelet-Driven Approach to Generative Modeling

Sapienza University of Rome

© 2025 Luigi Sigillo. All rights reserved

This thesis has been typeset by LaTeX and the Sapthesis class.

Author's email: luigi.sigillo@uniroma1.it

There's more to the picture than meets the eye

Acknowledgments

Ack on printed thesis.

Abstract

Deep generative modeling is revolutionizing visual data synthesis, from creative industry applications to tools for scientific discovery in fields like medical imaging and remote sensing. In image synthesis, deep learning models like Variational Autoencoders (VAEs), Generative Adversarial Networks (GANs), and Denoising Diffusion Probabilistic Models (DDPMs) help create novel, realistic, and controllable visual content. In specialized applications, these models can augment scarce data for medical analysis or enhance imagery for remote sensing tasks.

A crucial challenge in applying deep learning to image synthesis is overcoming the persistent trade-offs between resolution, fidelity, and computational efficiency. While generative models have advanced significantly, they often struggle with generalization across diverse domains and data modalities. Conventional methods tend to scale poorly to ultra-high resolutions (UHR), leading to artifacts like repeated structures or blurred textures. Furthermore, these models often apply uniform refinement processes across all spatial regions, disregarding local frequency variations and failing to optimally allocate supervision to areas of different visual complexity.

Wavelet transforms, particularly Discrete Wavelet Transforms (DWT) and their hypercomplex extension, Quaternion Wavelet Transforms (QWT), have shown promising results in multi-scale signal analysis. These methods operate by decomposing images into a hierarchy of frequency sub-bands, capturing both global structure and fine-grained details. This formulation endows models with the ability to leverage sparse representations and dimensionality reduction. However, their potential to reshape feature representation, inform conditioning, and adapt training objectives in a model-agnostic fashion had not been comprehensively exploited. In this thesis, we explore this concept and exploit the wavelet-driven learning paradigm to overcome the aforementioned shortcomings of traditional generative models. We leverage multi-scale analysis to make models inherently aware of frequency and spatial information.

We first design a structure-aware GAN, StawGAN, tailored for cross-domain infrared-to-RGB image translation. Building on this foundation, we develop specialized diffusion models for domain-specific tasks, including high-fidelity maritime image super-resolution and efficient EEG-to-image synthesis. Moving beyond these conventional generative approaches, we introduce a series of wavelet-driven architectures that explicitly incorporate multi-scale signal representations. Among them is QUAVE, a novel framework leveraging quaternion wavelet transforms (QWT) to enhance feature extraction and improve generalization in medical imaging. Expanding on these insights, we also pioneer several wavelet-based super-resolution models: a QWT-conditioned diffusion model, a metadata- and wavelet-aware architecture for satellite imagery, and a highly efficient hybrid framework, Wavelet Diffusion GAN, that combines the strengths of GANs and diffusion processes.

Finally, we address a direct extension of these works, focusing on high-fidelity synthesis. We culminate our investigation with a Latent Wavelet Diffusion (LWD) framework, a general and lightweight solution that enables existing latent diffusion and flow matching models to achieve UHR (up to 4K) synthesis without architectural modifications or additional inference costs.

Through extensive experiments across various generative tasks involving different domains and data modalities, we have thoroughly explored the wavelet-driven paradigm while addressing scenario-specific challenges in fidelity, efficiency, and generalization, advancing research in this field.

Contents

Acknowledgements iii

Abstract v

List of Acronyms ix

List of Figures xv

List of Tables xvii

1 Introduction 1
1.1 Deep Learning for Generative Image Modeling 1
1.2 The Vanguard of Generative Modeling and its Prevailing Limitations 2
1.3 A Wavelet-Driven Paradigm for Generative Modeling 3
1.4 The Proposed Wavelet-Centric Approach to Generative Modeling . . 4
1.5 Thesis Contributions 5
1.6 Thesis Organization 6

I Background and Foundations 9

2 Generative Models: A Deeper Dive 10
2.1 Variational Autoencoders (VAEs) 11
2.1.1 Theoretical Foundation and Variational Inference 11
2.1.2 The Reparameterization Trick and Auto-Encoding Variational Bayes 12
2.2 Generative Adversarial Networks (GANs) 13
2.2.1 Game-Theoretic Foundations and Adversarial Training 13
2.2.2 Training Dynamics and Stability Challenges 14
2.2.3 Theoretical Extensions and Variants 15
2.3 Denoising Diffusion Probabilistic Models (DDPMs) 16
2.3.1 Theoretical Framework and Non-Equilibrium Thermodynamics 16
2.3.2 Variational Lower Bound and Training Objectives 17
2.3.3 Discrete State Spaces and Advanced Formulations 18
2.4 Latent Diffusion Models (LDMs) 19
2.4.1 Latent Space Representation and Computational Efficiency . 19
2.4.2 Cross-Attention Mechanisms and Conditional Generation . . 20
2.4.3 Architectural Innovations and Scaling 20
2.5 Flow Matching Models 21
2.5.1 Continuous Normalizing Flows and Simulation-Free Training 21
2.5.2 Optimal Transport and Efficient Probability Paths 22

2.5.3 Applications and Scaling Properties 23
2.5.4 Comparative Analysis with Other Model Families 24

3 Generative Models for Images 26
3.1 Introduction to Generative Models for Image Synthesis 26
3.1.1 The Quest for Synthetic Photorealism and Diversity 27
3.1.2 Fundamental Properties of Image Data 27
3.1.3 Core Objectives and Applications of Generative Image Models 29
3.2 Challenges in Image Synthesis 30
3.3 Core Applications of Generative Image Synthesis 31
3.3.1 High-Resolution Image Synthesis and Super-Resolution . . . 31
3.3.2 Image-to-Image Translation 32
3.3.3 Text-to-Image Synthesis 32
3.3.4 Image Inpainting 33
3.4 Evaluation Paradigms for Generative Image Models 33
3.4.1 The Intricacies of Evaluating Generated Images 33
3.4.2 Reconstruction Metrics 34
3.4.3 Image Generation Metrics 35
3.4.4 Emerging Semantic Metrics 35
3.4.5 Human Preference Metrics 36

4 Wavelets and Quaternion Wavelets: The Power of Multi-Scale Analysis 37
4.1 Foundational Principles of Wavelet Transforms in Signal and Image Processing 38
4.1.1 Mathematical Foundations: Mother Wavelets and Transform Theory 38
4.1.2 Continuous vs. Discrete Wavelet Transforms 39
4.1.3 Two-Dimensional Wavelet Transforms for Image Analysis . . 40
4.1.4 Quadtree Structure and Sparse Representation Properties . . 40
4.1.5 Limitations and Advanced Extensions 41
4.2 Quaternions and their Application to Images 41
4.2.1 Quaternion Algebra 41
4.2.2 Learning in the Quaternion Domain 42
4.2.3 The Dual-Tree Quaternion Wavelet Transform (QWT): Construction, Shift-Invariance, and Directional Selectivity 43
4.2.4 Simpler Strategies: Quaternion Embedding of DWT Sub-bands 45
4.2.5 Strategic Use of QWT Sub-bands in Neural Models 46
4.3 Wavelets in Deep Learning and Generative Models 47
4.3.1 Wavelets in Deep Learning 47
4.3.2 Wavelets in Generative Models 48
4.4 A Multi-Scale Foundation for Generative Modeling 49

II Exploring the Generative Landscape: Applications and Challenges 50

5 Generative Adversarial Networks for Cross-Domain Translation 51
5.1 Introduction: GANs and the Infrared Translation Challenge 51
5.2 Structural-Aware Generative Adversarial Networks for Infrared Image Translation 53
5.2.1 Experimental evaluation 55

5.3 GAN Limitations, and the Shift to Diffusion . . . 57
5.3.1 Successes and Contributions . . . 57
5.3.2 Inherent Limitations and Emerging Challenges . . . 58
5.3.3 The Bridge to Diffusion . . . 58

6 The Potential of Diffusion Models: Domain-Specific SR and Cross-Modal Generation 59
6.1 The Diffusion Paradigm Shift . . . 60
6.2 Diffusion Models for Ship-Image Super Resolution . . . 61
6.2.1 ShipSpotting Dataset . . . 63
6.2.2 Experimental Results . . . 64
6.2.3 Ablation Study on Downstream Tasks . . . 65
6.3 Streamlined EEG-to-Image Generation with Latent Diffusion Models 67
6.3.1 Problem formulation . . . 67
6.3.2 Conditioning . . . 68
6.3.3 Training . . . 69
6.3.4 Experiments . . . 70
6.4 Conclusion: Achievements and the Path Forward . . . 73
6.4.1 Key Achievements . . . 73
6.4.2 Identifying Critical Limitations . . . 73
6.4.3 The Motivation for Multi-Scale Representations . . . 74

III The Wavelet Revolution in Generative Modeling 75

7 Learning to Generalize with Wavelet-Driven Representations 76
7.1 Introduction . . . 76
7.2 Grouping and Selecting Quaternion Wavelet Bands . . . 78
7.2.1 Experimental Evaluation . . . 80
7.3 Quaternion Wavelet Feature Extraction . . . 82
7.3.1 Pretraining . . . 83
7.3.2 Architecture . . . 84
7.3.3 Experimental Setup . . . 85
7.3.4 Results . . . 90
7.4 Conclusion . . . 91

8 Wavelets for Image Super-Resolution 93
8.1 Introduction . . . 94
8.2 Quaternion Wavelet-Conditioned Diffusion Models for Image Super-Resolution . . . 96
8.2.1 Quaternion Wavelet and Time-Aware Encoder . . . 96
8.2.2 Latent Diffusion Process . . . 96
8.2.3 Experimental Results . . . 97
8.3 Metadata, Wavelet, and Time Aware Diffusion Models for Satellite Image Super Resolution . . . 102
8.3.1 Feature Modulation . . . 102
8.3.2 MWT-Encoder . . . 102
8.3.3 Experimental Results . . . 104
8.3.4 Dataset details . . . 106
8.3.5 Training details . . . 106
8.3.6 Evaluation details . . . 107
8.3.7 Additional Qualitative Comparisons . . . 107

8.3.8 Ablation Study 108
8.4 Wavelet Diffusion GAN for Image Super-Resolution 109
8.4.1 Wavelet-based conditional diffusion scheme 110
8.4.2 Conditional backward diffusion 110
8.4.3 Optimization 111
8.4.4 Experimental results 111
8.5 Conclusion 114

9 High-Fidelity Synthesis Beyond Latent Diffusion 116
9.1 Introduction 116
9.2 Latent Wavelet Diffusion for 4K Image Synthesis 118
9.2.1 VAE Fine-Tuning with Scale-Consistency Loss 119
9.2.2 Reconciling Spectral Regularization and Saliency 120
9.2.3 Wavelet-Derived Frequency Saliency Maps 120
9.2.4 Adaptive Flow Matching with Frequency-Guided Masking . . 121
9.2.5 Experiments 122
9.3 Conclusion and Future Work 130

IV Synthesis and Future Directions 137

10 Discussion and Future Endeavours 138
10.1 Discussion 138
10.2 Future Works 139

Bibliography 141

List of Acronyms

ACC N-way Top-k Classification Accuracy

AE Autoencoder

BCI Brain-Computer Interface

CAE Convolutional Autoencoder

CFW Controllable Feature Wrapping

CNNs Convolutional Neural Networks

CT Computed Tomography

DDIM Denoising Diffusion Implicit Models

DDPM Denoising Diffusion Probabilistic Models

DiTs Diffusion Transformers

DSC Dice Similarity Coefficient

DSSIM Structural Dissimilarity Index Measure

DWT Discrete Wavelet Transform

EEG Electroencephalography

EFM Edge Feature Map

EMA Exponential Moving Average

fMRI Functional Magnetic Resonance Imaging

FID Fréchet Inception Distance

FSIM Feature Similarity Index

GAN Generative Adversarial Networks

GLCM Gray Level Co-occurrence Matrix

GSD Ground Sampling Distances

HFE High-Frequency Energy

HFEI High-Frequency Emphasis Index

HLFR High/Low Frequency Ratio

HR High-Resolution

IR Infrared Images

IS Inception Score

LDM Latent Diffusion Model

LPIPS Learned Perceptual Image Patch Similarity

LR Low-Resolution

LWD Latent Wavelet Diffusion

MAE Mean Absolute Error

mIoU Mean Intersection over Union

MRI Magnetic Resonance Imaging

MS-SSIM Multi-Scale Structural Similarity Index

MSE Mean Squared Error

MWT-Encoder Metadata- and Wavelet- and Time-aware Encoder

PSNR Peak Signal-to-Noise Ratio

QAE Quaternion Autoencoder

QCAE Quaternion Convolutional Autoencoder

QNNs Quaternion-valued Neural Networks

QU-Net Quaternion U-Net

QUAVE QUaternion wAVElet network

QWT Quaternion Wavelet Transform

RDR Ratio Difference from Real

RFEM Rough Feature Extraction Module

ROI Region of Interest

SC Scale-Consistency

SD Stable Diffusion

SFT Spatial Feature Transformations

SISR Single-Image Super-Resolution

SNR Signal-to-Noise Ratio

SR Super-Resolution

SSIM Structural Similarity Index

TIR Thermal Infrared

UHR Ultra-High-Resolution

VAE Variational Auto-Encoder

ViT Vision Transformer

WaveViT Wavelet Vision Transformer

WQS Wavelet Quality Score

WTM Wavelet Transform Module

List of Figures

2.1 Conceptual diagram of a VAE: an encoder network $q_\phi(z|x)$ maps an image x to a latent distribution (often Gaussian with mean $\mu(x)$ and std $\sigma(x)$), and a decoder $p_\theta(x|z)$ maps a latent sample z back to the data space. Training maximizes the ELBO (e.g. reconstruction minus KL divergence). 11
2.2 Illustration of GAN adversarial training dynamics. The generator G produces images from noise (left) and the discriminator D learns to distinguish them from real images (right). The objective is a minimax game whose equilibrium ideally corresponds to the generator matching the data distribution. 15
2.3 Illustration of the diffusion process: a data sample $\mathbf{x}_0$ is gradually corrupted by Gaussian noise to $\mathbf{x}_T$ (forward process), and a neural network learns the reverse denoising (reverse process). Each step of the reverse process resembles denoising autoencoding conditioned on the timestep. 17

4.1 Dual-tree quaternion wavelet transform (QWT). The input image is decomposed into sixteen sub-bands by interleaving low-pass and high-pass filters. 43
4.2 Filters for QWT: G_L and G_H are extracted from the wavelet db8, while F_L and F_H are Hilbert transforms of the previous ones. 44

5.1 Architecture of StawGAN. We show the part inside of the generator, to explore the shared connections between the two different flows. Indeed this is a translation from RGB to IR, but the same process holds for the contrary. 53
5.2 Random samples of StawGAN image modality translations on the DroneVehicle dataset. The first column refers to the original domain images, IR in the first row, and RGB in the second. The corresponding translations in the RGB domain are in the first row and subsequent columns, while IR is in the second one. Our approach produces better samples in both domains with respect to other state-of-the-art models. 56

6.1 Comparative sample images showcasing the super-resolution capabilities of different models. The visual assessments highlight variations in image quality, clarity, and detail enhancement achieved by each respective model in the upscale process. Ours is in the last column. . 62

6.2 Comparison between LR image and SR images reconstructed by other state-of-the-art methods and our proposed model. The red bounding box contains a zoomed-in part of the image in order to have a better understanding of the image resolution quality of our method compared to the others. 65
6.3 Comparison for precision scores on the Seaships dataset. We employ the models pre-trained on our dataset, thus performing zero-shot super resolution, improving the detection of a finetuned YOLOv7 [1] on Seaships. 66
6.4 Comparison of SSIM (a) and PSNR (b), which measure the quality of the reconstructed images. We confront those results against the classification accuracy reached with the generated images on the ShipsSpotting dataset. This comparison shows how the classical reconstruction metrics are not a measure of how well a model performs in a super-resolution scenario since the accuracy and the FID reached are lower concerning other models that have lower SSIM and PSNR. 67
6.5 **Guess What I Think (GWIT).** Outline of our streamlined framework which includes a projection function, a ControlNet adapter to handle EEG conditioning, a frozen LDM, and a frozen EEG image decoder to obtain a coarse-grained control. 68
6.6 Comparison for images generated by models trained on ThoughtViz [2] dataset. At the top, we report a random sample of the same class from the original dataset. 69
6.7 Comparison with the previous state-of-the-art method for images generated by models trained on all subjects of EEGCVPR40. At the top, we report a random sample of the same class from the original dataset. 69
6.8 Comparison of images generated with models trained on subject 4 (top three rows) and on all subjects (bottom four rows) of EEGCVPR40. 71
6.9 Comparison of images generated with models trained only on subject 4 of EEGCVPR40. Left: random sample from the dataset. 72

7.1 Random samples of TarGAN translations on the CHAOS dataset, where the first column refers to the source image. Red boxes highlight regions where the TarGAN distorts or blurs organs, while our method provides a superior translation. 80
7.2 Illustration of the QUAVE framework. Any medical image dataset can be processed in the quaternion domain via QUAVE, which carries out the QWT and the neural feature extraction. The training procedure can be then performed either in the real or in the quaternion domain depending on the user choice, producing a real-valued output. We integrate a quaternion-based framework in the analysis and synthesis of real-valued medical image datasets. 82
7.3 Detailed QUAVE architecture. Being a quaternion-based generalizable framework, the WTM enhances neural model performance when processing rich-of-detail real images, such as medical ones. Along with the inverse quaternion wavelet block, the WTM performs a salient feature extractor by leveraging quaternion algebra properties to achieve more detailed insights. During inference, the output of our framework can be then involved in any neural model either in the real or quaternion domain, improving performance in a variety of tasks (e.g., segmentation, reconstruction, and generation, among others). . 84

7.4 The Rough Feature Extraction Module (RFEM) is responsible for initial feature extraction before the Quaternion Wavelet Transform (QWT). It consists of a convolution layer and multiple residual groups designed to extract both shallow and deep features from the input image and does not apply QWT. This prepares the data with enriched initial features before entering the QWT, making the wavelet transform more effective. 85
7.5 Random samples of TarGAN image modality translations on the CHAOS dataset. The first column refers to the original domain images, while the others refer to the corresponding translation in the other domains. 87

8.1 Overview of ResQu Super-Resolution framework. We pre-trained QUAVE and trained only our encoder, the quaternion embeddings- and time- aware encoder. The other parts of the framework are frozen, which speeds up the overall training process. 95
8.2 Comparison of the LR input image with SR outputs generated by state-of-the-art methods and our proposed model on the DRealSR [3] and RealSR [4] datasets. The red bounding box highlights a zoomed-in region, illustrating the superior resolution and detail preservation achieved by our method compared to existing approaches. 98
8.3 Impact of the number of sampling steps on key evaluation metrics. Reducing the steps improves efficiency while maintaining performance, highlighting the inherent speed advantage of our model. 100
8.4 Visual comparison of diverse super-resolution results generated by changing the CFW. The images demonstrate the impact of our CFW enhanced by our encoder, producing more structured details. 101
8.5 Overview of MWT-Diff framework with our MWT-Encoder. 103
8.6 Qualitative results on 128x128 $\rightarrow$ 512x512 with fMoW [5]. 104
8.7 Additional results on 128x128 $\rightarrow$ 512x512 with fMoW [5]. Comparison between MWT-Diff and StableSR. 105
8.8 Qualitative results on Sentinel2-fMoW dataset [6]. The image includes a comparison of the LR Sentinel-2 input, the output of the model, and the corresponding HR fMoW images. 107
8.9 Qualitative comparison of super-resolution outputs with and without WaveViT embeddings. From left to right: Low-resolution input, MWT-Diff without WaveViT, MWT-Diff with WaveViT, and Ground Truth. 108
8.10 Method architecture and training scheme. In green our discriminator and in blue our conditional generator. x_0 undergoes forward diffusion in wavelet space and the resulting pure noise x_t gets concatenated to the low-res input x_{lr} to condition the generator for the backward diffusion. 110
8.11 Reverse diffusion process and inference: the model p_θ iteratively produces a more refined sample from x_t and x_{lr}. After T iterations, x'_0 is used to reconstruct the super-resolved image. 111
8.12 Qualitative comparison between our model, ESRGAN, SR3 and DiWa trained for 25k iteration steps on CelebA-HQ for the task of 16x16 $\rightarrow$ 128x128. 113
8.13 Results on Shipspotting [7]. 114

9.1 We propose Latent Wavelet Diffusion, achieving 4K image synthesis without architectural changes or additional inference cost to existing latent diffusion models. 117
9.2 (a) Temporal evolution of latent z_t, wavelet energy maps $A_{wavelet}$, and attention map M_t across diffusion timesteps. (b) Our wavelet-masked flow matching objective at a timestep t. The model computes a wavelet attention map $\mathbf{M}_t$ from latent z_t to modulate the prediction error between target velocity field $(\epsilon - z_0)$ and predicted velocity $v_\Theta(z_t, t, y)$. This focuses optimization on high-frequency regions with greater perceptual importance. While operations occur in latent space, decoded visualizations are shown for interpretability. 119
9.3 Normalized DCT amplitudes over zigzag frequency indices. VAE trained with the multi-scale loss reduces high-frequency energy in latents, aligning their spectrum with that of RGB images. 120
9.4 Visual comparison of 2K image generations. LWD demonstrates improved detail preservation in complex areas while avoiding over-sharpening or texture collapse. 122
9.5 4K images generated by LWD with different architectures. 125
9.6 4K generation of URAE vs LWD + URAE. Upper caption: *"Eiffel Tower was Made up of more than 2 million translucent straws to look like a cloud, with the bell tower at the top of the building, Michel installed huge foam-making machines in the forest to blow huge amounts of unpredictable wet clouds in the building's classic architecture."*. Lower caption: *"Barbarian woman riding a red dragon, holding a broadsword, in gold armour."* 126
9.7 2K generation of PixArt-Sigma-XL vs LWD + PixArt-Sigma-XL. Upper caption: *"An elderly man with a prominent, bushy beard and deep-set eyes wears a vibrant orange turban, his weathered face marked by lines of age and experience."*. Lower caption: *"A lone figure on a horse stands in a misty forest, gazing up at a tall, multi-tiered temple surrounded by towering trees and soft, diffused light. Steam rises from the rocks near a stream, creating an atmospheric scene of tranquility and mystery."* 129
9.8 4K generation of Sana vs LWD + Sana. Upper caption: *"A litter of golden retriever puppies playing in the snow. Their heads pop out of the snow, covered in."*. Lower caption: *"A curvy timber house near a sea, designed by Zaha Hadid, represent the image of a cold, modern architecture, at night, white lighting, highly detailed."* 131
9.9 2K generation of SD3-Diff4k-F16 vs LWD + SD3-F16. Upper caption: *"A serene landscape features a winding river, flanked by trees with autumn foliage, leading to a rustic wooden cabin with a corrugated roof, set against a softly blurred background."*. Lower caption: *"A grand interior featuring intricate stained glass windows, an elaborate rose window, ornate frescoes depicting biblical scenes, and elegant chandeliers illuminating the richly decorated walls and arches."* . . . 132
9.10 Images generated at 4K resolution with LWD+SANA. 133
9.11 Images generated at 4K resolution with LWD+URAE. 134
9.12 Images generated at 4K resolution with LWD+URAE. 135
9.13 Visual comparison of 4K image generations from LWD and competing baselines. 136

List of Tables

1.1 Schematic overview of the research contributions related to the thesis. 6

2.1 Comparative Overview of Major Generative Model Families 24

4.1 Comparative Overview of Fourier and Wavelet Transforms 39

5.1 Quantitative comparison on image modality translation. 57
5.2 Quantitative comparison on segmentation. 58

6.1 FID results obtained on the test set. 64
6.2 Overall precision and recall on Seaships dataset. The detection is done by a finetuned YOLOv7 [1] on Seaships. 66
6.3 Generation quality and semantic correctness of models trained on EEGCVPR40 (top: subject 4, bottom: all subjects) and ThoughtViz. 70
6.4 Ablation study for influence of EEG conditioning. 72
6.5 Ablation study for “drop" (during training) and “guess" (at inference time) modes on the multi-subject variant. 73

7.1 Selected sub-bands for each dataset. 79
7.2 Results (average scores over multiple runs) for reconstruction, segmentation and image modality translation. 81
7.3 Reconstruction results on the IXI dataset for vanilla autoencoders defined in the quaternion domain. 82
7.4 Results for task 3 image modality translation on the CHAOS dataset (Segmentation metrics) for real and quaternion-valued models. . . . 88
7.5 Results for task 1 reconstruction on the IXI dataset for real and quaternion-valued models. Average and standard deviation over multiple runs reported. Runs statistically significant with $p < 0.1$. 89
7.6 Results for task 2 segmentation on the Kvasir-SEG dataset for real and quaternion-valued models. Average and standard deviation over multiple runs reported. Runs statistically significant with $p < 0.1$. . 89
7.7 Results for task 3 image modality translation on the CHAOS dataset (image generation metrics) for real and quaternion-valued models. Average and standard deviation over multiple runs reported. Runs statistically significant with $p < 0.1$. 90
7.8 Ablation study for real and quaternion-valued models using different wavelets in QUAVE. 90

8.1 Quantitative comparison with state-of-the-art methods on both synthetic and real-world benchmarks. The best performance for each metric is in **bold**, while the second-best is underlined. 97

8.2 Metrics obtained with different versions of the CFW with our ResQu. 100
8.3 FID results obtained on the Shipspotting dataset presented in section 6.2.1. 102
8.4 Evaluation metrics comparison for the $128 \times 128 \rightarrow 512 \times 512$ super resolution task on fMoW dataset [5], and on Sentinel2-fMoW dataset [6]. 105
8.5 Evaluation metrics comparison for our model, ESRGAN, SR3 and DiWa trained for 25k iteration steps on CelebA-HQ for the task of 16x16 $\rightarrow$ 128x128. 112

9.1 Quantitative results on different metrics. The prompts are from the HPD [8] dataset. All images are at a resolution of 2048×2048. . . . 123
9.2 Quantitative benchmarks of latent diffusion models on Aesthetic-Eval at 2048×2048 and 4096×4096. GLCM Score measures high-frequency texture richness using gray-level co-occurrence matrices, while Compression Ratio assesses visual complexity via JPEG file size heuristics, both introduced in Diffusion-4K [9]. 124
9.3 Evaluation on ultra resolution (4096×4096) image generation task with [8] prompts. 127
9.4 Quantitative reconstruction results of VAEs with and without the Scale-Consistency Loss 9.2.1 on Aesthetic-4K [9] validation set. . . . 127
9.5 Comparison of frequency-sensitive metrics across different methods on Aesthetic-4K [9] validation set. 130

Chapter 1

Introduction

Contents

1.1	Deep Learning for Generative Image Modeling	1
1.2	The Vanguard of Generative Modeling and its Prevailing Limitations	2
1.3	A Wavelet-Driven Paradigm for Generative Modeling	3
1.4	The Proposed Wavelet-Centric Approach to Generative Modeling	4
1.5	Thesis Contributions	5
1.6	Thesis Organization	6

This chapter lays the foundation for the central argument of this thesis: that a deep and systematic integration of multi-scale wavelet theory offers a powerful pathway to overcoming persistent challenges in generative modeling. We begin in Section 1.1 by discussing the motivations and charting the evolving landscape of generative image synthesis. Following this, Section 1.2 presents a critical survey of the state-of-the-art, from Generative Adversarial Networks to Diffusion Models, and delineates the prevailing limitations in fidelity, resolution, efficiency, and controllability that motivate this research.

In response to these challenges, Section 1.3 introduces the core methodology of this work, multi-scale wavelet analysis. It makes the case for its principled application in modern generative architectures and, by reviewing prior work, identifies the critical research gap that this thesis aims to bridge. Building upon this, Section 1.4 articulates the novel wavelet-driven frameworks developed herein. Finally, Section 1.5 provides a consolidated summary of the distinct scientific contributions, and Section 1.6 outlines the organizational structure of the subsequent chapters.

1.1 Deep Learning for Generative Image Modeling

The aspiration to computationally replicate, manipulate, and ultimately understand the visual world represents a long-standing and profound challenge in artificial intelligence. Generative image synthesis, the discipline focused on creating novel, realistic, and diverse visual data from scratch or based on specified conditions, has rapidly evolved from a nascent field into a cornerstone of modern machine learning and computer vision [10, 11]. This quest is driven by dual imperatives: advancing scientific discovery and fostering technological innovation. It transcends mere

pixel manipulation, embodying the fundamental scientific challenge of learning and replicating the complex statistical distributions and intricate underlying structures that govern natural image formation. The overarching motivation is to develop computational systems that not only replicate the visual appearance of the world but also understand its semantic fabric and underlying physical principles. This enables the generation of images that are not only photorealistic, controllable, and coherent in content but also preserve semantic consistency, structural integrity, and perceptual fidelity across a multitude of visual domains. The works detailed in this thesis contribute to these broad goals.

The significance of generative image synthesis extends far beyond academic curiosity, with profound and transformative implications for a myriad of fields. These models have unlocked unprecedented capabilities, revolutionizing creative industries by providing artists and designers with novel tools for expression [12], and critically advancing scientific discovery through enhanced data visualization, simulation, and data augmentation. The latter is particularly vital in fields where real data is scarce or costly to acquire, such as in medical imaging (a context for research detailed in Chapter 7) or in specialized remote sensing applications like satellite image super-resolution (the focus of work in Chapter 8).

The landscape of deep generative models is dynamic and has seen a rapid evolution of paradigms. Initial breakthroughs were significantly driven by Variational Autoencoders (VAEs) [13], which learn a probabilistic mapping to a continuous latent space, and Generative Adversarial Networks (GANs) [10]. GANs, through their innovative adversarial training process, achieved significant milestones in generating sharp images and were explored for specific tasks such as domain-to-domain translation, as discussed in Chapter 5 concerning structural-aware infrared image translation. More recently, the field has been dominated by the ascendant paradigms of Denoising Diffusion Probabilistic Models (DDPMs) [11, 14] and their numerous variants. These methods, which learn to reverse a gradual noise-addition process, have set new benchmarks for high-fidelity image synthesis. The computational intensity of early diffusion models, however, spurred the development of Latent Diffusion Models (LDMs) [12]. LDMs perform the iterative denoising process within a compressed latent space, significantly enhancing efficiency and forming the architectural backbone for several investigations presented in this thesis, including contributions to cross-modal generation and advanced super-resolution techniques (Chapter 6). Concurrently, newer simulation-free training paradigms like Flow Matching models [15] are offering promising avenues for even greater training efficiency and scalability, a direction that aligns with the objectives of the high-resolution synthesis work detailed in Chapter 9.

1.2 The Vanguard of Generative Modeling and its Prevailing Limitations

The field of generative image modeling is characterized by a rapid and dynamic evolution of powerful architectures. **Generative Adversarial Networks (GANs)** [10], along with their conditional variants like Pix2PixHD [16] and CycleGAN [17], were a watershed, demonstrating an unparalleled ability to produce sharp, high-frequency details. Their aptitude for image-to-image translation tasks remains highly relevant, forming the basis for the structural-aware infrared image translation explored in Chapter 5. Despite their success, GANs are often hampered by unstable training dynamics and the persistent issue of mode collapse [18], limiting their reliability and the diversity of their outputs.

More recently, the vanguard has been claimed by **Denoising Diffusion Probabilistic Models (DDPMs)** [11] and their computationally efficient successors, **Latent Diffusion Models (LDMs)** [12]. These paradigms have set new standards for high-fidelity synthesis, offering stable training and superior sample diversity [19]. Their effectiveness in image restoration and enhancement is notable, with models like SR3 [20] and LDM-based approaches such as StableSR [21] serving as key benchmarks for the super-resolution contributions of this thesis.

Despite this remarkable progress, several profound and interconnected challenges persist, forming the primary impetus for the research presented in this dissertation. These are not minor hurdles, but fundamental limitations that prevent current models from achieving seamless, efficient, and controllable synthesis across all domains and scales.

- **Generalization and Domain Adaptation:** Models trained on one data distribution often falter when applied to another, requiring extensive retraining. This brittleness is particularly acute in specialized fields like medical imaging, where data is heterogeneous and scarce. This necessitates the development of robust feature representations that capture essential structure while remaining invariant to domain-specific styles, a core theme of Chapter 7.

- **Computational Efficiency:** The iterative nature of diffusion models, while effective, imposes a significant computational burden during training and inference. This high cost can limit their practical deployment and accessibility. Methodologies that reduce the number of sampling steps or the dimensionality of the processed data are therefore highly sought after and explored in Chapter 8.4.

- **Nuanced Conditioning and Controllability:** The ability to accurately guide image synthesis based on diverse and multimodal inputs, such as the low-resolution images used for super-resolution in Chapter 8 or the complex biomedical signals explored in Chapter 6.3, remains a key research avenue. Existing methods often rely on complex, task-specific encoders or alignment modules, indicating a need for more streamlined and potent conditioning strategies.

- **Resolution and Fidelity at Scale:** A significant and enduring grand challenge is the generation of ultra-high-resolution (UHR) images (e.g., 2K, 4K, and beyond) that maintain both global coherence and intricate, fine-grained fidelity. Naively scaling existing architectures often leads to prohibitive computational demands, excessive memory footprints, or a degradation in image quality [22, 23]. The development of lightweight, scalable solutions remains a frontier, a problem directly confronted in Chapter 9.

1.3 A Wavelet-Driven Paradigm for Generative Modeling

This thesis posits that a powerful, principled, yet historically under-exploited pathway to systematically addressing these intertwined challenges lies in the deep integration of multi-scale representations derived from **wavelet theory**. Wavelet transforms, a cornerstone of signal processing [24], provide a mathematical framework for decomposing a signal, such as an image, into a hierarchy of frequency sub-bands that are localized in both space and scale. This decomposition elegantly disentangles

coarse, low-frequency approximations (capturing global structure) from a series of high-frequency details (representing textures, edges, and other fine-grained features).

The intrinsic properties of wavelets render them exceptionally suitable for enhancing generative models. Their capacity for *sparse representation* allows for the compact encoding of image features; their *multi-scale nature* provides a natural framework for hierarchical processing and conditioning; and their simultaneous localization in *space and frequency* allows a model to reason about both what features are present and where. This thesis further extends this approach by exploring **Quaternion Wavelet Transforms (QWT)**, a hypercomplex generalization that offers a richer representational power by modeling inter-channel correlations and phase information, properties that are particularly beneficial for tasks demanding high structural fidelity (foundational work detailed in Chapter 7).

The integration of wavelets into generative models is not without precedent. Seminal works such as WaveDiff [25] and related approaches [26] have leveraged the Discrete Wavelet Transform (DWT) to accelerate unconditional diffusion models by performing the denoising process on the more compact wavelet coefficients. Other methods have incorporated wavelet-based losses or guidance mechanisms for specific objectives, such as UHR synthesis [9] or medical image analysis, which inspired some of the foundational research in this thesis.

However, a comprehensive review reveals a critical research gap. Prior to the contributions of this thesis, there has been a lack of *systematic, deep, and versatile integration* of advanced wavelet principles as a core, unifying mechanism across the full spectrum of modern generative architectures (GANs, DDPMs, LDMs, and Flow Matching models) and for a broad array of challenging synthesis tasks. Existing approaches have often employed wavelets in a circumscribed or ancillary manner, as a fixed preprocessing step, an auxiliary loss term, or within a single, task-specific model.

The central argument of this thesis is that the full potential of wavelets, to fundamentally reshape multi-scale feature representation, dynamically guide conditioning mechanisms, and adapt training objectives for enhanced fidelity and efficiency in a model-agnostic fashion, has remained largely untapped. This research seeks to unlock that potential, demonstrating that a strategically integrated, wavelet-driven paradigm offers a powerful and computationally efficient pathway to surmounting the persistent limitations of the field.

1.4 The Proposed Wavelet-Centric Approach to Generative Modeling

This dissertation introduces and validates a cohesive research program centered on the transformative potential of multi-scale wavelet representations when deeply integrated into the architecture and training of modern generative models. The central hypothesis posits that by leveraging the unique properties of DWT, QWT, and innovative wavelet-guided training and conditioning strategies, we can address the persistent challenges in generative image synthesis, namely resolution, fidelity, computational efficiency, controllability, and generalization, in a more unified, principled, and effective manner.

This research endeavors to move beyond superficial applications of wavelets, instead interweaving their analytical strengths into the core mechanisms of feature extraction, model conditioning, and loss formulation across a diverse range of generative architectures, including GANs, Diffusion Models, Latent Diffusion Models (LDMs), and Flow Matching paradigms.

The research journey documented herein commenced by establishing the versatility of generative models: Chapter 5 details StawGAN for structural-aware infrared-to-RGB translation. Chapter 6 presents "Ship in Sight" for maritime SR (including the ShipSpotting dataset contribution) and "Guess What I Think" (GWIT) for streamlined EEG-to-image generation using ControlNet-adapted LDMs [27].

We continue introducing foundational frameworks that harness the rich representational power of wavelet and quaternion wavelet transforms for enhanced feature representation and improved generalization, particularly within the demanding and data-variable domain of medical imaging. As detailed in Chapter 7, we pioneered a task- and model-agnostic methodology for producing salient features from medical images by strategically selecting and grouping informative QWT sub-bands. Building upon this, we also established a generalizable system that employs QWT within a neural network architecture to extract highly representative and robust features, demonstrably improving downstream medical image analysis and synthesis tasks.

These foundational wavelet-centric principles were subsequently adapted and extended to significantly advance the state-of-the-art in image super-resolution (SR) across various domains (Chapter 8).

The synthesis of these investigations culminates in the Latent Wavelet Diffusion (LWD) framework (Chapter 9).

The core novelty of this thesis lies in the systematic and multifaceted integration of wavelet transforms as a fundamental mechanism across the generative pipeline, offering a principled and efficient pathway to address complex challenges in image synthesis.

1.5 Thesis Contributions

This doctoral research makes several significant contributions to computer vision and generative modeling. The following narrative presents these achievements in an order that reflects the structure of the thesis, first detailing advancements made by applying and extending state-of-the-art models, and then introducing the core methodological innovations centered on our novel wavelet-driven paradigm.

The contributions detailed in Part II of the thesis begin by demonstrating the successful application of established generative architectures to challenging, real-world problems. Our research journey commences in Chapter 5 with the development of **StawGAN**, a GAN-based architecture specifically designed for translating infrared imagery while preserving critical structural details. This exploration into a specialized domain sets the stage for our subsequent work in Chapter 6, where we pivot to the more advanced paradigm of Latent Diffusion Models to address challenges in maritime imagery and cross-modal generation. This body of work includes the **"Ship in Sight"** project for high-fidelity ship image super-resolution, an effort supported by our curation and public release of the large-scale **ShipSpotting dataset**. In the same chapter, we also introduce the **GWIT ("Guess What I Think")** framework, which establishes a streamlined and effective method for EEG-to-image synthesis. Collectively, these initial contributions solve specific, complex problems while simultaneously motivating the need for the more advanced, wavelet-based techniques developed in the second part of this thesis.

The primary novelty of this dissertation, presented in Part III, lies in the systematic and deep integration of wavelet theory into modern generative models to establish a new, principled paradigm for efficient and high-fidelity image synthesis. The cornerstone of this innovation, detailed in Chapter 7, is our pioneering research into the use of **Quaternion Wavelet Transforms (QWT)** within deep learning

Table 1.1. Schematic overview of the research contributions related to the thesis.

Part II	**Chapter** 5	Presented at *2023 International Symposium on Circuits and Systems (ISCAS)* [28]
	Chapter 6	Presented at *2024 International Joint Conference on Neural Networks (IJCNN)* [7]
		Presented at *2025 IEEE International Conference on Acoustics, Speech, and Signal Processing Workshop (ICASSP)* [29]
Part III	**Chapter** 7	Published in *Signal Processing Letters* [30]
		Published in *Neurocomputing* [31]
	Chapter 8	Presented at *2025 International Joint Conference on Neural Networks (IJCNN)* [32]
		Presented at *2025 ICLR Workshop on Machine Learning for Remote Sensing (ML4RS)* [33]
		Presented at *2024 Italian Workshop on Neural Networks (WIRN)* [34]
	Chapter 9	Presented at *2026 Conference on International Conference on Learning Representations* [35]

architectures to learn highly robust and generalizable features. This foundational work demonstrates how to harness the hypercomplex properties of wavelets to capture rich phase and inter-channel information, particularly for challenging data domains like medical imaging.

With these principles established, we apply them in Chapter 8 to introduce a suite of advanced super-resolution models that leverage wavelet representations for superior performance. These include the hybrid **Wavelet Diffusion GAN**, the QWT-conditioned **ResQu**, and the context-aware **MWT-Diff** for satellite imagery. The synthesis of these investigations culminates in the **Latent Wavelet Diffusion (LWD)** framework, our flagship contribution presented in Chapter 9. LWD provides a model-agnostic and computationally efficient solution that enables existing latent diffusion and flow matching models to achieve ultra-high-resolution (up to 4K) synthesis without architectural modifications or additional inference overhead. This is accomplished through a novel synergy of scale-consistent VAE objectives and wavelet-derived saliency guidance.

To ensure the impact and verification of this research, reproducibility has been a guiding principle. All code, models, and experiments presented in this thesis are made publicly available for community use at https://github.com/luigisigillo.

1.6 Thesis Organization

The remainder of the thesis is organized into three main parts, with each part building upon the foundations established in the previous sections. The structure is designed to first establish the necessary theoretical background before presenting our core methodological contributions and applications in wavelet-driven generative modeling.

Part I provides essential background and foundational knowledge across three chapters, establishing the theoretical context required to understand the main challenges and opportunities in generative image modeling that our work addresses.

- **Chapter 2** introduces the landscape of generative models, providing a comprehensive overview of Variational Autoencoders (VAEs), Generative Adversarial Networks (GANs), Denoising Diffusion Probabilistic Models (DDPMs), Latent Diffusion Models (LDMs), and Flow Matching Models. The chapter establishes the theoretical foundations, training dynamics, and architectural principles that underpin modern generative modeling.
- **Chapter 3** focuses specifically on generative models for image synthesis, examining the unique challenges of image data, key application areas including super-resolution, text-to-image synthesis, inpainting, and image-to-image translation. The chapter also discusses evaluation metrics and success measures for image generation tasks.
- **Chapter 4** introduces the mathematical foundations of wavelet transforms and quaternion algebra. This chapter establishes the theoretical basis for multi-scale analysis through wavelets, quaternion representations of images, and their enhanced capabilities for capturing spatial-frequency information in deep generative models.

Part II explores the application landscape of generative models through two chapters, demonstrating both the capabilities and limitations of current approaches across different domains and modalities.

- **Chapter 5** presents our work on Generative Adversarial Networks for cross-domain translation, specifically introducing StawGAN for structural-aware infrared image translation. The chapter discusses both the successes and inherent limitations of GAN-based approaches, providing motivation for the transition to diffusion-based methods.
- **Chapter 6** demonstrates the power and potential of diffusion models through domain-specific applications, including ship image super-resolution and EEG-to-image generation. This chapter establishes the superior capabilities of diffusion models while identifying critical limitations that motivate our multi-scale wavelet approach.

Part III presents our core methodological contributions through three chapters, introducing the revolutionary integration of wavelets into generative modeling frameworks.

- **Chapter 7** introduces our approach to learning generalizable representations through wavelet-driven architectures. This chapter presents methods for grouping and selecting quaternion wavelet bands and demonstrates how quaternion wavelet networks can achieve superior generalization across different image domains.
- **Chapter 8** focuses specifically on wavelets for image super-resolution, presenting multiple innovative approaches including quaternion wavelet-conditioned diffusion models, metadata-wavelet-time aware diffusion models for satellite imagery, and wavelet diffusion GANs. The chapter demonstrates significant improvements in super-resolution quality and efficiency.

- **Chapter 9** extends beyond traditional latent diffusion approaches, introducing Latent Wavelet Diffusion for high-resolution synthesis. This chapter presents our methodology for enabling 4K image synthesis through VAE fine-tuning with scale-consistency loss, wavelet-derived frequency saliency maps, and adaptive flow matching with frequency-guided masking.

Finally, **Part IV** concludes the thesis with Chapter 10, providing a comprehensive synthesis of our contributions, key insights regarding the role of wavelets in modern generative models, discussion of limitations, and identification of promising future research directions in wavelet-based generative modeling.

Portions of this thesis are adapted from material published or currently under review in various journals and conferences. We present an overview of how the different research contributions are organized in Tab. 1.1. In addition, each chapter includes an introductory footnote with publication details when necessary.

Part I

Background and Foundations

Chapter 2

Generative Models: A Deeper Dive

Contents

2.1 Variational Autoencoders (VAEs) 11
2.1.1 Theoretical Foundation and Variational Inference 11
2.1.2 The Reparameterization Trick and Auto-Encoding Variational Bayes 12
2.2 Generative Adversarial Networks (GANs) 13
2.2.1 Game-Theoretic Foundations and Adversarial Training . . 13
2.2.2 Training Dynamics and Stability Challenges 14
2.2.3 Theoretical Extensions and Variants 15
2.3 Denoising Diffusion Probabilistic Models (DDPMs) . . . 16
2.3.1 Theoretical Framework and Non-Equilibrium Thermodynamics 16
2.3.2 Variational Lower Bound and Training Objectives 17
2.3.3 Discrete State Spaces and Advanced Formulations 18
2.4 Latent Diffusion Models (LDMs) 19
2.4.1 Latent Space Representation and Computational Efficiency 19
2.4.2 Cross-Attention Mechanisms and Conditional Generation 20
2.4.3 Architectural Innovations and Scaling 20
2.5 Flow Matching Models 21
2.5.1 Continuous Normalizing Flows and Simulation-Free Training 21
2.5.2 Optimal Transport and Efficient Probability Paths 22
2.5.3 Applications and Scaling Properties 23
2.5.4 Comparative Analysis with Other Model Families 24

We examine the theoretical foundations and practical implementations of modern generative models, focusing on the mathematical frameworks that enable these systems to learn complex data distributions and generate novel samples. The field has witnessed remarkable progress through the development of variational autoencoders, adversarial training paradigms, diffusion-based approaches, and flow-matching techniques, each contributing unique perspectives on the fundamental challenge of probabilistic modeling.

2.1 Variational Autoencoders (VAEs)

2.1.1 Theoretical Foundation and Variational Inference

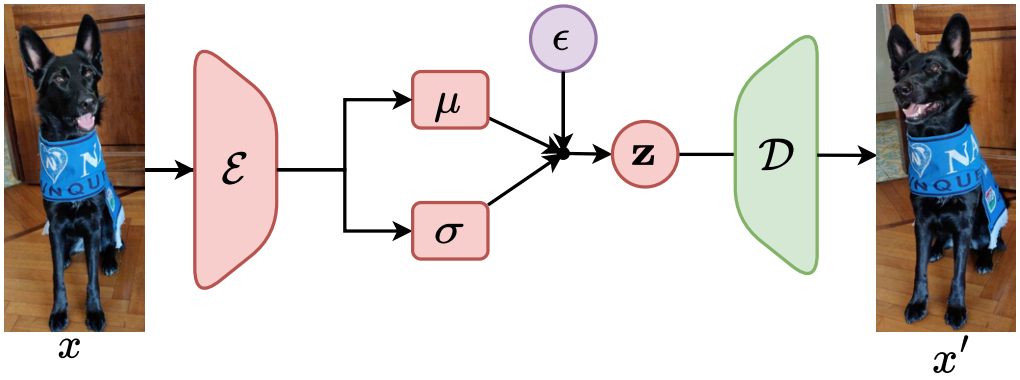


Figure 2.1. Conceptual diagram of a VAE: an encoder network $q_\phi(z|x)$ maps an image x to a latent distribution (often Gaussian with mean $\mu(x)$ and std $\sigma(x)$), and a decoder $p_\theta(x|z)$ maps a latent sample z back to the data space. Training maximizes the ELBO (e.g. reconstruction minus KL divergence).

Variational Autoencoders (VAEs) represent a principled and powerful approach to learning deep latent variable models, grounded in the framework of probabilistic graphical models and variational inference [13, 36]. In a VAE, we posit a generative model

$$p_\theta(\mathbf{x}, \mathbf{z}) = p_\theta(\mathbf{x}|\mathbf{z})p(\mathbf{z})$$

, where $\mathbf{x}$ represents the observed high-dimensional data (e.g., images) and $\mathbf{z}$ is a low-dimensional latent code drawn from a prior distribution $p(\mathbf{z})$ (often a standard Gaussian). The distribution $p_\theta(\mathbf{x}|\mathbf{z})$ is a decoder, typically parameterized by a neural network, which maps the latent code $\mathbf{z}$ to the data space of $\mathbf{x}$.

A central challenge in such models is that the true posterior distribution $p_\theta(\mathbf{z}|\mathbf{x})$ is generally intractable to compute. VAEs address this by introducing a parametric inference model, also known as an encoder or recognition network, $q_\phi(\mathbf{z}|\mathbf{x})$, to approximate the true posterior [36]. This encoder is also commonly realized as a deep neural network. The training of VAEs is performed by maximizing the Evidence Lower Bound (ELBO) on the marginal log-likelihood of the data, $\log p_\theta(\mathbf{x})$:

$$\mathcal{L}(\theta, \phi; \mathbf{x}) = \mathbb{E}_{\mathbf{z}\sim q_\phi(\mathbf{z}|\mathbf{x})}[\log p_\theta(\mathbf{x}|\mathbf{z})] - D_{KL}(q_\phi(\mathbf{z}|\mathbf{x}) \,\|\, p(\mathbf{z}))$$

This objective function consists of two terms: the first is the expected negative reconstruction error, which encourages the decoder to accurately reconstruct the data $\mathbf{x}$ from the latent representation $\mathbf{z}$ sampled from the approximate posterior. The second term is the Kullback-Leibler (KL) divergence between the approximate posterior $q_\phi(\mathbf{z}|\mathbf{x})$ and the prior $p(\mathbf{z})$, acting as a regularization term that encourages the learned latent space to conform to the structure of the prior.

Maximizing the ELBO jointly optimizes the generative model parameters θ (decoder) and the inference model parameters ϕ (encoder) [13]. This objective naturally arises from variational Bayesian inference, as the ELBO lower-bounds $\log p_\theta(\mathbf{x})$ by an amount equal to $D_{KL}(q_\phi(\mathbf{z}|\mathbf{x}) \,\|\, p_\theta(\mathbf{z}|\mathbf{x}))$ (see Figure 2.1 for a graphical model representation). Consequently, maximizing the ELBO drives $q_\phi(\mathbf{z}|\mathbf{x})$ to become

a better approximation of the true posterior and simultaneously tunes $p_\theta(\mathbf{x}|\mathbf{z})$ to effectively model the data distribution.

The learned latent space Z can capture salient, low-dimensional structures within the data (e.g., digit identity or facial attributes), which enables applications in representation learning, conditional generation, image denoising, and semi-supervised tasks [13]. A key advantage of VAEs is the amortization of inference costs [36]. Instead of performing iterative optimization to infer latent variables for each data point, the encoder learns a global mapping from observations $\mathbf{x}$ to the parameters of the approximate posterior $q_\phi(\mathbf{z}|\mathbf{x})$, allowing for efficient inference on large datasets and for new, unseen data points.

2.1.2 The Reparameterization Trick and Auto-Encoding Variational Bayes

A critical innovation that rendered VAEs practical for training via gradient-based methods is the *reparameterization trick* [13, 37]. The expectation term in the ELBO involves sampling $\mathbf{z}$ from $q_\phi(\mathbf{z}|\mathbf{x})$, which is a stochastic operation that would typically prevent gradients from flowing back to the parameters ϕ of the encoder. The reparameterization trick reformulates the sampling process by expressing the latent variable $\mathbf{z}$ as a deterministic function of the input $\mathbf{x}$, the parameters ϕ, and an auxiliary noise variable $\boldsymbol{\epsilon}$ with a fixed distribution $p(\boldsymbol{\epsilon})$. For instance, if $q_\phi(\mathbf{z}|\mathbf{x})$ is a Gaussian distribution $\mathcal{N}(\mathbf{z}; \boldsymbol{\mu}_\phi(\mathbf{x}), \boldsymbol{\sigma}^2_\phi(\mathbf{x})\mathbf{I})$, then $\mathbf{z}$ can be reparameterized as:

$$\mathbf{z} = \boldsymbol{\mu}_\phi(\mathbf{x}) + \boldsymbol{\sigma}_\phi(\mathbf{x}) \odot \boldsymbol{\epsilon}, \quad \text{where } \boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$$

Here, $\boldsymbol{\mu}_\phi(\mathbf{x})$ and $\boldsymbol{\sigma}_\phi(\mathbf{x})$ are outputs of the encoder network. This transformation decouples the stochasticity from the network parameters, allowing gradients of the ELBO with respect to θ and ϕ to be computed using standard backpropagation. This technique yields an unbiased, often low-variance, estimator of the ELBO gradient, effectively integrating stochastic gradient variational Bayes (SGVB) principles into deep learning models [13].

This reparameterization enables the *Auto-Encoding Variational Bayes* (AEVB) algorithm [13, 37]. In a typical minibatch update, the AEVB algorithm proceeds as follows:

1. For each data sample $\mathbf{x}$ in the minibatch, the encoder computes the parameters (e.g., mean $\boldsymbol{\mu}_\phi(\mathbf{x})$ and standard deviation $\boldsymbol{\sigma}_\phi(\mathbf{x})$) of the approximate posterior $q_\phi(\mathbf{z}|\mathbf{x})$.
2. A sample $\boldsymbol{\epsilon}$ is drawn from the auxiliary noise distribution $p(\boldsymbol{\epsilon})$ (e.g., $\mathcal{N}(\mathbf{0}, \mathbf{I})$).
3. The latent variable $\mathbf{z}$ is computed using the reparameterization: $\mathbf{z} = g_\phi(\mathbf{x}, \boldsymbol{\epsilon})$ (e.g., $\mathbf{z} = \boldsymbol{\mu}_\phi(\mathbf{x}) + \boldsymbol{\sigma}_\phi(\mathbf{x}) \odot \boldsymbol{\epsilon}$).
4. The decoder reconstructs $\hat{\mathbf{x}}$ from $\mathbf{z}$, and the reconstruction loss term $\log p_\theta(\mathbf{x}|\mathbf{z})$ is computed.
5. The KL divergence $D_{KL}(q_\phi(\mathbf{z}|\mathbf{x}) \,\|\, p(\mathbf{z}))$ is computed (often analytically if both distributions are Gaussian).
6. The ELBO (or its negative) is calculated, and gradient steps are taken with respect to both θ and ϕ.

This end-to-end training procedure is fundamental to the success of VAEs with high-dimensional continuous latent spaces. In practice, networks are often deep convolutional architectures for image data, incorporating techniques like batch normalization and non-linear activation functions (e.g., ReLU, or softplus for standard deviations). Notably, VAE training relies on a single, differentiable objective function and does not require adversarial training procedures, often leading to more stable training dynamics compared to some other classes of generative models.

2.2 Generative Adversarial Networks (GANs)

2.2.1 Game-Theoretic Foundations and Adversarial Training

Generative Adversarial Networks (GANs) introduce a fundamentally different paradigm for generative modeling compared to VAEs, formulating the learning process as a two-player minimax game between two neural networks: a generator G and a discriminator D [10]. The generator network, $G(\mathbf{z}; \theta_G)$, maps samples $\mathbf{z}$ from a simple prior distribution $p(\mathbf{z})$ (e.g., a multivariate Gaussian) to the data space, producing synthetic samples $G(\mathbf{z})$. The discriminator network, $D(\mathbf{x}; \theta_D)$, acts as a binary classifier, tasked with distinguishing between real data samples $\mathbf{x}$ drawn from the true data distribution $p_{\text{data}}(\mathbf{x})$ and these "fake" samples produced by G.

The training objective seeks a Nash equilibrium of this game. The discriminator D is trained to maximize its accuracy in identifying real versus generated samples (assigning high scores to real data and low scores to generated data), while the generator G is simultaneously trained to produce samples that are increasingly realistic, thereby "fooling" the discriminator into classifying them as real. This adversarial dynamic drives both networks to improve their respective capabilities. The classic GAN objective function is expressed as:

$$\min_{\theta_G} \max_{\theta_D} V(D, G) = \mathbb{E}_{\mathbf{x} \sim p_{\text{data}}(\mathbf{x})}[\log D(\mathbf{x}; \theta_D)] + \mathbb{E}_{\mathbf{z} \sim p(\mathbf{z})}[\log(1 - D(G(\mathbf{z}; \theta_G); \theta_D))]$$

Goodfellow et al. [10] demonstrated that for a sufficiently expressive D and G, this adversarial process effectively minimizes the Jensen-Shannon (JS) divergence between the true data distribution $p_{\text{data}}(\mathbf{x})$ and the distribution implicitly defined by the generator $p_G(\mathbf{x})$. In the ideal scenario, with infinite model capacity and an optimal discriminator, the unique equilibrium is achieved when $p_G(\mathbf{x}) = p_{\text{data}}(\mathbf{x})$, at which point the discriminator cannot distinguish between real and generated samples, i.e., $D(\mathbf{x}) \equiv \frac{1}{2}$ everywhere.

GANs are particularly well-suited for high-dimensional data modalities such as images, where both G and D are commonly implemented as deep convolutional neural networks (CNNs). For instance, G might employ a series of transposed convolutions to upsample a low-dimensional latent vector $\mathbf{z}$ into an image, while D typically uses a standard convolutional architecture to classify images. Unlike VAEs, GANs do not explicitly compute a likelihood function $p_G(\mathbf{x})$. Instead, the discriminator provides a rich, adaptive training signal to the generator via gradients backpropagated through D. Furthermore, GANs avoid the need for Markov Chain Monte Carlo (MCMC) methods or the explicit parameterization of an approximate posterior, as all gradients flow deterministically through G and D [10]. The general GAN training procedure is outlined in Algorithm 1. In practice, the number of discriminator updates per generator update can vary.

Algorithm 1 GAN Training (minimax objective)

Require: real data $\{x_i\}$, prior $p(z)$, Generator $G(\cdot;\theta_G)$, Discriminator $D(\cdot;\theta_D)$

1: **for** each training iteration **do**
2: Sample real batch $\{x^{(i)}\} \sim p_{\text{data}}$ and latent batch $\{z^{(i)}\} \sim p(z)$.
3: Update discriminator by ascending its stochastic gradient:
$\nabla_{\theta_D}\big[\frac{1}{m}\sum_i \log D(x^{(i)}) + \log(1 - D(G(z^{(i)})))\big]$.
4: Sample new latent batch $\{z^{(i)}\} \sim p(z)$.
5: Update generator by descending its gradient (or equivalently ascending):
$\nabla_{\theta_G}\big[\frac{1}{m}\sum_i \log(1 - D(G(z^{(i)})))\big]$ (often replaced with $\nabla_{\theta_G} \log D(G(z))$ to improve gradients).
6: **end for**

2.2.2 Training Dynamics and Stability Challenges

Despite their conceptual elegance and empirical successes, the practical training of GANs is fraught with challenges related to stability and convergence [38]. The alternating minimax optimization occurs over a non-convex (for G) and non-concave (for D) objective landscape [39], which can lead to complex dynamics including oscillations, divergence, or limit cycles rather than straightforward convergence. Several pathological behaviors are commonly observed:

- **Mode Collapse:** The generator learns to produce only a limited subset of the modes present in the true data distribution, failing to capture its full diversity. This often occurs if G finds a few samples that maximally fool the current D.
- **Vanishing Gradients:** If the discriminator becomes too proficient and confident in its classifications, the gradients passed to the generator can become vanishingly small, stalling its learning process.
- **Training Instability:** The loss functions of G and D may oscillate without clear signs of convergence, often due to an imbalance in the learning rates or capacities of the two networks.

Insights from control theory have been explored to model and stabilize GAN training dynamics, viewing the process as a dynamical system in function space and designing closed-loop control mechanisms, often via regularization terms that constrain discriminator behavior [40].

To address these challenges, a plethora of techniques have been developed. These include modifications to the loss function (e.g., least-squares GANs, hinge loss), architectural adjustments (e.g., skip connections, careful use of normalization layers), and specialized training heuristics (e.g., one-sided label smoothing, minibatch discrimination [41]).

A significant theoretical and practical breakthrough was the introduction of Wasserstein GANs (WGANs) [18]. Arjovsky et al. proposed replacing the JS divergence with the Earth Mover's (Wasserstein-1) distance, W_1. The WGAN objective is formulated as:

$$\min_{\theta_G} \max_{\theta_D \in \mathcal{W}} \mathbb{E}_{\mathbf{x}\sim p_{\text{data}}(\mathbf{x})}[D(\mathbf{x};\theta_D)] - \mathbb{E}_{\mathbf{z}\sim p(\mathbf{z})}[D(G(\mathbf{z};\theta_G);\theta_D)]$$

where $\mathcal{W}$ denotes the set of 1-Lipschitz functions. The discriminator (termed "critic" in WGANs) aims to estimate the Wasserstein distance. This formulation provides more stable gradients even when the generator and data distributions have disjoint

supports, mitigating mode collapse and improving training stability. The Lipschitz constraint was initially enforced via weight clipping, but this was found to sometimes lead to pathological behavior. Subsequently, Gulrajani et al. proposed WGAN with Gradient Penalty (WGAN-GP), which enforces the Lipschitz constraint by adding a penalty term to the critic's loss based on the norm of its gradients with respect to its inputs [42]. This approach further stabilized training and allowed for deeper architectures.

Other notable stabilization techniques include Two-Time-Scale Update Rules (TTUR), where G and D use different learning rates; spectral normalization, which constrains the Lipschitz constant of network weights and is widely applied to both G and D; and various forms of normalization like batch or instance normalization. More recent developments such as classifier-free guidance and progressive growing of GANs (starting training at low resolution and gradually increasing it) have also contributed to improved convergence and sample fidelity [43]. Figure 2.2 illustrates the general adversarial dynamic. Despite these advancements, comprehensive theoretical guarantees for GAN convergence remain largely elusive, and their understanding often relies on empirical analysis and heuristics.

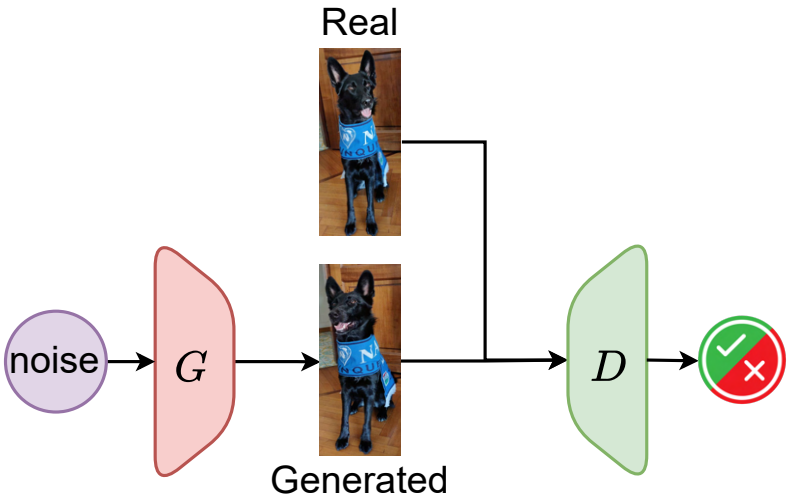


Figure 2.2. Illustration of GAN adversarial training dynamics. The generator G produces images from noise (left) and the discriminator D learns to distinguish them from real images (right). The objective is a minimax game whose equilibrium ideally corresponds to the generator matching the data distribution.

2.2.3 Theoretical Extensions and Variants

The foundational GAN framework has spurred a vast ecosystem of variants and theoretical extensions designed to enhance stability, control generation, improve sample quality, or adapt GANs to specific tasks [44].

Conditional GANs (cGANs): These extend the basic GAN by incorporating auxiliary conditioning information $\mathbf{y}$ (e.g., class labels, text, or other images) into both the generator $G(\mathbf{z}, \mathbf{y})$ and the discriminator $D(\mathbf{x}, \mathbf{y})$ [45]. This allows for controlled generation of samples belonging to specific categories or possessing desired attributes.

Information-Theoretic GANs (InfoGAN): InfoGAN aims to learn disentangled and interpretable latent representations in an unsupervised manner [46]. It

achieves this by maximizing the mutual information between a subset of the latent variables (latent codes c) and the generated samples $G(\mathbf{z}, c)$, effectively encouraging these codes to capture salient semantic features of the data.

Architectural Innovations and High-Resolution Synthesis: Significant improvements in image quality and resolution have been driven by architectural advancements.

- **Progressive GANs (ProGAN):** Karras et al. [43] introduced a training methodology where both generator and discriminator networks are grown progressively, starting with low-resolution images and gradually adding layers to handle finer details. This approach stabilized training for high-resolution synthesis and improved output quality.
- **Self-Attention GANs (SAGAN):** By incorporating self-attention mechanisms, SAGANs enable the generator and discriminator to model long-range dependencies across different image regions, which is crucial for capturing global structures [47].
- **Style-Based GANs (StyleGAN):** StyleGAN architectures [48] introduced novel techniques for controlling the style of generated images at different levels of detail by injecting the latent code $\mathbf{z}$ (mapped through an intermediate learned space W) into multiple layers of the generator via adaptive instance normalization. This often leads to superior disentanglement and perceptual quality.
- **PatchGAN Discriminators:** Used in architectures like Pix2Pix [16], PatchGAN discriminators classify local image patches as real or fake, rather than the entire image. This encourages sharpness and local realism and can be more computationally efficient for high-resolution images.

These extensions and architectural refinements continue to advance the capabilities of GANs, particularly in generating photorealistic images and scenes that often surpass other generative methods in visual fidelity.

2.3 Denoising Diffusion Probabilistic Models (DDPMs)

2.3.1 Theoretical Framework and Non-Equilibrium Thermodynamics

Denoising Diffusion Probabilistic Models (DDPMs) represent a powerful class of deep generative models that draw inspiration from concepts in non-equilibrium thermodynamics [11, 49]. The core idea involves a two-stage process: a fixed forward diffusion process that progressively corrupts data by adding noise over a sequence of timesteps, and a learned reverse denoising process that aims to invert this corruption, transforming noise back into data.

The forward diffusion process is defined as a Markov chain $q(\mathbf{x}_{1:T}|\mathbf{x}_0)$ that gradually perturbs an initial data sample $\mathbf{x}_0 \sim p_{\text{data}}(\mathbf{x})$ over T discrete timesteps. At each timestep t, a small amount of Gaussian noise is added to $\mathbf{x}_{t-1}$ to produce $\mathbf{x}_t$. This transition is typically defined as:

$$q(\mathbf{x}_t|\mathbf{x}_{t-1}) = \mathcal{N}(\mathbf{x}_t; \sqrt{1-\beta_t}\mathbf{x}_{t-1}, \beta_t\mathbf{I})$$

where $\{\beta_t\}_{t=1}^T$ is a predefined variance schedule that controls the magnitude of noise added at each step. These β_t values are usually small, such that the conditional

distribution $q(\mathbf{x}_t|\mathbf{x}_{t-1})$ remains close to a Gaussian centered at $\mathbf{x}_{t-1}$. As $T \to \infty$, and with an appropriate schedule, the distribution of $\mathbf{x}_T$ approaches a standard isotropic Gaussian distribution $\mathcal{N}(\mathbf{0}, \mathbf{I})$, irrespective of the initial data distribution $\mathbf{x}_0$. This process can be intuitively understood as the data distribution "melting" or equilibrating into a simple, tractable noise distribution. An important property is that $\mathbf{x}_t$ can be sampled directly from $\mathbf{x}_0$ for any t using $\mathbf{x}_t = \sqrt{\bar{\alpha}_t}\mathbf{x}_0 + \sqrt{1-\bar{\alpha}_t}\boldsymbol{\epsilon}$, where $\alpha_t = 1 - \beta_t$, $\bar{\alpha}_t = \prod_{s=1}^{t} \alpha_s$, and $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$.

The generative power of DDPMs lies in the reverse process,

$$p_\theta(\mathbf{x}_{0:T}) = p(\mathbf{x}_T) \prod_{t=1}^{T} p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t),$$

which learns to invert the diffusion. Starting from a sample $\mathbf{x}_T \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$, the reverse process iteratively denoises it to produce a sample $\mathbf{x}_0$ that resembles the true data. Each reverse transition $p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t)$ is parameterized by a neural network, typically a U-Net architecture denoted $\boldsymbol{\epsilon}_\theta(\mathbf{x}_t, t)$, which is trained to predict the noise component $\boldsymbol{\epsilon}$ that was added to obtain $\mathbf{x}_t$ from $\mathbf{x}_{t-1}$ (or, equivalently, to predict $\mathbf{x}_{t-1}$ or $\mathbf{x}_0$).

This iterative refinement allows the model to capture complex data distributions. The overall diffusion and denoising process is illustrated in Figure 2.3.

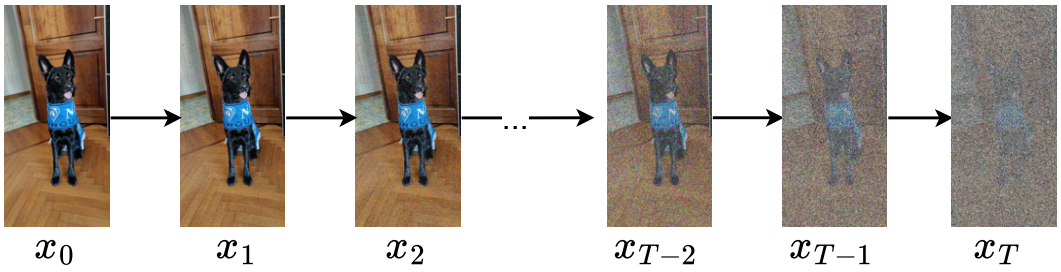


Figure 2.3. Illustration of the diffusion process: a data sample $\mathbf{x}_0$ is gradually corrupted by Gaussian noise to $\mathbf{x}_T$ (forward process), and a neural network learns the reverse denoising (reverse process). Each step of the reverse process resembles denoising autoencoding conditioned on the timestep.

The training of the denoiser $\boldsymbol{\epsilon}_\theta(\mathbf{x}_t, t)$ is closely related to score matching [14]. The score function is defined as the gradient of the log probability density, $\nabla_{\mathbf{x}_t} \log p(\mathbf{x}_t)$. It can be shown that predicting the noise $\boldsymbol{\epsilon}$ in the DDPM framework is equivalent to estimating the score function of the noisy data distribution at each noise level t.

2.3.2 Variational Lower Bound and Training Objectives

The training of DDPMs is guided by optimizing a variational lower bound (VLB) on the negative log-likelihood $-\log p_\theta(\mathbf{x}_0)$, analogous to VAEs [11]. The ELBO for diffusion models can be expressed as:

$$\mathcal{L}_{VLB} = \mathbb{E}_q \left[\log p_\theta(\mathbf{x}_0|\mathbf{x}_1) + \sum_{t=2}^{T} \log \frac{p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t)}{q(\mathbf{x}_{t-1}|\mathbf{x}_t, \mathbf{x}_0)} - \log q(\mathbf{x}_T|\mathbf{x}_0) \right]$$

This can be rewritten as a sum of terms:

$$\mathcal{L}_{VLB} = \mathbb{E}_q \left[D_{KL}(q(\mathbf{x}_T|\mathbf{x}_0) \,\|\, p_\theta(\mathbf{x}_T)) + \sum_{t=2}^{T} D_{KL}(q(\mathbf{x}_{t-1}|\mathbf{x}_t, \mathbf{x}_0) \,\|\, p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t)) - \log p_\theta(\mathbf{x}_0|\mathbf{x}_1) \right] \quad (2.1)$$

Ho et al. [11] showed that by parameterizing

$$p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t) = \mathcal{N}(\mathbf{x}_{t-1}; \boldsymbol{\mu}_\theta(\mathbf{x}_t, t), \boldsymbol{\Sigma}_\theta(\mathbf{x}_t, t))$$

where $\boldsymbol{\mu}_\theta(\mathbf{x}_t, t)$ depends on the noise prediction $\boldsymbol{\epsilon}_\theta(\mathbf{x}_t, t)$ and fixing the variance $\boldsymbol{\Sigma}_\theta(\mathbf{x}_t, t)$ to be time-dependent constants (e.g., $\sigma_t^2\mathbf{I} = \beta_t\mathbf{I}$), most terms in the VLB can be computed analytically.

This leads to a simplified objective function which predominantly involves training the network $\boldsymbol{\epsilon}_\theta(\mathbf{x}_t, t)$ to predict the noise $\boldsymbol{\epsilon}$ added at timestep t:

$$\mathcal{L}_{\text{simple}}(\theta) = \mathbb{E}_{t\sim\mathcal{U}(1,T),\mathbf{x}_0\sim p_{\text{data}}(\mathbf{x}_0),\boldsymbol{\epsilon}\sim\mathcal{N}(\mathbf{0},\mathbf{I})}\left[\|\boldsymbol{\epsilon} - \boldsymbol{\epsilon}_\theta(\sqrt{\bar{\alpha}_t}\mathbf{x}_0 + \sqrt{1-\bar{\alpha}_t}\boldsymbol{\epsilon}, t)\|^2\right]$$

This objective is a weighted sum of denoising score matching terms at different noise levels. Optimizing this simple mean squared error (MSE) loss has been empirically found to yield excellent results and is computationally efficient. Kingma et al. [50] provided further analysis, showing connections to continuous-time diffusion processes and demonstrating that the VLB can be nearly invariant to the noise schedule in continuous time, allowing for schedule optimization.

DDPMs, particularly with U-Net based architectures incorporating ResNet blocks and self-attention mechanisms, scale effectively to high-resolution image synthesis and have achieved sample quality competitive with or exceeding state-of-the-art GANs on many benchmarks [11]. Unlike typical GANs, DDPMs are likelihood-based models, allowing for density estimation and applications like anomaly detection. A primary drawback has historically been the slow sampling process, as generating a single sample requires sequentially evaluating the model T times (often $T = 1000$ or more). However, significant research has focused on accelerating sampling through techniques such as Denoising Diffusion Implicit Models (DDIMs) [51], knowledge distillation, and optimized SDE solvers. A general DDPM sampling procedure is outlined in Algorithm 2.

Algorithm 2 DDPM Sampling (reverse diffusion)

1: $x_T \sim \mathcal{N}(0, I)$
2: **for** $t = T, T-1, \ldots, 1$ **do**
3: $\quad\epsilon_\theta = \epsilon_\theta(x_t, t)$ (predict noise)
4: $\quad\tilde{\mu} = \frac{1}{\sqrt{\alpha_t}}\left(x_t - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}}\epsilon_\theta\right)$
5: $\quad x_{t-1} = \tilde{\mu} + \sigma_t z$, where $z \sim \mathcal{N}(0, I)$ and σ_t is the chosen variance.
6: **end for**
7: **return** x_0 as a generated sample.

2.3.3 Discrete State Spaces and Advanced Formulations

Further theoretical advancements and alternative formulations have enriched the understanding and capabilities of diffusion models. These include:

- **Stochastic Differential Equations (SDEs):** [14] formulated diffusion processes in continuous time using SDEs, providing a unified view that encompasses various discrete-time DDPMs and score-based models. This perspective allows for more flexible noise schedules and sampling procedures.

- **Schrödinger Bridges and Optimal Transport:** Connections have been drawn between diffusion models and the Schrödinger bridge problem, which seeks the most likely stochastic evolution between two distributions. This links diffusion models to concepts in optimal transport theory [52].
- **Flow Matching:** More recent work on flow matching (see Section 2.5) generalizes diffusion models by directly learning a vector field (a velocity field) that transports samples from a noise distribution to the data distribution, offering alternative training objectives and potentially more efficient sampling.

In summary, diffusion models provide a robust and theoretically grounded framework for generative modeling, characterized by iterative refinement. Their applicability to both continuous and discrete data, coupled with ongoing research into advanced formulations and faster sampling, positions them as highly versatile and powerful tools in the landscape of modern generative AI.

2.4 Latent Diffusion Models (LDMs)

2.4.1 Latent Space Representation and Computational Efficiency

Latent Diffusion Models (LDMs) represent a significant advancement in diffusion-based generative modeling, primarily addressing the substantial computational demands associated with applying diffusion processes directly to high-dimensional data such as images [12]. The core innovation of LDMs is to perform the iterative diffusion and denoising process within a lower-dimensional learned latent space, rather than in the pixel space. This strategy leverages the representational power of autoencoders to compress data into compact latent representations, where the diffusion process can be executed much more efficiently, thereby reducing computational costs and memory requirements while preserving, and often enhancing, the quality of generated samples.

The LDM framework typically comprises three main components [12]:

1. An **encoder** (E): This network maps input data $\mathbf{x}$ from the high-dimensional pixel space to a lower-dimensional latent representation $\mathbf{z}_0 = E(\mathbf{x})$.
2. A **latent diffusion model** (p_θ): This model operates entirely within the learned latent space. It defines a forward diffusion process $q(\mathbf{z}_{1:T}|\mathbf{z}_0)$ that progressively adds noise to $\mathbf{z}_0$ over T timesteps, and a corresponding learned reverse denoising process $p_\theta(\mathbf{z}_{0:T})$ that transforms a noise vector $\mathbf{z}_T \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ back into a clean latent representation $\hat{\mathbf{z}}_0$.
3. A **decoder** (D): This network reconstructs the final data sample $\hat{\mathbf{x}} = D(\hat{\mathbf{z}}_0)$ from the denoised latent representation $\hat{\mathbf{z}}_0$.

The encoder and decoder components are often realized as parts of a pre-trained Variational Autoencoder (VAE) or a similar autoencoding architecture, which is trained to achieve good reconstruction quality and provide a semantically rich and stable latent space. This two-stage approach, first compression, then latent-space diffusion, allows the diffusion model to focus on the essential semantic and structural aspects of the data, while the decoder handles the reconstruction of fine-grained details.

By operating in this compressed latent space (e.g., with spatial resolutions 8x or 16x smaller than the original image), the number of neural operations per

denoising step is drastically reduced. This leads to significant gains in sampling speed and memory efficiency, enabling high-resolution image synthesis with orders of magnitude less compute compared to pixel-space diffusion models [12]. For example, the influential Stable Diffusion model is a notable realization of an LDM, employing a VAE for perceptual compression and a U-Net architecture for the latent diffusion process. This approach strikes an effective balance between computational tractability and the preservation of visual fidelity, achieving state-of-the-art results in image generation.

2.4.2 Cross-Attention Mechanisms and Conditional Generation

A key strength of LDMs is their remarkable aptitude for conditional generation, facilitated by the integration of sophisticated attention mechanisms, particularly cross-attention, into the U-Net architecture of the latent diffusion model [12]. These mechanisms allow the denoising process to be guided by various forms of conditioning information $\mathbf{c}$, such as text descriptions, class labels, semantic segmentation maps, or other image modalities.

The integration of transformer architectures and their attention mechanisms within the LDM framework enables complex and fine-grained conditioning schemes. In text-to-image synthesis, for instance, textual prompts are typically encoded into semantic embeddings using powerful pre-trained language models (e.g., CLIP). These embeddings are then injected into the latent U-Net via cross-attention layers at multiple stages of the denoising network. This allows the model to selectively attend to relevant parts of the conditioning input at each generation step, effectively aligning the generated latent features with the semantic concepts conveyed by the condition. The result is a highly flexible generative framework capable of producing diverse outputs that accurately reflect the conditioning information. This has enabled a wide array of applications, including text-guided image synthesis, image editing (e.g., inpainting and outpainting), style mixing, and super-resolution, all achievable by providing different types of context inputs.

Furthermore, the cross-attention mechanism in LDMs can operate at multiple scales within the U-Net architecture. Lower-resolution feature maps in the U-Net can attend to conditioning information to guide global layout and composition, while higher-resolution feature maps can focus on finer details and textures in alignment with the condition. This hierarchical conditioning has been instrumental in achieving coherent and high-fidelity synthesis, particularly for high-resolution text-to-image generation tasks, as exemplified by models like Stable Diffusion and Imagen [12, 53]. The theoretical underpinning for conditional LDMs extends the variational lower bound (ELBO) framework to incorporate the conditioning variable $\mathbf{c}$, while preserving the core iterative refinement properties of diffusion models.

2.4.3 Architectural Innovations and Scaling

The continued evolution of LDMs has been marked by significant architectural innovations and advancements in scaling, further enhancing generation quality and applicability to diverse tasks and datasets. The U-Net architecture, which forms the backbone of the latent denoising model, typically incorporates design elements from state-of-the-art vision research, including residual convolutional blocks, group or layer normalization, and self-attention modules, particularly at mid-to-coarse feature resolutions, to capture long-range spatial dependencies. Transformer-style blocks and multi-head attention mechanisms have also been effectively adapted for operation on convolutional feature maps within these U-Nets. Techniques originally

developed for stabilizing GAN training, such as spectral normalization or gradient checkpointing, are also sometimes employed to manage the training of large U-Net models.

A major advantage of the latent-space formulation is its resource efficiency, which facilitates scaling LDMs to larger model sizes and training them on massive datasets (e.g., containing millions or billions of image-text pairs) [12]. Since the computationally intensive diffusion process occurs in a significantly smaller latent space, the channel width or depth of the U-Net can be increased substantially before encountering memory or computational bottlenecks. This has enabled the development of LDMs with billions of parameters that can still generate samples within seconds on modern GPUs, often outperforming pixel-space diffusion models of similar parameter counts due to this efficiency.

Several advanced techniques have emerged to further improve LDM performance and capabilities:

- **Classifier-Free Guidance (CFG):** This technique has become a standard for enhancing adherence to conditioning information (e.g., text prompts) while maintaining sample diversity. During inference, CFG involves performing two forward passes of the denoising network, one conditioned and one unconditioned (or conditionally dropped out), and then extrapolating in the direction of the conditioned prediction.
- **Cascaded and Multi-Stage Architectures:** To achieve ultra-high-resolution synthesis (e.g., 4K resolution and beyond), cascaded LDM systems or layered denoising approaches are employed. These typically involve an initial LDM generating a lower-resolution latent representation, followed by one or more stages of super-resolution diffusion models (which may also be LDMs) that progressively refine the output to higher resolutions [12].
- **Specialized Fine-tuning:** Techniques such as wavelet-based fine-tuning have been explored to better preserve high-frequency details during generation, particularly for very high-resolution outputs, while aiming to maintain computational efficiency [9].

These architectural and methodological innovations underscore the adaptability and scalability of LDMs, positioning them as a cornerstone technology in modern generative AI, particularly for complex, controllable, and high-resolution image synthesis scenarios.

2.5 Flow Matching Models

Flow Matching represents a significant advancement in generative modeling that unifies and extends both normalizing flow and diffusion model methodologies. This framework provides a simulation-free approach to training Continuous Normalizing Flows (CNFs) while maintaining theoretical rigor and enabling novel probability path designs that improve both training efficiency and sample quality.

2.5.1 Continuous Normalizing Flows and Simulation-Free Training

Flow matching models generalize diffusion and flow approaches by directly learning continuous-time flows without requiring expensive simulation procedures

during training [15]. A continuous normalizing flow (CNF) is defined by an ordinary differential equation (ODE)

$$\frac{dz}{dt} = v_\theta(z, t)$$

that transports a noise distribution to the data distribution. Traditional CNF training employs the instantaneous change-of-variables formula through neural ODEs, necessitating the solution of ODEs during the training process, which is computationally expensive and limits scalability.

Flow matching offers a revolutionary simulation-free alternative that eliminates these computational bottlenecks while preserving the theoretical guarantees of flow-based models. Instead of backpropagating through an ODE solver, the method directly regresses the vector field $v_\theta(z, t)$ to match a specified probability path. The theoretical foundation rests upon the observation that CNFs can be trained by learning the vector field that transforms a simple base distribution into the target data distribution without explicit simulation.

The flow matching objective is formulated as minimizing the expected squared error between the model vector field and the true conditional vector field:

$$\mathcal{L}_{FM}(\theta) = \mathbb{E}_{t,\mathbf{x}_1,\mathbf{x}_0}\left[\|v_\theta(t, \mathbf{x}_t) - u_t(\mathbf{x}_t|\mathbf{x}_1)\|^2\right]$$

where $\mathbf{x}_t$ is sampled from the conditional probability path between noise $\mathbf{x}_0$ and data $\mathbf{x}_1$.

In this framework, one specifies a conditional path $\pi_t(x|z)$ from $z \sim p(z)$ (simple base distribution) to $x \sim p_{\text{data}}(x)$. For diffusion paths, π_t follows a Gaussian conditional structure; for optimal transport paths, π_t interpolates deterministically. Training minimizes the expected squared difference between v_θ and the true velocity field that transports $p(z)$ into p_{data} along π_t. Crucially, this yields a simple regression loss without divergence terms and accommodates any convenient path choice.

Flow matching subsumes diffusion models as a special case when employing diffusion paths, but also enables novel path designs such as optimal transport geodesics. The method demonstrates superior stability compared to traditional diffusion training while allowing for alternative paths that yield faster training convergence and more efficient sampling procedures.

2.5.2 Optimal Transport and Efficient Probability Paths

A key theoretical insight of flow matching is the flexibility to choose probability paths that connect base and target distributions, with optimal transport (OT) providing particularly advantageous path designs. In the OT interpretation, one seeks the most efficient path to transform one distribution into another, minimizing the expected transport cost and resulting in more direct trajectories.

Flow matching demonstrates how to parameterize OT paths using neural ODEs, specifically by choosing π_t as linear interpolation of quantile maps to yield displacement interpolation paths. The mathematical formulation leverages established results from optimal transport theory, where the displacement interpolation can be computed efficiently for Gaussian base distributions, providing a practical alternative to the complex stochastic paths typically employed in diffusion models.

The integration of optimal transport concepts satisfies a form of the Benamou–Brenier formulation, effectively solving geodesic transport in probability space using neural networks. This connection opens new theoretical directions and suggests flows that mirror natural transformations in the data domain. For vision

tasks, this translates to flows that smoothly interpolate object features and geometric structures.

Empirically, OT-based flow matching models demonstrate significant advantages: the learned ODEs require fewer solver steps and admit larger timesteps due to simpler and more direct velocity fields compared to diffusion-based approaches. This leads to faster sampling procedures while maintaining or improving sample quality. The straight-line nature of optimal transport paths provides computational efficiency and robust generalization properties.

The flexibility of flow matching also enables hybrid approaches that interpolate between diffusion and OT paths, or combine multiple path types, creating versatile frameworks that can be tailored to specific problem domains and computational constraints.

2.5.3 Applications and Scaling Properties

Flow-based models and CNFs possess unique advantages for computer vision and other generative modeling applications due to their likelihood-based formulation. Unlike GANs, they enable exact or near-exact density evaluation, making them suitable for tasks such as image compression through bits-back coding, anomaly detection, and probabilistic modeling where likelihood estimates are crucial.

Flow matching models have demonstrated state-of-the-art performance on large-scale image datasets including FFHQ and ImageNet, achieving superior log-likelihood scores and competitive FID metrics compared to diffusion models of similar architectural complexity [15]. The method has shown particular effectiveness in domains requiring high-quality continuous generation, including image synthesis, audio generation, and scientific modeling applications.

In terms of computational efficiency, while traditional flows required numerical ODE integration at inference time, flow matching often yields closed-form solutions or very simple ODEs, particularly for optimal transport paths. This enables fast one-shot sampling or requires minimal integration steps, providing significant speed advantages over traditional diffusion approaches while maintaining generation quality.

Architecturally, CNFs and flow matching models typically employ similar UNet backbones as diffusion models but replace time-embedding and diffusion-specific components with continuous vector field networks. The models can incorporate conditional information through concatenation or affine coupling mechanisms for tasks such as image-to-image translation and text-guided generation [54].

The scaling properties are particularly promising: flow matching models can accommodate arbitrarily large image sizes through multi-scale architectural designs and multiresolution flow approaches. Recent developments explore combining flow matching with compression techniques and downsampling flows to handle megapixel images efficiently. The cascaded architectures enabled by flow matching allow for sophisticated multi-scale generation procedures that maintain efficiency while scaling to high-resolution outputs.

The simulation-free training paradigm addresses fundamental computational bottlenecks that have historically limited the practical application of flow-based models, enabling large-scale generative modeling that was previously computationally prohibitive. The theoretical properties of flow matching, particularly the unified treatment of continuous and discrete data, position these models as versatile tools for diverse generative modeling challenges.

2.5.4 Comparative Analysis with Other Model Families

Flow-based models, including those trained via Flow Matching (FM), offer a distinctive set of advantages and limitations when compared to variational autoencoders (VAEs), generative adversarial networks (GANs), and diffusion models. One of their primary strengths lies in their ability to compute exact log-likelihoods. This property provides a rigorous theoretical framework for both model evaluation and direct likelihood-based optimization, which stands in contrast to the approximate objectives used in VAEs and the implicit distributions of GANs.

Table 2.1. Comparative Overview of Major Generative Model Families

Model Family	Core Principle	Key Features	Strengths	Challenges
VAEs	Probabilistic latent modeling Maximize ELBO	Encoder-Decoder Reparameterization trick	Stable training Smooth latent space Principled framework	Often blurry samples Restrictive priors
GANs	Minimax game between Generator and Discriminator	Generator and Discriminator Networks	High perceptual quality Sharp, high-fidelity images	Training instability Mode collapse Difficult to evaluate
DDPMs	Iteratively reverse a fixed noise (diffusion) process	U-Net for noise/score prediction Learned reverse Markov chain	SOTA sample quality High diversity Stable training	Slow and costly sampling Computationally intensive
Flow Matching	Learn a continuous velocity field to map noise to data Simulation-free training via regression	Continuous velocity field ODE-based formulation Paired with DiT backbones	Robust and scalable training Eliminates fixed noise schedules Architecturally flexible	Performance is dependent on the quality of the VAE latent space

A further advantage is the explicitly defined, invertible mapping between latent and data spaces. This not only facilitates tractable inference but also ensures that generation and reconstruction are inherently symmetric processes, something not guaranteed in VAEs or GANs. Additionally, the training process for flow-based models is often more stable, as it avoids the adversarial dynamics that can lead to mode collapse or instability in GANs.

Flow Matching introduces specific enhancements to this paradigm. By reframing generative modeling as a regression problem between vector fields, FM enables simulation-free training of continuous normalizing flows (CNFs). It has been shown to offer improved training stability over traditional score matching in diffusion-based methods and opens avenues for efficient sampling and better generalization, especially when optimal transport-inspired paths are employed.

Nonetheless, flow-based models are not without their challenges. Traditional normalizing flows, in particular, impose strict architectural constraints due to the need for invertible transformations with efficiently computable Jacobian determinants. These constraints can limit the expressiveness of the underlying neural networks, especially when compared to the flexible designs permissible in GANs or diffusion models. Moreover, computing Jacobian determinants for high-dimensional data can be computationally demanding, although this is often mitigated through carefully

designed structures like coupling layers or autoregressive transformations.

Historically, flow-based models have also lagged behind GANs in terms of perceptual quality. While they provide elegant probabilistic foundations, their samples were often less sharp or realistic. However, recent innovations, including Flow Matching, are rapidly narrowing this gap, making flow-based models increasingly competitive in high-fidelity image synthesis.

Taken together, flow-based generative models, particularly those leveraging Flow Matching, represent an evolving and influential direction in the landscape of deep generative modeling. They offer a compelling balance between theoretical rigor and practical performance, bridging the gap between the statistical clarity of likelihood-based approaches and the empirical success of adversarial and diffusion-based models.

Table 2.1 summarizes these comparisons, presenting a high-level overview of the fundamental trade-offs among major generative model families. This comparative perspective is crucial for understanding both the historical context and the future potential of flow-based methods. It reinforces the idea that no single framework universally dominates; rather, the choice of model depends critically on the application's demands, resource constraints, and the desired trade-off between sample quality, training stability, and interpretability.

Chapter 3

Generative Models for Images

Contents

3.1 Introduction to Generative Models for Image Synthesis 26
3.1.1 The Quest for Synthetic Photorealism and Diversity . . . 27
3.1.2 Fundamental Properties of Image Data 27
3.1.3 Core Objectives and Applications of Generative Image Models 29
3.2 Challenges in Image Synthesis 30
3.3 Core Applications of Generative Image Synthesis 31
3.3.1 High-Resolution Image Synthesis and Super-Resolution . 31
3.3.2 Image-to-Image Translation 32
3.3.3 Text-to-Image Synthesis 32
3.3.4 Image Inpainting 33
3.4 Evaluation Paradigms for Generative Image Models . . 33
3.4.1 The Intricacies of Evaluating Generated Images 33
3.4.2 Reconstruction Metrics 34
3.4.3 Image Generation Metrics 35
3.4.4 Emerging Semantic Metrics 35
3.4.5 Human Preference Metrics 36

This chapter establishes the theoretical and practical foundations of generative models for image synthesis, analyzing their unique requirements, technical challenges, and evaluation paradigms. The explosive growth of image generation capabilities in recent years has been driven by fundamental advances in deep learning architectures and training methodologies, yet significant challenges remain in achieving photorealistic quality, precise controllability, and robust evaluation.

3.1 Introduction to Generative Models for Image Synthesis

The field of generative modeling has witnessed remarkable advancements, particularly in the domain of image synthesis. These models endeavor to learn the underlying patterns and structures within vast datasets of images to generate novel, realistic visual content. This introductory section outlines the fundamental motivations, historical context, core objectives, and the overall structure of this chapter, which aims to provide a thorough grounding in generative models for images.

3.1.1 The Quest for Synthetic Photorealism and Diversity

At its core, generative modeling for images seeks to create algorithms that can learn the statistical distribution of a given set of training images, typically natural images, and subsequently synthesize new samples that are perceptually indistinguishable from this training set [55]. This endeavor is not merely about replicating existing images but about capturing the essence of the visual world, its textures, structures, and semantic content, to produce entirely novel instances. The ambition extends beyond simple mimicry; a crucial goal is the generation of diverse samples that reflect the vast richness and variability inherent in visual data [56].

Achieving true photorealism necessitates grappling with the immense complexity of high-dimensional image data, where subtle nuances in lighting, texture, and geometric arrangement collectively contribute to a convincing visual experience. The challenge lies in developing models that are expressive enough to capture these intricate details and the long-range dependencies that define coherent scenes, while also being trainable with finite data and computational resources. The continuous negotiation between model expressivity, training tractability, and the ability to capture the complex, high-dimensional, and perceptually nuanced nature of visual data has shaped the evolution of generative models for images.

3.1.2 Fundamental Properties of Image Data

Understanding the inherent characteristics of image data is paramount for developing effective generative models. Natural images are far from random assortments of pixel values; they possess rich statistical regularities, complex structures, and are processed by the human visual system in specific ways that must be considered in model design.

Natural Image Statistics and Spectral Properties

Natural images exhibit strong statistical regularities that distinguish them from arbitrary pixel configurations. A cornerstone of these regularities is the $1/f^{\alpha}$ power law, which describes the average behavior of the spatial frequency content in natural scenes [57]. This law states that the average power spectrum, $S(f)$, of natural images tends to fall off with spatial frequency f according to $S(f) \propto 1/f^{\alpha}$, where the exponent α is typically close to 2 [57]. This empirical observation implies a scale-invariant structure in natural images, with significant energy concentrated in the lower spatial frequencies, corresponding to coarser structures, and progressively less energy in higher frequencies, corresponding to finer details.

The human visual system (HVS) appears to be highly adapted to these statistical properties. Images that conform to the $1/f^2$ power spectrum are generally perceived as more natural and are processed more efficiently by the HVS [57]. This adaptation has profound implications for generative modeling: for synthetic images to appear realistic, they should ideally mimic these natural statistics. Beyond second-order statistics, higher-order statistics are essential for capturing more complex image features. The distribution of image gradients in natural scenes is typically heavy-tailed, meaning that while most gradient magnitudes are small (corresponding to smooth regions), there is a non-negligible probability of encountering very large gradient magnitudes (corresponding to sharp edges or strong textures). This property, often referred to as sparsity in the gradient domain, is exploited in various image processing tasks and must be learned by generative models to achieve photorealism.

High-Dimensionality and the Manifold Hypothesis

Image data is inherently high-dimensional, creating significant challenges for generative modeling. A modest 256×256 pixel RGB image resides in a space of 196,608 dimensions, while a 1024×1024 RGB image has over 3 million dimensions. Modeling probability distributions directly in such astronomically high-dimensional spaces presents the formidable challenge known as the "curse of dimensionality" [58]. This principle dictates that the volume of the space increases so rapidly with dimension that available data becomes sparsely distributed, making it exceedingly difficult to estimate densities without exponentially increasing amounts of data and computational power.

However, natural images, despite their high ambient dimensionality, are believed to occupy or lie close to a low-dimensional manifold embedded within the high-dimensional ambient space [59]. This "manifold hypothesis" posits that the intrinsic degrees of freedom of natural images are substantially lower than their raw pixel-level dimensionality. For instance, the set of all possible images of human faces is constrained by underlying factors like identity, pose, expression, and illumination, which likely form a lower-dimensional structure.

This perspective has enabled the success of latent variable models such as VAEs and GANs, which learn mappings from lower-dimensional latent spaces to the high-dimensional image space [60]. Prominent approaches like Latent Diffusion Models (LDMs) operate in compressed latent spaces, typically learned by Variational Autoencoders, to improve computational efficiency [12]. However, this compression can sometimes result in the loss of fine-grained semantic detail from the original image.

Spatial Structure and Hierarchical Organization

Images are characterized by significant spatial structure that generative models must effectively capture. Natural images exhibit substantial spatial redundancy, with nearby pixels tending to be highly correlated [61]. This local correlation is what Convolutional Neural Networks (CNNs) exploit through their use of convolutional filters operating on local receptive fields, making them inherently well-suited for image generation tasks.

Visual information in images is organized hierarchically across multiple scales; low-level features (edges, simple textures) aggregate to form mid-level features (object parts, complex textures), which compose high-level semantic structures (whole objects, scenes). This hierarchical nature has led to the development of progressive growing architectures in GANs [43, 62] and multi-resolution diffusion processes that explicitly model scale hierarchies [62].

While CNNs excel at modeling local interactions, capturing long-range dependencies between spatially distant pixels remains challenging due to limited effective receptive fields [63]. These long-range dependencies are crucial for ensuring global coherence, such as maintaining consistent illumination across scenes or generating plausible completions for large occluded areas. This challenge is particularly pronounced in ultra-high-resolution (UHR) image synthesis, where scaling models trained on lower-resolution data can lead to artifacts such as repeated structures or spatial inconsistencies [64, 65].

Perceptual Considerations and Color Representation

Human perception of image quality is not based on simple pixel-wise comparisons but relies on complex interactions of color, texture, and structural organization.

Standard RGB color representations are convenient but not perceptually uniform, meaning numerical differences don't correspond to equal perceived differences. Perceptually uniform color spaces like CIELAB separate lightness from chromaticity components, which can be advantageous for models that need to treat luminance and color information distinctly. Advanced models often perform color transformations or operate in hybrid spaces to better align with human vision characteristics [66–68].

The non-Euclidean nature of image manifolds becomes particularly apparent in high-dimensional spaces, where traditional distance metrics like the L_2 norm often fail to capture perceptual similarity. This has necessitated the development of specialized loss functions and evaluation metrics [69, 70] such as the Structural Similarity Index (SSIM), perceptual loss functions, and metrics like LPIPS that aim to quantify perceptual quality and align with human judgments [71].

3.1.3 Core Objectives and Applications of Generative Image Models

The primary objective of a generative image model is to learn an implicit or explicit representation of the probability density function, $p_{\text{data}}(x)$, over a set of real-world images x. Once this distribution is learned, the model, typically through its generator component G, can synthesize novel samples $x' \sim p_g(x)$, where $p_g(x)$ is the learned generator distribution that approximates $p_{\text{data}}(x)$ [55]. The goal is for $p_g(x)$ to be as close as possible to $p_{\text{data}}(x)$, such that generated samples are both realistic and diverse.

Beyond the fundamental task of synthesizing novel data samples, generative image models have found a wide array of applications, underscoring their utility and versatility. These applications include, but are not limited to:

- **Semantic Image Editing**: Modifying specific attributes of an image in a meaningful way, such as changing facial expressions, age, or adding/removing objects, often by manipulating the learned latent space of the model [55].
- **Data Augmentation**: Generating synthetic data to expand training sets for other machine learning models, which can be particularly useful in scenarios where labeled data is scarce [55].
- **Style Transfer**: Applying the artistic style of one image to the content of another [55].
- **Image Super-Resolution (SR)**: Reconstructing a high-resolution (HR) image from a low-resolution (LR) input. Generative models, especially GANs and diffusion models, have shown great promise in producing perceptually convincing SR results with fine texture details [71].
- **Image Inpainting**: Filling in missing or corrupted regions of an image in a visually plausible manner [72].
- **Representation Learning**: The features learned by generative models, particularly within their discriminators or encoders, can serve as powerful representations for downstream discriminative tasks such as image classification and retrieval [55].
- **Creative Content Generation**: Assisting artists and designers in creating novel visual content, exploring new aesthetics, and automating certain aspects of the creative process.

The expanding range of these applications signifies a notable transition for generative image models, moving from primarily academic explorations of generative capability to becoming powerful tools with considerable real-world impact. This shift inherently drives research focus towards aspects beyond raw generative power, such as controllability, computational efficiency, interpretability, and the ethical considerations surrounding the generation and use of synthetic media [48]. As these models become more integrated into various workflows and products, the demand for robust, reliable, and deployable solutions intensifies, creating a dynamic feedback loop where practical application needs increasingly shape research priorities in the field. The heavy-tailed distribution of natural image patches, with frequent smooth regions punctuated by rare edge features, creates imbalanced learning challenges that affect both model training and sampling strategies [73]. This progression from simpler statistical models to sophisticated deep neural networks illustrates a persistent drive to bridge the gap between algorithmic generation and the richness of human visual perception.

3.2 Challenges in Image Synthesis

Generating high-quality images that are realistic, diverse, and controllable poses a unique set of requirements and technical challenges. These stem from the high dimensionality of image data, the complexity of natural image statistics, the nuances of human visual perception, and the computational demands of sophisticated generative models. This section explores these critical aspects.

A primary challenge is achieving photorealism, where generated images are indistinguishable from real photographs. This demands high fidelity, meaning the images must be sharp and free of artifacts, while also exhibiting realistic textures and fine details, from the roughness of stone to individual strands of hair. Traditional pixel-wise loss functions, such as Mean Squared Error (MSE), often fail here, promoting overly smooth or blurry results by averaging plausible high-frequency details [66]. To overcome this, the field has shifted towards perceptual loss functions, which measure similarity in a feature space learned by deep neural networks rather than in pixel space. By minimizing the feature-space distance between a generated and a real image, these losses encourage perceptual similarity in content and style, as exemplified by the SRGAN model and the Learned Perceptual Image Patch Similarity (LPIPS) metric [66, 71, 74].

However, photorealism is not just a matter of local texture; it also depends on the global coherence and structural integrity of the image. Modeling long-range dependencies, the relationships between distant pixels, is crucial but difficult for standard CNNs with their limited receptive fields. To address this, advanced models incorporate mechanisms like self-attention and Transformers, which provide a global receptive field, or Fast Fourier Convolutions (FFCs) that efficiently capture global context [63, 75]. Furthermore, realism hinges on coherently synthesizing features across multiple scales, from coarse layouts to fine-grained details. Architectures like the U-Net are highly effective, using skip connections to merge deep, semantic information with shallow, spatial details [16, 76]. Similarly, progressive generation techniques build up images from low to high resolution, allowing the model to learn coarse structures first before refining details.

Beyond realism, a key objective is controllability, the ability to direct the synthesis process. This is typically achieved through conditional generation, where the model is guided by auxiliary information. For instance, Conditional GANs (cGANs) feed this information to both the generator and discriminator to ensure the output matches a

given condition [16]. Likewise, conditional diffusion models integrate external data into the denoising process, often via cross-attention, to steer the generation [12]. An alternative paradigm for control is offered by models like StyleGAN, which enables intuitive manipulation of attributes such as pose and identity by operating on a disentangled latent space [48].

3.3 Core Applications of Generative Image Synthesis

3.3.1 High-Resolution Image Synthesis and Super-Resolution

Generating images at high resolutions (e.g., 1024×1024 pixels and beyond) presents substantial computational and memory challenges [77]. As image resolution increases, the number of pixels grows quadratically, leading to a dramatic increase in the dimensionality of the data and the computational load for generative models. A central task within high-resolution synthesis is Image Super-Resolution (SR), which aims to restore high-resolution (HR) images from their low-resolution (LR) counterparts. The task is inherently ill-posed, as a single LR input can correspond to multiple valid HR solutions, making the reconstruction highly ambiguous [78, 79]. The degradation process is often modeled as:

$$\tilde{\mathbf{x}} = (\mathbf{x} \otimes k) + n, \tag{3.1}$$

where $\tilde{\mathbf{x}}$ is the LR image, $\mathbf{x}$ is the HR image, k represents the degradation kernel, $\otimes$ is a convolution-like operation, and n is a noise term.

Deep Learning-Based Approaches

The advent of deep learning marked a turning point for SR. Early pioneering models like SRCNN [79–81] applied Convolutional Neural Networks (CNNs) to the task. Subsequent architectures such as EDSR [82] and RCAN [83] demonstrated the power of deeper and more complex CNNs in learning the mapping from LR to HR space. These models improved performance by optimizing network architecture, for instance by removing unnecessary modules like Batch Normalization [82] or using residual dense blocks to leverage hierarchical features [84]. However, a primary limitation of standard CNNs is their local receptive field, which can hinder their ability to capture global semantic context and long-range dependencies.

To address this, transformer-based models were introduced. The Vision Transformer (ViT) [85] and its successor for restoration tasks, SwinIR [86], leverage attention mechanisms to model long-range dependencies efficiently, setting new benchmarks in SR performance.

Generative and Perceptual-Driven Models

A paradigm shift occurred with the use of generative models, which focus on producing plausible and perceptually convincing details rather than just minimizing pixel error. SRGAN [66] was a landmark, pioneering the use of a GAN with a perceptual loss for photo-realistic SR. This inspired a new family of models, including ESRGAN [87] and Real-ESRGAN [88], which extended the approach to handle more complex real-world degradations.

Other notable generative strategies include PULSE [89] which offers a self-supervised approach that searches the latent space of a pre-trained generative model to find HR images that downscale correctly, avoiding the need for paired LR-HR

training data. SR3 [90] and other diffusion-based models perform SR through an iterative refinement process, often achieving state-of-the-art results by gradually denoising an image conditioned on the LR input. To manage the computational cost of high-resolution generation, common strategies include progressive growing, cascaded refinement [48], and operating in a compressed latent space, as exemplified by Latent Diffusion Models (LDMs) [77].

3.3.2 Image-to-Image Translation

Image-to-image translation is a core task in conditional generation that involves translating an input image from a source domain to a target domain while preserving its core content. The challenge lies in learning a mapping function that can, for example, convert a semantic segmentation map into a photorealistic image, transform a day scene into a night scene, or change an artistic style.

Key approaches are distinguished by their data requirements:

- **Paired Image Translation**: When aligned pairs of images from the source and target domains are available, models can learn a direct mapping. Pix2pix [16] is a foundational cGAN-based framework for this purpose, widely used for tasks like converting sketches or architectural labels into photos.

- **Unpaired Image Translation**: A more challenging scenario arises when paired data is unavailable. CycleGAN [17] famously addresses this by introducing a cycle consistency loss. It simultaneously trains two generator-discriminator pairs to learn the forward mapping (A $\rightarrow$ B) and the reverse mapping (B $\rightarrow$ A), enforcing that translating an image to the target domain and back should restore the original image.

3.3.3 Text-to-Image Synthesis

Perhaps the most high-profile challenge in modern generative AI is text-to-image synthesis: generating a novel, coherent, and realistic image from a natural language description. This task requires a profound synergy between natural language understanding and image synthesis. The model must not only parse complex text prompts but also translate the resulting semantic concepts, including objects, attributes, relations, and artistic styles, into a visually faithful pixel representation.

The breakthrough in this area came from combining powerful pre-trained language models with high-fidelity generative models, particularly diffusion models. Modern text-to-image architectures like Stable Diffusion, DALL-E 2, and Imagen [77] work by:

1. Encoding the input text prompt into a rich numerical embedding using a text encoder (e.g., a CLIP variant).

2. Feeding this text embedding into the generative model to guide the image synthesis process. In diffusion models, this is typically achieved via cross-attention layers within the U-Net architecture, allowing the model to "pay attention" to relevant parts of the prompt at each step of the denoising process.

The success of these models, particularly Latent Diffusion Models (LDMs) [12], which perform the computationally intensive diffusion in a compressed latent space, has made high-resolution text-to-image generation widely accessible.

3.3.4 Image Inpainting

Image inpainting, the task of filling in missing regions of an image, presents unique challenges that have led to specialized architectures. While early methods relied on patch-based synthesis [72], deep learning models have achieved significant advances.

Key challenges include handling free-form masks and borrowing features from distant regions to fill large holes. Innovative deep learning techniques include:

- **Contextual Attention Layer**: This mechanism explicitly allows the network to borrow or copy feature information from known, distant spatial locations to generate consistent textures and structures [72].
- **Gated Convolutions**: Gated convolutions learn a dynamic feature gating mechanism that allows the network to distinguish between valid and masked pixels, ignoring invalid information and preventing artifacts [72].
- **LaMa (Large Mask Inpainting)**: This model [75] uses Fast Fourier Convolutions (FFCs) to gain an image-wide receptive field, enabling coherent completion of very large missing regions.
- **RePaint**: This approach leverages a pre-trained unconditional DDPM for inpainting by iteratively denoising the known regions while resampling the unknown regions, ensuring harmony between original and generated content [91].

In summary, the challenges in image synthesis are deeply interconnected. The pursuit of higher resolution [77] complicates the modeling of long-range dependencies [63] and strains computational resources. Efforts to enhance sample diversity [55] must be balanced against the need for high fidelity [71]. A recurring bottleneck, especially for CNNs, is the local receptive field. Innovations like attention mechanisms [63] and FFCs [75] are direct responses to this limitation, aiming to equip models with a more global understanding of image content.

3.4 Evaluation Paradigms for Generative Image Models

Evaluating the performance of generative image models is a notoriously difficult yet crucial aspect of research and development in this field. Unlike supervised learning tasks, where clear ground truth labels allow for straightforward error computation, assessing the quality of generated images involves judging multifaceted attributes such as realism, diversity, coherence, and, for conditional models, alignment with input conditions. This section reviews the common quantitative metrics and human-centric approaches used for evaluation, discussing their strengths, limitations, and the ongoing quest for more comprehensive and reliable assessment methods.

3.4.1 The Intricacies of Evaluating Generated Images

The core difficulty in evaluating generative models stems from the ill-posed nature of "image quality." There is no single, universally agreed-upon definition of what constitutes a "good" generated image. Desirable properties often include:

- **Fidelity (Realism/Quality)**: Individual generated samples should be sharp, detailed, and perceptually indistinguishable from real images from the target domain. They should not contain noticeable artifacts.

- **Diversity (Variety)**: The generative model should be able to produce a wide variety of samples that cover the full spectrum of variations present in the true data distribution, avoiding mode collapse.
- **Global Coherence**: Images should be structurally and semantically consistent as a whole.
- **Conditional Consistency**: If the model is conditional (e.g., text-to-image), the generated samples must accurately reflect the given condition.

A significant challenge is that quantitative metrics, while offering objectivity and ease of computation, often fail to fully capture human perception of these qualities [71]. An image might score well on a particular metric but still appear unnatural or contain subtle flaws to a human observer. Conversely, a perceptually excellent image might score poorly on certain metrics, especially those based on pixel-wise comparisons if it represents a novel but plausible variation.

3.4.2 Reconstruction Metrics

The Peak Signal-to-Noise Ratio (PSNR) and Structural Similarity Index (SSIM) emerged as foundational metrics in early image processing, rooted in information theory and human vision modeling. PSNR computes the logarithmic relationship between maximum possible pixel intensity and mean squared error:

$$\text{PSNR} = 10 \cdot \log_{10} \left(\frac{\text{MAX}_I^2}{\text{MSE}} \right)$$

where MAX_I represents maximum pixel value (typically 255 for 8-bit images) [92, 93]. While computationally efficient, PSNR's limitations became apparent with the rise of generative models. The metric fails to account for perceptual relevance - two images with identical PSNR can exhibit dramatically different visual quality due to PSNR's blindness to structural and semantic distortions [92, 94].

SSIM addressed these limitations through a multi-component approach comparing luminance, contrast, and structure:

$$\text{SSIM}(x, y) = \frac{(2\mu_x\mu_y + C_1)(2\sigma_{xy} + C_2)}{(\mu_x^2 + \mu_y^2 + C_1)(\sigma_x^2 + \sigma_y^2 + C_2)}$$

where μ represents local means, σ variances, and C stabilization constants [95]. SSIM's window-based computation better aligns with human vision's localized processing, but inherits limitations in evaluating generative outputs where structural hallucinations are intentional [96].

Recent analyses reveal fundamental flaws in both metrics for modern applications:

- PSNR correlates poorly (-0.2 to 0.3 Spearman) with human judgments of generative quality [92]
- SSIM exhibits saturation effects for high-quality images and fails to capture semantic distortions [95, 96]
- Both metrics assume pixel-wise alignment, making them unsuitable for evaluating diverse outputs in tasks like inpainting or style transfer [74, 93]

3.4.3 Image Generation Metrics

The limitations of classical metrics drove the development of feature-space measures using deep neural networks. The Inception Score (IS) leverages an ImageNet-pretrained Inception-v3 network to compute:

$$\text{IS} = \exp\left(\mathbb{E}_x D_{KL}(p(y|x) \parallel p(y))\right)$$

where $p(y|x)$ is class distribution per image and $p(y)$ marginal distribution across samples [97]. While IS rewards diversity and class specificity, it suffers from dataset bias (favors ImageNet-like images) and fails to compare against real data distributions [70, 97].

Fréchet Inception Distance (FID) [98] addressed these limitations by comparing multivariate Gaussians fitted to real and generated features:

$$\text{FID} = \|\mu_r - \mu_g\|^2 + \text{Tr}(\Sigma_r + \Sigma_g - 2(\Sigma_r\Sigma_g)^{1/2})$$

where μ, Σ are means and covariances of Inception-v3 [99] pool3 features.

FID demonstrates better correlation with human judgment, but remains sensitive to feature extraction details and sample size [70].

Learned Perceptual Image Patch Similarity (LPIPS) [74] introduced a paradigm shift by using deep features from pretrained networks (VGG, AlexNet) to compute perceptual distance:

$$\text{LPIPS}(x, y) = \sum_l \frac{creswell2018generative}{H_l W_l} \sum_{h,w} \|w_l \odot (f_l(x)_{h,w} - f_l(y)_{h,w})\|_2^2$$

where f_l represents layer activations and w_l learned channel weights. LPIPS correlates better with human judgments than SSIM while enabling cross-domain comparisons.

3.4.4 Emerging Semantic Metrics

The rise of vision-language models enabled metrics to evaluate semantic alignment. CLIPScore [100] measures image-text compatibility using CLIP's embedding space:

$$\text{CLIPScore} = \max(100 \cdot \cos(E_i, E_t), 0)$$

where E_i, E_t are CLIP image/text embeddings.

QualiCLIP [101] extends this concept through quality-aware alignment, using contrastive ranking of degraded images to predict perceptual quality without human labels.

For no-reference quality assessment, MANIQA (Multi-dimension Attention Network for IQA) [102] introduces a vision transformer architecture with dual attention branches:

1. Transposed Attention Block models channel interactions
2. Scale Swin Transformer Block captures multi-scale spatial relationships

MANIQA architecture's patch-weighted prediction head enables interpretable quality analysis at multiple scales.

3.4.5 Human Preference Metrics

Recent metrics directly optimize for human perceptual preferences. HPSv2 (Human Preference Score v2) [8] evaluates text-to-image models across four stylistic dimensions: Animation, Concept Art, Painting, and Photography.

Using a diverse dataset of 798k human judgments, HPSv2 trains a style-specific scoring function.

PickScore [103] instead leverages 851k human comparisons from the Pick-a-Pic dataset to fine-tune CLIP for preference prediction:

$$\text{PickScore}(I, P) = \text{CLIP}(I) \cdot \text{CLIP}(P) + \beta \cdot \text{CLIP}(I) \cdot \text{CLIP}_{\text{quality}}$$

where P is prompt embedding and $\text{CLIP}_{\text{quality}}$ a quality-specific projection.

These preference-based metrics mark a paradigm shift towards human-aligned evaluation, though challenges remain in cultural bias mitigation and compositional reasoning. Current research focuses on multi-dimensional preference modeling and sample-efficient adaptation to niche domains.

Chapter 4

Wavelets and Quaternion Wavelets: The Power of Multi-Scale Analysis

Contents

4.1 Foundational Principles of Wavelet Transforms in Signal and Image Processing 38
- 4.1.1 Mathematical Foundations: Mother Wavelets and Transform Theory 38
- 4.1.2 Continuous vs. Discrete Wavelet Transforms 39
- 4.1.3 Two-Dimensional Wavelet Transforms for Image Analysis 40
- 4.1.4 Quadtree Structure and Sparse Representation Properties 40
- 4.1.5 Limitations and Advanced Extensions 41

4.2 Quaternions and their Application to Images 41
- 4.2.1 Quaternion Algebra 41
- 4.2.2 Learning in the Quaternion Domain 42
- 4.2.3 The Dual-Tree Quaternion Wavelet Transform (QWT): Construction, Shift-Invariance, and Directional Selectivity 43
- 4.2.4 Simpler Strategies: Quaternion Embedding of DWT Subbands 45
- 4.2.5 Strategic Use of QWT Sub-bands in Neural Models . . . 46

4.3 Wavelets in Deep Learning and Generative Models . . . 47
- 4.3.1 Wavelets in Deep Learning 47
- 4.3.2 Wavelets in Generative Models 48

4.4 A Multi-Scale Foundation for Generative Modeling . . . 49

THIS chapter provides a comprehensive foundation for understanding wavelet transforms and their quaternion extensions, establishing the theoretical groundwork necessary for advanced image processing and representation learning applications. The multi-scale nature of wavelet analysis offers unique advantages in capturing both local and global image features, while quaternion extensions provide enhanced representational capacity for color and multi-channel data processing.

4.1 Foundational Principles of Wavelet Transforms in Signal and Image Processing

Wavelet analysis has emerged as one of the most powerful tools in signal and image processing, offering a mathematically rigorous framework for multi-resolution analysis that simultaneously captures both spatial (or temporal) and frequency information [104]. This dual localization stands in stark contrast to the Fourier transform, which, while offering precise frequency decomposition, discards spatial localization details [104]. The fundamental advantage of wavelets lies in their ability to provide a more nuanced view by capturing information across various scales and locations, making them particularly well-suited for analyzing signals and images that are inherently non-stationary and contain transient features.

4.1.1 Mathematical Foundations: Mother Wavelets and Transform Theory

The mathematical foundation of wavelet theory rests on the concept of a mother wavelet $\psi(t)$, which must satisfy specific admissibility conditions that are fundamental to the wavelet's effectiveness as an analytical tool [24]. These conditions include:

- **Zero mean condition:** $\int \psi(t)dt = 0$, or equivalently in the frequency domain, $\Psi(0) = 0$
- **Finite energy condition:** $\int |\psi(t)|^2 dt < \infty$ (square integrability)

The zero-mean property is not merely a mathematical formality but is fundamental to the wavelet's role as a "detail" extractor. Because the wavelet integrates to zero, it is sensitive to fluctuations and changes around a local average, rather than to constant or slowly varying components. When convolved with a signal, it effectively highlights deviations from local smoothness, which correspond to the signal's details. This naturally leads to a sparse representation where information is concentrated in a few significant coefficients.

From the mother wavelet, a complete wavelet dictionary $\{\psi_{u,s}\}$ is constructed through dilation (scaling) and translation (shifting) operations:

$$\psi_{u,s}(t) = \frac{1}{\sqrt{s}}\psi\left(\frac{t-u}{s}\right)$$

where $s > 0$ is the scale parameter controlling the dilation (inversely related to frequency), and $u \in \mathbb{R}$ is the translation parameter determining the spatial localization [24]. The normalization factor $\frac{1}{\sqrt{s}}$ ensures that the energy of each wavelet function remains constant across all scales, preserving the mathematical integrity required for accurate signal reconstruction.

The wavelet transform of a function $f(t)$ is obtained by computing its inner products with the members of this wavelet family:

$$\mathcal{W}_f(u,s) = \langle f, \psi_{u,s}\rangle = \int f(t)\psi^*_{u,s}(t)dt$$

where $\psi^*_{u,s}(t)$ is the complex conjugate of $\psi_{u,s}(t)$ [104].

4.1.2 Continuous vs. Discrete Wavelet Transforms

When the scale parameter s and translation parameter u vary continuously, this defines the **Continuous Wavelet Transform (CWT)**. The CWT provides a highly redundant yet exceptionally rich representation of the signal's time-frequency content, making it ideal for detailed signal analysis, identifying subtle frequency shifts over time, and precisely localizing transient events [105–107]. The scale parameter in CWT is often discretized more finely, typically using exponential scales with a base like $2^{1/v}$, where v is the number of "voices per octave."

For practical computational purposes, particularly in digital signal processing, the scale and translation parameters are discretized, leading to the **Discrete Wavelet Transform (DWT)**.

Commonly, dyadic sampling is employed, where $s = 2^j$ and $u = k2^j$ for integers j (scale level) and k (translation index) [105, 107]. This dyadic decomposition corresponds to an efficient implementation using filter banks, known as the Mallat algorithm, which iteratively applies pairs of high-pass and low-pass filters to the signal.

The choice between CWT and DWT reflects a fundamental trade-off: CWT offers analytical richness through redundancy, while DWT offers computational efficiency and compactness. The DWT, especially its decimated version, prioritizes computational efficiency and can provide a non-redundant representation when using orthogonal wavelets. However, this efficiency comes at the cost of shift-variance [1]. Table 4.1 provides a comprehensive comparison of these transforms.

Table 4.1. Comparative Overview of Fourier and Wavelet Transforms

Feature	**Fourier Trans.**	**CWT**	**DWT**
Basis Functions	Global, non-localized sinusoids	Localized wavelets (scaled & translated)	Localized wavelets (dyadically scaled & translated)
Time/Freq. Localization	Poor time, good frequency	Good time, good frequency	Good time, good frequency
Stationarity Assumption	Assumes stationarity	Well-suited for non-stationary signals	Well-suited for non-stationary signals
Redundancy	Non-redundant (for DFT)	Highly redundant	Non-redundant (orthogonal)
Shift-Invariance	Shift in time $\rightarrow$ phase shift in freq.	Approximately shift-invariant	Shift-variant
Computational Complexity	$O(N \log N)$ (FFT)	Higher, depends on scale/translation density	$O(N)$ (Fast Wavelet Transform)
Typical Use Cases	Stationary signal analysis, filtering	Detailed time-frequency analysis, transient detection	Compression, denoising, fast algorithms

[1] www.mathworks.com/help/wavelet/gs/continuous-and-discrete-wavelet-transforms.html

4.1.3 Two-Dimensional Wavelet Transforms for Image Analysis

The extension of wavelet theory to two-dimensional signals, particularly images, involves constructing bivariate wavelet bases through tensor products of one-dimensional wavelets and scaling functions [108, 109]. This approach preserves the essential properties of univariate wavelets while enabling the analysis of spatial relationships and directional features inherent in image data.

Given a 1D scaling function $\phi(t)$ and a 1D mother wavelet $\psi(t)$, three fundamental types of 2D wavelets are formed, each capturing different orientational information:

- $\psi^{(1)}(x, y) = \phi(x)\psi(y)$: Low-pass in x-direction, high-pass in y-direction (detects horizontal edges)
- $\psi^{(2)}(x, y) = \psi(x)\phi(y)$: High-pass in x-direction, low-pass in y-direction (detects vertical edges)
- $\psi^{(3)}(x, y) = \psi(x)\psi(y)$: High-pass in both directions (detects diagonal edges)

Additionally, a 2D scaling function $\phi(x, y) = \phi(x)\phi(y)$ captures the low-frequency approximation of the image.

The two-dimensional discrete wavelet transform (2D-DWT) decomposes an image into four subbands at each level [108]:

- **LL (Low-Low):** The approximation subband containing low-frequency information in both horizontal and vertical directions
- **LH (Low-High):** Detail subband containing horizontal features (low-frequency horizontally, high-frequency vertically)
- **HL (High-Low):** Detail subband containing vertical features (high-frequency horizontally, low-frequency vertically)
- **HH (High-High):** Detail subband containing diagonal features (high-frequency in both directions)

This decomposition process can be iteratively applied to the LL subband to achieve multi-level decomposition, creating a hierarchical representation at progressively coarser scales.

4.1.4 Quadtree Structure and Sparse Representation Properties

The multi-resolution decomposition naturally gives rise to a quadtree structure, where each wavelet coefficient at a coarser scale j serves as a parent to four coefficients at the next finer scale $j - 1$ [110]. This hierarchical organization mirrors the inherent spatial relationships in natural images, where large-scale structures typically encompass finer-scale details within their spatial extent.

The quadtree structure has proven invaluable for image compression algorithms, such as JPEG2000 [111]. Significant wavelet coefficients (those with large absolute values, typically corresponding to edges or important textural features) tend to propagate through the scales, creating predictable patterns of importance across the quadtree. This property, known as energy compaction, allows for efficient representation of image information using relatively few coefficients.

The sparse representation properties of wavelets are fundamental to their effectiveness in image processing. In typical natural images, edge information, while

perceptually crucial, constitutes only a small fraction of the total image content. Wavelet coefficients exhibit large absolute values primarily when edges or other singularities intersect the support regions of the corresponding wavelets [110]. This sparsity enables:

- **Efficient compression:** Many coefficients can be discarded or heavily quantized with minimal perceptual loss
- **Effective denoising:** Small coefficients (likely noise) can be thresholded while preserving significant features
- **Feature extraction:** Multi-scale and multi-orientation information for image recognition and analysis

4.1.5 Limitations and Advanced Extensions

Despite their power, standard separable 2D-DWT transforms have limitations in capturing true diagonal information. The HH subband tends to mix responses from features oriented at +45° and -45°, sometimes leading to checkerboard artifacts and suboptimal directional selectivity [108, 112]. This limitation has driven the development of more advanced wavelet transforms.

Recent developments have explored learnable and adaptive approaches that move beyond fixed wavelet bases [113]. Wavelet-driven masked image modeling leverages the multi-scale nature of wavelet decomposition to create more efficient training procedures for self-supervised learning models. The integration of wavelet transforms with modern deep learning architectures has opened new possibilities for enhanced image representation, including learnable wavelet designs implemented through filterbank autoencoders.

These advanced extensions, including the Dual-Tree Complex Wavelet Transform and Quaternion Wavelet Transforms, address the directional limitations while maintaining the beneficial properties of traditional wavelets.

4.2 Quaternions and their Application to Images

In this section, we expound on basic concepts of quaternion algebra, the foundations of quaternion neural networks and how they can be involved in a variety of studies. Indeed, real-valued data can be easily processed with quaternion-valued models by encapsulating input dimensions into the components of a quaternion. Similarly, a quaternion framework can be handily integrated into a real-valued analysis.

4.2.1 Quaternion Algebra

The quaternion set of numbers $\mathbb{H}$ belongs to the class of Clifford algebras, being a four-dimensional associative, normed, and non-commutative division algebra over real numbers. While complex numbers are defined by a real-valued component and one imaginary unit, quaternion numbers comprise two additional imaginary components, being defined as:

$$q = q_0 + q_1\hat{\imath} + q_2\hat{\jmath} + q_3\hat{\kappa}, \tag{4.1}$$

whereby $q_c, c \in \{0, 1, 2, 3\}$ are real-valued coefficients and $\hat{\imath}, \hat{\jmath}, \hat{\kappa}$ imaginary units, which comply with the properties:

1. $\hat{\imath}^2 = \hat{\jmath}^2 = \hat{\kappa}^2 = -1$;
2. $\hat{\imath} \times \hat{\jmath} = \hat{\kappa}, \hat{\jmath} \times \hat{\kappa} = \hat{\imath}, \hat{\kappa} \times \hat{\imath} = \hat{\jmath}$;
3. $\hat{\imath} \times \hat{\jmath} \neq \hat{\jmath} \times \hat{\imath}, \hat{\jmath} \times \hat{\kappa} \neq \hat{\kappa} \times \hat{\jmath}, \hat{\kappa} \times \hat{\imath} \neq \hat{\imath} \times \hat{\kappa}$,

where $\times$ is the vector product in R^3. An alternative quaternion representation can be derived considering the magnitude and the three phases:

$$q = |q| e^{\hat{\imath}\theta_1} e^{\hat{\jmath}\theta_2} e^{\hat{\kappa}\theta_3}, \tag{4.2}$$

in which $|q|$ is the amplitude of the quaternion and $(\theta_1, \theta_2, \theta_3) \in [\pi, \pi) \times [-\pi/2, \pi/2) \times [-\pi/4, \pi/4]$ are the phase angles of it. The latter can be computed as follows:

$$\begin{cases} \theta_1 = \arctan\left(\frac{2q_0 \cdot q_2 + q_1 \cdot q_3}{q_0^2 + q_1^2 - q_2^2 - q_3^2}\right) \\ \theta_2 = \arctan\left(\frac{q_0 \cdot q_1 + q_2 \cdot q_3}{q_0^2 - q_1^2 + q_2^2 - q_3^2}\right) \\ \theta_3 = \frac{1}{2}\arctan\left(2\left(q_0 \cdot q_3 - q_3 \cdot q_1\right)\right). \end{cases} \tag{4.3}$$

The quaternion domain is endowed with a norm

$$|q| = \sqrt{q_0^2 + q_1^2 + q_2^2 + q_3^2},$$

a conjugate

$$q^* = q_0 - q_1\hat{\imath} - q_2\hat{\jmath} - q_3\hat{\kappa},$$

the operations of scalar multiplication

$$q \cdot p = q_0 p_0 + q_1 p_1 \hat{\imath} + q_2 p_2 \hat{\jmath} + q_3 p_3 \hat{\kappa},$$

and associative multiplication of elements.

Due to the vector product non-commutative property, the Hamilton product has been introduced for properly modelling imaginary unit interplays [114]:

$$q \otimes p = \begin{bmatrix} q_0 & -q_1 & -q_2 & -q_3 \\ q_1 & q_0 & -q_3 & q_2 \\ q_2 & q_3 & q_0 & -q_1 \\ q_3 & -q_2 & q_1 & q_0 \end{bmatrix} \begin{bmatrix} p_0 \\ p_1 \\ p_2 \\ p_3 \end{bmatrix}. \tag{4.4}$$

4.2.2 Learning in the Quaternion Domain

Quaternion neural networks (QNNs) are built upon the Hamilton product for multiplying and convolving weights with input data [115, 116]. More precisely, a real-valued fully connected layer with weights matrix $\mathbf{W}$ and bias $\mathbf{b}$ has the following expression:

$$\mathbf{y} = \mathbf{W}\mathbf{x} + \mathbf{b}, \tag{4.5}$$

where the real-valued input $\mathbf{x}$ is modelled into the real-valued output $\mathbf{y}$. A quaternion fully connected (QFC) layer involves quaternion characters and its definition is based on the Hamilton product between the weights matrix, arranged in the form of (4.4), and the quaternion input as:

$$\mathbf{y} = \mathbf{W} \otimes \mathbf{x} + \mathbf{b}. \tag{4.6}$$

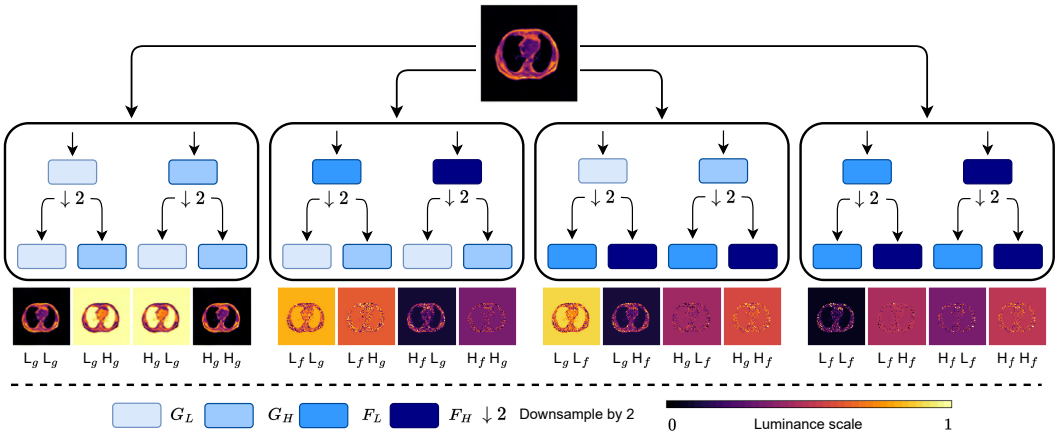


Figure 4.1. Dual-tree quaternion wavelet transform (QWT). The input image is decomposed into sixteen sub-bands by interleaving low-pass and high-pass filters.

The QNNs formulation through the Hamilton product has two key advantages. First, since the quaternion weight matrix is composed of four sub-matrices $\mathbf{W}_c$, with $c \in \{0, 1, 2, 3\}$, each with 1/16 the parameters of the complete matrix $\mathbf{W}$, and these sub-matrices are reused to build the final weight matrix $\mathbf{W}$ according to (4.4), QNNs save 75% of free parameters with respect to real-valued counterparts. Second, due to such sharing of weight sub-matrices, each parameter is multiplied by each dimension of the input (e.g., by each channel of RGB images, or of multichannel signals), thus capturing complex relationships among input dimensions and preserving their correlations [117,118]. This allows QNNs to gain comparable results when processing multidimensional data despite the lower number of free parameters. However, due to the four-dimensional nature of quaternion numbers, QNNs are mainly limited to multidimensional signals, while their extension to one-channel and monodimensional data is still an open problem.

4.2.3 The Dual-Tree Quaternion Wavelet Transform (QWT): Construction, Shift-Invariance, and Directional Selectivity

A more advanced and integrated approach to combining wavelets and quaternions is the Quaternion Wavelet Transform (QWT), particularly its dual-tree implementation. The QWT is a natural extension of the Dual-Tree Complex Wavelet Transform (DT-CWT) [119]. The DT-CWT itself was developed to overcome the primary drawbacks of the standard DWT: shift-variance and poor directional selectivity [112]. It achieves approximate shift-invariance and improved directional selectivity by using two parallel DWT filter banks designed such that one set is approximately the Hilbert transform of the other, resulting in complex-valued wavelet coefficients that provide both magnitude (energy) and phase (local shift) information.

The QWT extends this concept further by employing four real DWTs (or two DT-CWTs in parallel). It comprises four quaternion wavelet transforms that are built by means of a real DWT and its three Hilbert transforms applied along the x-axis, y-axis, and xy-axes (diagonally) [31,119]. This structure results in the decomposition of an input image into four quaternion wavelet sub-bands, which collectively consist of 16 real sub-bands at each decomposition level. Similar to the 2D-DWT, the QWT yields a quaternion scaling function ϕ_q and three quaternion wavelets $\psi_q^D, \psi_q^V, \psi_q^H$ capturing diagonal, vertical, and horizontal features, respectively. These are formally

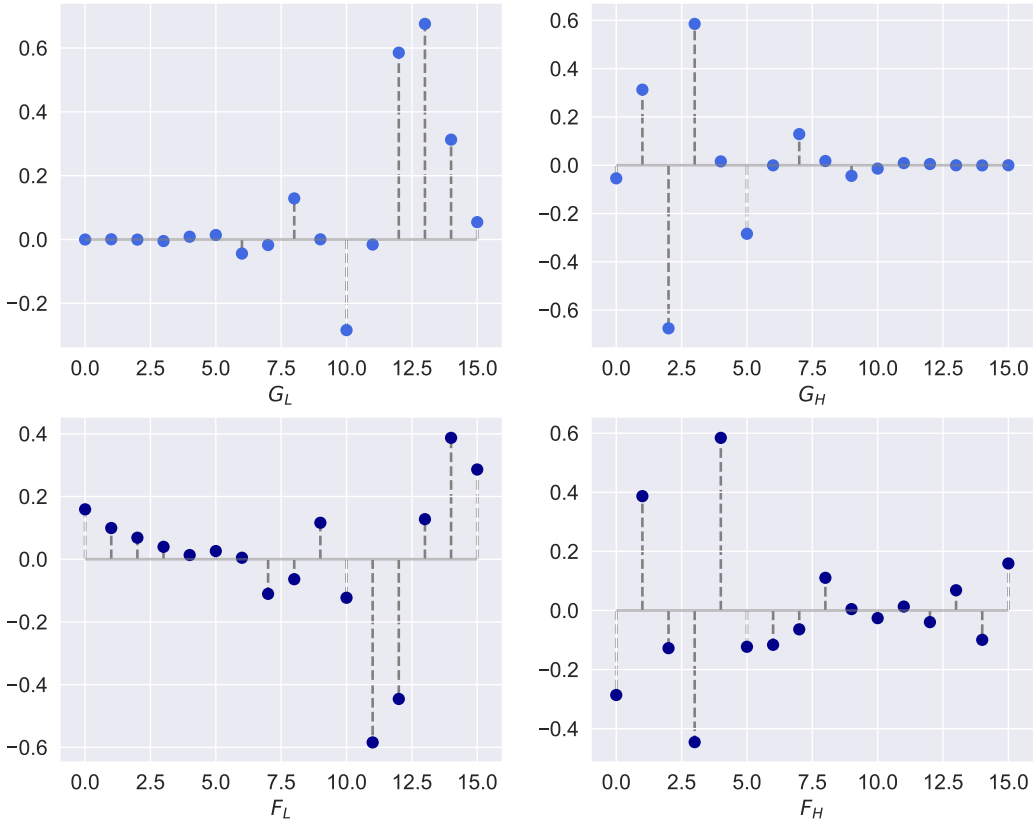


Figure 4.2. Filters for QWT: G_L and G_H are extracted from the wavelet db8, while F_L and F_H are Hilbert transforms of the previous ones.

defined as:

$$\begin{aligned}
\phi_q &= \phi_g(x)\phi_g(y) + \phi_f(x)\phi_g(y)\hat{\imath} + \phi_g(x)\phi_f(y)\hat{\jmath} + \phi_f(x)\phi_f(y)\hat{\kappa} \\
\psi_q^V &= \psi_g(x)\phi_g(y) + \psi_f(x)\phi_g(y)\hat{\imath} + \psi_g(x)\phi_f(y)\hat{\jmath} + \psi_f(x)\phi_f(y)\hat{\kappa} \\
\psi_q^H &= \phi_g(x)\psi_g(y) + \phi_f(x)\psi_g(y)\hat{\imath} + \phi_g(x)\psi_f(y)\hat{\jmath} + \phi_f(x)\psi_f(y)\hat{\kappa} \\
\psi_q^D &= \psi_g(x)\psi_g(y) + \psi_f(x)\psi_g(y)\hat{\imath} + \psi_g(x)\psi_f(y)\hat{\jmath} + \psi_f(x)\psi_f(y)\hat{\kappa}
\end{aligned} \tag{4.7}$$

where g represents a real-valued filter (e.g., a low-pass or high-pass filter from a DWT, such as Daubechies db8 [120]), and f is its corresponding Hilbert transform. More concretely, the QWT of an image can be computed by combining four filters. An asymmetric, orthogonal and biorthogonal Daubechies wavelet with 8 vanishing moments (db8) can be considered, following [120]. From the db8, the decomposition low-pass (G_L) and high-pass filters (G_H) are extracted, and then the Hilbert transform is applied to them to obtain the filters F_L and F_H (see Fig. 4.2). The use of Hilbert transform filter pairs ensures that the real and imaginary components of the quaternion wavelets are approximately analytic, which is key to achieving the desired properties. Once G_L, G_H, F_L and F_H are available, the QWT is calculated by combining filters and downsampling operations according to the dual-tree structure shown in Fig. 4.1.

The QWT exhibits several important properties:

1. **Approximate Shift-Invariance:** The squared magnitude of the quaternion

wavelet coefficients is designed to be non-oscillatory with respect to shifts in the input signal. This means the transform is approximately shift-invariant, a crucial property for robust feature extraction in image processing, as it makes the representation less sensitive to small translations of the input image [112, 119].

2. **Superior Directional Selectivity:** The QWT offers significantly better directional selectivity than the standard DWT, capable of distinguishing features at a richer set of orientations (e.g., $\pm 15°, \pm 45°, \pm 75°$, and potentially more, depending on the filter design and interpretation of the 16 sub-bands) [112].
3. **Rich Phase Information:** The quaternion representation provides three phase components per coefficient. These phases can encode local shifts in two dimensions as well as information about edge orientation and local texture [119, 121, 122].

The QWT can be viewed as a more complete extension of the 1D analytic signal concept to 2D compared to the DT-CWT. While DT-CWT yields complex coefficients (one real and one imaginary part), QWT provides quaternion coefficients (one real and three imaginary parts). This richer structure allows for a more nuanced representation of 2D features, maintaining shift-invariance and offering an even richer framework for directional analysis through its 16 real sub-bands and three phase components.

4.2.4 Simpler Strategies: Quaternion Embedding of DWT Sub-bands

The integration of wavelet transforms with quaternion algebra offers advanced approaches for extracting salient characteristics from images. These methods aim to overcome limitations inherent in traditional real-valued wavelet transforms, particularly concerning shift-invariance and the explicit representation of phase information, by leveraging the richer mathematical structure of quaternions.

A pragmatic approach to combine the strengths of Discrete Wavelet Transforms (DWT) and quaternion algebra involves encapsulating the standard 2D-DWT sub-bands into a quaternion structure. The 2D-DWT characterizes a 2D signal by involving tensor products of 1D-DWTs over the two dimensions, resulting in the scaling function $\phi(x)\phi(y)$ and three wavelets $\psi(x)\psi(y)$, $\phi(x)\psi(y)$, and $\psi(x)\phi(y)$ that extract diagonal, horizontal and vertical features, respectively [119, 123]. This decomposition yields a low-frequency approximation sub-band (LL) and three high-frequency detail sub-bands (LH for horizontal features, HL for vertical features, and HH for diagonal features) at each decomposition level [31].

While this decomposition is effective for multi-scale analysis, the standard DWT suffers from several limitations. First, it lacks translation invariance, meaning that small shifts in the input image can cause significant changes in wavelet coefficients. Second, it does not explicitly provide phase information, which is crucial for describing the spatial layout and structural details of image content [31, 112, 124]. Indeed, phase information may be crucial in medical images that could contain artifacts due to patient movements or inaccurate equipment.

To address these shortcomings and create a more robust representation, the four real-valued sub-bands obtained from a 2D-DWT at a specific level can be mapped to the components of a single quaternion. A common embedding is:

$$q = \mathrm{LL} + \mathrm{LH}\hat{\imath} + \mathrm{HL}\hat{\jmath} + \mathrm{HH}\hat{\kappa}$$

Once this quaternion representation q is formed, its properties, such as the magnitude $|q|$ and the three phase angles $(\theta_1, \theta_2, \theta_3)$, can be exploited following the magnitude and phases characterization: $q = |q|e^{\hat{\imath}\theta_1}e^{\hat{\jmath}\theta_2}e^{\hat{\kappa}\theta_3}$

This embedding strategy offers a clever and practical way to harness some of the advantages of quaternion algebra, such as the holistic representation of the four components and access to a multi-component phase, without the full implementation complexity of a native Quaternion Wavelet Transform. It leverages existing, well-understood DWT algorithms while providing enhanced feature engineering. The magnitude of this quaternion, $|q|$, provides a measure of energy across all four sub-bands that tends to be more shift-invariant than the energies of individual sub-bands. The three phase angles encode the relative orientations and spatial relationships between the features captured by the LL, LH, HL, and HH sub-bands, thereby capturing some of the phase information missing from the standard DWT output and allowing for a joint representation where inter-dependencies between sub-bands are naturally modeled within the quaternion structure [124]. Such manipulated data can then be fed into real, quaternion, or hypercomplex-valued neural networks without needing additional preprocessing.

4.2.5 Strategic Use of QWT Sub-bands in Neural Models

The rich output of the QWT, comprising 4 low-frequency approximation sub-bands (ϕ_q) and 12 high-frequency sub-bands ($\psi_q^V, \psi_q^H, \psi_q^D$) for a total of 16 real sub-bands per decomposition level, presents both opportunities and challenges when used as input to neural networks.

Some earlier studies, often employing shallower classifier models, reported improved classification accuracy by using only a subset of these sub-bands, specifically the four low-frequency components corresponding to the real parts of the quaternion sub-bands, or the four components of the quaternion approximation sub-band ϕ_q [125, 126]. The rationale is that these low-frequency components contain the majority of the image's energy and structural information, while high-frequency details might be noisy or less discriminative for certain classification tasks when processed by models with limited capacity. Indeed, shallow models may not have enough capacity to handle very complex inputs such as high-frequency wavelet details.

However, it is argued that deep neural networks, with their substantially greater capacity for learning complex representations, may benefit from more detailed inputs. Neural networks are more complex systems and more detailed inputs can further enhance model expressiveness, improving performance. For tasks beyond simple classification, such as image reconstruction, segmentation, or generation, high-frequency details are often critical. Deep networks are, in principle, more capable of sifting through complex, high-dimensional inputs and extracting meaningful features from the interplay of all 16 QWT sub-bands, including subtle high-frequency characteristics and their phase relationships.

Nevertheless, directly feeding all 16 real sub-bands into a neural model can introduce challenges. On one hand, high-frequency features are less informative with respect to low ones, and involving the sixteen sub-bands as input to a neural model may undermine its performance due to redundant features, forcing the model to distinguish between relevant and uninformative channels before effectively learning a proper representation. Furthermore, QNNs inherently require four-channel inputs to be encapsulated into a quaternion, thus it is not suitable to directly involve the sixteen sub-bands. Therefore, a specific mapping strategy is necessary if all 16 real sub-bands are to be utilized with a QNN. This might involve grouping the

sub-bands spatially, by orientation type, or by employing quaternion-valued pooling or dimensionality reduction layers designed to respect the quaternion algebraic structure.

A proper combination of low and high sub-bands can better represent input images while avoiding redundancy. Learning-based approaches to this challenge represent a promising direction, aiming to extract features from only the most informative sub-bands of the QWT. This suggests that an adaptive or learned selection and weighting strategy is beneficial for effectively harnessing the QWT's rich output in deep learning models, offering a data-driven method of determining feature importance to resolve the dilemma of how best to utilize the comprehensive information provided by the QWT as we will further explore in Section 7.3.

4.3 Wavelets in Deep Learning and Generative Models

Recent advancements have pushed wavelet theory beyond fixed bases, integrating adaptive and learnable approaches with modern deep learning architectures. This fusion harnesses the multi-scale analytical power of wavelets while leveraging the potent feature learning capabilities of neural networks, leading to more efficient and expressive image representations. The integration addresses fundamental challenges in generative modeling, training efficiency, and representation quality while maintaining the mathematical rigor and interpretability of traditional wavelet analysis.

4.3.1 Wavelets in Deep Learning

A significant evolution in this field is the shift from pre-defined wavelet families to data-driven systems that are either adaptive or fully learnable. Frameworks have been developed, often as specialized layers or filterbank autoencoders, that learn optimal wavelet bases tailored for specific tasks end-to-end [127, 128, 128]. These architectures embed key properties of the Fast Discrete Wavelet Transform, such as its cascade algorithm structure and conjugate quadrature filter constraints, directly into the network. This allows them to learn both the wavelet filters and adaptive coefficient processing operations, like novel hard-thresholding activations, as part of the optimization process. This adaptivity finds powerful application in areas like self-supervised learning, as seen in Wavelet-Driven Masked Image Modeling (WaMIM) [113].

WaMIM addresses the inefficiency of standard masked image modeling by operating in the sparse wavelet domain instead of reconstructing dense pixels. By creating reconstruction targets from multi-level wavelet coefficients, it achieves enhanced efficiency and fosters more direct hierarchical feature learning. This approach not only reduces computational overhead but also leads to faster convergence and improved performance on downstream tasks like classification and object detection.

Beyond creating new representations, wavelets are also being integrated into modern deep learning architectures to address inherent computational bottlenecks, most notably within Transformers. Several prominent strategies have emerged to mitigate the quadratic complexity of the self-attention mechanism.

One approach, exemplified by Wavelet Space Attention (WavSpA), retains the attention mechanism but performs its operations within the wavelet coefficient space [129]. This involves a three-stage process: a forward wavelet transform projects the input into multi-resolution bases, attention operations are conducted in

this structured space to better capture long-range dependencies, and a backward transform reconstructs the final representation.

Another paradigm, found in the Wavelet Vision Transformer (WaveViT), uses wavelets to create a more efficient and information-preserving attention operation [130]. Instead of using standard pooling methods that irreversibly discard information, WaveViT employs the Discrete Wavelet Transform (DWT) for a lossless down-sampling of the key and value tensors within the self-attention module. This reduces computational complexity without losing the high-frequency details essential for reconstruction.

All those methodologies leverage the linear time complexity of the Fast Wavelet Transform to enable Transformers to scale efficiently to the long sequences required for high-resolution images, video processing, and extensive documents.

4.3.2 Wavelets in Generative Models

Operating directly in the high-dimensional pixel space presents significant hurdles for deep generative models, making it difficult to capture long-range dependencies and manage computational costs. The integration of wavelet transforms offers a powerful alternative, addressing fundamental challenges in training efficiency, sampling speed, and representation quality. This approach leverages the inherent multi-scale nature of wavelets to decompose images into a more structured and manageable format, creating more efficient and effective generative frameworks.

The core advantage of this technique lies in the wavelet transform's ability to decompose an image into a multi-resolution representation. This separation allows a model to simultaneously learn an image's global structure and overall layout from the low-frequency "approximation" coefficients, while independently modeling the fine details, edges, and textures contained within the high-frequency "detail" coefficients. This hierarchical organization is highly conducive to progressive, coarse-to-fine generation schemes, where the model first establishes the basic composition and then iteratively adds detail. Furthermore, this decomposition can significantly improve training stability, as predicting the often sparse and well-behaved wavelet coefficients is a more constrained and numerically stable task than predicting raw pixel values directly.

This principle is effectively demonstrated in autoregressive models, as shown in approaches that convert images into wavelet coefficient sequences for more efficient sequential modeling [110]. Such methods exploit the predictable statistical properties of the wavelet domain. For instance, significant coefficients often form quadtree structures, where a large-magnitude coefficient at a coarse scale predicts the presence of significant coefficients in corresponding spatial locations at finer scales. This inherent regularity makes the sequence easier to model, leading to more efficient learning and providing natural termination points for the generation process. Consequently, this approach dramatically reduces the computational resources required compared to modeling dense, less predictable pixel grids.

Moreover, the inherent sparsity of the wavelet representation, where a vast majority of coefficients are zero or near-zero for natural images, allows the model to focus its capacity on the small subset of critical coefficients that define the image's primary structure and content. This focus ensures that high generation quality is preserved, all while benefiting from a more compact and efficient representation.

4.4 A Multi-Scale Foundation for Generative Modeling

This chapter has established the theoretical groundwork for the central hypothesis of this thesis: that the integration of multi-scale analysis via wavelet transforms, particularly when augmented with the holistic processing power of quaternion algebra, offers a superior foundation for deep generative modeling of images. We have journeyed from the foundational principles of the wavelet transform, which dissects images into a sparse and structured hierarchy of features, to the algebraic elegance of quaternions, which provide a natural framework for preserving the inherent inter-channel correlations within this multi-band representation.

The limitations of conventional generative models, as surveyed in Chapters 2 and 3, often stem from their operation in the raw pixel space, a domain that is high-dimensional, densely correlated, and structurally agnostic. The Quaternion Wavelet Transform (QWT) directly confronts this inefficiency. It provides a representational space where global structure and fine-grained detail are explicitly disentangled and where the relationships between directional sub-bands are encoded as a single, coherent quaternion entity. This is not merely a change of basis, but a paradigm shift towards a representation that is primed for efficient learning, imbued with the inductive biases of scale-invariance and directional selectivity that deep neural networks would otherwise have to learn from scratch.

With this powerful representational toolkit in hand, we are now positioned to confront the practical challenges of generative modeling. The subsequent chapters will transition from theory to empirical validation. In the next part of this thesis (part II), we will explore the performance and expose the critical limitations of state-of-the-art GAN and diffusion architectures on demanding tasks such as cross-domain translation and specialized super-resolution. These investigations will serve to motivate, with concrete evidence, the pressing need for the more robust representations developed herein. Subsequently, Part III will present the core contributions of this work, systematically demonstrating how injecting the quaternion wavelet framework into modern generative architectures can overcome these identified shortcomings. We will show how this approach enhances generalization, improves conditioning for high-fidelity super-resolution, and ultimately enables image synthesis at resolutions previously unattainable by standard latent diffusion models. This chapter, therefore, concludes the foundational part of our study, setting the stage for a comprehensive exploration of the wavelet revolution in generative modeling.

Part II

Exploring the Generative Landscape: Applications and Challenges

Chapter 5

Generative Adversarial Networks for Cross-Domain Translation

Contents

5.1 Introduction: GANs and the Infrared Translation Challenge 51

5.2 Structural-Aware Generative Adversarial Networks for Infrared Image Translation 53

5.2.1 Experimental evaluation 55

5.3 GAN Limitations, and the Shift to Diffusion 57

5.3.1 Successes and Contributions 57

5.3.2 Inherent Limitations and Emerging Challenges 58

5.3.3 The Bridge to Diffusion 58

The exploration of generative models across diverse and challenging domains serves as a crucial stepping stone toward understanding their fundamental capabilities and limitations. Before delving into novel architectural enhancements, it is essential to thoroughly investigate how state-of-the-art approaches perform when confronted with complex real-world scenarios. This chapter[1] examines the application of Generative Adversarial Networks to cross-domain image translation, specifically focusing on the challenging task of infrared-to-RGB translation,a problem that reveals both the power and constraints of current generative modeling paradigms.

5.1 Introduction: GANs and the Infrared Translation Challenge

Recently, image translation models based on Generative Adversarial Networks (GANs) [131] have demonstrated outstanding performances [132], particularly in tasks such as colorizing images from grayscale or infrared sources [133, 134]. GANs comprise two sub-networks: a generator that creates images in a target domain from input images in a source domain, and a discriminator that distinguishes between

[1]The content of this Chapter is adapted from the material published in [28]

real and generated samples. The training objective ensures that the discriminator cannot differentiate between authentic and synthesized images.

Infrared images (IR) prove invaluable in low-light conditions where standard cameras fail to capture critical details such as vehicles, animals, or pedestrians. However, thermal infrared (TIR) images, treated as single-channel entities, exhibit very low contrast that, while revealing hidden nocturnal details, significantly impairs human perception [135]. Conversely, well-established models [136–139] designed for RGB images benefit from rich annotations and proven architectures. Rather than adapting these algorithms to the infrared domain, translating IR images back to the RGB space presents a compelling alternative approach.

The complexity of this translation task is further emphasized by recent approaches [140, 141] demonstrating that grayscale colorization lacks unique solutions,a challenge amplified in the IR domain where models must predict chrominance without luminance information. Existing grayscale colorization methods [68, 142] typically employ CIE Lab color space, predicting only the a and b channels while preserving the lightness component L. To address IR-specific challenges, some approaches [143] decompose the problem using separate generators for each RGB channel.

Since translated images frequently serve downstream tasks such as classification or object detection [134], preserving image details and contours becomes paramount for subsequent analysis. Recent work [144] addresses this through structured gradient alignment loss ensuring edge consistency, though requiring explicit edge detection via Canny operators [145].

While Chapter 4 explored the theoretical power of multi-scale representations, this chapter first establishes a performance baseline using generative models that operate primarily in the pixel space. This allows us to empirically identify the precise challenges, such as preserving fine structural details, that advanced representations are poised to solve.

This investigation reveals a fundamental limitation of standard generative approaches: while capable of producing visually plausible translations, they often struggle with structural consistency and fine-grained detail preservation,particularly crucial for cross-domain scenarios. These observations motivate our exploration of enhanced architectural designs and, ultimately, the wavelet-based methodologies presented in subsequent chapters.

Our contribution here introduces Structured-Aware Generative Adversarial Networks (StawGAN), inspired by TarGAN [146] for medical image synthesis. StawGAN employs a single generator for both IR and RGB images while enhancing target quality through interconnected flows for image and segmented target processing. The network incorporates structural awareness via Structural Dissimilarity Index Measure (DSSIM) loss, encouraging spatial smoothness in generated images. This approach bridges paired and unpaired translation paradigms, demonstrating generalizability while leveraging task-specific losses when available.

Through comprehensive experiments on the DroneVehicle dataset [147], we demonstrate improvements over state-of-the-art metrics, validating enhanced target generation quality. More importantly, this work establishes the foundation for understanding the inherent challenges in cross-domain generation that subsequently drive the development of more sophisticated, wavelet-enhanced approaches detailed in later chapters.

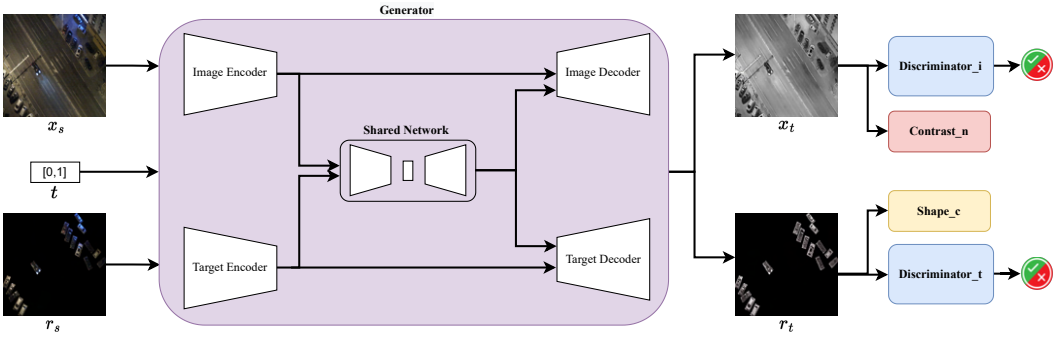


Figure 5.1. Architecture of StawGAN. We show the part inside of the generator, to explore the shared connections between the two different flows. Indeed this is a translation from RGB to IR, but the same process holds for the contrary.

5.2 Structural-Aware Generative Adversarial Networks for Infrared Image Translation

To introduce our approach, we briefly review the recent Target-Aware Generative Adversarial Network (TarGAN) [146]. This network translates images from a source domain to a target domain while paying particular attention to the region of interest, passed as a segmentation map during the translation, this model is originally made for a medical scenario where this attention to the details can be crucial. The input data for the generator model are the image in the source domain (s), the segmented image, and the one-hot vector representing the target domain (t), the output will be the translated image and its segmentation:

$$G(x_s, r_s, t) \rightarrow (x_t, r_t)$$

The key to enhancing the generation of vehicles lies in the shared connections between the two flows of the model, composed of two encoder-decoder architectures, one for the whole image and another for the segmented one, while between the two flows there are shared layers to combine information for a better translation. As the original TarGAN, we involve a Shape Controller S that takes as input the fake target image r_t and generates a binary mask representing the foreground area of the generated images. Additionally, we introduce a novel Contrast Network C, to help the generator in choosing the right contrast, sharpness, and gamma factors.

For training, we employ adversarial loss, which is the canonical loss to train adversarial networks, and a modality classification loss to force the target modality. It is applied on the generator G, on the discriminator D_x, and on the discriminator D_r.

$\mathcal{L}^r_{\text{cls_}x}$ refers to the loss of real images, thus used for the discriminators, while $\mathcal{L}^f_{\text{cls_}x}$ to the fake ones and used on the generator G. Since some losses are the same for the image and the target, in the next equations, z could represent an image x or a target image r:

$$\mathcal{L}^r_{\text{cls_}z} = \mathbb{E}_{z_s,s}\left[-\log D_{\text{cls_}z}\left(s \mid z_s\right)\right] + \lambda_u \mathbb{E}_{z_t,s'}\left[-\log D_{\text{cls_}z}\left(s' \mid z_t\right)\right], \tag{5.1}$$

$$\mathcal{L}^f_{\text{cls_}z} = \mathbb{E}_{z_t,t}\left[-\log D_{\text{cls_}z}\left(t \mid z_t\right)\right]. \tag{5.2}$$

We use the shape consistency loss [148] since the structural shape of original images could sometimes slightly differs from one of the translated samples, so the shape loss tries to contain this behavior by:

$$\mathcal{L}_{\text{shape_}r} = \mathbb{E}_{r_t, b^x}\left[\left\|b^x - S\left(r_t\right)\right\|_2^2\right]. \tag{5.3}$$

We introduce the structural information in the model by involving the Structural Dissimilarity Index Measure (DSSIM) [149] in the optimization process of the proposed network, so to improve the estimation of both the luminance and chrominance pixels of TIR images. The DSSIM loss defines a region, called window size, where we are predicting the luminance and chrominance of the image, and it is defined as:

$$\mathcal{L}_{\text{ssim}} = \frac{1 - \text{SSIM}(x_{\text{paired}} - x_t)}{2}, \tag{5.4}$$

where x_{paired} is the original image but in the same domain of x_t, which is instead the translated sample. Since we apply also the reconstruction loss [150] aims at preserving the original characteristics of the source images, we employ it with the DSSIM loss as suggested in [144]. We are using the reconstruction for both the target and the entire image, while DSSIM is only for the latter:

$$\mathcal{L}_{\text{rec_}z} = \mathbb{E}_{z_s, z_s'}\left[\left\|z_s - z_s'\right\|_1\right] + \lambda_{\text{ssim}}\mathcal{L}_{\text{ssim}} \tag{5.5}$$

We employ crossing loss to force the generator to pay attention to the target area when generating the translated image. Denoting with y the target area label of x_s, we have:

$$\mathcal{L}_{\text{cross}} = \mathbb{E}_{x_t, r_t, y}\left[\left\|x_t \cdot y - r_t\right\|_1\right], \tag{5.6}$$

To train the Contrast Network and the Generator, we involve the L_1 loss that takes the fake image, with the applied factors, and the paired image, serving in reducing the production of more gray-like and more colorful images [151]. The complete objective function takes therefore the following form, with hyperparameters lambda to control the weights of the loss functions:

$$\begin{aligned}
\mathcal{L}_{D_{(z)}} &= -\mathcal{L}_{\text{adv_}(z)} + \lambda_{\text{cls}}^r \mathcal{L}_{\text{cls_}(z)}^r, \\
\mathcal{L}_G &= \mathcal{L}_{\text{adv_}(z)} + \lambda_{\text{cls}}^f \mathcal{L}_{\text{cls_}(z)}^f + \lambda_{\text{rec}}\mathcal{L}_{\text{rec_}(z)} \\
&\quad + \lambda_{\text{cross}}\mathcal{L}_{\text{cross}} + \lambda_{L_1}\mathcal{L}_{L_1}, \\
\mathcal{L}_{G,S} &= \mathcal{L}_{\text{shape_}(r)}, \\
\mathcal{L}_C &= \lambda_{L_1}\mathcal{L}_{L_1}.
\end{aligned} \tag{5.7}$$

Therefore, thanks to the DSSIM loss in (5.5), the proposed model is aware of structural information producing better-shaped objects.

The encoder-decoder couples in the generator are composed of different building blocks: the convolutional one, with the usage of convolution, instance normalization, and Leaky ReLU, and the upsampling one involving the same operation as the previous one but with an upsampling operation done by interpolation at the beginning. For the discriminator, differently from the previous model, we are using two identical models but as separate entities, one determines if an image is fake or real and the other does the same, but for the segmented one, furthermore being a conditional GAN, they determine even if the domain is being correctly translated. After each convolutional layer of the discriminators, we adopted a layer of spectral normalization [152] to further stabilize training. The only difference between the two discriminators regards

the final layer. The first one involves kernel size 3, and stride and padding 1. The other one, instead, varies the dimension depending on the number of target modalities and the kernel depending on the image size. The Shape Controller, instead, is built with convolutional blocks introduced before, while different from the discriminator the upsampling is done via transposed convolutions. Finally, the Contrast Network is composed of a convolutional layer, a max pooling, another convolutional layer, and finally three fully connected layers. We pass the fake image in this model, that predicts the desired values mentioned above. We show an overview of the complete architecture in Fig. 5.1

5.2.1 Experimental evaluation

Dataset

To test our image modality translation task we utilize the DroneVehicle [147] dataset. It contains $56,878$ samples taken by drones divided between half RGB and half infrared image. The data comes with annotations with oriented bounding boxes, labeled with five different categories ordered respectively by a number of occurrences in the dataset: car, truck bus, van, and freight car. The image size is 840×712. The difference between RGB and IR images is in the number of channels for our analysis, considering the IR as grayscale images with only one channel. For instance, IR images are divided into near-infrared, closer to the visible spectrum, and thermal infrared, which is useful in dark light conditions of the emitted energy from objects. The dataset we consider is composed of the latter type of IR. For our experiments we employ all the data resizing images at 256×256, obtaining approximately 17000 samples for the training dataset, and 3000 for the validation one. The only preprocess we do is the removal of the darkest samples, around 2000 samples. The images are normalized between -1 and 1. For our analysis is essential to have the segmentation masks of the targets, for this aim we built a dataset of masks starting from the bounding boxes provided by the original dataset. The segmentation masks are divided also into different colors depending on the class of the target, it is not of interest to our model, but can be for future work involving this dataset.

Metrics

To perform an objective evaluation, we compute several metrics: the Fréchet inception distance (FID) [98], the Inception Score (IS) [41], the structural similarity index (SSIM) [149], and the PSNR. To evaluate the results of segmentation instead, we use the Dice similarity coefficient (DSC) [153], the Segmentation score (S-Score) [154], and the mean absolute error (MAE). DSC quantifies how much two sets are similar averaging the size of their intersection. S-score sees how the shape of a segmentation behaves after the translation in another domain. For each fake segmented image, S-score is computed with the DSC between the generated segmented image and the ground truth of the real image. $DSC(SegNet(s), g)$ where s is the generated image, and g is the original image. Here $SegNet$ is the segmentation part of the generator, so given in input the yet translated image we are taking only the segmented image as output and then evaluating with the ground truth segmented image.

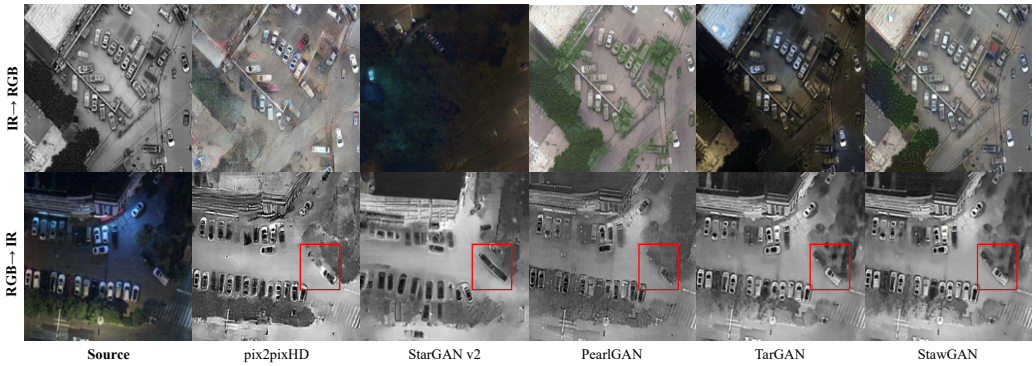


Figure 5.2. Random samples of StawGAN image modality translations on the DroneVehicle dataset. The first column refers to the original domain images, IR in the first row, and RGB in the second. The corresponding translations in the RGB domain are in the first row and subsequent columns, while IR is in the second one. Our approach produces better samples in both domains with respect to other state-of-the-art models.

Experimental Setup

StawGAN is implemented with PyTorch and trained on an NVIDIA Tesla V100 GPU, experiments last two and a half days. We tune different hyperparameters to successfully train this model. We set the lambda of the cycle loss $\lambda_{\text{cycle}} = 10$, the one of the cross loss to $\lambda_{\text{cross}} = 25$, while all the others are equal to 1. We start to involve the DSSIM loss only after epoch 10. There are different learning rates for the discriminators and the other networks. The optimizer we use is Adam [155] with $(\beta_1, \beta_2) = (0.5, 0.9)$. The learning rate for the discriminator is 0.0003 while for the other models is 0.0002, both during the training start to decrease linearly to zero after half of the epochs, which is set to 50. The batch size is set to 12. Due to the faster convergence of the target flow of the network, we train it only during even epochs number.

Results

Translation We compare our method with different models that reached state-of-the-art performance on other datasets. We test pix2pixHD [156] used for unpaired image-to-image translation, StarGAN v2 [157] interesting for the characteristics of extracting the style code from the scene and not injecting it with a fixed label, PearlGAN [144] introduced for this specific task of translation from IR to RGB and finally the original version of TarGAN. Figure 5.2 shows the comparison on randomly chosen samples. For the translation $IR \rightarrow RGB$ the results of PearlGAN are visually poor considering because of the green color of the buildings, supposing that it interprets them as vegetation. Regarding the translation $RGB \rightarrow IR$ the results of StarGAN v2 are slightly blurred, mostly on the targets, PearlGAN does not translate correctly all the cars, and some of them are ghosted. TarGAN and pix2pixHD samples are visually comparable to our results even though our method produces more realistic colors, but still, if we focus on the red boxes in Fig.5.2, we notice a concrete difference. From RGB to IR pix2pixHD loses the details of the target car, while TarGAN shows black holes and paler. For the opposite translation, both the color and the shape of the StawGAN image are more appealing with respect to pix2pixHD. Furthermore, TarGAN loses information on the targets, since the image is darker, making it difficult to distinguish the shape and the type of vehicle.

In the end, the proposed StawGAN produces high-definition samples with realistic colorization and well-shaped and defined targets. Table 5.1 reports objective metrics results for the image modality translation task. From this analysis, we can confirm that the FID of pix2pixHD is higher with respect to our approach that scores the second-best value, but our method surpasses it in all the other metrics. Those results confirm our theoretical statements, and prove the effectiveness of the proposed method with respect to the state of the art, clearly showing the significance of the proposed StawGAN.

Table 5.1. Quantitative comparison on image modality translation.

Model	FID↓	IS↑	PSNR↑	SSIM↑
pix2pixHD [156]	<u>0.0259</u>	**4.2223**	11.2101	0.2125
StarGAN v2 [157]	0.4476	2.7190	11.2211	<u>0.2297</u>
PearlGAN [144]	0.0743	<u>3.9441</u>	10.8925	0.2046
TarGAN [146]	0.1177	3.4285	<u>11.7085</u>	0.2382
StawGAN	**0.0119**	3.5163	**11.8251**	**0.2453**

Segmentation We perform also a secondary analysis on the segmentation capability of the proposed model with respect to the TarGAN, which is the only model from the one we compare that performs segmentation too. The comparison shows that our model gains better performance with respect to the TarGAN. Since segmented images have in both domains large black zones, the translation task results easier to be accomplished. Therefore, the segmentation metrics are evaluating mainly the generation of concrete target shapes. In Tab. 5.2 we report objective metrics results for this task.

5.3 GAN Limitations, and the Shift to Diffusion

In this chapter, we introduced StawGAN, an image translation model that enhances cross-domain synthesis through structural consistency. The proposed method generates corresponding images in IR or RGB domains from input images while utilizing masked targets and exploiting shared connections in the generator architecture, thereby improving translation quality. This neural approach strengthens edge preservation in translations, making it particularly valuable for downstream tasks such as object detection or classification. Importantly, our method demonstrates flexibility by operating effectively with both paired and unpaired datasets.

5.3.1 Successes and Contributions

Our experimental results on the DroneVehicle dataset [147] demonstrate that StawGAN achieves state-of-the-art performance across multiple evaluation metrics, validating the effectiveness of structural awareness in cross-domain translation. The integration of DSSIM loss successfully enforces spatial smoothness while preserving critical structural details, addressing a key challenge in infrared-to-RGB translation. The dual-flow architecture, processing both original images and segmented targets through interconnected pathways, represents a meaningful advancement in target-aware generation.

These achievements underscore the remarkable capabilities of GANs in complex image-to-image translation tasks. The ability to learn sophisticated mappings between vastly different visual domains, from thermal infrared signatures to natural RGB representations, demonstrates the fundamental power of adversarial training paradigms.

Table 5.2. Quantitative comparison on segmentation.

Model	DSC↑	S-Score↑	MAE↓
TarGAN [146]	79.34	85.08	0.0115
StawGAN	**84.27**	**88.61**	**0.0081**

5.3.2 Inherent Limitations and Emerging Challenges

However, this investigation also illuminates several fundamental limitations inherent to GAN-based approaches that extend beyond our specific application. During extensive experimentation, we observed recurring challenges that plague the broader GAN framework: training instability requiring careful hyperparameter tuning, occasional mode collapse leading to limited output diversity, and difficulties in achieving fine-grained control over generated content quality.

More critically, the adversarial training process, while powerful, often struggles with consistent generation of high-frequency details, a limitation particularly evident in cross-domain scenarios where structural fidelity is paramount. The discriminator's binary classification objective, though effective for overall realism, lacks the nuanced guidance necessary for preserving intricate spatial relationships and fine-textured regions.

Additionally, GANs exhibit limited controllability during inference, making it challenging to guide generation toward specific desired characteristics without extensive architectural modifications or additional conditioning mechanisms. This constraint becomes particularly problematic when dealing with diverse target domains or when seeking to balance multiple competing objectives simultaneously.

5.3.3 The Bridge to Diffusion

These observed limitations align with broader challenges recognized throughout the generative modeling community, prompting a significant paradigm shift toward diffusion-based approaches. Unlike the adversarial framework's winner-takes-all dynamics, diffusion models offer inherently more stable training through their progressive denoising formulation, eliminating concerns about discriminator-generator equilibrium.

The iterative refinement process characteristic of diffusion models provides superior control over generation quality and enables more precise conditioning mechanisms, addressing the fine-detail control limitations we encountered with GANs. Furthermore, the probabilistic foundation of diffusion models offers theoretical guarantees about coverage and mode representation that adversarial approaches cannot provide.

This recognition of GAN limitations, coupled with the emerging promise of diffusion models, motivated our subsequent research direction. The challenges encountered in achieving consistent structural preservation and fine-grained detail control in cross-domain translation directly informed our exploration of diffusion-based methodologies, as detailed in the following chapters.

The transition from adversarial to diffusion-based approaches represents more than a mere architectural choice, it reflects a fundamental shift toward more stable, controllable, and theoretically grounded generative modeling. This evolution sets the stage for our comprehensive investigation into diffusion model applications and, ultimately, their enhancement through wavelet-based innovations that address the remaining challenges in high-fidelity image synthesis.

Chapter 6

The Potential of Diffusion Models: Domain-Specific SR and Cross-Modal Generation

Contents

6.1 The Diffusion Paradigm Shift 60
6.2 Diffusion Models for Ship-Image Super Resolution . . . 61
6.2.1 ShipSpotting Dataset 63
6.2.2 Experimental Results 64
6.2.3 Ablation Study on Downstream Tasks 65
6.3 Streamlined EEG-to-Image Generation with Latent Diffusion Models 67
6.3.1 Problem formulation 67
6.3.2 Conditioning 68
6.3.3 Training 69
6.3.4 Experiments 70
6.4 Conclusion: Achievements and the Path Forward 73
6.4.1 Key Achievements 73
6.4.2 Identifying Critical Limitations 73
6.4.3 The Motivation for Multi-Scale Representations 74

HAVING demonstrated both the capabilities and fundamental limitations of GAN-based approaches in cross-domain translation, our research trajectory naturally evolved toward exploring the emerging paradigm of diffusion models. While GANs had shown remarkable potential, their inherent training instabilities and mode collapse issues motivated the investigation of more stable and controllable generative frameworks. Diffusion models, with their progressive denoising formulation and superior training dynamics, presented a compelling alternative for tackling challenging domain-specific applications.

This chapter[1] examines two distinct yet complementary investigations that showcase the versatility and power of diffusion models across vastly different application domains. First, we explore their application to maritime image super-resolution, a

[1]The content of this Chapter is adapted from the material published in [7, 29]

specialized domain where traditional approaches struggle with the unique characteristics of ship imagery. Subsequently, we push the boundaries further by investigating diffusion models' potential in the even more challenging cross-modal domain, attempting to translate complex neural signals directly into visual representations.

6.1 The Diffusion Paradigm Shift

Diffusion Models for Specialized Super-Resolution

Our exploration begins with maritime surveillance applications, where traditional super-resolution approaches face unique challenges. Unlike conventional domains that primarily focus on natural or face images [89,158], maritime surveillance presents several critical factors including long distances, limited sensor capabilities, adverse weather conditions, and the inherent mobility of ships that prevent the acquisition of high-quality images [159–162].

The enhancement of low-resolution ship images reveals potential defects, anomalies, or critical details that might be missed in original pictures, enabling more effective identification and quality control measures. These aspects prove critically important for applications including ship detection [163], classification [164], and tracking [165]. The complexity arises from the wide variety of ship categories, each possessing distinct characteristics, from cargo ships capable of transporting different container types to vessels exhibiting diverse small and highly detailed elements requiring enhancement, including portholes, railings, crests, text, and armaments.

The super-resolution process must yield high-quality results to mitigate risks of misidentification, distinguishing, for instance, between containers and armaments. Unlike natural images where smooth edges are expected, ships comprise steel and other components with sharp, well-defined features that must be preserved to minimize artifacts in the super-resolution process.

Generative models, particularly diffusion models [51], have emerged as promising solutions by effectively enhancing low-resolution images while preserving visual fidelity [166]. Their ability to hallucinate missing details from low-quality source images has proven beneficial for image-to-image translation problems [28, 167–170]. The unsupervised nature of these models allows enhancement without relying on large annotated datasets [87, 89, 158], with various studies demonstrating diffusion-based single-image super-resolution across different input types, including hyperspectral images [171].

Cross-Modal Neural-to-Visual Translation

Building upon our understanding of diffusion models' capabilities in specialized image domains, we then ventured into the significantly more challenging realm of cross-modal generation. The advancement of Brain-Computer Interfaces (BCIs) through understanding how the human brain represents the world represents a central challenge in neurocognitive research, with potential to revolutionize healthcare, education, and entertainment [172].

While early neurocognitive works demonstrated that brain waves retain information about object structures from visual cues [173], the reconstruction of visual stimuli from brain signals had previously stagnated due to limitations of earlier generative methods such as GANs. The emergence of diffusion models has reinvigorated this field, though most successful approaches have focused on functional MRI (fMRI) signals [174–176] due to their high spatial resolution.

However, fMRI's cost-prohibitive equipment and lack of portability make it unsuitable for real-time BCI systems [177]. EEG offers a compelling alternative, capturing electrical brain activity through scalp electrodes with high temporal resolution while remaining portable, non-invasive, and low-cost [178]. Yet EEG presents inherent challenges: high susceptibility to noise resulting in very low signal-to-noise ratios, with artifacts frequently caused by electrode misplacement or body movement.

Existing EEG-to-image studies face three main limitations: extensive preprocessing requiring domain expertise [179], reliance on outdated generative models like GANs [180, 181], and complex frameworks combining various alignment losses [182], pretraining on large datasets [183], and captioning networks [177].

Chapter Overview and Contributions

This chapter presents two complementary investigations that demonstrate diffusion models' remarkable adaptability. Our StableShip-SR approach leverages pre-trained foundation models like Stable Diffusion [12], introducing a novel class- and time-aware encoder that exploits prior knowledge from text-to-image models for maritime super-resolution. Simultaneously, our streamlined EEG-to-image framework utilizes ControlNet [27] for the first time in EEG-based image generation, requiring minimal preprocessing while surpassing state-of-the-art methods.

These investigations serve a dual purpose within our research narrative: they demonstrate the superior capabilities of diffusion models over GANs in challenging scenarios while simultaneously revealing the computational and representational limitations that motivate our subsequent exploration of wavelet-enhanced approaches. The success and remaining challenges observed in these applications directly inform the methodological innovations presented in the following chapters.

6.2 Diffusion Models for Ship-Image Super Resolution

Motivated by the research presented in [184], our proposed model capitalizes on the foundational knowledge derived from a pre-trained Stable Diffusion model [12], based on a U-Net as a denoiser to perform the latent diffusion process.

A conditional generation process can be directly derived from an unconditional generation one by using direct conditioning in the diffusion process, as done in [90]: the latent representation of the degraded input image obtained from an encoder is then used as guidance for the generation process

Moreover, we exploit the text conditioning given by the design of Stable Diffusion, using CLIP [185] to encode the text and provide embeddings to our model. We use this representation as an additional conditioning of the denoising process, but only in the training phase. This is necessary to allow the network to be more responsive to the textual information represented by the class and name of the ship, since they are known a priori, being present as metadata in our dataset. This conditioning is not provided in the inference phase, as we want this information to be derived online through the classifier applied to low-resolution images, as analyzed below. We chose in advance a set of five different parametric prompts to improve the generation process of ship images through the textual information. We randomly pick one of the prompts during the training process and populate it with consistent information about the ship to be processed, i.e. the ship name and the category.

The core of our proposed architecture is the class- and time- aware encoder. Indeed, it provides the diffusion model with additional information on the class of



Figure 6.1. Comparative sample images showcasing the super-resolution capabilities of different models. The visual assessments highlight variations in image quality, clarity, and detail enhancement achieved by each respective model in the upscale process. Ours is in the last column.

the ship represented in the image to be reconstructed. This information is extracted by a classifier directly from the low-resolution image to be super resoluted.

As a classifier, we utilize a ResNet-50 [186] model and finetune it on our proposed dataset ShipSpotting. This pre-training is done before the training of StableShip-SR, consequently, this part of the architecture is frozen during the whole training. The classifier training is performed using as input high-resolution images to improve the accuracy of the prediction since it will affect the generation of the images during the denoising process.

We employ the class- and time- aware encoder output to modulate the intermediate feature maps within the residual blocks of the U-Net through spatial feature transformations (SFT) [187]. This slight modification is essential to obtain higher-quality images than other state-of-the-art super-resolution models, as proven in

experimental results. Following the proposed method in [184], we also integrate the temporal information in the encoder, enhancing the overall qualitative result of the generated images. Indeed, at early timesteps of the process, when the generated image is still poor in terms of information and consequently the resulting signal-to-noise ratio is low, having stronger conditioning allows the diffusion process to have more helpful guidance. This gives a better understanding of what information needs to be reconstructed, as also in-depth explored in [188]. By contrast, at higher timesteps when the diffusion process has already reconstructed much of the image information, i.e., when you have higher signal-to-noise ratio values, the conditioning given by the low-resolution image may be softer, leaving the diffusion process with the finest guidance. Such time awareness is permitted by the fact that the noise schedule is a hyperparameter set beforehand so that the signal-to-noise ratio values at each timestep of the training phase are known. Besides the aforesaid temporal conditioning allows balancing the use of the low-resolution image in the diffusion process.

We concatenate the two tensors of class and timestep embeddings as shown in Fig. 6.5 resulting in one vector $b \in \mathbb{R}^{1024}$ and then we encode it with the latents of the image $z \in \mathbb{R}^{h \times w \times c}$ obtained by the encoder $z = \mathcal{E}(x)$, where $x \in \mathbb{R}^{3 \times H \times W}$ is the low-resolution image. The encoder $\mathcal{E}$ and the decoder $\mathcal{D}$ are frozen and are part of the VAE used in [12], as well as the neural backbone used to perform the latent diffusion process, i.e. a time-conditional U-Net. The class- and time- aware encoder δ_θ follows the same architecture as the U-Net encoder thus providing conditioning embeddings to the conditional denoising autoencoder ϵ_θ at different scales. The domain-specific encoder $\tau(y)$, in our case, is a frozen CLIP, where y is the text conditioning. The complete objective function we aim to learn is represented in (6.1).

$$L = \mathbb{E}_{\mathcal{E}(x),y,\epsilon\sim\mathcal{N}(0,1),t,c}\left[\|\epsilon - \epsilon_\theta\left(z_t, \delta_\theta(c,t,z), \tau(y)\right)\|_2^2\right]. \tag{6.1}$$

We optimize $\delta_\theta(c,t,z)$ where (c,t) are the classes and the timesteps embeddings respectively, z is the output of the VAE encoder while z_t is its noisy version.

6.2.1 ShipSpotting Dataset

Given the specific task of super resolution for ship images, we create a dataset from scratch, in this way, we can ensure that our super-resolution model is trained on ship images that are relevant and representative. It contains ship images that exhibit diversity in terms of ship types, sizes, orientations, and environmental conditions. This diversity is crucial for training an effective super-resolution model capable of handling various real-world scenarios. We devote significant attention to constructing a dataset that prioritized generality, ensuring its applicability beyond our specific use case. Our objective is to develop a dataset that does not impose unnecessary restrictions and can be effectively utilized for multiple purposes. The endeavor extended beyond gathering the necessary images for our super-resolution task since we carefully collected additional metadata, making this dataset suitable for other potential problems of interest, such as ship classification.

To construct such a dataset, a straightforward approach was scraping images from the web. The main source for our dataset is ShipSpotting[2], which serves as a repository for user-uploaded images, hosting a vast collection of ship images, amounting to approximately 3 million. Furthermore, for each image, valuable

[2] www.shipspotting.com

supplementary information is available, such as the type of the ship, and present and past names.

Next, we made sure that as many images as possible were collected in our dataset, since in deep learning, the quantity of training data directly influences the quality of results. A larger volume of data enables models to generalize more effectively. Thus we scrape all the images and as a result, the dataset comprises a total of 1.517.702 samples. Motivated by [189] we exclude many classes of ships from our final analysis and concentrate on the more common and valuable for a real scenario use case. The total number of different classes is 20 and the ship categories included are Bulkers, Containerships, Cruise ships, Dredgers, Fire Fighting Vessels, Floating Sheerlegs, General Cargo, Inland, Livestock Carriers, Passenger Vessels, Patrol Forces, Reefers, Ro-ro, Supply ships, Tankers, Training ships, Tugs, Vehicle Carriers, Wood Chip Carriers. The total amount of samples after this class selection is 507.918. To obtain the high-resolution version, we mainly employ a center crop technique, ensuring that the essential details of the image are preserved. Conversely, for the low-resolution counterpart, the high-resolution images were degraded with different techniques following the preprocessing steps given in [190], effectively reducing its resolution. Furthermore, to create a reference image for comparison with the output of the model, we apply a standard upscaling super-resolution technique to the low-resolution image. The obtained refined dataset is optimized and prepared for training the super-resolution models.

6.2.2 Experimental Results

We conduct a challenging super-resolution task of 8x scaling, going from 64×64 pixel low-resolution images to 512×512 pixel high-resolution images. We chose this scaling factor to evaluate the ability of the model to handle finer details and accurately recreate them in the output with a very small image as input.

Models were trained on a machine equipped with an NVIDIA Quadro RTX 8000. After completing the model training, we performed inference on a test set consisting of 1000 images. We analyze the objective metrics PSNR, SSIM, and FID.

Table 6.1. FID results obtained on the test set.

Model	FID ↓
Low Resolution (LR)	66.05
SwinIR [86]	50.79
SR3 [90]	16.59
StableSR [184]	12.41
StableShip-SR (Ours)	**11.72**

A low FID value indicates that the embedding representations of both images are similar, capturing not only general features but also finer details. The images produced by our model are overall more realistic compared to their counterparts, which explains the proximity of their embedding representations, as we can observe from Fig. 6.2 and results reported in table 6.1. By examining the images, we observe specific characteristics that provide insights into the results. The images generated by Swin-IR and SR3 have a cartoonish appearance, consisting of large areas of color with minimal detail. Each block of color has a well-defined shape, resulting in an artificial aesthetic for the overall image. SwinIR also struggles to accurately reproduce background elements such as water and trees, leading to a slight blurring effect that reduces overall quality.

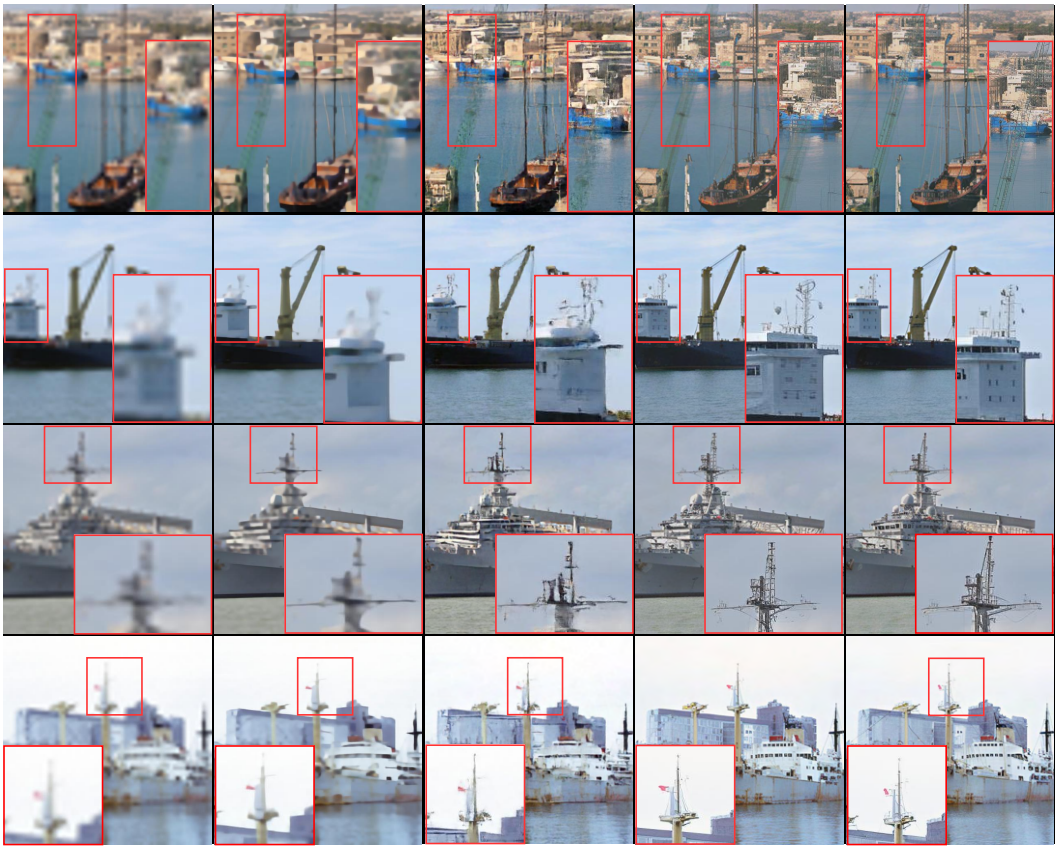


Figure 6.2. Comparison between LR image and SR images reconstructed by other state-of-the-art methods and our proposed model. The red bounding box contains a zoomed-in part of the image in order to have a better understanding of the image resolution quality of our method compared to the others.

In contrast, our approach can achieve a better reconstruction of background elements and details, such as portholes and antennas as shown in the third row of Fig. 6.2. Moreover, StableShip-SR produces images without the blurred effect present in other approaches. This is likely due to a more realistic approach to color reconstruction, capturing changes in lighting conditions and reflections on water surfaces. It is important to note that the reconstruction of StableSR is not perfect, and careful examination of smaller details reveals artifacts, as we can observe from the first row of 6.2.

Another drawback of SR3 is its long inference time, taking $\approx$ 2 minutes to generate each sample compared to the other models few seconds, this is due to the diffusion process being done on the pixel space and not in the latent one.

With this analysis we prove StableShip-SR to be a superior model for this specific task, providing higher-quality high-resolution images and more realistic representations compared to the others.

6.2.3 Ablation Study on Downstream Tasks

To show the effectiveness and the need for super-resolution images to drastically improve downstream tasks, we conduct experiments in object detection and

Table 6.2. Overall precision and recall on Seaships dataset. The detection is done by a finetuned YOLOv7 [1] on Seaships.

Model	Precision ↑	Recall ↑
Low Resolution (LR)	0.536	0.313
SwinIR [86]	0.536	0.313
SR3 [90]	0.569	0.316
StableSR [184]	0.615	0.415
StableShip-SR (Ours)	**0.647**	**0.439**

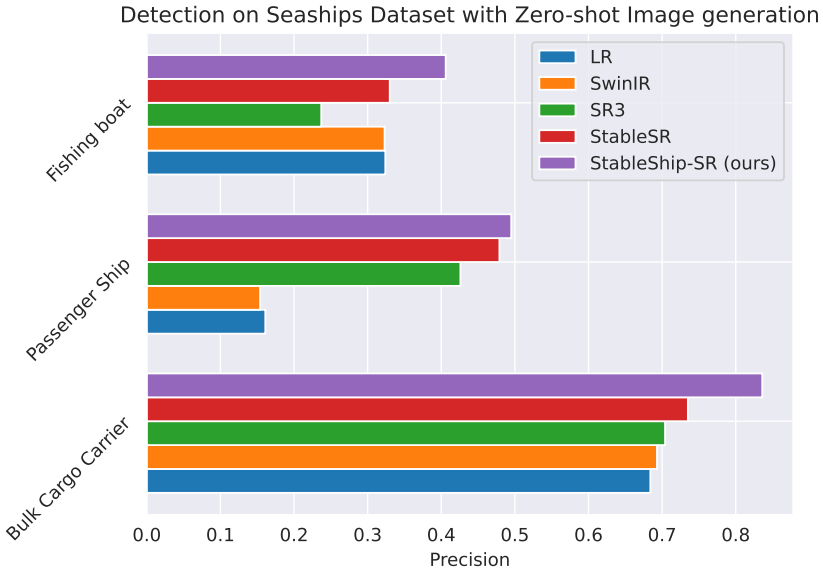


Figure 6.3. Comparison for precision scores on the Seaships dataset. We employ the models pre-trained on our dataset, thus performing zero-shot super resolution, improving the detection of a finetuned YOLOv7 [1] on Seaships.

classification tasks with different images.

We employ pre-trained neural models in order to have only to fine-tune them on ship datasets.

Since SeaShip, the ship dataset introduced by [163], also provides bounding boxes, we can test our super-resolution approach in the object detection task, we choose the YOLOv7 [1] architecture as the object detector, moreover with this model, it is also possible to perform classification tasks.

Since we are also interested in evaluating the model on our dataset based on ShipSpotting, we fine-tuned the well-known ResNet architecture [186] on ship images as the classifier.

The results for the detection and classification task performed on the SeaShips dataset are reported in Fig. 6.3. SwinIR [86] achieves higher PSNR and SSIM values, indicating a higher fidelity between the original and reconstructed images. However, we performed many experiments that point out that those results are useless when it comes to using those images for any other subtasks. Moreover, when considering the FID metric (lower is better), diffusion models outperform SwinIR, and this is outlined also by a subjective evaluation of the results reported in Fig. 6.2. These results may seem counterintuitive since we expect that if the original and generated images were similar with low perceptual differences, their embedding

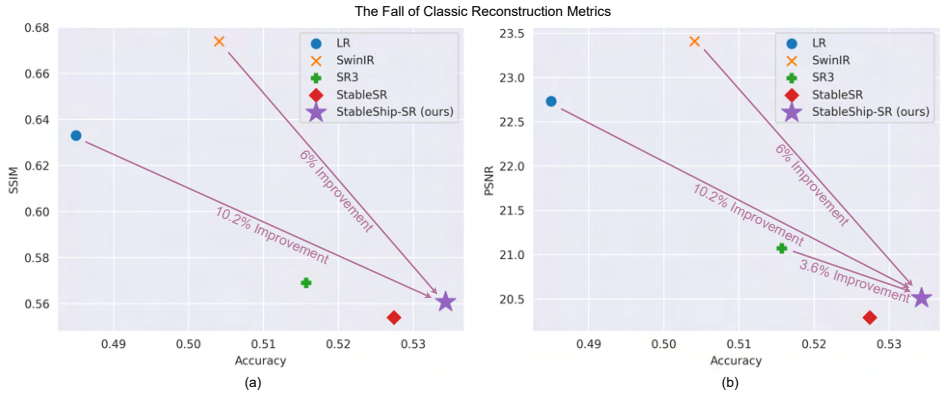


Figure 6.4. Comparison of SSIM (a) and PSNR (b), which measure the quality of the reconstructed images. We confront those results against the classification accuracy reached with the generated images on the ShipsSpotting dataset. This comparison shows how the classical reconstruction metrics are not a measure of how well a model performs in a super-resolution scenario since the accuracy and the FID reached are lower concerning other models that have lower SSIM and PSNR.

representations computed by FID would also be close. However, this is not the case. This behavior aligns with the findings described in [90], where PSNR and SSIM may not accurately represent the quality of a generated image, because they focus on the exact comparison between pixels while generative models may change pixel content without changing their meaning. Indeed the graph in Fig. 6.4 shows how a higher PSNR does not bring benefit when it comes to a classification task. Furthermore, using an upsampled low-resolution image as an input for a ship classifier, we will obtain a worse result in terms of accuracy compared to the one obtained with the image with a lower PSNR used in our method.

6.3 Streamlined EEG-to-Image Generation with Latent Diffusion Models

We propose a simple framework, GWIT, comprising ControlNet [27] as a mechanism for controlling a latent diffusion model [12] using EEG signals. We chose ControlNet as it provides a highly effective and parameter-efficient method for injecting novel spatial or abstract conditioning into large, pre-trained diffusion models without requiring costly retraining of the entire foundation model.

6.3.1 Problem formulation

Let $\{\mathbf{y}, \mathbf{x}\}$ be a pair from the dataset, where $\mathbf{y} \in \mathbb{R}^{C \times L}$ is an EEG signal with C channels and L time steps, and $\mathbf{x} \in \mathbb{R}^{H_x \times W_x \times 3}$ is the corresponding image with height H_x, width W_x, and 3 channels. The overall framework involves processing the image $\mathbf{x}$ with the LDM in the standard manner, while the ControlNet adapter handles the input conditioning EEG $\mathbf{y}$, as it is able to adapt diverse modalities for conditioning [27, 175, 191]. Additionally, we add a coarse-grained control $\mathbf{c}_l$, i.e., a caption of the form "Image of [label]", following the approach of [179]. The label is predicted by a pretrained (frozen) EEG image decoder. The framework is then trained following the original ControlNet formulation, where the weights of the UNet encoder are copied and trained in an efficient way.

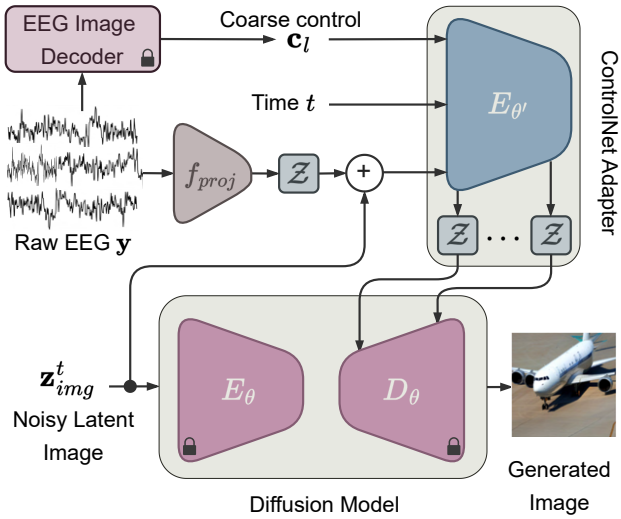


Figure 6.5. Guess What I Think (GWIT). Outline of our streamlined framework which includes a projection function, a ControlNet adapter to handle EEG conditioning, a frozen LDM, and a frozen EEG image decoder to obtain a coarse-grained control.

6.3.2 Conditioning

The image is processed by the (pretrained) LDM, where it is mapped into a latent code $\mathbf{z}_{img} \sim E^{\text{VAE}}(\mathbf{x})$ via the VAE stochastic encoder, with $\mathbf{z}_{img} \in \mathbb{R}^{H_z \times W_z \times D}$, where H_z and W_z are the spatial dimensions and D the number of latent channels. Then, noise is progressively added through a forward Gaussian process indexed by $t \in [0, T]$ to obtain $\mathbf{z}^t_{img}$. The EEG is mapped into the same latent space of $\mathbf{z}_{img}$ via a simple 1D convolutional neural network and reshape operations, i.e., $\mathbf{z}_{eeg} = f_{proj}(\mathbf{y})$, where $f_{proj} : \mathbb{R}^{C \times L} \to \mathbb{R}^{H \times W \times D}$. Then, $\mathbf{z}_{eeg}$ is passed thorugh a 1×1 zero convolution layer and summed with $\mathbf{z}^t_{img}$, i.e., the control input to ControlNet is $\mathbf{c}_{eeg} = \mathbf{z}^t_{img} + \mathcal{Z}(\mathbf{z}_{eeg})$, with $\mathcal{Z}$ the zero-initilized convolution. We apply zero convolutions following the approach of ControlNet, as they prevent harmful noise from affecting the LDM backbone during the initial training steps [27]. Then, the control input is processed by the ControlNet block. The adapter is defined as a trainable copy of the encoder of the underlying UNet architecture of the LDM which implements the diffusion model. That is, assuming the LDM with a UNet backbone formed by an encoder E_θ and a decoder D_θ, the ControlNet adapter is defined as an encoder with a trainable copy of the weights $E_{\theta'}$. Thus, given input image $\mathbf{x}$, the corresponding noisy latent code $\mathbf{z}^t_{img}$, and a set of conditions including the time step t, the coarse control $\mathbf{c}_l$ and the EEG control $\mathbf{c}_{eeg}$, they are processed by the ControlNet $E_{\theta'}$ adapter as:

$$\begin{aligned} \text{ControlNet}(\mathbf{c}_{eeg}, \mathbf{c}_l, t) &= E_{\theta'}(\mathbf{c}_{eeg}, \mathbf{c}_l, t) \\ &= E_{\theta'}(\mathbf{z}^t_{img} + \mathcal{Z}(\mathbf{z}_{eeg}), \mathbf{c}_l, t) \\ &= E_{\theta'}(\mathbf{z}^t_{img} + \mathcal{Z}(f_{proj}(\mathbf{y})), \mathbf{c}_l, t). \end{aligned} \tag{6.2}$$

When training on multiple subjects we include a linear layer $S(\mathbf{y}, s)$ that encodes information on the subject by feeding it the subject id $s \in \mathbb{N}$, as done in [191, 192]. In this case, the EEG is processed as $f_{proj}(S(\mathbf{y}, s))$.

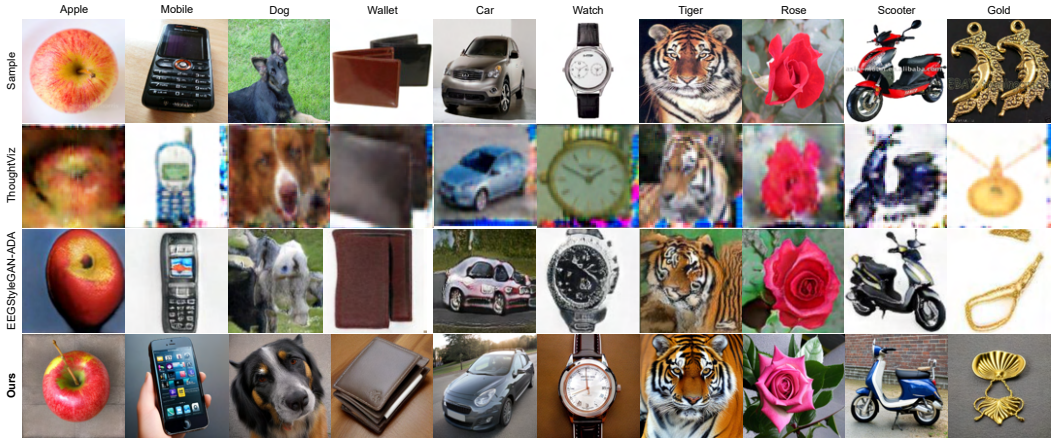


Figure 6.6. Comparison for images generated by models trained on ThoughtViz [2] dataset. At the top, we report a random sample of the same class from the original dataset.



Figure 6.7. Comparison with the previous state-of-the-art method for images generated by models trained on all subjects of EEGCVPR40. At the top, we report a random sample of the same class from the original dataset.

6.3.3 Training

The training loss used by ControlNet is identical to the original LDM loss, with two key differences: it includes an additional task-specific input condition, in this case, the EEG signal, and the UNet backbone remains frozen. Only the weights of the ControlNet adapter $E_{\theta'}$ and f_{proj} are updated as follows:

$$\mathcal{L} = \mathbb{E}_{\mathbf{z}^t_{img}, \mathbf{z}_{\text{eeg}}, \mathbf{c}_l, t, \epsilon \sim \mathcal{N}(0,1)} ||\epsilon - \epsilon_\theta(\mathbf{z}^t_{img}, \mathbf{z}_{\text{eeg}}, \mathbf{c}_l, t)||^2_2, \tag{6.3}$$

where ϵ_θ is implemented by the UNet backbone and learns to predict the progressively added noise to the image. In this way, it can decode the image using the VAE-based decoder, i.e., the final generated image is $\hat{\mathbf{x}} = D^{\text{VAE}}(\hat{\mathbf{z}}^0_{img})$, where $\hat{\mathbf{z}}^0_{img}$ is the denoised sample of $\hat{\mathbf{z}}^t_{img} \sim \mathcal{N}(0, 1)$.

Table 6.3. Generation quality and semantic correctness of models trained on EEGCVPR40 (top: subject 4, bottom: all subjects) and ThoughtViz.

	Model	IS ↑	FID ↓	ACC ↑
EEGCVPR40	DreamDiffusion [183]	—	—	0.45
	BrainVis [182]	31.52	121.02	0.49
	GWIT (Ours)	**33.32** (+6.34%)	**80.47** (-33.51%)	**0.91** (+85.71%)
	Brain2Image-GAN [193]	5.07	—	—
	NeuroVision [194]	5.15	—	—
	Improved-SNGAN [195]	5.53	—	—
	DCLS-GAN [196]	5.64	—	—
	NeuroImagen [177]	33.50	—	0.85
	EEGStyleGAN-ADA [181]	10.82	174.13	—
	GWIT (Ours)	**33.87** (+1.1%)	**78.11** (-55.14%)	**0.91** (+7%)
ThoughtViz	AC-GAN [197]	4.93	—	—
	ThoughtViz [2]	5.43	—	—
	NeuroGAN [198]	6.02	—	—
	EEG2Image [199]	6.78	—	—
	EEGStyleGAN-ADA [181]	9.23	109.49	—
	GWIT (Ours)	**13.76** (+49.07%)	**66.33** (-39.41%)	**0.787**

6.3.4 Experiments

Datasets

We evaluate our method with the EEGCVPR40 [200, 201] and ThoughtViz [202] benchmark datasets. The first contains EEG recordings from 6 subjects who were shown 50 images for each of 40 classes from the ImageNet dataset [203]. EEGs were recorded for 0.5s at 1 kHz following the 128-channel system. We used the official training, validation, and test sets. Instead, ThoughtViz contains recordings of EEG of 10s sampled at 128Hz with 14 electrodes from 23 participants. The samples were split into chunks of 32 time steps with overlap [2]. We employed the subset of EEGs relative to images ranging in 10 classes. For preprocessing, we apply only standardization following [181] and directly employ raw EEG data.

Implementation

For the projection f_{proj} we use 1D convolutional layers with (320, 640, 1280, 2560) channels and strides (5, 2, 2, 2). Finally, we apply padding and reshaping to map the EEG conditioning to the same dimension as the latent image. We employ Stable Diffusion [12] as LDM, using the weights of the 2.1 version, and the LSTM EEG imade decoder proposed in [181]. The model, i.e., the ControlNet adapter and f_{proj} are trained for 100 epochs with Adam and a learning rate of $1e-5$. During training, we drop the coarse control for half of the samples to further emphasize the EEG conditioning [27]. Lastly, for sampling the generated image we employ the guess mode of ControlNet which enforces the model to prioritize the EEG conditioning over the coarse control [27].

Metrics

We evaluate the generated images in terms of generation quality with Fréchet inception distance (FID) and inception score (IS). Moreover, we evaluate the semantic



Figure 6.8. Comparison of images generated with models trained on subject 4 (top three rows) and on all subjects (bottom four rows) of EEGCVPR40.

accuracy of generated images with N-way Top-k classification accuracy (ACC) [174, 177, 182]. This metric evaluates whether the original visual cue and the generated image are assigned to the same class by a pretrained ImageNet classifier, i.e., a ViT [182]. Finally, we employ the Learned Perceptual Image Patch Similarity (LPIPS) to measure the similarity between original and generated images.

Results

For the EEGCVPR40 dataset, we train two variants of our method, i.e., a single and multi-subject model. For the first, the model is trained on EEGs corresponding to subject 4 [182, 183]. Instead, the multi-subject model is trained on the whole dataset, as is done in the ThoughtViz experiments. The quantitative evaluation of these results is reported in Tab. 6.3. Our streamlined approach GWIT achieves state-of-the-art results in every scenario, in both generation quality and semantic correctness. In particular, we improve the semantic accuracy by a great margin, i.e., by 85.71% and 7% for single and multi-subject variants, respectively. This result is highly significant as this metric directly measures if the images generated from EEG signals are semantically correct with respect to the original visual cues, which is the most crucial aspect. Finally, in Fig. 6.8 and Fig. 6.9 we present the qualitative results on EEGCPVR40, while in Fig. 6.6 we show results on ThoughtViz. The images generated with our method demonstrate superior quality compared to other models. Moreover, we attain semantic accuracy with respect to original visual cues

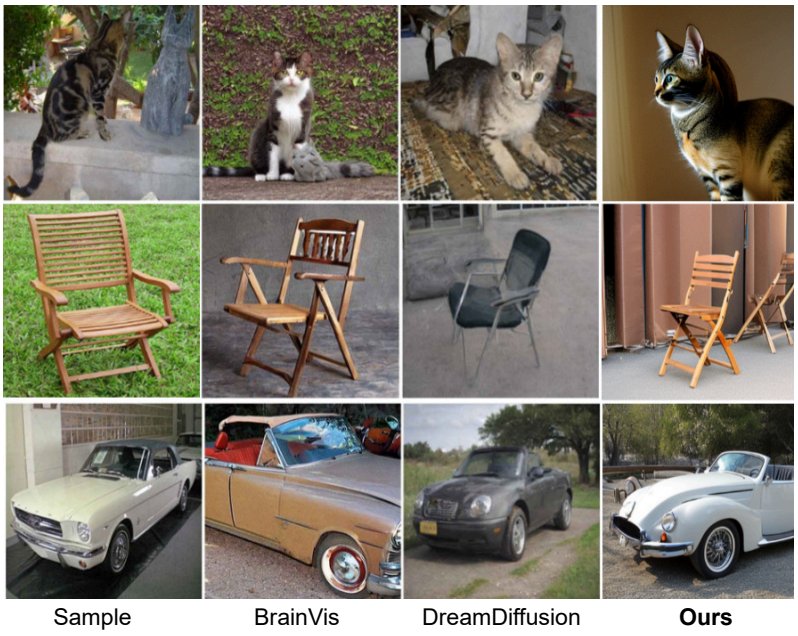


Figure 6.9. Comparison of images generated with models trained only on subject 4 of EEGCVPR40. Left: random sample from the dataset.

Table 6.4. Ablation study for influence of EEG conditioning.

Model	EEG Control	LPIPS ↓
GWIT-Only coarse	✗	0.811
GWIT-Subject 4	✓	0.772
GWIT-All Subjects	✓	0.770

as observed by comparing the generated images with samples of the original dataset in Fig. 6.6 and Fig. 6.9.

Ablations

In Tab. 6.4 we conduct an ablation study to investigate the influence of the EEG control. Specifically, we demonstrate that images generated with EEG conditioning reach a lower LPIPS, i.e., they are closer to the “ground truth" image, compared to images generated with only coarse control. This shows that the EEG actually guides the diffusion model to generate images semantically closer to the original visual stimuli. Lastly, in Tab. 6.5 we investigate how the “drop" and “guess" modes utilized during training and inference respectively allow to prioritize the EEG conditioning. Lower LPIPS and higher accuracy indicate that the model is actually following the EEG conditioning, as they are related to distance and semantic correctness with respect to the original visual cue. We find the guess mode to be much more influential than the drop mode, however also the latter leads to an improvement.

Table 6.5. Ablation study for “drop" (during training) and “guess" (at inference time) modes on the multi-subject variant.

Model	Drop	Guess	LPIPS ↓	ACC ↑
GWIT	✗	✗	0.774	0.59
	✗	✓	0.770	0.90
	✓	✗	0.781	0.71
	✓	✓	0.771	0.91

6.4 Conclusion: Achievements and the Path Forward

This chapter has demonstrated the remarkable versatility and effectiveness of diffusion models across diverse domain-specific applications. Through our investigations into ship image super-resolution and EEG-to-image generation, we have established that carefully tailored diffusion approaches can achieve state-of-the-art performance in highly specialized scenarios.

6.4.1 Key Achievements

Our StableShip-SR model represents a significant advancement in maritime image enhancement, consistently producing images with heightened realism that align closely with human perceptual capabilities. The comprehensive evaluation across standard and non-standard metrics, including downstream task performance, validates the superiority of our approach for ship super-resolution applications. The introduction of a meticulously curated dataset containing over 500,000 samples across 20 distinct ship classes provides a valuable resource for the research community and industry applications.

Similarly, our streamlined EEG-to-image generation framework demonstrates the power of diffusion models in cross-modal synthesis. By leveraging ControlNet adapters with minimal preprocessing and efficient fine-tuning, we achieved state-of-the-art performance while moving toward real-time brain-computer interface applications, a significant step beyond the complex frameworks typically required for such challenging tasks.

These successes underscore the fundamental strengths of diffusion models: their stable training dynamics, superior mode coverage, and exceptional conditioning capabilities make them ideally suited for complex, domain-specific generation tasks.

6.4.2 Identifying Critical Limitations

However, our investigations have also revealed several persistent challenges that limit the full potential of current diffusion-based approaches. While effective, achieving high resolutions remains computationally prohibitive with standard diffusion architectures. The iterative denoising process, though stable, scales poorly with resolution, creating a significant barrier to practical high-fidelity applications.

Furthermore, despite achieving impressive overall quality, capturing fine-grained textures and multi-scale details in super-resolution tasks can still be substantially improved. Our ship SR experiments revealed that while global structure and realism are well-preserved, intricate surface details and complex geometric patterns often lack the sharpness and precision demanded by critical applications such as maritime surveillance or automated inspection systems.

Additionally, handling diverse data types and improving representational power remains a key challenge for complex conditioning scenarios. The EEG-to-image generation task highlighted the difficulty of creating robust representations that can effectively bridge vastly different modalities while maintaining generalization across subjects and experimental conditions.

6.4.3 The Motivation for Multi-Scale Representations

These challenges point toward a fundamental limitation in current diffusion model architectures: their predominantly pixel-space or simple latent-space operations lack the sophisticated multi-scale awareness necessary for optimal performance across resolution scales and texture complexities. The computational bottleneck in high-resolution synthesis, the difficulty in preserving fine-grained details, and the challenges in handling diverse conditioning modalities all suggest a common underlying issue: the need for more efficient and expressive representational frameworks.

The challenges identified here, from the computational cost of generating 512×512 ship images to the subtle artifacts remaining on ship hulls (as seen in Figure 6.2) and the noisy, multi-scale nature of EEG signals, all point to the need for a more fundamentally efficient and expressive representation. The research presented in the following part III will argue that wavelet transforms provide exactly this. We will systematically investigate how their integration can overcome the computational and representational constraints of standard diffusion models, enabling higher-fidelity and more controllable generation across diverse applications.

Part III

The Wavelet Revolution in Generative Modeling

Chapter 7

Learning to Generalize with Wavelet-Driven Representations

Contents

7.1 Introduction 76
7.2 Grouping and Selecting Quaternion Wavelet Bands . . . 78
7.2.1 Experimental Evaluation 80
7.3 Quaternion Wavelet Feature Extraction 82
7.3.1 Pretraining 83
7.3.2 Architecture 84
7.3.3 Experimental Setup 85
7.3.4 Results 90
7.4 Conclusion 91

The preceding chapter explored the capabilities of Latent Diffusion models for domain-specific image generation, particularly focusing on high-resolution reconstruction and cross-modal synthesis such as EEG-to-image generation. While powerful, such approaches often benefit from robust and informative input representations, especially when tackling challenges of fidelity, domain adaptation, and data scarcity inherent in generative modeling. This chapter[1] transitions to investigate the potential of wavelet-driven representations, specifically leveraging the mathematical elegance of Quaternion Wavelet Transforms (QWT), to foster enhanced generalization in learning systems.

7.1 Introduction

As introduced in Chapter 4, wavelets offer rich multi-scale representations, and quaternion algebra provides a framework for holistically processing multi-channel data, making their combination a promising avenue for developing sophisticated feature extractors.

The challenge of generalizability in neural networks, particularly when faced with data from diverse sources or acquired under varying conditions, remains a significant research frontier [204]. This issue is pronounced in domains where methodological

[1]The content of this Chapter is adapted from the material published in [30, 31]

standards are still evolving, leading to substantial dataset variability. While deep learning has made substantial inroads in various analytical tasks [205–208], the capacity of models to generalize beyond their training distribution without extensive retraining is crucial for their widespread adoption and practical impact. Often, images, especially those rich in detail, undergo preprocessing steps like downscaling due to computational constraints, potentially discarding critical information [209]. This underscores the need for methods that can extract salient and comprehensive features from the outset.

In this context, techniques that can distill essential information from raw data are invaluable. Wavelet transforms, by decomposing signals into different frequency sub-bands, provide both approximations (low-frequency) and fine-grained details (high-frequency) [210, 211]. The Quaternion Wavelet Transform (QWT) extends these benefits, offering improved shift-invariance and a richer set of analytical components [212, 213]. The adoption of quaternion algebra itself is proliferating in neural network design, recognized for its efficacy in preserving multidimensional data structures [214–216].

This chapter delves into two distinct but related research contributions that harness the power of QWT to construct highly informative and generalizable image representations. These frameworks are demonstrated within the demanding context of medical image analysis, which serves as an ideal and challenging testbed for general-purpose methods. The high demand for detail fidelity, the presence of subtle yet critical textures, and significant variability across acquisition modalities are all challenges that a truly robust feature extractor must overcome, making it a perfect domain to validate techniques intended for broader applicability. The core innovation of these frameworks lies in their task- and model-agnostic nature, suggesting their potential to enhance various downstream applications, including those discussed in subsequent chapters on super-resolution and high-fidelity synthesis.

The first framework [30] focuses on producing salient features by systematically grouping and selecting QWT sub-bands. This process aims to identify the most enlightening set of bands by learning their statistical dependencies to avoid redundancy and by scoring their informativeness. The resulting representative features can then serve as a more complete input to any neural architecture.

The second framework [31] introduces a novel approach that integrates QWT feature extraction with a neural learning component. It first extracts diverse sub-bands via QWT and then employs a neural model to weigh and select the most representative set of these features. This creates an enhanced input representation suitable for various image processing tasks.

The primary aim of this chapter is to showcase how these QWT-driven strategies can lead to representations that are not only rich in information but also inherently more generalizable across different tasks, models, and potentially data domains. This exploration lays further groundwork for leveraging wavelet-based techniques in advanced generative models, as will be explored in Chapters 8 and 9.

The rest of the chapter is organized as follows: Section 7.2 describes the GROUSE framework and its statistical feature selection process. Section 7.3 introduces the QUAVE architecture and its integration with real and quaternion-valued networks. Finally, Section 7.4 summarizes our findings and sets the stage for wavelet-based generative models introduced in the next chapters.

7.2 Grouping and Selecting Quaternion Wavelet Bands

A common approach in literature for selecting real-valued wavelet features is involving low-frequency sub-bands that have been proved to improve classification accuracy [125]. A similar method proposes to compute the energy for each sub-band and then get the top k features with the highest energy [217]. However, these methods aim at selecting separately the more informative or discriminative sub-bands, without exploring whether they contain redundant information or enough details when combined together. Indeed, while low-frequency bands may be the most informative and obtain the highest energy values, they may miss crucial details and high-frequency characteristics suitable for proper signal analysis and synthesis. Moreover, previous approaches implicitly assume independence among sub-bands that does not hold in practice. For this reason, when choosing the best sub-bands to feed as input to the model, the selection method has to take into account both the informativeness of the single sub-band and the total information brought by the k sub-bands combined together. To this end, we take inspiration from an approach that builds clusters for real DWT tree sub-bands considering the dependence among them and then perform an evaluation of individual sub-bands in order to improve the accuracy of texture classification [218]. We propose a quaternion wavelet transform sub-band GROUping and SElection (GROUSE) approach that completely operates in a quaternion fashion while being easy to be incorporated into any real-valued analysis. The algorithm first discovers the structure of the statistical dependence among QWT sub-bands, split them in clusters based on the dependence value, then an intra-cluster evaluation of wavelet features is performed by computing the energy of each element, selecting the sub-band with the highest energy as a candidate for the final set. While the original approach [218] computes the DWT sub-band selection on the training set and then involves the discovered sub-bands for texture classification on the test set only, we insert a selection method in the framework that preprocesses each input image, and then, with the preprocessed data, we perform model training. This helps the model to take advantage of the QWT extracted features during training and consequently improve the performance at the testing stage. The potential of our approach is that it does not require labels or additional information, so it can be involved both in the training and testing stages, as preferred by the user.

Energy function

Different energy functions are involved in the literature, however, in order to be consistent with the method adapted for sub-band selection [218], we employ the squared magnitude, computed as:

$$E_{i,j}(\mathbf{s}_{i,j}) = \sum_{h=1}^{H} \sum_{w=1}^{W} [\mathbf{s}_{i,j,(h,w)}]^2 \tag{7.1}$$

whereby H and W are the height and width of the sub-band $\mathbf{s}_j$ of the image $\mathbf{X}_i$. In this way, the energy of a sub-band is regarded as a second-order polynomial in terms of the pixel values of the considered image.

Statistical dependency

The optimal result of the statistical dependency discovering analysis is that sub-bands from the same cluster are strongly dependent, while sub-bands coming

Table 7.1. Selected sub-bands for each dataset.

Dataset	Selected Sub-bands
IXI	L_gL_g, L_fL_g, H_fH_g, L_fH_f
Kvasir-SEG	L_gL_g, L_fL_g, H_fH_g, L_fH_f
CHAOS CT	L_gL_g, H_gL_g, H_gH_g, L_fL_g
CHAOS T1	L_gL_g, L_fL_g, L_fH_g, H_gL_f
CHAOS T2	L_gL_g, L_fL_g, H_gL_f, H_gH_f

from diverse clusters are independent. If the clustering algorithm converges to this solution, then it is possible to build a meaningful representation of the input by selecting just the most representative sub-band from each cluster. As a metric for estimating the amount of dependency of a sub-band with respect to a set of features we involve the mutual information (I) [218]. Indeed, the mutual information quantifies how much information the set of features contains about the sub-band, being zero when the two objects are independent. The higher the value of I, the better the estimation of the marginal distribution of the sub-band will be. However, the estimate may be inaccurate in case of high-dimensional distributions such as multiple sub-bands and complex input data such as images thus, for this reason, a many-to-one scalar mapping is involved, assuming that the set of features in a cluster provides information about the distribution S of the selected sub-band $\mathbf{s}$ through a function $T = f(\mathbf{X}_1, ..., \mathbf{X}_N)$. At this point, the mutual information can be directly computed as $\mathrm{I}(S;T)$ and remains open the choice of the function $f(\cdot)$. From the data-processing theorem [219] we know that the true mutual information $\mathrm{I}(S;\{\mathbf{X}, ..., \mathbf{X}_N\})$ is the upper bound of our approximation and that the equality holds if and only if T is the sufficient statistics for the set of data $\{\mathbf{X}_1, ..., \mathbf{X}_N\}$. Therefore, the function $f(\cdot)$ has to be chosen to maximize the estimate $\mathrm{I}(S;T)$ while remaining convenient in terms of computation. To this end, involving a linear model of the type $T = f(\mathbf{X}_1, ..., \mathbf{X}_N) = \sum_{i=1}^{N} \mathbf{W}_i \mathbf{X}_i$, with $\mathbf{W}_i = 1/N$, has been proved to match the required statistical properties being the unbiased estimate of the mean of $\mathbf{X}_i$ [218, 220]. At this point, the mutual information can be computed as:

$$\mathrm{I}(S;T) = \sum_{s} \sum_{t} P_{S,T}(\mathbf{s}, \mathbf{t}) \log \frac{P_{S,T}(\mathbf{s}, \mathbf{t})}{P_S(\mathbf{s}) P_T(\mathbf{t})}, \tag{7.2}$$

in which $P_{S,T}(\mathbf{s}, \mathbf{t})$ is the joint probability distribution while $P_S(\mathbf{s})$ and $P_T(\mathbf{t})$ are the marginal probability distributions of the selected sub-band and of the unbiased mean estimate T. A formal definition of the GROUSE algorithm follows.

Algorithm

Given a dataset of images $\{\mathbf{X}_1, ..., \mathbf{X}_N\}$, the proposed quaternion wavelet sub-band GROUping and SElection (GROUSE) algorithm can be formalized as:

1. Compute the QWT as described in Section 4.2 and extract the sixteen sub-bands $\mathbf{s}_{i,1}, ..., \mathbf{s}_{i,16}$ for each image $\mathbf{X}_i$.

2. For each sub-band $\mathbf{s}_{1:N,j}$, with $j = 1, ..., 16$, estimate the marginal probability density function of the energy $E_j(\mathbf{s}_{1:N,j})$ in (7.1) using a Gaussian kernel.

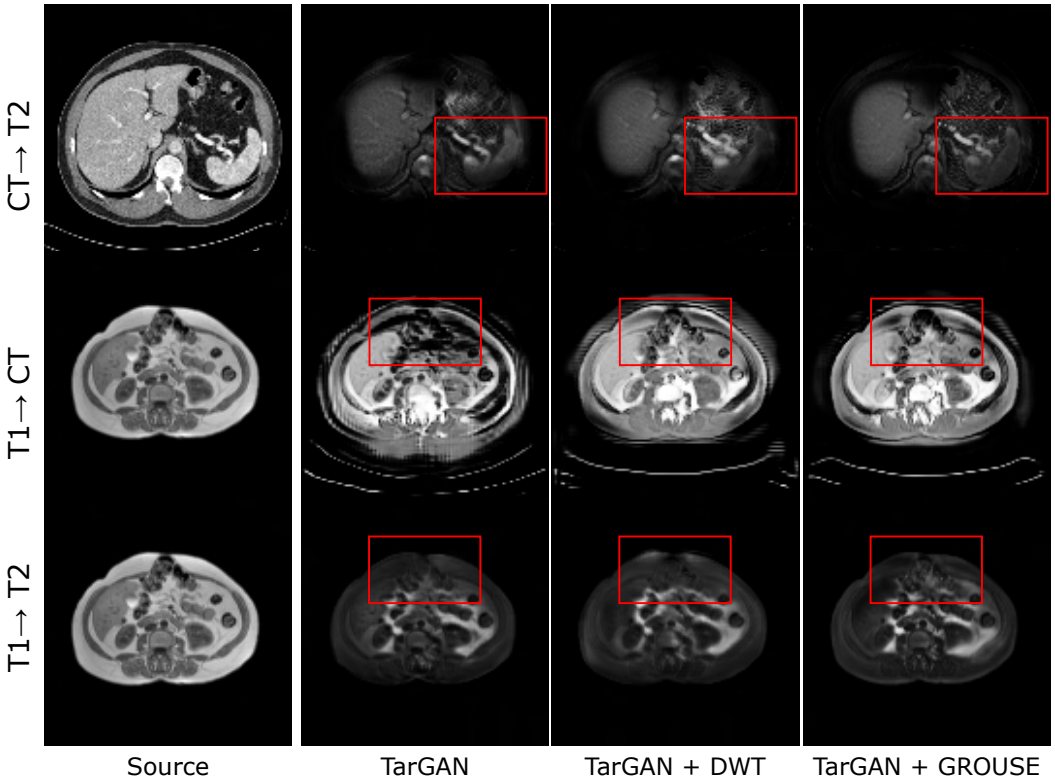


Figure 7.1. Random samples of TarGAN translations on the CHAOS dataset, where the first column refers to the source image. Red boxes highlight regions where the TarGAN distorts or blurs organs, while our method provides a superior translation.

3. Apply the k-Means algorithm involving the mutual information in (7.2) as distance metric and generate the set of clusters $C = \{C_1, ..., C_k\}$. In our experiments, to properly fit the input also in QNNs, we select $k = 4$.

4. From each cluster in C, select the sub-band with the highest energy value, i.e. the most informative for the given partition.

5. Involve the set of selected sub-bands as input to the neural model.

We repeat the algorithm 20 times for more robust results.

7.2.1 Experimental Evaluation

In this section, we validate our theoretical claims on different datasets and for various tasks. We consider the IXI dataset, which comprises brain magnetic resonance (MR) images (T1 and T2) of dimension 256×256 collected from three different hospitals, and divided in 866 for training and 290 for testing. Then, we take into account the Kvasir-SEG dataset that is composed of 1000 256×256 images of colon polyps, whose 75% is involved for training. Finally, we take into consideration the CHAOS dataset that is composed of CT and MR abdominal scans sliced and resized at 128×128 resolution, resulting in 3315 samples for training and 829 for testing.

We run the GROUSE algorithm for each dataset, so to obtain the four most representative sub-bands specific to the given data. For the multimodal CHAOS

Table 7.2. Results (average scores over multiple runs) for reconstruction, segmentation and image modality translation.

Dataset	Recontruction	Method	SSIM↑	MSE↓	FID↓
IXI	AE	Grayscale	0.856	0.0040	0.178
		DWT	0.846	0.0044	0.201
		GROUSE	**0.923**	**0.0030**	**0.122**
Dataset	Segmentation	Method	DSC↑	mIoU↑	MAE↓
Kvasir-SEG	U-Net [221]	RGB	0.740	0.614	0.073
		DWT	0.745	0.626	0.075
		GROUSE	**0.749**	**0.631**	**0.071**
	U-Net++ [222]	RGB	0.773	0.638	0.066
		DWT	0.776	0.647	0.066
		GROUSE	**0.782**	**0.655**	**0.065**
Dataset	Translation	Method		FID [146]↓	
CHAOS	TarGAN [146]	Grayscale		0.0955	
		DWT		0.0908	
		GROUSE		**0.0897**	

dataset, we run the sub-band selection algorithm specific for each modality, as the CT, T1, and T2 are unbalanced in the dataset and are quite different from each other so the salient features may differ from one modality to another. Table 7.1 shows the GROUSE selected sub-bands for each dataset and modality. It is worth noting that the GROUSE algorithm always selects the low-frequency L_gL_g band, as it is the most representative of the input and the one with the highest energy, and it helps in building a good approximation of the data sample. Fine-grained details are then captured by the three selected high-frequency bands that provide the information to create a precise representation of the sample.

We perform tests in multiple scenarios, involving reconstruction, segmentation, and image modality translation tasks. As baselines we consider a vanilla convolutional autoencoder (AE) for reconstruction, a U-Net [221] and a more advanced U-Net++ [222] for segmentation, and the recent large-scale Target-aware generative adversarial networks (TarGAN) [146] for the image modality translation task. The TarGAN is used to translate medical images from a reference domain to a target one, paying more attention to the ROI in the form of a segmentation map of the chosen organ. While there are two different discriminators, one for the image and one for the ROI target, the generator can handle both at the same time. It is composed of two branches, the first one for the whole image and the second one for the region of interest. It takes as input the original image, the four wavelets associated with it, the ROI, and the output target modality, producing as output an image in the translated domain. For each model, we involve the same hyperparameters setting and network architecture of the original works. Table 7.2 shows the objective scores for the three tasks, where for the reconstruction task we compute the structural similarity index (SSIM), the mean squared error (MSE), and the Fréchet Inception distance (FID), for the segmentation task the dice score coefficient (DSC), the mean intersection over union (mIoU) and the mean absolute error (MAE), while for the translation task we consider a specific FID proposed for medical images in [146], which rewards images that translate better meaningful features. We report the average scores over multiple runs with different seeds. As it is evident from Tab. 7.2, models endowed

Table 7.3. Reconstruction results on the IXI dataset for vanilla autoencoders defined in the quaternion domain.

Dataset	Reconstruction	Method	SSIM↑	MSE↓	FID↓
IXI	QAE	Channel repet. [117]	0.900	0.0036	0.092
		DWT	0.665	0.0072	**0.084**
		GROUSE	**0.915**	**0.0031**	**0.084**

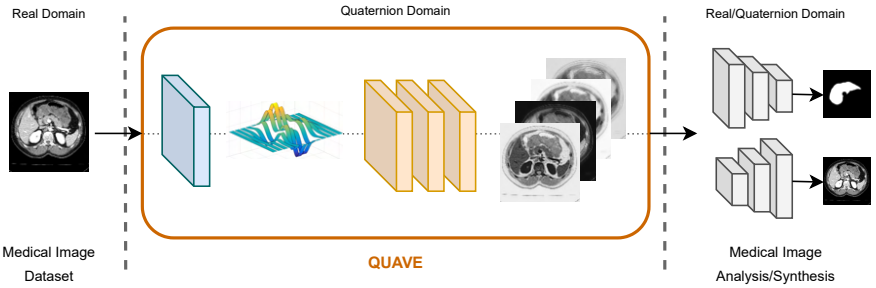


Figure 7.2. Illustration of the QUAVE framework. Any medical image dataset can be processed in the quaternion domain via QUAVE, which carries out the QWT and the neural feature extraction. The training procedure can be then performed either in the real or in the quaternion domain depending on the user choice, producing a real-valued output. We integrate a quaternion-based framework in the analysis and synthesis of real-valued medical image datasets.

with the proposed framework clearly outperform baseline networks that directly involve grayscale or RGB images, proving that our GROUSE approach can produce meaningful features to be involved in neural network training. Additionally, we show that our framework can be employed in various medical image analysis and synthesis scenarios that comprise different tasks and datasets without requiring specific operations or modifications to either data or models. A visual comparison of the translation task in Fig. 7.1 underlines the superiority of the TarGAN that involves features provided by our framework. The bounding boxes prove that our model can correctly translate the ROI of the given image while the original fails. Therefore, the task-agnostic GROUSE can be easily generalized to multimodality datasets, while improving the performance of large-scale models too.

Additionally, we also extend our approach to networks defined in the quaternion domain in order to show how our framework can generalize QNNs adoption in onedimensional problems. Indeed, for such tasks, the common approach is to replicate the single channel four times to fit the desired quaternion dimension, so adding useless and redundant information [117]. We prove that our GROUSE framework enhances the performance of QNNs providing a representative set of sub-bands, as Table 7.3 shows for the reconstruction task, where the quaternion autoencoder (QAE) with GROUSE data far exceeds the DWT approach and the common channel repetition one [117].

7.3 Quaternion Wavelet Feature Extraction

While GROUSE demonstrates the effectiveness of statistical feature selection for identifying the most informative QWT sub-bands, its approach relies on predefined statistical measures and heuristic grouping strategies. This limitation becomes

apparent when dealing with complex data distributions or when the optimal feature combinations are not easily captured by conventional statistical dependencies. Moreover, GROUSE's selection process, while systematic, does not adapt to the specific requirements of different downstream tasks or learn from the data itself.

In this section, to address these limitations, we introduce QUAVE, a learning-based framework that goes beyond statistical selection by employing neural networks to automatically discover the most informative representations from the sixteen QWT sub-bands. Unlike GROUSE's static approach, QUAVE learns to dynamically weigh and combine sub-bands through a trainable process, enabling it to capture complex, non-linear relationships between features that statistical measures might miss. The key innovation of QUAVE lies in its ability to learn optimal feature combinations directly from data while simultaneously solving a downstream task. We build QUAVE upon the recently demonstrated effectiveness of wavelets in various tasks, such as image fusion methods [223] involving multimodal image fusion using wavelets and multiple features. Among these methods, wavelets are particularly effective in enhancing super-resolution by enriching output images with improved fine-grain details [34, 224, 225]. For this reason, we propose a neural model that leverages quaternion wavelets to perform super-resolution, using this task as a vehicle for learning the most informative feature representations. The super-resolution objective serves as a natural training signal that encourages the model to extract and preserve the most salient details from the QWT sub-bands. We propose to leverage transfer learning, wherein QUAVE is trained to perform a super-resolution task on a generic medical dataset (e.g., IXI) and then is involved in downstream tasks with the knowledge it has acquired during training. This approach ensures that the learned feature representations are not only informative for the training task but also generalizable to other applications.

7.3.1 Pretraining

We pretrain the model with the objective of enhancing resolution by reconstructing high-quality images with enriched detail through the quaternion wavelet and feature refinement techniques like those in [225]. QUAVE pretraining combines a content loss $\mathcal{C}_{loss}$ and a wavelet-based loss $\mathcal{W}_{loss}$ component to enforce finer structure and detail preservation in the predictions of the model. The content loss is an L1 loss between the generated image and the ground truth image, which penalizes direct differences in pixel values, while the wavelet loss captures differences in frequency information. This is because the ground truth image x_{gt} and the prediction x_{pred} are transformed into the wavelet domain. The wavelet loss is computed as the L1 loss between these transformed representations $wavelet_{gt}$ and $wavelet_{pred}$. The final loss function combines the content and wavelet losses, with α controlling the influence of the wavelet loss.

$$\mathcal{L} = \mathcal{C}_{loss} + \alpha \cdot \mathcal{W}_{loss} \tag{7.3}$$

Once reached the convergence, the pretrained QUAVE model is utilized to extract the features for a downstream task of interest, such as reconstruction, segmentation, or image translation. Extracting the features entails employing the pretrained QUAVE model and, as depicted in Fig. 7.3, selecting the weights from the second-last layer, prior to the inverse wavelet transform.

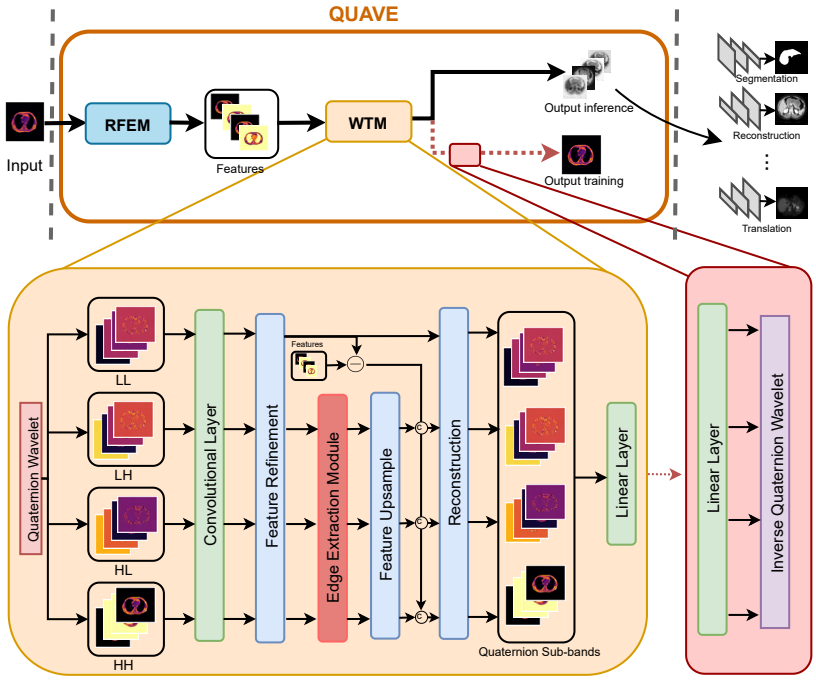


Figure 7.3. Detailed QUAVE architecture. Being a quaternion-based generalizable framework, the WTM enhances neural model performance when processing rich-of-detail real images, such as medical ones. Along with the inverse quaternion wavelet block, the WTM performs a salient feature extractor by leveraging quaternion algebra properties to achieve more detailed insights. During inference, the output of our framework can be then involved in any neural model either in the real or quaternion domain, improving performance in a variety of tasks (e.g., segmentation, reconstruction, and generation, among others).

7.3.2 Architecture

The architecture of QUAVE consists of two principal modules, as shown in Fig. 7.3. The first one is the rough feature extraction module (RFEM), displayed in Fig. 7.4, which provides the necessary features for performing the quaternion wavelet transform. The residual groups in RFEM are structured like in [83] and visualized in Fig. 7.4. Note that to improve the clarity of the figure, the residual connections among the layers are not reported. Subsequently, the wavelet transform module (WTM) takes the features extracted by the RFEM and applies the QWT to those. Moving forward, the module operates in four groups: one low-frequency group and three high-frequency groups. Before entering the feature refinement block, we reduce the dimensionality of the image features by a factor of four using a convolutional layer. Inside the WTM, the Edge Extraction Module focuses on high-frequency information and captures essential edge details of the image. The module utilizes all the high-frequency sub-bands (HL, LH, HH) to create an edge feature map (EFM), which consolidates all high-frequency information to enhance edge detail. The EFM

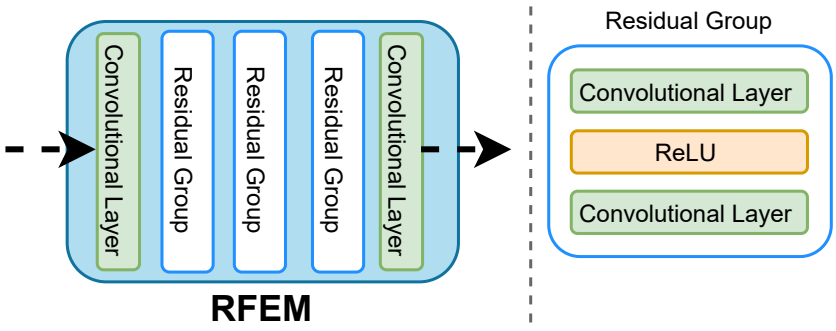


Figure 7.4. The Rough Feature Extraction Module (RFEM) is responsible for initial feature extraction before the Quaternion Wavelet Transform (QWT). It consists of a convolution layer and multiple residual groups designed to extract both shallow and deep features from the input image and does not apply QWT. This prepares the data with enriched initial features before entering the QWT, making the wavelet transform more effective.

is computed with the L2 Norm:

$$\text{EFM} = ||F_x||_2 = \sqrt{|F_{LH}|^2 + |F_{HL}|^2 + |F_{HH}|^2} \tag{7.4}$$

where F_x represents the high-frequency sub-bands ($F_x \in \{F_{LH}, F_{HL}, F_{HH}\}$), and EFM denotes the edge feature map. Next, the edge information is incorporated back into each frequency sub-band through concatenation.

Two linear layers are inserted before the inverse quaternion wavelet transform to extract the enriched features of the sub-bands, which outputs the final super-resolution image. Those layers allow us to retrieve the desired features from QUAVE, representing four sub-bands of the quaternion wavelet, that are considered the most informative by the neural model, as they activate the neurons that yield the best performance.

During the training phase of a downstream task of interest (e.g. segmentation), the original data flows into the pretrained QUAVE that performs inference and outputs the four most representative sub-bands. This enhanced representation is then fed as input to the downstream model to perform the task, as shown in the upper part of Fig. 7.3. Interestingly, the features representing the four sub-bands of the quaternion wavelet can naturally represent a quaternion number, and this perfectly fits on quaternion-valued neural networks.

7.3.3 Experimental Setup

To demonstrate the generalizability of our framework, we evaluate QUAVE across the same range of tasks and benchmark datasets introduced in Section 7.2.1, including reconstruction on IXI, segmentation on Kvasir-SEG, and image modality translation on CHAOS. This allows for a direct comparison of the learning-based QUAVE approach against the statistical GROUSE method and other baselines. For each task, we utilize the same baseline neural architectures (e.g., CAE, U-Net, TarGAN) to isolate and fairly assess the improvements conferred by the QUAVE-generated feature representations. We conduct analysis on two algebraic domains: first, we

set up experiments to prove that our framework can enhance the performance of real-valued neural networks. When using the term *real-valued*, we are referring to neural networks defined in the domain of real numbers, therefore not relying on quaternion numbers and thus not involving quaternion-valued layers, to which instead we refer as *quaternion-valued*. Second and most importantly, we perform evaluations with quaternion-valued models to show that QUAVE outperforms other preprocessing techniques and it is flexible to be adopted in different domains, further improving QNNs generalizability. Moreover, we consider diverse types of input images. We take into account medical grayscale and RGB images, and multimodal images, proving that the proposed method can be generalized to any kind of input, thus being data-agnostic and not suffering from domain shifting.

Description of the Tasks

Reconstruction

As a first step of our evaluation, we take into account the IXI dataset[2] for the reconstruction task. It comprises magnetic resonance images (MRI) collected from three different hospitals in London, and it is organized as follows: T1, T2, PD-weighted images, MRA images and diffusion-weighted images. We consider the combination of the multimodal T1 and T2 images of 581 and 578 patients respectively. Each 3D MRI scan is stored in NIFTI format and is composed of a stack of 2D slices that together represent the full volumetric structure of the brain. For example, a typical T1-weighted volume comprises approximately 130–150 axial slices, while a T2-weighted volume consists of around 112–136 axial slices. In our experiments, for the purpose of the reconstruction task, we extract a single, representative 2D slice per patient. This approach simplifies the reconstruction task by not processing the entire 3D volume or a large number of slices.

After this step, we obtain 1156 images from the combination of T1 and T2 ones, each of size 256×256. We employ 866 of them for training and 290 for validation.

Segmentation

The second set of experiments comprises the segmentation task on the Kvasir-SEG dataset [226]. This dataset contains snapshots of polyps with manually annotated masks verified by an experienced gastroenterologist. The 1000 samples range from 332×487 to 1920×1072 pixels, and we resize them to 256×256. We involve 750 samples for training and the remaining part for validation. Interestingly, and differently from previous datasets, the Kvasir-SEG dataset contains RGB images with three channels. Therefore, we train the real-valued model with original three-channel samples, while we apply the DWT and the QWT computed on the corresponding grayscaled image.

Image Modality Translation

The third set of experiments concerns image modality translation on the CHAOS - Combined (CT-MR) Healthy Abdominal Organ Segmentation dataset. It contains CT and MR scans from unpaired abdominal image series, the ground truth is generated with the auxiliary participation of multiple experts' annotations and introduced for the first time in [227]. The data comes from 80 patients from the Department of Radiology of Dokuz in Turkey, half of them have done only a single

[2] https://brain-development.org/ixi-dataset

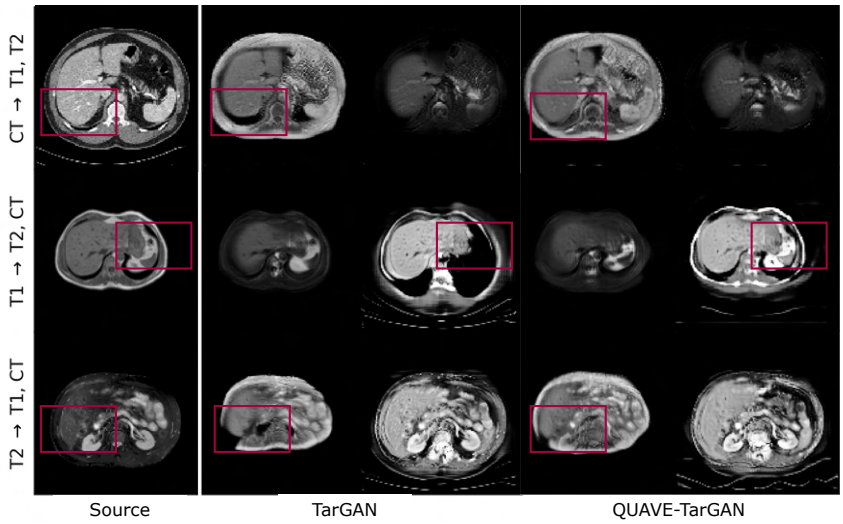


Figure 7.5. Random samples of TarGAN image modality translations on the CHAOS dataset. The first column refers to the original domain images, while the others refer to the corresponding translation in the other domains.

CT scan while the others went through MR scans. For our experiments we have used all the data resizing the slices at 128×128, obtaining approximately 4144 samples, we randomly select 80% from all the modalities to be the training dataset, and the remaining as test one.

QUAVE in Combination with Real- and Quaternion-Valued Models

We involve different models in our tests. For the reconstruction task, we consider a vanilla convolutional autoencoder (CAE) [117], whereby the encoder is composed of three convolutional layers $[28, 64, 128]$ with kernels $[8, 3, 4]$ for the IXI dataset. The decoder has a mirrored structure with transposed convolutions. The quaternion CAE (QCAE) [117] has the same backbone while involving quaternion convolutions instead of real-valued ones. For each reconstruction experiment, we set the learning rate of the Adam optimizer to 0.0001, the batch size to 4, and the number of epochs to 10, repeating the learning process for multiple runs with different seeds.

Concerning the segmentation task, we consider a U-Net [221] and a more advanced U-Net++ [222] in our experiments, as they are widespread models for this scenario. In addition, quaternion counterparts for these architectures are available and have been validated in prior work, which enables us to perform a direct and fair comparison between real-valued and quaternion-valued approaches within our QUAVE framework. Our primary focus is on demonstrating the generalizability and benefits of integrating QUAVE with existing segmentation networks. Thus, using U-Net and U-Net++ allows us to clearly highlight the improvements brought by our framework rather than solely pursuing the highest possible segmentation accuracy. The U-Net encoder comprises four blocks, each with two convolutions, batch normalization, and ReLU activation function, while max pooling is interleaved between each block. The number of filters increases going deeper, being $[32, 64, 128, 256]$, each with kernel sizes 3 and stride 1. The bottleneck adds a further block with 512 filters. Then, the information is passed to the decoder, which has the same architecture as the encoder except for the replacement of transposed convolutions instead of max pooling operations. A final convolutional layer refines the resulting segmentation map with 1 channel. Similar

Table 7.4. Results for task 3 image modality translation on the CHAOS dataset (Segmentation metrics) for real and quaternion-valued models.

Model	DSC↑	S-Score↑	mIoU↑	MAE↓
TarGAN [146]	87.53 ± 0.6474	76.23 ± 9.5988	0.7812 ± 0.0090	0.0489 ± 0.0007
QUAVE-TarGAN	$\mathbf{90.53 \pm 4.3608}$ (+3.4%)	$\mathbf{84.41 \pm 4.2848}$ (+10.7%)	$\mathbf{0.8320 \pm 0.0676}$ (+6.5%)	$\mathbf{0.0487 \pm 0.0004}$ (+0.4%)
QTarGAN	$\mathbf{83.56 \pm 1.1758}$	76.85 ± 1.2373	0.7246 ± 0.0158	0.0492 ± 0.0011
QUAVE-QTarGAN	83.02 ± 0.7664 (−0.6%)	$\mathbf{78.90 \pm 1.1821}$ (+2.6%)	$\mathbf{0.7360 \pm 0.0096}$ (+1.6%)	$\mathbf{0.0488 \pm 0.0005}$ (+0.8%)

to the QCAE, the quaternion U-Net (QU-Net) replaces convolutions and transposed convolutions with quaternion-valued counterparts. The more advanced U-Net++ has a similar structure to the original one, inserting however dense connections that aggregate intermediate feature maps. Every hyperparameter is set according to [222]. In these experiments, as for the reconstruction task, we set the learning rate to 0.0005 for the U-Net and to 0.0003 for the U-Net++, as suggested in [222], with Adam optimizer and the batch size to 4. We train all the models for 50 epochs, repeating the learning process for multiple runs with different seeds.

Regarding the image modality translation task, we consider the recent Target-Aware Generative Adversarial Network (TarGAN) [146]. The TarGAN translates images from a source domain to a target domain while paying particular attention to the region of interest passed as a segmentation map during the translation. In the TarGAN framework, segmentation information is explicitly incorporated into the I2I translation process. Specifically, given an image x_s from the source modality s and its corresponding target area label y, a binarization operation $y \cdot x_s$ is applied to isolate the target area, yielding r_s, an image containing only the target region. TarGAN then translates both the full input image x_s and the extracted target area r_s into their corresponding representations in the target modality t, denoted as $G(x_s, r_s, t) \rightarrow (x_t, r_t)$. To accomplish this, TarGAN employs a double-stream generator with a shared middle block and two pairs of encoder-decoders: one branch processes the entire image x_s, while the other focuses exclusively on the target area r_s. The generator takes as input the original image and the four wavelets associated with it, the target area of interest, and the output target modality, generating in the output the translated image in the desired domain. This network comprises two branches, the first one for the whole image and the second one for the region of interest. A single branch is composed of an initial encoder, then a set of blocks shared with the other branch, and finally, a decoder to output the desired image. This network stacks convolutional blocks that employ a convolution layer then the instance normalization, and finally a Leaky ReLU activation function. The upsampling blocks instead involve upsampling layers by interpolation suddenly a convolution operation, and, as in the previous block, finally the instance normalization with Leaky ReLU. The discriminator has a similar structure, for each detail regarding the model structure and hyperparameters we refer to the original paper [146]. Overall, this model comprises 125 million parameters, being the largest in our experiments. In this case, we introduce the quaternion-valued counterparts as the quaternion TarGAN has not been introduced yet in literature.

Metrics

As objective evaluation metrics in the reconstruction task, we compute the structural similarity index (SSIM), the mean squared error (MSE), and the Fréchet inception distance (FID). While SSIM measures image degradation through structural information, FID estimates how the distributions of real and reconstructed images are far from each other. Indeed, FID is traditionally used in generative tasks

Table 7.5. Results for task 1 reconstruction on the IXI dataset for real and quaternion-valued models. Average and standard deviation over multiple runs reported. Runs statistically significant with $p < 0.1$.

Model	Params	SSIM↑	MSE↓	FID↓
CAE [117]	305k	0.9915 ± 0.0079	0.0006 ± 0.0001	0.0046 ± 0.0035
QUAVE-CAE	305k	$\mathbf{0.9915 \pm 0.0051}$ (+0%)	$\mathbf{0.0006 \pm 0.0001}$ (+0%)	$\mathbf{0.0036 \pm 0.0006}$ (+21.7%)
QCAE [117]	78k	0.9504 ± 0.0027	0.0015 ± 0.0001	0.0293 ± 0.0024
QUAVE-QCAE	78k	$\mathbf{0.9842 \pm 0.0022}$ (+3.5%)	$\mathbf{0.0010 \pm 0.0001}$ (+33.3%)	$\mathbf{0.0126 \pm 0.0034}$ (+57.1%)

Table 7.6. Results for task 2 segmentation on the Kvasir-SEG dataset for real and quaternion-valued models. Average and standard deviation over multiple runs reported. Runs statistically significant with $p < 0.1$.

Model	Params	DSC↑	mIoU↑	MAE↓
U-Net [221]	7.8M	0.7474 ± 0.0595	$\mathbf{0.6546 \pm 0.0166}$	$\mathbf{0.0653 \pm 0.0030}$
QUAVE-U-Net	7.8M	$\mathbf{0.7726 \pm 0.0025}$ (+4.1%)	0.6394 ± 0.0094 (−2.3%)	0.0666 ± 0.0015 (-2%)
U-Net++ [222]	9.2M	0.7796 ± 0.0046	0.6379 ± 0.0147	0.0664 ± 0.0016
QUAVE-U-Net++	9.2M	$\mathbf{0.8000 \pm 0.0107}$ (+2.3%)	$\mathbf{0.6729 \pm 0.0109}$ (+5.4%)	$\mathbf{0.0614 \pm 0.0024}$ (+7.5%)
QU-Net [117]	1.95M	0.7668 ± 0.0116	0.6293 ± 0.0203	0.0689 ± 0.0025
QUAVE-QU-Net	1.95M	$\mathbf{0.7728 \pm 0.0102}$ (+0.8%)	$\mathbf{0.6382 \pm 0.0159}$ (+1.4%)	$\mathbf{0.0663 \pm 0.0027}$ (+3.7%)
QU-Net++ [117]	2.3M	0.7857 ± 0.0078	0.6496 ± 0.0223	0.0656 ± 0.3496
QUAVE-QU-Net++	2.3M	$\mathbf{0.7912 \pm 0.0020}$ (+0.7%)	$\mathbf{0.6526 \pm 0.0039}$ (+0.4%)	$\mathbf{0.0636 \pm 0.0003}$ (+3.1%)

where paired metrics are unavailable. In reconstruction tasks, its ability to capture differences in the distribution of high-level image features can offer complementary insights into reconstruction fidelity, particularly in terms of global structure and perceptual coherence. By including it alongside SSIM and MSE, we ensure a more comprehensive evaluation of our method ability to reconstruct high-quality images while maintaining both local and global features. For evaluating the results of the segmentation task, instead, we compute the Dice similarity coefficient (DSC), the mean intersection over union (mIoU), and the mean absolute error (MAE). In order to perform a robust evaluation of the image modality translation task, we employ both a standard FID and an FID proposed in the original work [146]. The standard FID metric measures the distance between the distributions of feature vectors extracted from real and generated images using a pretrained network, commonly the Inception v3 model [99], which is trained on generic (non-medical) images. In the classic implementation, feature vectors from the last pooling layer of the pretrained Inception model are used to calculate the mean and covariance for both the real and generated images. Instead, the FID implemented by TarGAN [146] incorporates specific adaptations tailored for the task of translating medical images. To compute the FID, the feature vectors are extracted from the last layer of the encoder of a pre-trained Models Genesis [228] network and not the Inception v3 model like in the classic FID. Since Models Genesis [228] is based on 3D medical volume data, it is crucial that the input to this network also has a 3D context. In our image translation task, only 2D translations are performed. To provide the required 3D context for feature extraction, we form 3D volume patches by concatenating every 16 consecutive axial slices, following the receipt in [228]. The choice of 16 is based on a balance between capturing sufficient volumetric context and maintaining computational efficiency. Importantly, we ensure that the slices are anatomically consistent, i.e. they are consecutive so that the resulting 3D patch reflects the true spatial continuity of the anatomy. This ensures that the features extracted by Models Genesis are meaningful and appropriate for FID computation. The average FID score across all modality pairs (e.g., CT → T1w, T2w → T1w) is reported. Performance differences may exist between individual translation pairs, for example, translating from CT to

Table 7.7. Results for task 3 image modality translation on the CHAOS dataset (image generation metrics) for real and quaternion-valued models. Average and standard deviation over multiple runs reported. Runs statistically significant with $p < 0.1$.

Model	FID [146]↓	FID↓	IS↑
TarGAN [146]	0.0912 ± 0.0002	0.7009 ± 0.1181	3.1879 ± 0.2172
QUAVE-TarGAN	$\mathbf{0.0893 \pm 0.0025}$ (+2.1%)	$\mathbf{0.6295 \pm 0.0233}$ (+10.1%)	$\mathbf{3.3452 \pm 0.0185}$ (+4.9%)
QTarGAN	$\mathbf{0.0881 \pm 0.0021}$	0.6513 ± 0.0144	2.7842 ± 0.2628
QUAVE-QTarGAN	0.0902 ± 0.0024 (−2.3%)	$\mathbf{0.6253 \pm 0.0860}$ (+3.9%)	$\mathbf{3.0228 \pm 0.0524}$ (+8.5%)

Table 7.8. Ablation study for real and quaternion-valued models using different wavelets in QUAVE.

Data	Model	Wavelet	SSIM↑	MSE↓	FID [146]↓	FID↓	DSC↑	mIoU↑	MAE↓
IXI	QUAVE-CAE	DWT	0.991 ± 0.002	0.001 ± 0.001	-	0.010 ± 0.001	-	-	-
		QWT	**0.991** ± 0.005	**0.001** ± 0.001	-	**0.003** ± 0.001	-	-	-
	QUAVE-QCAE	DWT	0.949 ± 0.015	0.001 ± 0.001	-	0.025 ± 0.003	-	-	-
		QWT	**0.984** ± 0.002	**0.001** ± 0.001	-	**0.012** ± 0.003	-	-	-
KVASIR	QUAVE-U-Net	DWT	-	-	-	-	**0.775** ± 0.002	**0.643** ± 0.015	0.066 ± 0.001
		QWT	-	-	-	-	0.772 ± 0.002	0.639 ± 0.009	**0.066** ± 0.001
	QUAVE-QU-Net	DWT	-	-	-	-	0.760 ± 0.015	0.624 ± 0.028	0.068 ± 0.004
		QWT	-	-	-	-	**0.772** ± 0.010	**0.638** ± 0.015	**0.066** ± 0.002
	QUAVE-U-Net++	DWT	-	-	-	-	0.781 ± 0.011	0.649 ± 0.016	0.064 ± 0.003
		QWT	-	-	-	-	**0.798** ± 0.010	**0.672** ± 0.010	**0.061** ± 0.002
	QUAVE-QU-Net++	DWT	-	-	-	-	0.782 ± 0.019	0.645 ± 0.026	0.065 ± 0.004
		QWT	-	-	-	-	**0.791** ± 0.001	**0.652** ± 0.003	**0.063** ± 0.002
CHAOS	QUAVE-TarGAN	DWT	-	-	0.092 ± 0.002	0.684 ± 0.040	0.888 ± 0.201	0.802 ± 0.031	0.048 ± 0.001
		QWT	-	-	**0.089** ± 0.002	**0.629** ± 0.023	**0.905** ± 0.436	**0.832** ± 0.067	**0.048** ± 0.001
	QUAVE-QTarGAN	DWT	-	-	**0.087** ± 0.001	0.649 ± 0.069	**0.836** ± 0.012	0.724 ± 0.017	0.050 ± 0.001
		QWT	-	-	0.090 ± 0.002	**0.625** ± 0.086	0.830 ± 0.076	**0.736** ± 0.009	**0.048** ± 0.001

T1w typically yields a higher FID score than translating from T1w to CT, owing to inherent contrast differences between these modalities. However, our analysis shows that despite these individual variations, the average FID remains a robust indicator of the overall quality and consistency of the translation process. In this context, the averaged metric effectively captures the general performance of our approach across all modality pairs, demonstrating that our method handles varying degrees of translation difficulty competently.

7.3.4 Results

In this section, we present the selected quaternion sub-bands we employ in our tests and the empirical evaluation we accomplish for validating our theoretical claims. We perform multiple runs with different seeds for each experiment and we report in the tables the average score and the standard deviation for each metric. We initialized the neural networks weights (both the 2D convolutional and the linear layers) using the well-known Kaiming and He initialization with a uniform distribution and the biases to zero.

Reconstruction on IXI dataset

Table 7.5 shows the average and the standard deviation over different runs of the objective metrics results for the IXI dataset. The models that employ the proposed preprocessing (QUAVE) enhance the scores with both real and quaternion-valued models. These results prove that QUAVE can boost model learning without requiring modification or manipulation of models or any kind of data augmentation. Indeed, features extracted with the QWT capture salient characteristics of the image that are crucial for proper reconstruction.

Segmentation on Kvasir-SEG dataset

Table 7.6 reports instead objective metrics results for the segmentation task on the Kvasir-SEG dataset with RGB images. It is worth noting that the Kvasir-SEG dataset comprises RGB images, thus the real-valued U-Net is trained with three-channel color samples. Therefore, we prove that our technique, QUAVE, increases model performance also with respect to color images. Hence, QUAVE can be generalized to grayscale and color images as well, improving the accuracy of involved networks.

Image modality translation on CHAOS dataset

Table 7.7 and Table 7.4 report objective metrics results for the image modality translation task on the CHAOS dataset with both image generation and segmentation metrics. The segmentation results presented in Table 7.4 provide insight into the quality of structural preservation in the translated images. This aspect is inherently linked to the TarGAN architecture and methodology.

As in previous experiments, we apply the proposed QUAVE to data for extraction, to both the original TarGAN [146] and a quaternion version referred to as QTargan. To further stabilize GAN training, we also added to the original network the Spectral Normalization to limit the discriminator sensitivity and ensure a more stable training process [152]. We can observe that both the DWT and QUAVE improve the performance with respect to the original network. The FID [146] proposed in the original TarGAN work tends to reward the images that best translate meaningful features in medical images, and our approach QUAVE beats the original processing model. Even if we consider the classical FID, QUAVE performs best among the original work. We also show this effect on a visual comparison among the predicted translation in Fig. 7.5 on randomly chosen samples. We display the superiority of the translation of QUAVE using red boxes. The boxes help to indicate the artifacts in the translation committed by the other methods. To conclude, we show that the data-agnostic QUAVE can be easily generalized to multimodality datasets, improving the performance of large-scale models too.

Ablation study

We adopt the QWT for our framework, consequently, we show an ablation study instead of using the DWT in QUAVE. This is possible with only a slight change in the output size of the final layers of QUAVE. From the scores in Table 7.8, we can observe that both the DWT and QUAVE approaches have a high impact on model performance, improving scores with respect to baselines. However, the best scores are, again, achieved by the models with the QUAVE method with QWT wavelet, proving its effectiveness in most of all the different tasks.

7.4 Conclusion

This chapter has presented two foundational contributions aimed at integrating frequency-aware and hypercomplex representations into neural image modeling pipelines. GROUSE introduced a model-agnostic and task-independent method for extracting and selecting salient quaternion wavelet sub-bands based on statistical dependencies. By identifying and preserving only the most informative components, GROUSE enables efficient and generalizable feature integration for any neural network architecture.

Building on this, the QUAVE framework extended these principles into a flexible, data-agnostic design that enables the use of quaternion-valued learning even on grayscale or single-channel datasets. QUAVE generalizes the preprocessing pipeline using quaternion wavelet transforms and demonstrates strong performance gains across diverse tasks and model types. Moreover, it showcases how quaternion and hypercomplex neural networks can be made compatible with standard datasets, broadening their applicability beyond niche high-dimensional domains.

The generalizable, information-rich representations developed by GROUSE and QUAVE are not merely academic exercises; they are poised to directly address the critical limitations of diffusion models identified in the previous chapter. By providing a more efficient and expressive feature space, these wavelet-driven techniques offer a clear path to mitigating the prohibitive computational cost of high-resolution synthesis and improving the preservation of fine-grained details. The following chapters will build directly on this foundation, demonstrating how integrating these advanced representations into latent diffusion architectures can unlock new capabilities in high-fidelity and controllable image generation.

Chapter 8

Wavelets for Image Super-Resolution

Contents

8.1 Introduction 94
8.2 Quaternion Wavelet-Conditioned Diffusion Models for Image Super-Resolution 96
8.2.1 Quaternion Wavelet and Time-Aware Encoder 96
8.2.2 Latent Diffusion Process 96
8.2.3 Experimental Results 97
8.3 Metadata, Wavelet, and Time Aware Diffusion Models for Satellite Image Super Resolution 102
8.3.1 Feature Modulation 102
8.3.2 MWT-Encoder 102
8.3.3 Experimental Results 104
8.3.4 Dataset details 106
8.3.5 Training details 106
8.3.6 Evaluation details 107
8.3.7 Additional Qualitative Comparisons 107
8.3.8 Ablation Study 108
8.4 Wavelet Diffusion GAN for Image Super-Resolution . . 109
8.4.1 Wavelet-based conditional diffusion scheme 110
8.4.2 Conditional backward diffusion 110
8.4.3 Optimization 111
8.4.4 Experimental results 111
8.5 Conclusion 114

THIS chapter presents three super-resolution applications: (1) a quaternion wavelet-aware encoder for latent diffusion models that enhances detail preservation built on the quaternion wavelets framework from the previous chapter 7.3, (2) a wavelet-integrated framework for satellite image super-resolution targeting geospatial applications, and (3) a wavelet-based Diffusion GAN that significantly reduces computational costs while maintaining quality. These contributions demonstrate how wavelet-driven strategies can advance deep learning-based super-resolution across different domains and efficiency requirements.

8.1 Introduction

Building directly upon the concepts introduced in the preceding chapter 7, which detailed quaternion wavelets and the QUAVE framework for their efficient utilization, the first section 8.2 presented in this chapter [1] pioneers the application of these novel quaternion wavelet representations within a latent diffusion model for image super-resolution. This work introduces a quaternion wavelet- and time-aware encoder, designed to dynamically leverage quaternion wavelet embeddings during the denoising process. This approach aims to achieve superior fine-grained detail preservation and perceptual quality by effectively conditioning state-of-the-art generative models, such as Stable Diffusion (SD) [12].

Following this exploration of quaternion wavelets in SR, the chapter delves into broader applications and further innovations using wavelet transforms in the context of super-resolution. Section 8.3 shifts focus to the challenging domain of satellite image super-resolution, where the need for HR imagery is ever-growing for applications like environmental monitoring and urban planning [229]. It proposes a robust framework that integrates wavelet transforms with conditional latent diffusion models to address the unique characteristics of satellite imagery [230], emphasizing the reconstruction of fine details crucial for geospatial analysis.

Section 8.4 tackles another critical aspect of SR: computational efficiency and speed. Standard diffusion models, especially those operating in pixel space [90], can be slow. While latent diffusion models [12] offer improvements, this work introduces a novel wavelet-based conditional Diffusion GAN approach, drawing inspiration from efforts to accelerate generative models like Diffusion GAN [231] and WaveDiff [25] (for unconditional generation). By leveraging the Discrete Wavelet Transform (DWT) at both image and feature levels, a method that allows for spatial dimension reduction, in line with principles like the Nyquist Rule, and enhanced detail capture, this method aims to significantly reduce training and inference times while maintaining, and even surpassing, state-of-the-art image quality, particularly demonstrated in tasks like face super-resolution.

Finally, Section 8.5 synthesizes the findings, discusses broader implications, and outlines directions for future research. Together, the three contributions in this chapter underscore the versatility and potency of wavelet-driven strategies in advancing the field of image super-resolution. They explore distinct facets, from the novel application of quaternion wavelets and specialized domain adaptation to significant improvements in computational performance, collectively highlighting how multi-scale wavelet analysis can enrich and optimize deep learning-based SR frameworks.

Image Super-Resolution

Image Super-Resolution (SR) stands as a critical and enduring challenge in computer vision, pivotal for a multitude of applications ranging from medical diagnostics to remote sensing and security [33, 232]. The objective of reconstructing high-resolution (HR) images from their low-resolution (LR) counterparts is not merely a technical pursuit but a practical imperative, as the quality of input images significantly impacts the efficacy of subsequent analytical tasks such as object detection, segmentation, and classification [233]. Despite considerable research advancements, achieving SR that faithfully recovers intricate details and realistic textures, particularly at high magnification factors, remains an active area of

[1] The content of this Chapter is adapted from the material published in [32–34]

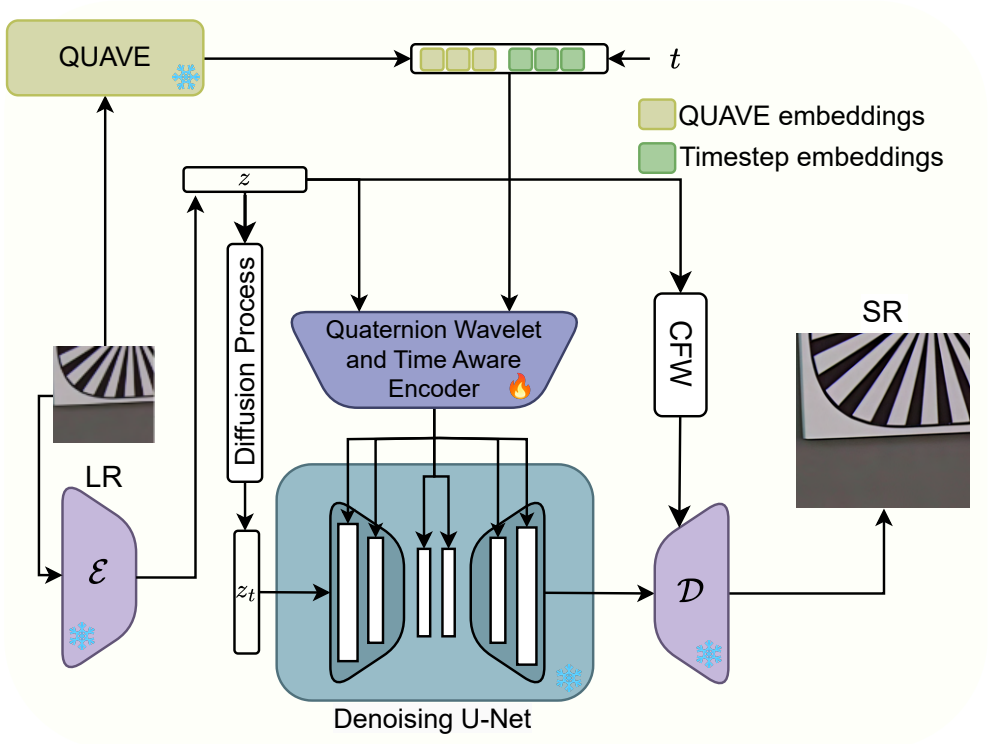


Figure 8.1. Overview of ResQu Super-Resolution framework. We pre-trained QUAVE and trained only our encoder, the quaternion embeddings- and time- aware encoder. The other parts of the framework are frozen, which speeds up the overall training process.

investigation.

The landscape of SR methodologies has been dramatically reshaped by the advent of deep learning. While traditional techniques, such as interpolation and frequency-domain transformations, laid essential groundwork, they often struggled to generate perceptually rich results, frequently leading to artifacts and blurring [234]. Convolutional Neural Networks (CNNs), Vision Transformers (ViTs), and more recently, generative models, particularly diffusion models [235,236], have consistently pushed the boundaries of SR performance. These models have demonstrated remarkable capabilities in generating coherent and high-fidelity images, often by iteratively refining noise into structured visual information, and have been successfully applied to SR tasks [90, 237, 238].

Within this evolving domain, there's a burgeoning interest in synergizing diffusion models with wavelet-based techniques [26]. Wavelets, with their inherent strengths in multi-resolution analysis and feature decomposition, offer a complementary approach to capturing image information across different scales and frequencies. This integration, as seen in works like [25], which use wavelets to adaptively handle frequency components, and other tailored SR approaches [34, 239], promises to enhance the ability of SR models to reconstruct both global structures and fine local details.

8.2 Quaternion Wavelet-Conditioned Diffusion Models for Image Super-Resolution

In this section, we describe our proposed method, which leverages the generative prior of SD [12] and introduces a novel quaternion wavelet and time-aware encoder for image super-resolution. By integrating quaternion wavelet embeddings and temporal conditioning into the diffusion process, our approach achieves high-fidelity reconstructions while preserving domain-specific details.

8.2.1 Quaternion Wavelet and Time-Aware Encoder

Quaternion Wavelets Embeddings. Directly utilizing all sixteen QWT sub-bands as input to a neural model introduces redundancy and inefficiency [30]. To address these challenges, we employ QUAVE, introduced in chapter 7.3, a learning-based approach that extracts only the features from the most informative sub-bands of the QWT, which utilizes the dual-tree quaternion wavelet transform (QWT) [119].

Our quaternion wavelet, and time aware encoder leverages quaternion wavelet embeddings, obtained from QUAVE, to provide multidimensional representations that capture low- and high-frequency information essential for super-resolution. Our encoder processes LR input images to extract the enhanced features. These features, enriched by information from both the frequency and spatial domains, are injected into the denoising process via spatial feature transformations (SFT) [187].

During the diffusion process, the QUAVE embeddings together within the timestep embeddings are passed through the encoder and used to modulate the intermediate feature maps of the U-Net via SFT. This ensures that the model incorporates both global and local details, enhancing the fidelity and realism of the reconstructed HR image.

Time-Aware Embeddings. In addition to quaternion wavelet embeddings, the encoder incorporates temporal information by embedding the timestep t into the conditioning mechanism. Early in the denoising process, when the latent representation has low SNR and contains significant noise, the encoder provides strong guidance using the enhanced features derived from QUAVE. This ensures that the structural integrity of the image is preserved during the initial stages of generation. This is because, as discovered in [188], when the SNR increases and the latent representation becomes more refined, the encoder reduces its influence, allowing the diffusion model to focus on fine-grained details. This adaptive behavior is crucial for achieving a balance between preserving global structure and enhancing local textures. The known noise schedule of the diffusion process enables the encoder to adjust its conditioning strength dynamically at each timestep, ensuring optimal guidance throughout the super-resolution pipeline.

8.2.2 Latent Diffusion Process

The conditioning process is defined by concatenating the QUAVE and timestep embedding tensors into a unified vector $b \in \mathbb{R}^{1024}$, as depicted in Fig. 8.1. This vector is processed alongside the latent representation $z \in \mathbb{R}^{h \times w \times c}$ of the input low-resolution image $x \in \mathbb{R}^{3 \times H \times W}$, where $z = \mathcal{E}(x)$ is obtained via the encoder $\mathcal{E}$.

For our ResQu we leverage the pre-trained variational autoencoder (VAE) from [12], with both the encoder $\mathcal{E}$ and decoder $\mathcal{D}$ kept frozen. Additionally, we employ their time-conditional U-Net backbone for the latent diffusion process.

To enhance the conditioning mechanism, we introduce the encoder δ_θ, designed to mirror the U-Net encoder architecture. This structure enables the extraction of

Table 8.1. Quantitative comparison with state-of-the-art methods on both synthetic and real-world benchmarks. The best performance for each metric is in **bold**, while the second-best is underlined.

	Metrics	RealSR [240]	BSRGAN [241]	FeMaSR [242]	R-ESRGAN+ [88]	ResShift [243]	LDM [12]	StableSR [184]	Ours
DIV2K-Val	PSNR ↑	24.62	24.58	22.97	24.29	24.53	20.58	23.26	**25.21**
	SSIM ↑	0.5970	0.6269	0.5887	0.6372	0.7323	0.5762	0.5726	**0.645**
	LPIPS ↓	0.5276	0.3351	0.3126	**0.3112**	0.4406	0.3199	0.3114	0.4161
	FID ↓	49.49	44.22	35.87	37.64	49.16	26.47	**24.44**	25.43
RealSR	PSNR ↑	25.56	24.70	23.58	24.33	24.79	22.26	23.55	**26.45**
	SSIM ↑	0.7390	**0.7651**	0.7132	0.7456	0.7423	0.6462	0.7080	0.7627
	LPIPS ↓	0.3570	0.2713	0.2937	**0.2524**	0.3134	0.3159	0.3002	0.3215
DRealSR	PSNR ↑	27.79	26.18	24.56	25.82	27.87	23.39	24.85	**29.69**
	SSIM ↑	0.8148	0.8028	0.7569	0.8052	0.8056	0.7448	0.7536	**0.8211**
	LPIPS ↓	0.3938	**0.2929**	0.3157	0.2818	0.5408	0.3379	0.3284	0.3332

conditioning embeddings at multiple scales, which are subsequently incorporated into the conditional denoising autoencoder ϵ_θ, facilitating a more refined and hierarchical feature representation. We optimize the following objective:

$$L = \mathbb{E}_{\mathcal{E}(x),y,\epsilon\sim\mathcal{N}(0,1),t,c}\left[\|\epsilon - \epsilon_\theta(z_t, \delta_\theta(c,t,z))\|_2^2\right]. \tag{8.1}$$

By integrating quaternion wavelet embeddings and temporal guidance, our method effectively leverages the generative prior of SD for ship image super-resolution. This combination ensures high-fidelity reconstructions, enhanced detail preservation, and robustness across diverse scenarios.

8.2.3 Experimental Results

Experimental Setup

All experiments were conducted using an NVIDIA RTX A6000 GPU with 48GB of VRAM. We employ the AdamW optimizer with a learning rate of 5e-05. The batch size was set limited to 6 to balance memory efficiency and stable convergence. The model was trained for 14k iterations, starting from the released checkpoints of StableSR [184]. We perform super-resolution going from a resolution of 128x128 to 512x512, thus with a x4 factor. QUAVE was pre-trained separately for 45k steps on a combined dataset consisting of DIV2K [244], Flickr2K [245], and OutdoorSceneTraining (OST) [246], ensuring robust feature extraction across diverse image domains.

Quantitative Evaluation

We evaluated our model's performance across three benchmark datasets: DRealSR [3], RealSR [4], and DIV2K, each posing unique challenges for super-resolution. The quantitative results, summarized in Table 8.1, demonstrate the superior performance of our approach. We assess our super-resolution model using PSNR and SSIM (calculated on the Y channel in YCbCr space), FID, and LPIPS, ensuring a comprehensive pixel-level accuracy and perceptual quality evaluation. PSNR quantifies reconstruction fidelity by measuring pixel-wise differences, but it fails to align with human perception. SSIM improves upon PSNR by incorporating structural similarity, yet it remains limited in capturing perceptual realism. Conversely, FID evaluates realism by comparing feature distributions in a learned embedding space, effectively capturing perceptual discrepancies. LPIPS, leveraging deep network activations, provides a more human-aligned similarity measure. These perceptual metrics better reflect image quality, particularly in adversarial and generative models, where traditional pixel-wise approaches fall short [90].



Figure 8.2. Comparison of the LR input image with SR outputs generated by state-of-the-art methods and our proposed model on the DRealSR [3] and RealSR [4] datasets. The red bounding box highlights a zoomed-in region, illustrating the superior resolution and detail preservation achieved by our method compared to existing approaches.

On the DIV2K [244] dataset, our model exhibited particularly strong performance, achieving a higher SSIM and PSNR while having a competitive FID and LPIPS score. These results indicate the effectiveness of our method in maintaining both structural fidelity and perceptual quality, even on high-resolution images with diverse textures.

For the RealSR [4] and DRealSR [3] datasets, which include real-world degradations, our approach demonstrated strong generalization capabilities, achieving higher PSNR and SSIM which also indicates that our model effectively preserves fine-grained textures, leading to perceptually superior reconstructions.

Qualitative Analysis

To further evaluate the effectiveness of our approach, we conduct a qualitative comparison of super-resolved images, presented in Figure 8.2. The visual results underscore several key advantages of our method over existing techniques.

Our model excels in reconstructing intricate structural details and fine textures. The quaternion wavelet-based conditioning effectively preserves sharp edges while mitigating artifacts commonly observed in baseline methods. This is particularly

evident in the Lincoln portrait (the second to last image in the grid in Fig. 8.2), where our approach retains subtle facial features, such as fine wrinkles and hair textures, with minimal over smoothing. Compared to alternative methods, our model produces a more naturalistic representation, avoiding excessive blurring or unnatural sharpness.

Additionally, our method demonstrates enhanced robustness in reconstructing complex high-frequency textures, such as those present in the lynx's fur (the first row in the grid in Fig. 8.2). While traditional SR methods often introduce noise or fail to recover the natural stochastic patterns in such textures, our approach maintains structural coherence, effectively capturing both coarse and fine-scale fur patterns. The reduction in visual artifacts and improved textural realism contribute to a more perceptually pleasing output.

A particularly challenging aspect of super-resolution lies in handling textual elements, as fine details must be recovered with high fidelity to maintain readability. Our approach significantly outperforms previous methods in this regard, as demonstrated in the writing of "Gartner" text example (the fifth row in the grid of Fig. 8.2), where letter edges remain crisp, and subtle font characteristics are preserved without aliasing or distortion. Traditional methods tend to introduce blurring or edge artifacts, whereas our model retains sharpness and legibility, highlighting its advantages in text-based image enhancement applications.

Overall, our experimental findings validate the effectiveness of the proposed method in achieving state-of-the-art super-resolution performance. The ability to balance structural fidelity and perceptual quality across diverse content types, ranging from human portraits and natural textures to text elements, demonstrates its broad applicability in real-world image enhancement tasks.

Ablation Studies

Ablation Study on the number of sampling steps. For a fair comparison, we set the number of sampling steps to 200, following the configuration of StableSR [184]. Subsequently, we explored reducing this number and observed that our model benefits from fewer sampling steps. Figure 8.3 illustrates the impact of this reduction on key evaluation metrics. We discover that reducing the number of sampling steps introduces a tradeoff among the evaluation metrics. Specifically, we observe an increase in PSNR and SSIM, which suggests improved pixel-wise and structural similarity to the ground truth. However, LPIPS also increases, indicating a decline in perceptual quality. This suggests that while fewer steps enhance fidelity in terms of traditional similarity metrics, they may compromise perceptual realism. Therefore, the choice of sampling steps depends on the desired balance between pixel-wise accuracy and perceptual quality in practical applications.

Ablation Study on the Controllable Feature Wrapping. To thoroughly evaluate the impact of different Controllable Feature Wrapping (CFW) implementations, we conduct a comparative analysis between our custom-trained CFW module and the one introduced by StableSR [184]. The training pipeline generates synthetic LR-HR pairs specifically designed to enhance the ability of the module to maintain realistic textures and natural image statistics.

Table 8.2 presents quantitative results across different metrics for two model variants: our model with custom-trained CFW and our model utilizing CFW module pretrained by StableSR. The improvements are particularly notable in areas with complex geometrical patterns and sharp edges, suggesting that our training approach effectively leverages the structural information captured by the quaternion wavelet features, this is confirmed by the SSIM and PSNR metrics.

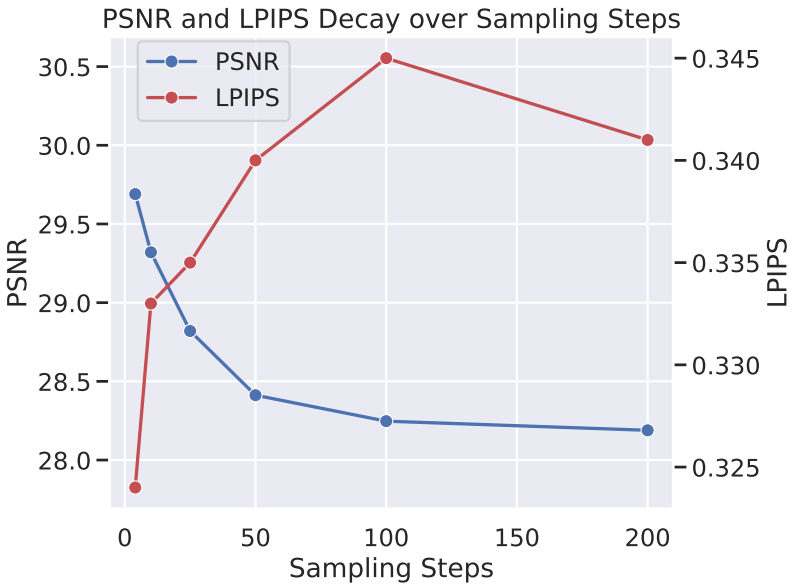


Figure 8.3. Impact of the number of sampling steps on key evaluation metrics. Reducing the steps improves efficiency while maintaining performance, highlighting the inherent speed advantage of our model.

Table 8.2. Metrics obtained with different versions of the CFW with our ResQu.

Dataset	CFW	PSNR ↑	SSIM ↑	FID ↓
DIV2K	StableSR	23.34	0.562	**24.52**
	Ours	**23.61**	**0.591**	25.42
RealSR	StableSR	23.26	0.702	**129**
	Ours	**24.98**	**0.714**	134
DrealSR	StableSR	27.94	0.734	151
	Ours	**28.18**	**0.755**	**151**

These results suggest that the choice of CFW implementation should be guided by specific application requirements. When structural fidelity is paramount, our custom-trained CFW provides superior results. In Figure 8.4 there is a visual comparison of the output generated with the different CFW.

Ablation Study on ShipSpotting Dataset. To assess the generalization capability of our model and its applicability to domain-specific tasks without retraining, we conduct a zero-shot evaluation on the ShipSpotting dataset [7]. This dataset comprises diverse maritime vessel images captured under varying conditions, presenting unique super-resolution challenges due to complex textures, fine structural details, and significant variations in scale.

As shown in Table 8.3, despite not being explicitly trained on this dataset, our model achieves performance comparable to fully trained approaches. Notably, ResQu attains SSIM and PSNR scores that closely match those of specialized models while maintaining a competitive FID score. This demonstrates its ability to generalize effectively to unseen domains, highlighting its robustness in handling domain-specific scenarios without requiring retraining.

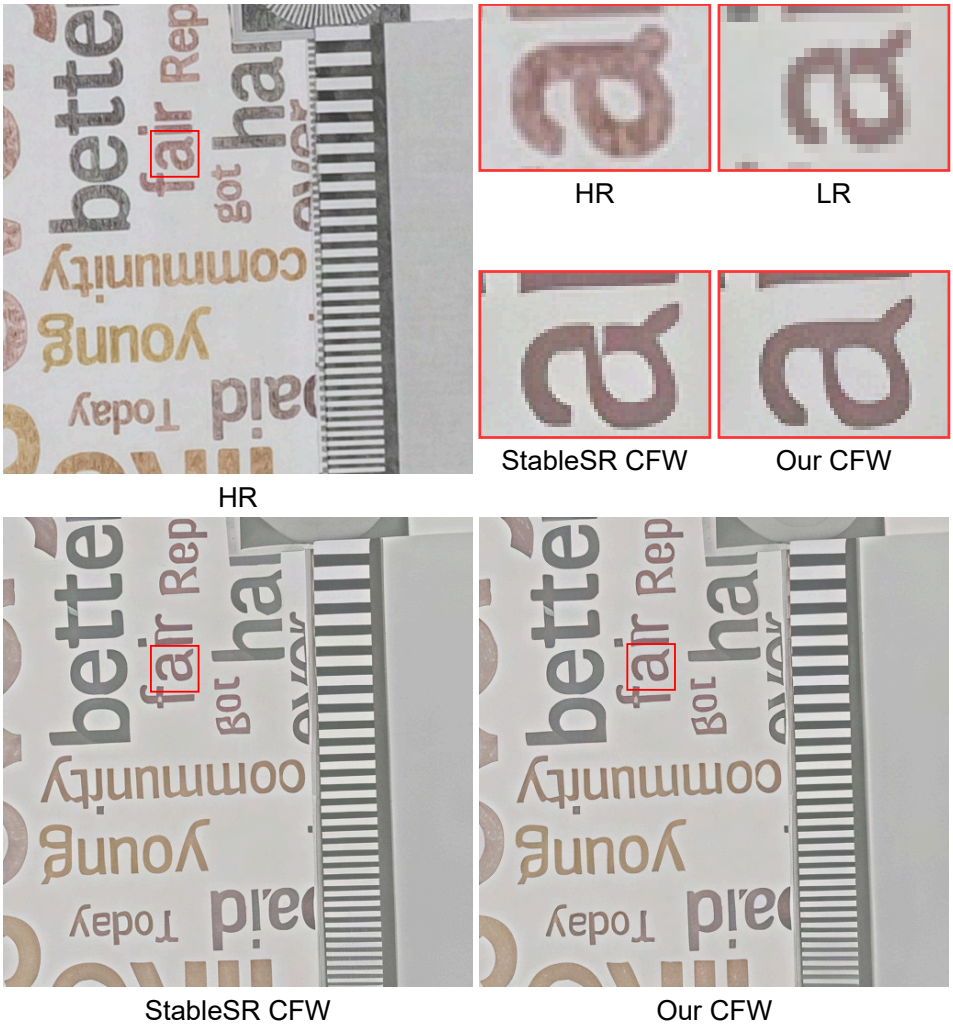


Figure 8.4. Visual comparison of diverse super-resolution results generated by changing the CFW. The images demonstrate the impact of our CFW enhanced by our encoder, producing more structured details.

The improvements are particularly evident in the reconstruction of intricate maritime structures, such as ship rigging, antenna arrays, and deck equipment, where traditional methods often struggle. Our quaternion-based representation captures essential phase relationships, ensuring sharper edges and enhanced texture preservation, which are critical for maritime imagery. Additionally, the multi-scale conditioning mechanism enables our model to consistently recover both large-scale structural elements (e.g., hull details) and fine-grained features (e.g., navigation equipment and water surface textures).

These findings validate the versatility of our approach, demonstrating that it not only achieves state-of-the-art results on its primary benchmark datasets but also extends effectively to domain-specific applications in a zero-shot setting.

Table 8.3. FID results obtained on the Shipspotting dataset presented in section 6.2.1.

	Model	SSIM ↑	PSNR ↑	FID ↓
Full Training	SR3 [90]	0.569	21.07	16.59
	StableSR [184]	0.554	20.29	<u>12.41</u>
	StableShip-SR [7]	0.561	20.51	**11.72**
Zero-Shot	ResQu	**0.611**	**22.01**	14.63

Concluding Remarks on Quaternion Wavelet Conditioning

The ResQu framework successfully demonstrates that conditioning a latent diffusion model on learned quaternion wavelet features leads to significant improvements in super-resolution quality, particularly in preserving fine details and textures. Having established the power of these complex hypercomplex representations for enhancing general-purpose SR, our focus now shifts to a different challenge: adapting these principles to highly specialized data domains. The following section explores how a similar wavelet-aware conditioning strategy can be tailored for satellite imagery, a domain where contextual metadata is as crucial as the image content itself.

8.3 Metadata, Wavelet, and Time Aware Diffusion Models for Satellite Image Super Resolution

8.3.1 Feature Modulation

The conditional generation approach for SR with diffusion models leverages the latent representation of degraded input images, extracted through an encoder, to guide the generation process.

As we have seen in the previous section 8.2, to enhance the guidance during generation, in StableSR [21], an additional encoder is introduced to extract multi-scale features $\{F^n\}_{n=1}^N$ from the degraded LR image. These extracted features were subsequently used to modulate the intermediate feature maps, $\{F_{\text{dif}}^n\}_{n=1}^N$, within the residual blocks of Stable Diffusion (SD) [12] via spatial feature transformations (SFT) introduced by [246].

During fine-tuning, the weights of the SD model are frozen, and only the encoder and the SFT layers are trained. This approach ensured the preservation of the generative prior inherent in SD while allowing structural information from the LR image to be incorporated effectively.

Additionally, the capacity of the model to integrate textual information is improved by utilizing CLIP-encoded text embeddings during training. This enhancement facilitates a more nuanced understanding of both the image category and the country metadata embedded in the dataset.

8.3.2 MWT-Encoder

We introduce a metadata- and wavelet- and time-aware encoder (MWT-Encoder) as the core of the baseline model architecture inspired by the work presented in chapter 6.2 that introduced a class- and time-aware encoder, and [247], which highlighted the importance of metadata (e.g., timestamp, latitude, longitude) associated with satellite images.

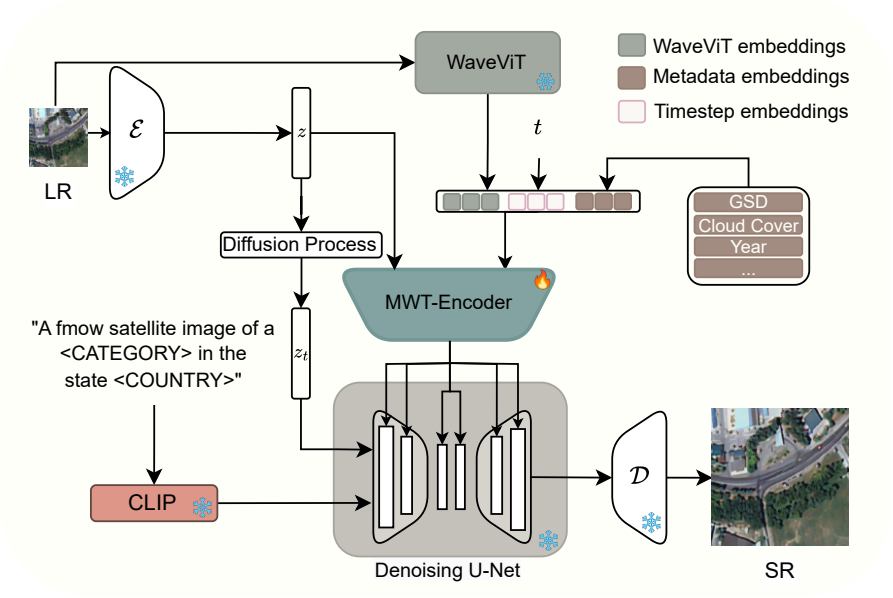


Figure 8.5. Overview of MWT-Diff framework with our MWT-Encoder.

In this context, each image $x \in \mathbb{R}^{3\times H\times W}$ is paired with numerical metadata $K \in \mathbb{R}^M$, where M represents the number of metadata items. Following the methodology introduced in StableSR, we incorporate temporal information into the encoder using a time-embedding layer. We encode metadata, as done in [247], using the sinusoidal timestep embedding. This is commonly employed in diffusion models since incorporating metadata directly into text captions using pre-trained CLIP models often leads to the discretization of continuous values, which introduces inaccuracies due to CLIP's limitations in encoding numerical information [248]. The benefits of this integration were analyzed from a signal-to-noise ratio (SNR) perspective and validated both quantitatively and qualitatively.

Recent works show that the use of wavelets can improve SR within diffusion models [34, 249]. Our idea is to take advantage of the wavelets, not to perform diffusion in frequency space [25], but instead to extract features from the wavelet transform of the satellite images through the pre-training of a Wavelet Vision Transformer (WaveViT) [130].

WaveViT consists of four stages, each comprising a patch embedding layer, a series of Wavelet blocks, and subsequent feed-forward layers. In the initial phase of the architecture, discrete wavelet transforms (DWT) are applied to the input images. This operation decomposes the images into multiple frequency components, enabling the model to simultaneously capture both low-frequency (global structure) and high-frequency (fine detail) information. This is particularly critical for preserving the intricate textures and details commonly found in satellite imagery. After pre-training, which extracts robust feature representation leveraged throughout the network to perform the super-resolution task, we integrate WaveViT in our architecture, concatenating the embeddings along with the timesteps and metadata embeddings.

The concatenated metadata tensors $m \in \mathbb{R}^{1024}$, wavelet embeddings $w \in \mathbb{R}^{1024}$, and timestep embeddings $t \in \mathbb{R}^{1024}$ yield a vector $b \in \mathbb{R}^{3072}$, which is combined with the latent representation of the image $z \in \mathbb{R}^{h\times w\times c}$. The latent z is obtained from the encoder $z = \mathcal{E}(x)$, where $x \in \mathbb{R}^{3\times H\times W}$ is the low-resolution image. Both the

encoder $\mathcal{E}$ and the decoder $\mathcal{D}$ are frozen components of the VAE used in SD [12], alongside the time-conditional U-Net responsible for performing the latent diffusion process.

The MWT-Encoder δ_θ generates conditioning embeddings that guide the denoising process of the conditional auto-encoder ϵ_θ at multiple scales. These embeddings are incorporated following the architecture of the U-Net. The complete objective function for training the model is defined as:

$$L = \mathbb{E}_{E(x),y,\epsilon\sim\mathcal{N}(0,1),m,w,t}\left[\|\epsilon - \epsilon_\theta(z_t, \delta_\theta(m, w, t, z), \tau(y))\|_2^2\right] \tag{8.2}$$

In equation 8.2, $\tau(y)$ represents the domain-specific encoder, implemented using a frozen CLIP model, which processes the conditioning caption y. The variable m corresponds to the metadata embeddings, w is the wavelet embedding, t is the timestep embedding, z is the latent representation of the low-resolution image extracted by the VAE encoder, and z_t is the latent representation after the addition of noise during the diffusion process.

The objective is to optimize the MWT-Encoder $\delta_\theta(m, w, t, z)$ to provide effective conditioning embeddings. These embeddings ensure that the denoising process generates HR images that faithfully capture the information encoded in both the input metadata and the conditioning captions. The final MWT-Diff model used for the super-resolution task is illustrated in Figure 8.5 within our MWT-Encoder.

8.3.3 Experimental Results



Figure 8.6. Qualitative results on 128x128 → 512x512 with fMoW [5].

We evaluate our model on two datasets. The first dataset is a subset of the fMoW dataset [5], used for the super-resolution task of 4x scaling, where 128x128 pixel low-resolution images are upscaled to 512x512 pixel high-resolution images. The second dataset is a subset of the Sentinel2-fMoW dataset [6], employed for the super-resolution task on hyperspectral images. For inference we adopt the DDPM [11] sampling strategy utilizing 200 timesteps.

The results underscore the superior performance of our model across both the fMoW and hyperspectral image super-resolution tasks, as presented in Table 8.4. By integrating our MWT-Encoder, the model significantly enhances its ability to

	Model	FID ↓	LPIPS ↓
fMoW	Low Resolution	114.38	0.756 +/- 0.004
	StableSR	53.85	0.345 +/- 0.002
	MWT-Diff	**53.07** (-1.44%)	**0.336** +/- 0.002 (-2.61%)
Sentinel2-fMoW	WorldStrat [230]	426.7	0.736 ± 0.092
	MSRResNet [250]	286.5	0.783 ± 0.081
	DBPN [251]	278.2	0.750 ± 0.052
	Pix2Pix [16]	196.3	0.643 ± 0.045
	SatDiffMoE [252]	115.6	0.606 ± 0.044
	DiffusionSat [247]	102.9	0.638 ± 0.034
	ControlNet [27]	102.3	0.644 ± 0.034
	MWT-Diff	**98.1** (-4.11%)	**0.555** ± 0.003 (-8.41%)

Table 8.4. Evaluation metrics comparison for the $128 \times 128 \rightarrow 512 \times 512$ super resolution task on fMoW dataset [5], and on Sentinel2-fMoW dataset [6].

reconstruct intricate features and capture fine-grained details within the images. MWT-Diff consistently outperforms its counterparts, including traditional CNN-based methods and other diffusion models, as evidenced by its superior performance in both FID and LPIPS metrics.



Figure 8.7. Additional results on 128x128 → 512x512 with fMoW [5]. Comparison between MWT-Diff and StableSR.

8.3.4 Dataset details

Functional Map of the World (fMoW)

The metadata accompanying each image of the fMoW dataset [5] provides contextual information, including location, time, sun angles, physical sizes, and other attributes, facilitating a more informed analysis of the objects within the images.

The original fMoW dataset consists of over 1 million satellite images collected from more than 200 countries, each annotated with at least one bounding box corresponding to one of 63 categories. For the purpose of conducting a 4x scaling super-resolution task, transforming 128×128 pixel LR images into 512×512 pixel HR images we select a balanced subset of 100k samples from the original dataset, ensuring proportional representation of all 63 classes.

To generate the HR images, a resize operation to 512×512 pixels followed by a center crop was applied, preserving the critical details of the images. The LR images were derived from the HR counterparts by degrading them using various techniques outlined Real-ESRGAN [88], effectively reducing their resolution while simulating realistic degradations.

Additionally, to establish a baseline for comparison with the super-resolution model outputs, a standard upscaling technique was applied to the LR images, creating reference images for evaluation. The resulting refined dataset is optimized and prepared for training and benchmarking super-resolution models.

Functional Map of the World - Sentinel-2 (fMoW-Sentinel)

The dataset [6] comprises temporal image series collected by the Sentinel-2 satellite, corresponding to the geographic locations of the original fMoW dataset. These images were captured at multiple time points and are categorized into 62 distinct types of building and land use, following the classification scheme of the fMoW dataset.

The fMoW-Sentinel dataset consists of Sentinel-2 images with a 10-meter spatial resolution, created from cloud composites aggregated over 90-day intervals. Each image includes 13 spectral bands from the Sentinel-2 surface reflectance dataset, offering a comprehensive multispectral representation. Building on previous works [6, 252], we use a derived dataset named fMoW-Sentinel-fMoW-RGB. This dataset pairs Sentinel-2 images (10m–60m ground sampling distance, GSD) with high-resolution fMoW-RGB images (0.3m–1.5m GSD) at each of the original fMoW locations.

A key distinction from SatdiffMoE is the handling of high-resolution images. While it uses bounding boxes provided in the metadata to extract relevant areas and crop them into 512×512 patches, the fMoW high-resolution images in this work were resized in their entirety to 512×512 pixels.

For each high-resolution fMoW image, we select one corresponding Sentinel-2 image from the same location. Similar to SattdiffMoE, only the RGB bands of the Sentinel-2 images were retained. These RGB images were then resized to match the dimensions of the fMoW-RGB images, ensuring consistency in resolution for downstream tasks.

8.3.5 Training details

The MWT-Diff model is trained for eight and four epochs, respectively. Further training beyond these limits resulted in worsening FID scores. The Adam optimizer was employed, with a learning rate set to 5e-5. An instance of the WaveViT-B model

was pre-trained and evaluated directly on the fMoW dataset. This evaluation was conducted on 1,000 images from the fMoW validation set.

8.3.6 Evaluation details

In line with the methodology presented by [252], the evaluation of the WaveViT is further refined by focusing on the first 100 images from the validation set across six specific classes: Airport, Amusement Park, Car Dealership, Crop Field, Educational Institution, and Electric Substation. We generate a synthetic set of captions for each class of the fMoW dataset using the pattern: "A fMoW satellite image of a class_type". Each image is paired with these synthetic captions and input into the model under evaluation. The model then predicted the most relevant caption associated with each image.

8.3.7 Additional Qualitative Comparisons

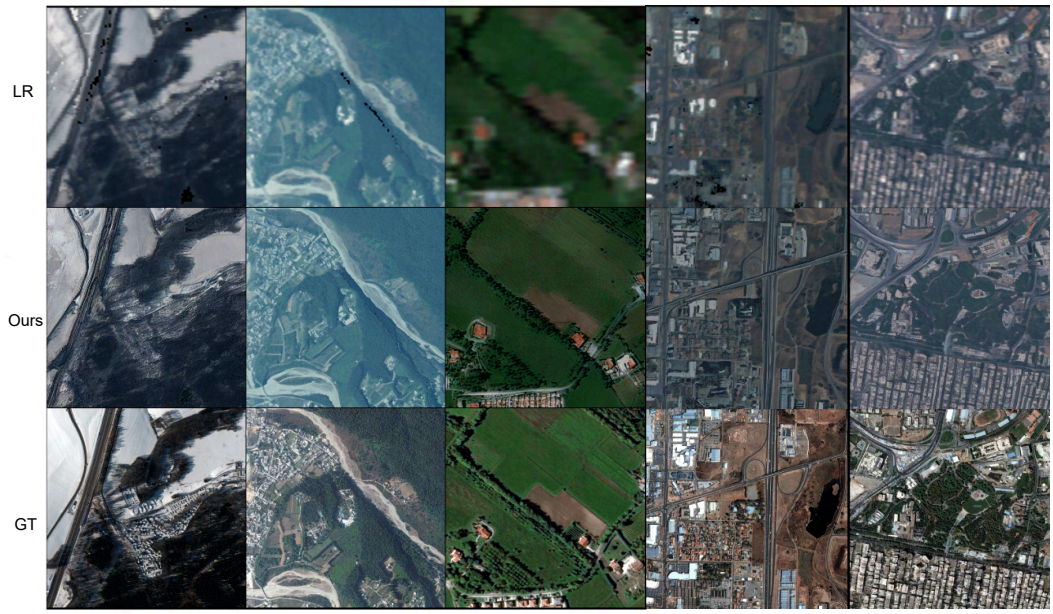


Figure 8.8. Qualitative results on Sentinel2-fMoW dataset [6]. The image includes a comparison of the LR Sentinel-2 input, the output of the model, and the corresponding HR fMoW images.

Figure 8.8 illustrates qualitative results on the Sentinel2-fMoW dataset [6], showcasing the low-resolution Sentinel-2 input, the super-resolved output generated by MWT-Diff, and the corresponding high-resolution fMoW reference image. Figure 8.7 depicts an additional comparison on the standard fMoW dataset [5], where MWT-Diff is directly compared against the StableSR baseline.

The results demonstrate consistent qualitative improvements across datasets. In the Sentinel2-fMoW case, MWT-Diff successfully reconstructs fine-grained structures despite the challenging resolution gap between Sentinel-2 inputs and fMoW ground truth. Meanwhile, in the fMoW comparison, MWT-Diff outputs exhibit clearer boundaries and more coherent textures than those produced by StableSR, indicating the effectiveness of our metadata- and frequency-aware conditioning mechanisms.

These examples highlight the model's generalization capability and the robustness of its generative priors across different spatial resolutions and sensor characteristics.

8.3.8 Ablation Study

We conduct a series of ablation experiments to evaluate the individual contributions and trade-offs of key components in the MWT-Diff framework. This analysis explores the impact of removing conditioning modules and examines design choices that affect the model's complexity, efficiency, and applicability.

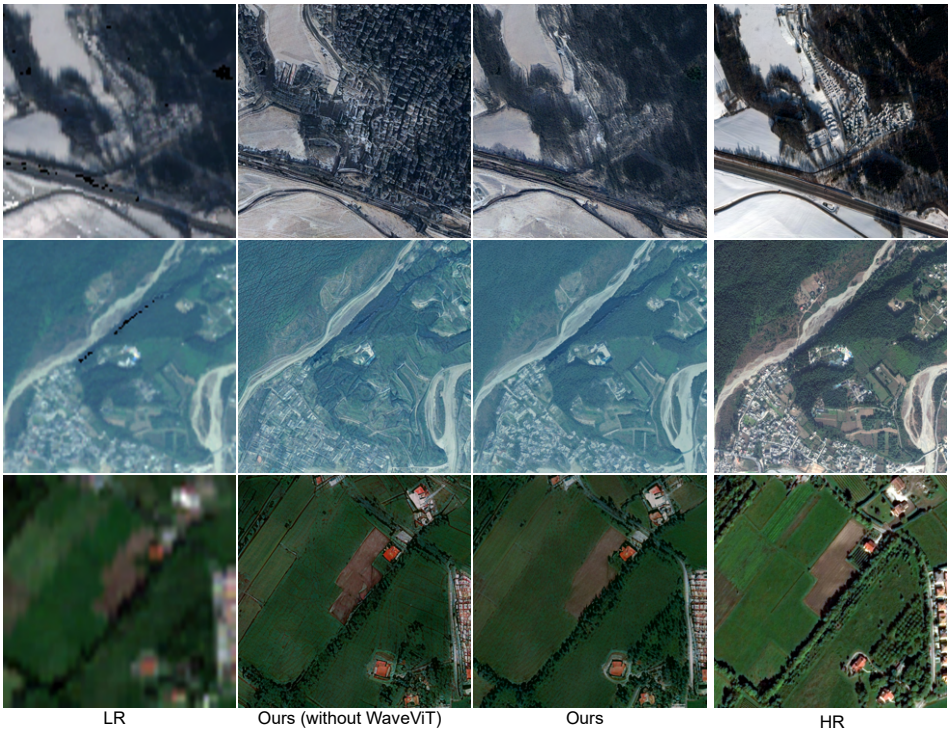


Figure 8.9. Qualitative comparison of super-resolution outputs with and without WaveViT embeddings. From left to right: Low-resolution input, MWT-Diff without WaveViT, MWT-Diff with WaveViT, and Ground Truth.

CLIP-based Text Encoder

To assess the necessity of the CLIP-based caption encoder, we removed this component from the architecture and retrained the model. The goal was to investigate whether this simplification would reduce computational overhead or accelerate inference. Empirical results indicated no measurable improvement in sampling speed. On the contrary, a slight degradation in performance was observed across FID and LPIPS metrics. These findings suggest that while the CLIP encoder introduces additional complexity, it provides marginal gains in perceptual quality and semantic alignment, which justify its inclusion in the final model.

Comparison with StableSR and Implications for Real-World Deployment

We further examined the computational cost of MWT-Diff relative to StableSR. Despite the integration of additional embeddings (metadata, wavelet, and temporal

encodings), the inference time remained comparable. This parity is attributed to the lightweight nature of the MWT-Encoder and the use of a single-image input pipeline. Notably, when compared to methods that requires multi-image fusion [230, 247, 252] from different timestamps, MWT-Diff demonstrates superior practicality. The absence of complex temporal alignment and fusion procedures enables more straightforward deployment in real-world scenarios where temporally-aligned sequences may not be available.

Impact of RGB-Only Input

Our current implementation utilizes only the RGB bands of the Sentinel-2 imagery. This choice was made to maintain a lean architecture, facilitating efficient training and inference while enhancing the usability of the model in operational environments with limited multispectral data. Although RGB data suffice for a broad range of super-resolution tasks, we recognize the potential advantages of incorporating additional spectral channels. Future work will investigate the integration of multispectral and hyperspectral inputs to assess their impact on reconstruction accuracy and downstream utility, particularly in applications such as vegetation monitoring and land use classification.

Effect of Wavelet-Based Embeddings (WaveViT)

To evaluate the contribution of frequency-aware representations, we conducted an experiment disabling the WaveViT component. Figure 8.9 shows a qualitative comparison of super-resolved samples with and without WaveViT-based embeddings.

From visual inspection, it is evident that the inclusion of WaveViT significantly enhances the perceptual sharpness and texture fidelity of the generated images. Samples generated with WaveViT appear less blurry and less noisy, exhibiting structural characteristics that more closely resemble the ground truth. These results confirm that multi-scale frequency features, as captured by WaveViT, play a critical role in recovering fine details and reducing generative artifacts.

Concluding Remarks on Domain-Specific Conditioning

The MWT-Diff model illustrates the adaptability of wavelet-based conditioning, proving its effectiveness in a specialized domain like satellite imaging, where metadata and frequency analysis are paramount. While the ResQu and MWT-Diff frameworks demonstrate significant gains in image fidelity, they inherit the computational demands of large-scale diffusion models. Acknowledging this practical limitation, the final investigation in this chapter tackles the critical issue of efficiency. We now pivot from enhancing quality to accelerating the process, exploring how operating entirely within the wavelet domain within a GAN framework can drastically reduce training and inference costs without sacrificing generative performance.

8.4 Wavelet Diffusion GAN for Image Super-Resolution

In this section, we propose a methodology that aims to overcome the inherent challenges of image super-resolution, specifically the protracted inference and training periods resulting from the utilization of a diffusion model. This is achieved through the incorporation of the wavelet transform within a conditional Diffusion GAN architecture [231]. We allow larger step sizes in the reverse diffusion process by using

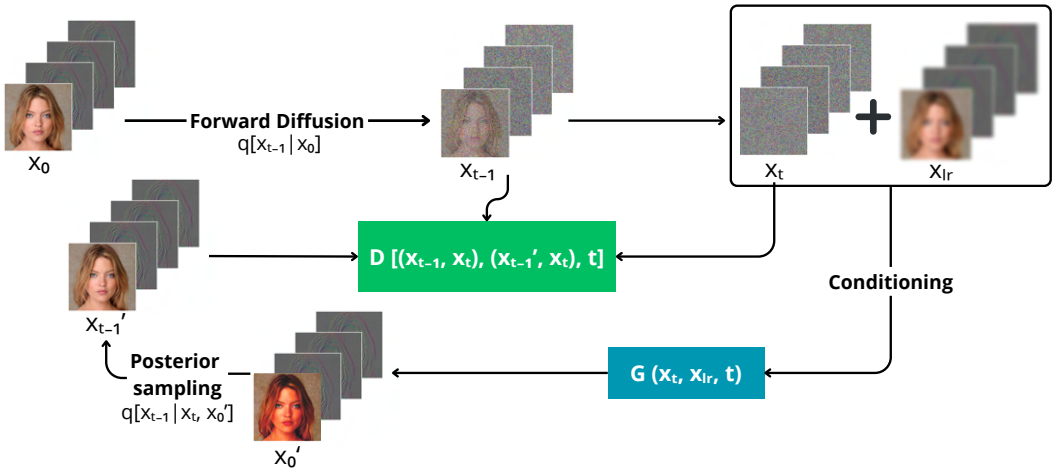


Figure 8.10. Method architecture and training scheme. In green our discriminator and in blue our conditional generator. x_0 undergoes forward diffusion in wavelet space and the resulting pure noise x_t gets concatenated to the low-res input x_{lr} to condition the generator for the backward diffusion.

a GAN. We use the low-resolution images as conditioning input to the model, but we perform the whole diffusion process in wavelet space instead of pixel space.

8.4.1 Wavelet-based conditional diffusion scheme

Figure 8.10 illustrates our method. The forward diffusion process follows the traditional definition [11], but it is done in wavelet space instead of pixel space: noise gets progressively added to x_0, the wavelet decomposition of the high-resolution input, until only pure Gaussian noise x_t is remaining. This pure Gaussian noise will be concatenated to x_{lr}, the wavelet decomposition of the low-resolution input image, to model the backward diffusion process.

During the reverse process, our generator, given x_t and x_{lr} will produce an unperturbed prediction x'_0 of the initial high-res image, which will be used to take from the tractable posterior distribution the perturbed sample x'_{t-1}. In inference time, this sampling process will be iterated for a predefined number of timesteps T (≤ 8) to complete the denoising process and acquire the super-resolved image.

The role of the discriminator during the training phase is to distinguish between real pairs (x_{t-1}, x_t, t) and fake pairs (x'_{t-1}, x_t, t).

8.4.2 Conditional backward diffusion

Instead of the traditional reverse diffusion approach which requires several thousands of steps to properly denoise the image, the GAN utilization in our architecture enables us to do bigger steps in the diffusion process (Figure 2). This assumption implies that a Gaussian distribution cannot approximate the reverse process but instead, we use a multimodal distribution, which will be implicitly modeled by the time-dependent GAN generator $G(x_t, x_{lr}, t)$. In this formulation, the model does not directly predict x_{t-1}. Instead, it predicts the clean image x_0 and uses the known diffusion process to obtain x_{t-1}. Specifically x_t is the noisy image at timestep t; x_{lr} is obtained by first scaling the low-resolution input to the target resolution via bicubic interpolation, and then applying a discrete wavelet transform to decompose it into multiple frequency sub-bands; $f_\theta(x_t, x_{lr}, t)$ is the

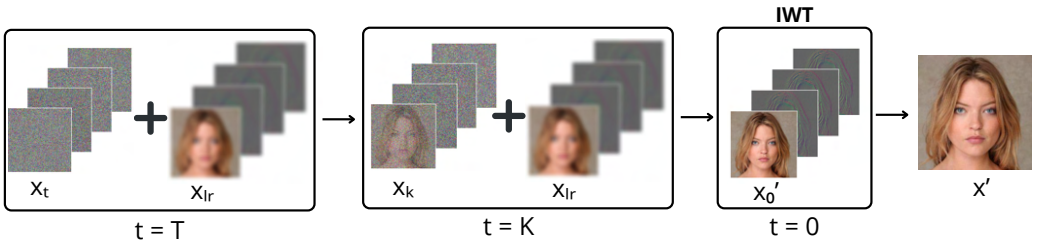


Figure 8.11. Reverse diffusion process and inference: the model p_θ iteratively produces a more refined sample from x_t and x_{lr}. After T iterations, x'_0 is used to reconstruct the super-resolved image.

neural network (the "denoising generator") that estimates the clean image x_0. Once x_0 is estimated, the probability distribution $p_\theta(x_{t-1} \mid x_t, x_{lr})$ is defined using the standard posterior from the forward diffusion process, but with x_0 replaced by the network's estimate $f_\theta(x_t, x_{lr}, t)$. Formally:

$$p_\theta(x_{t-1} \mid x_t, x_{lr}) := q\big(x_{t-1} \mid x_t,\, x_0 = f_\theta(x_t, x_{lr}, t)\big). \tag{8.3}$$

In other words, the model learns to reconstruct the clean image x_0 at each timestep, and the actual sampling step to get x_{t-1} follows the known diffusion posterior with this estimated x_0.

8.4.3 Optimization

Given the time-dependent discriminator $D(x_{t-1}, x_t, t) : \mathbb{R}^N \times \mathbb{R}^N \times \mathbb{R} \to [0, 1]$, and the generator $G(x_t, x_{lr}, t)$, the training process is set up in an adversarial manner to match the conditional GAN generator $p_\theta(x_{t-1}|x_t, x_{lr})$ and the posterior distribution $q(x_{t-1}|x_t, x_0)$, optimizing both the generator G and the discriminator D using adversarial losses:

$$L^D_{adv} = -\log(D(x_{t-1}, x_t, t)) + \log(D(x'_{t-1}, x_t, t)) \tag{8.4}$$

$$L^G_{adv} = -\log(D(x'_{t-1}, x_t, t)) \tag{8.5}$$

Furthermore, following WaveDiff [25], we use a reconstruction term to fully exploit the wavelet sub-bands. This term serves a dual purpose: preventing the loss of frequency information and maintaining the consistency of wavelet sub-bands. The formulation involves an L1 loss between the generated image and its ground-truth counterpart:

$$L_{rec} = \|x'_0 - x_0\| \tag{8.6}$$

The total objective of the generator will be a linear combination of the adversarial and the reconstruction losses:

$$L^G = L^G_{adv} + \lambda L_{rec} \tag{8.7}$$

8.4.4 Experimental results

Evaluation Under Practical Constraints. To demonstrate the method's efficiency and suitability for real-world applications, this study was conducted under practical resource constraints. The goal was to achieve state-of-the-art results without relying on massive computational budgets, thereby focusing on the model's

performance in scenarios where training and inference time are critical factors. We evaluate the model for face super-resolution to demonstrate our approach's effectiveness, comparing it to several other state-of-the-art approaches. Finally, we perform a small ablation study on a real-world dataset of ships, which we presented in chapter 6.2.1.

Training details

We train all the models for 25k iteration steps.

Dataset Following [20], we experiment on CelebA-HQ [73], a large-scale face attributes collection of 30k celebrity images. We employ bicubic interpolation to produce LR-HR image pairs, resizing all images to match the task of 8x magnification factor 16x16 → 128x128.

Generator architecture We employed a U-Net-like architecture for the generator: our implementation is based on WaveDiff [25], but we remove the conditioning on the latent variable z since it is not needed for our purposes. The wavelet information is integrated into the feature space of the generator: instead of using the standard downsampling and upsampling blocks, frequency-aware blocks and frequency-aware residual connections are employed to enhance the generator's awareness of high-frequency components, resulting in sharper and more detailed images.

Experimental setup We trained our model on the CelebA-HQ (16, 128) dataset with learning rates set to 2×10^{-4} and 1×10^{-4} for the generator and discriminator, respectively. The generator architecture uses base channels of 64, with a channel multiplier following the pattern [1, 2, 2, 2, 4] and employs two ResNet [186] blocks per level. The discriminator consists of 6 layers, and we utilize 2 timesteps in the reverse diffusion process for super-resolution. For optimization, we employ the Adam optimizer with decay rates $\beta_1 = 0.5$ and $\beta_2 = 0.9$, along with an Exponential Moving Average (EMA) with a decay factor of 0.9999. Training was conducted over 25k iteration steps using a batch size of 64 and lazy regulation with a value of 10 to ensure stable training dynamics.

Quantitative results

Metric	ESRGAN [250]	SR3 [20]	DiWa [253]	**Ours**
PSNR ↑	<u>21.13</u>	14.65	13.68	**23.38**
SSIM ↑	<u>0.59</u>	0.42	0.13	**0.68**
LPIPS ↓	<u>0.082</u>	0.365	0.336	**0.061**
FID ↓	**20.8**	99.4	270	<u>47.2</u>
Runtime	0.04s	60.3s	34.7s	<u>0.12s</u>
Parameters	31M	98M	92M	<u>57M</u>

Table 8.5. Evaluation metrics comparison for our model, ESRGAN, SR3 and DiWa trained for 25k iteration steps on CelebA-HQ for the task of 16x16 → 128x128.

Inference time and parameters We also opted to compare the models in terms of inference time and parameters to show how powerful and lightweight our method is compared to the other baselines.

Table 8.5 demonstrates that, even with a limited number of iterations ($25k$), our method achieves excellent performances and surpasses the other baselines in the 16x16 → 128x128 task. Notably, even though our model has a significantly



Figure 8.12. Qualitative comparison between our model, ESRGAN, SR3 and DiWa trained for 25k iteration steps on CelebA-HQ for the task of 16x16 → 128x128.

lower number of parameters compared to the diffusion baselines, we still achieve substantial performance improvements, thus delivering state-of-the-art results with considerably reduced computational complexity and inference time, showing that our approach is suitable for real-time image super-resolution applications.

Qualitative results

Due to the significant magnification factor (8x), there are numerous possible outputs. Therefore, an exact match with the reference image is not expected. Figure 8.13 reveals that our model excels in enhancing image details, providing a superior visual fidelity compared to the ESRGAN [250], SR3 [20] and DiWa [253] counterparts. The generated images showcase a significant reduction in artifacts, preserving finer details and avoiding common distortions often observed in all of the other baselines. More specifically, the traditional diffusion baselines suffer from annoying crosshatch artifacts, along with color shift artifacts, where the color distribution of the reconstructed image does not correspond with that of the target image, as noted in [254]. Notably, even with a number as low as 25k iteration steps, our method provides state-of-the-art results and a high degree of visual fidelity, further proving its suitability for real-time Image Super-Resolution applications and tasks demanding high-fidelity output. We conducted an extensive ablation



Figure 8.13. Results on Shipspotting [7].

study utilizing a real-world dataset of ships, specifically the Shipspotting dataset introduced in chapter 6.2.1, to evaluate the performance of our super-resolution framework under practical conditions. This dataset, comprising diverse ship images captured in real-world scenarios, provided a robust testbed for assessing the ability of our model to recover fine-grained details, preserve structural integrity, and generalize across varying environmental conditions. This rigorous evaluation underscores the practical applicability of our approach in real-world settings, particularly in domains such as maritime surveillance and remote sensing, where high-resolution imagery is critical.

8.5 Conclusion

This chapter presented three complementary approaches that push the boundaries of image super-resolution by leveraging wavelet and quaternion representations within diffusion and diffusion-GAN-based frameworks. Across the presented methods, the integration of multi-scale frequency features and structured priors proved critical for achieving state-of-the-art performance, enhancing both perceptual quality and fine-detail preservation.

The Quaternion Wavelet-Conditioned Diffusion Model introduced a hybrid latent diffusion architecture enhanced with hypercomplex signal representations. This model demonstrated robust performance across challenging conditions and datasets, particularly in preserving fine-scale structure and textural details, thus setting a new benchmark in super-resolution.

The Metadata and Time Aware Diffusion Model for satellite image super-resolution showcased the power of context-aware conditioning. By integrating wavelet decompositions with metadata and temporal cues, it achieved improved image fidelity and realism, offering utility in high-impact application areas such as urban planning and environmental monitoring.

Finally, the Wavelet Diffusion GAN combined fast, efficient training with high-resolution output quality, thanks to the use of discrete wavelet transforms for

frequency-aware modeling. It addressed key limitations in computational cost and inference time while maintaining strong generative performance, making it well-suited for scenarios with resource constraints.

Together, these methods demonstrate a key thesis of this work: that wavelet-based modeling is not a single solution, but a versatile toolkit for enhancing generative models. We have shown how it can improve detail fidelity (ResQu), adapt to specialized domains (MWT-Diff), and drastically improve efficiency (WadiganSR). Having assembled and validated these crucial components, we are now equipped to tackle our most ambitious goal. The following chapter 9 will synthesize these learnings into a single, unified framework designed to overcome the final frontier of generative modeling: efficient, ultra-high-resolution synthesis.

Chapter 9

High-Fidelity Synthesis Beyond Latent Diffusion

Contents

9.1 Introduction 116
9.2 Latent Wavelet Diffusion for 4K Image Synthesis 118
9.2.1 VAE Fine-Tuning with Scale-Consistency Loss 119
9.2.2 Reconciling Spectral Regularization and Saliency 120
9.2.3 Wavelet-Derived Frequency Saliency Maps 120
9.2.4 Adaptive Flow Matching with Frequency-Guided Masking 121
9.2.5 Experiments 122
9.3 Conclusion and Future Work 130

HIGH-RESOLUTION image synthesis remains a core challenge in generative modeling, with existing approaches facing a fundamental trade-off between computational efficiency and visual quality preservation. Methods achieving high visual quality typically require substantial resources, while efficient approaches sacrifice fine details essential for convincing high-resolution imagery.

This chapter[1] presents *Latent Wavelet Diffusion (LWD)*, a lightweight framework enabling any latent diffusion model to scale to ultra-high-resolution generation without additional computational overhead. By integrating wavelet-based analysis into the latent diffusion training process, LWD achieves remarkable improvements in both efficiency and quality for high-resolution synthesis tasks.

9.1 Introduction

Diffusion models have become a dominant paradigm in generative modeling, achieving state-of-the-art results in tasks such as image synthesis [29, 255–257], super-resolution [7, 20, 258], and image editing [259, 260]. Foundational methods like Denoising Diffusion Probabilistic Models (DDPM) [11] and Denoising Diffusion Implicit Models (DDIM) [261] have enabled increasingly powerful variants. Latent Diffusion Models (LDMs) [12] improve efficiency by operating in a learned latent space, while recent architectures such as Diffusion Transformers (DiTs) [23, 262] further enhance scalability and modeling capacity.

[1]The content of this Chapter is adapted from the material under review in [35]

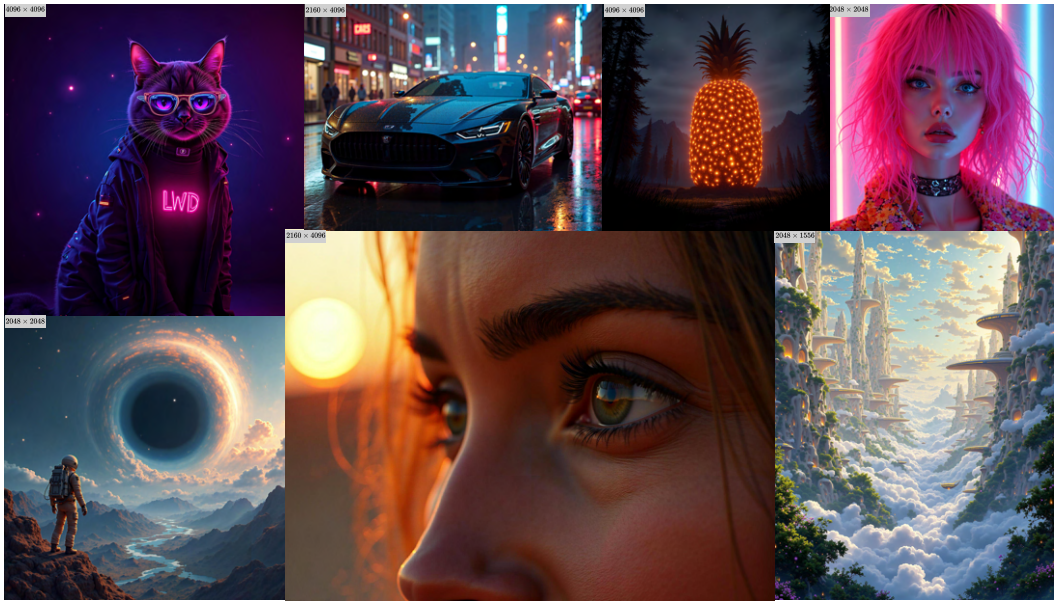


Figure 9.1. We propose Latent Wavelet Diffusion, achieving 4K image synthesis without architectural changes or additional inference cost to existing latent diffusion models.

Despite recent advances, generating ultra-high-resolution (UHR) images at resolutions of 2K to 4K and beyond remains a significant challenge. Simply scaling models trained on lower-resolution data often fails to generalize, leading to repeated structures, blurred textures, and spatial inconsistencies [64, 65]. Naive alternatives, such as hierarchical generation pipelines and post-hoc super-resolution techniques [65, 263], typically produce oversmoothed outputs. To overcome these limitations, several strategies have been explored. One direction involves direct UHR training or fine-tuning [264, 265], although this typically demands extensive computational resources, access to proprietary high-resolution datasets [266], and substantial GPU memory for large model backbones [54, 262]. Other efforts focus on architectural modifications to improve long-range dependency modeling [267], or optimization techniques that enhance the quality of latent representations [268].

Across these varied approaches, a common limitation persists: most methods treat all spatial regions uniformly during generation, applying the same refinement process to areas with vastly different visual complexity. This uniformity disregards local frequency variation, failing to distinguish between smooth regions and areas rich in textures, edges, or semantic structure. The consequences are twofold. First, computation is wasted on low-detail areas that require minimal refinement. Second, high-detail regions are not sufficiently supervised, leading to artifacts or loss of fidelity. These issues stem from both architectural and algorithmic constraints: latent representations often lack the structural granularity required for UHR synthesis, and current diffusion training objectives do not incorporate spatial adaptivity into the denoising process. Together, these limitations present a core bottleneck for perceptually accurate ultra-high-resolution image generation.

In this chapter, we propose *Latent Wavelet Diffusion* (LWD), a general and modular framework that introduces frequency-sensitive spatial supervision into the latent denoising process of diffusion models. LWD is motivated by the observation that different regions of an image exhibit varying levels of structural complexity and perceptual importance. While some areas contain intricate textures or sharp

edges, others are homogeneous or low in detail. Our goal is to exploit this spatial heterogeneity by allocating greater learning signal to regions with high visual complexity, and reducing supervision in low-detail areas. Importantly, LWD achieves this adaptivity without modifying the underlying architecture of the diffusion model, making it broadly applicable across model families.

The LWD framework consists of three key components:

1. A **spectrally-aware VAE fine-tuning objective** that improves the structure and diffusability of latent representations. By encouraging semantic consistency and frequency regularity, this objective enhances the suitability of latent spaces for high-resolution generation. It serves as the foundation for the subsequent components of LWD.

2. A **wavelet-derived spatial saliency map**, computed via a discrete wavelet transform (DWT) applied to the latent representation. This map aggregates the local energy of high-frequency subbands (LH, HL, HH) and is normalized to highlight spatial regions with strong structural detail. These saliency maps are interpretable, data-driven, and require no additional training, providing a principled measure of spatial importance directly from the signal.

3. A **time-dependent masking strategy** that leverages the frequency-based saliency maps to modulate the training loss. At each spatial location and timestep, a binary mask determines whether the denoising loss is applied. Regions with high wavelet energy receive supervision across more timesteps, while simpler areas are updated less frequently. This mechanism introduces spatial adaptivity into the learning process and improves the fidelity of fine-scale detail.

LWD is compatible with a broad range of latent diffusion models, including both classical diffusion and flow-matching formulations. Because it operates solely through the training objective, LWD can be seamlessly integrated into existing pipelines without requiring architectural changes or additional inference cost. We demonstrate its flexibility and effectiveness by applying it to several state-of-the-art latent diffusion models and evaluating performance on ultra-high-resolution image synthesis (2K to 4K). Experimental results show that LWD consistently enhances perceptual quality, improves FID scores, and better preserves fine-grained detail, all without increasing inference complexity.

9.2 Latent Wavelet Diffusion for 4K Image Synthesis

Latent Wavelet Diffusion (LWD) introduces frequency-aware supervision into latent diffusion models by coupling signal-driven saliency analysis with adaptive training. Our key insight is that structural complexity in images is unevenly distributed in space, yet most denoising models refine all positions equally. LWD addresses this by modulating the supervision schedule based on local frequency content, improving detail fidelity without increasing computational cost.

LWD comprises two sequential stages. First, we fine-tune a variational autoencoder (VAE) using a scale-consistency objective to obtain spectrally stable and semantically meaningful latent representations. This step is performed independently and prepares the latent space for subsequent frequency-based modulation.

In the second stage, we fine-tune a latent diffusion model (e.g., Flux) using a modified flow-matching objective that integrates frequency-guided supervision. This stage incorporates three tightly coupled components: (1) extraction of wavelet-based spatial saliency maps from latent codes; (2) construction of a time-dependent mask that adapts the training signal based on local frequency energy; and (3)

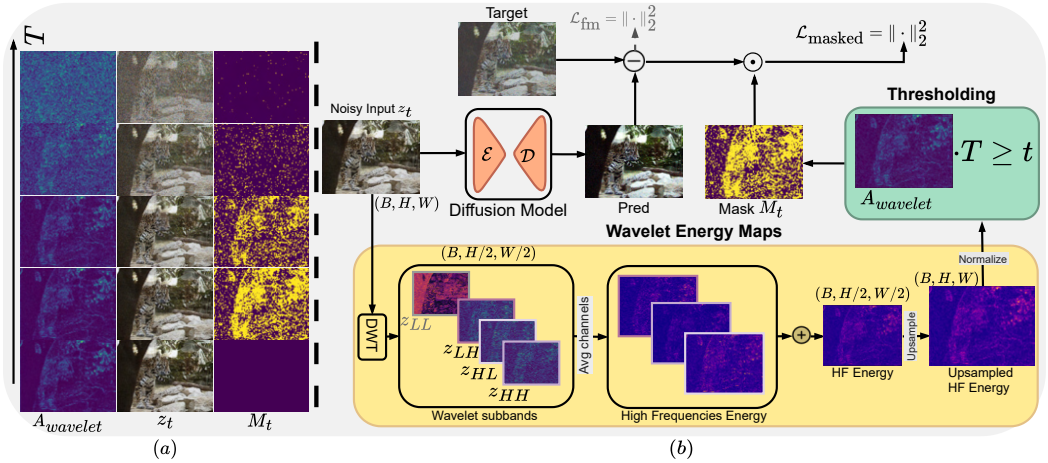


Figure 9.2. (a) Temporal evolution of latent z_t, wavelet energy maps $A_{wavelet}$, and attention map M_t across diffusion timesteps. (b) Our wavelet-masked flow matching objective at a timestep t. The model computes a wavelet attention map $\mathbf{M}_t$ from latent z_t to modulate the prediction error between target velocity field $(\epsilon - z_0)$ and predicted velocity $v_\Theta(z_t, t, y)$. This focuses optimization on high-frequency regions with greater perceptual importance. While operations occur in latent space, decoded visualizations are shown for interpretability.

application of this mask to modulate the training loss dynamically across spatial positions and timesteps. Together, these mechanisms introduce spatial adaptivity into the denoising process, directing learning resources toward detail-rich regions. All components are model-agnostic and can be integrated into standard latent diffusion pipelines without architectural changes.

9.2.1 VAE Fine-Tuning with Scale-Consistency Loss

High-resolution generation places unique demands on the latent space: it must retain both semantic structure and spectral coherence across scales. To ensure this, we fine-tune the variational autoencoder (VAE) using a multi-resolution reconstruction objective that regularizes frequency content while preserving perceptual fidelity.

Formally, let $z = E(x)$ be the latent encoding of image x, x_{down} a downsampled version of x, and z_{down} the downsampled version of the latent z. Our loss combines four components:

$$\mathcal{L}_{\text{VAE}} = \underbrace{\|D(z) - x\|_2^2}_{\text{Reconstruction}} + \alpha \underbrace{\|D(E(z_{\text{down}})) - x_{\text{down}}\|_2^2}_{\text{Scale Consistency}} + \beta \underbrace{D_{\text{KL}}(q(z \mid x) \,\|\, p(z))}_{\text{Latent Regularization}} + \lambda \underbrace{\mathcal{L}_{\text{LPIPS}}(D(z), x)}_{\text{Perceptual Loss}} \tag{9.1}$$

Here, we incorporate a scale-consistency term [269, 270] that encourages the VAE to maintain structural coherence across resolution scales. This regularization naturally suppresses high-frequency artifacts while ensuring robust encoding of essential structural information in z, which is critical for the downstream wavelet decomposition in LWD.

Unlike recent approaches that inject frequency conditioning directly into the encoder [9, 34], our formulation decouples signal regularization and generation: we first sculpt the latent space to exhibit desirable frequency properties, and then

use this structure to guide the training of the diffusion model. This preserves architectural modularity while enabling effective frequency-aware supervision.

9.2.2 Reconciling Spectral Regularization and Saliency

A natural question arises from the apparent tension between our use of a multi-scale VAE loss to suppress high-frequency components in the latent space, and our later use of high-frequency energy to guide the adaptive masking process. While at first glance these two strategies may appear contradictory, they in fact serve complementary and sequential roles in enforcing a more structured and semantically meaningful generative process.

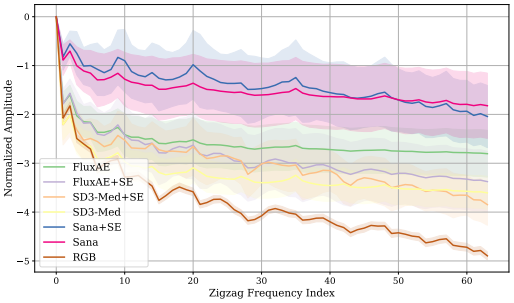


Figure 9.3. Normalized DCT amplitudes over zigzag frequency indices. VAE trained with the multi-scale loss reduces high-frequency energy in latents, aligning their spectrum with that of RGB images.

The multi-scale loss used to fine-tune the VAE does not aim to eliminate all high-frequency content. Rather, it penalizes frequency components that are inconsistent across scales, often indicative of reconstruction noise, aliasing artifacts, or signal components that do not correspond to real-world structures in the input image. In this context, high-frequency suppression is not a blanket filtering of detail, but a targeted regularization of spectral instability, with the goal of aligning the latent spectral distribution more closely with that of clean, natural images.

Crucially, it is precisely this filtered latent space that makes frequency-based saliency meaningful. Once the latent tensor has been regularized to remove spurious high-frequency noise, the remaining high-frequency energy, extracted via wavelet decomposition, is more strongly correlated with visually salient features such as edges, object boundaries, fine textures, and local geometric complexity. In other words, by denoising the latent representation, the VAE enhances the signal-to-noise ratio of the wavelet saliency mechanism. This enables the diffusion process to make effective use of frequency-aware masking, prioritizing spatial refinement in regions that contain real, informative structure, rather than noise.

9.2.3 Wavelet-Derived Frequency Saliency Maps

To guide spatial supervision based on structural complexity, we extract saliency maps from latent representations using localized frequency analysis. Given a latent tensor $z \in \mathbb{R}^{C\times H\times W}$, we apply a single-level Discrete Wavelet Transform (DWT), producing four subbands:

$$\mathrm{DWT}(z) \to \{z_{LL}, z_{LH}, z_{HL}, z_{HH}\}, \tag{9.2}$$

where z_{LL} captures low-frequency approximations and $\{z_{LH}, z_{HL}, z_{HH}\}$ encode directional high-frequency detail.

We compute the localized high-frequency energy as:

$$E(i,j) = \frac{1}{C}\sum_{c}\left[(z_{LH}^{c,i,j})^2 + (z_{HL}^{c,i,j})^2 + (z_{HH}^{c,i,j})^2\right], \tag{9.3}$$

where (i, j) denotes spatial position and $c \in \{1, \ldots, C\}$ indexes feature channels. The resulting energy map $E \in \mathbb{R}^{H/2 \times W/2}$ is bilinearly upsampled and min-max normalized per sample to obtain the final saliency map $A_{\text{wavelet}} \in [0, 1]^{H \times W}$.

The decision to use only high-frequency subbands is deliberate: while z_{LL} encodes coarse spatial content critical for reconstruction, it offers limited information about local complexity. In contrast, the aggregated energy of directional subbands effectively approximates local gradient magnitude, similar to classical edge detectors, highlighting regions with rich structural detail such as textures, contours, and transitions.

This map serves as a proxy for local structural richness. Unlike learned attention mechanisms based on semantic similarity (e.g., DINO [271]), our approach is deterministic and directly derived from signal properties. This approach is theoretically supported by prior work demonstrating that wavelet detail coefficients correspond to structural singularities and perceptual boundaries [272]. While we refer to A_{wavelet} as an "attention map" for interpretability, it is best understood as a frequency-aware saliency measure.

Notably, our VAE fine-tuning objective (Eq. 9.1) helps suppress high-frequency artifacts and stabilize spectral behavior. This preprocessing step ensures that high-frequency activations captured by the DWT correspond to meaningful structure rather than noise, thereby improving the utility of A_{wavelet} for guiding spatial supervision.

9.2.4 Adaptive Flow Matching with Frequency-Guided Masking

We adopt a continuous-time flow-matching formulation [15, 262] for training the latent diffusion model. Given a target latent z_0 and noise sample $\epsilon \sim \mathcal{N}(0, I)$, we define interpolated samples as:

$$z_t = (1 - t)\, z_0 + t\, \epsilon, \quad t \in [0, 1], \tag{9.4}$$

and supervise the predicted velocity field $v_\Theta(z_t, t, y)$, conditioned on text y, via:

$$\mathcal{L}_{\text{fm}} = \|(\epsilon - z_0) - v_\Theta(z_t, t, y)\|_2^2. \tag{9.5}$$

To incorporate frequency-awareness into the training objective, we apply spatially adaptive masking based on the wavelet saliency map A_{wavelet}. Specifically, for each location (i, j), we define a time-dependent binary mask:

$$M_t(i, j) = \begin{cases} 1 & \text{if } T \cdot (A_{\text{wavelet}}(i, j) + \ell) \geq t \\ 0 & \text{otherwise} \end{cases}, \tag{9.6}$$

where T is the total number of training steps and $\ell \in (0, 1)$ sets a lower bound on refinement. This ensures that all regions receive at least ℓT steps of supervision, while high-frequency regions benefit from extended refinement.

While selective spatial supervision has been investigated using transformer attention [273], our wavelet-derived saliency offers a fundamentally different perspective based on signal processing principles rather than learned semantic features, offering computational advantages and a more generalizable solution.

The final masked loss becomes:

$$\mathcal{L}_{\text{masked}} = \|M_t \odot [(\epsilon - z_0) - v_\Theta(z_t, t, y)]\|_2^2, \tag{9.7}$$

where $\odot$ denotes element-wise multiplication. This formulation focuses learning capacity on detail-rich regions, improves fidelity in high-frequency content, and



Figure 9.4. Visual comparison of 2K image generations. LWD demonstrates improved detail preservation in complex areas while avoiding over-sharpening or texture collapse.

does so without increasing inference complexity. Crucially, this mechanism operates purely at the objective level and is compatible with any latent diffusion model using a flow-based or score-based trajectory.

9.2.5 Experiments

Experimental Setup

Datasets. We evaluate LWD on two ultra-resolution datasets. Aesthetic-4K [9] is a curated 4K benchmark with GPT-4o-generated captions and high visual quality. LAION-High-Res is a filtered subset of LAION-5B [274], from which we sample 50K 2K-resolution and 20K 4K-resolution image-caption pairs. These two datasets differ in both visual and linguistic distributions, allowing us to assess both generation fidelity and generalization under caption variance.

Implementation Details. We implement LWD using PyTorch and PyTorch-wavelets [275] for Haar-based DWT. VAE fine-tuning is performed for 60K steps with a batch size of 4 and a learning rate of 1×10^{-5} on NVIDIA A100 GPUs. The frequency-guided loss is applied during the training of the architectures we tested. We set the masking lower bound $\ell = 0.3$, selected via ablation to ensure each spatial location receives at least 30% of refinement steps. Training requirements varied across architectures. LWD+URAE (Flux-based) [54, 276] required approximately 4 hours and 2k iterations at batch size 1 for the 2048 version, while the 4096 version needed 24 hours for 2k iterations, significantly fewer than the original work, which suggests 10k iterations. Diffusion4K (SD3-based) [9, 262] was trained for 48 hours at 10k iterations (half the originally required iterations) with batch size 8 for the 2048 version, maintaining identical parameters for the 4096 version except with batch size 4. SANA [277] models (both 2048 and 4096 versions) were trained for 24

Table 9.1. Quantitative results on different metrics. The prompts are from the HPD [8] dataset. All images are at a resolution of 2048 × 2048.

Model	FID ↓	LPIPS ↓	MAN-IQA ↑	QualiCLIP ↑	HPSv2.1 ↑	PickScore ↑
SDEdit [260]	35.59	0.6456	0.3736	0.4480	<u>30.92</u>	22.86
I-Max [278]	36.28	0.6750	0.3641	0.4139	30.62	**23.02**
PixArt-Sigma-XL [265]	36.58	0.6801	0.2949	0.4438	30.66	<u>22.92</u>
Sana-1.6B [277]	35.75	0.7169	0.3666	0.5796	<u>30.42</u>	22.83
Lumina-Image 2.0 [279]	54.96	0.6445	0.3663	0.4567	23.08	21.15
FLUX-1.dev [54]	37.58	<u>0.6371</u>	**0.4110**	**0.5468**	28.73	22.68
URAE [276]	<u>35.25</u>	0.6717	0.4076	<u>0.5423</u>	**31.15**	22.41
LWD + URAE	**32.88**	**0.6336**	<u>0.4099</u>	0.5356	28.78	22.43

hours, achieving 33k iterations (one-third of the originally suggested iterations) with batch sizes of 2 and 1, respectively. PixArt-Σ [265] required only 24 hours and 1.5k iterations at batch size 2, compared to the suggested 15k iterations.

Evaluation Metrics. To comprehensively evaluate ultra-resolution text-to-image generation, we report multiple complementary metrics across three dimensions: image quality (Fréchet Inception Distance (FID), LPIPS), semantic alignment (CLIP-Score [100], QualiCLIP [101]), and perceptual quality (MAN-IQA [102], HPSv2.1 [8], PickScore [103]). Note that lower values indicate better performance for FID and LPIPS, while higher values are preferred for all other metrics. We assess texture fidelity with the Gray Level Co-occurrence Matrix (GLCM) Score [9] and fine-grained detail preservation using JPEG Compression Ratio to measure high-frequency compressibility as a perceptual quality proxy.

Quantitative and Qualitative Results for 2K and 4K Synthesis

2K Results. Table 9.1 and the top block of Table 9.2 demonstrate consistent improvements from integrating LWD across multiple backbone models. On the HPD prompt dataset [8], LWD reduces FID by up to 7% and LPIPS up to 6% while also achieving comparable MAN-IQA and QualiCLIP, indicating improved semantic alignment and perceptual quality. Moreover, on the Aesthetic dataset [9], these gains are observed across diverse architectures, reinforcing the generality and model-agnostic nature of our approach.

4K Results. On the Aesthetic-4K (the bottom block of Table 9.2) and HPD prompt dataset [8] (can be found in the supplementary materials), LWD outperforms baselines such as URAE and PixArt-Sigma, particularly in metrics like FID, CLIP-Score and Aesthetics. The improvements are especially pronounced in regions with fine structural detail, such as hair, foliage, or architectural elements, highlighting LWD's ability to scale effectively to ultra-high resolutions. These results suggest that frequency-aware supervision provides meaningful guidance even in challenging high-frequency regimes where baseline methods often struggle. LWD also achieves the highest GLCM score when paired with SD3-F16, and substantially improves Sana's performance on both GLCM and compression ratio, indicating stronger fine-detail retention and texture fidelity, without compromising overall quality.

Figures 9.4, 9.5, and 9.13 compare outputs from LWD and baseline models under identical prompts. LWD consistently renders sharper textures in high-frequency regions, such as fabric, skin, and hair, while preserving global structure.

Zoomed-in comparisons highlight improved reconstruction of fine details (e.g., hair strands, eyelashes, small objects) that are often blurred or omitted by baselines. These results suggest that frequency-aware masking not only enhances local precision but does so in a context-sensitive manner, avoiding over-sharpening or

Table 9.2. Quantitative benchmarks of latent diffusion models on Aesthetic-Eval at 2048×2048 and 4096×4096. GLCM Score measures high-frequency texture richness using gray-level co-occurrence matrices, while Compression Ratio assesses visual complexity via JPEG file size heuristics, both introduced in Diffusion-4K [9].

	Model	FID ↓	CLIPScore ↑	Aesthetics ↑	GLCM Score ↑	Compression Ratio ↓
2K	SD3-F16 [262]	43.82	31.50	5.91	0.75	**11.23**
	SD3-Diff4k-F16 [9]	40.18	34.04	5.96	**0.79**	11.73
	LWD + SD3-F16	**38.74**	**34.94**	**6.17**	0.74	11.99
	PixArt-Sigma-XL [265]	39.13	35.02	6.43	0.79	13.66
	LWD + PixArt-Sigma-XL	**36.14**	35.21	6.27	**0.87**	**6.05**
	Sana-1.6B [277]	**32.06**	35.28	6.15	**0.93**	**24.01**
	LWD + Sana-1.6B	34.30	**35.58**	**6.23**	0.78	27.34
4K	SD3-F16 [262]	-	33.12	5.97	0.73	11.97
	SD3-Diff4k-F16 [9]	-	33.41	5.97	0.70	11.90
	LWD + SD3-F16	-	**34.08**	**6.03**	**0.77**	**12.27**
	Sana-1.6B [277]	-	34.40	6.14	0.39	48.36
	LWD + Sana-1.6B	-	**34.59**	**6.21**	**0.60**	**32.62**

artifacts. This indicates that LWD effectively balances fine detail refinement with structural coherence. More full-resolution results can be found in Appendix 9.2.5 and supplementary materials.

To assess the efficacy of our proposed Latent Wavelet Diffusion (LWD), we conduct a detailed evaluation focusing on the challenging 4K resolution (4096×4096). **Quantitative Results** Table 9.3 presents the quantitative comparison on 4K image generation. We evaluate the generated images using several key metrics: MAN-IQA [102] and QualiCLIP [101], which assess perceptual image quality and alignment with textual prompts, respectively. Additionally, we compute the GLCM (Gray-Level Co-occurrence Matrix) score as a measure of texture complexity and detail richness in the generated high-resolution outputs. Finally, we report the Compression Ratio, indicating the compressibility of the generated images, which can be indicative of redundancy or lack of fine details.

Our LWD-enhanced URAE demonstrates competitive performance across all evaluated metrics. Notably, it achieves the highest MAN-IQA score and GLCM score, suggesting superior perceptual quality and richer textural details compared to the baselines. Furthermore, our LWD + URAE achieves a favorable Compression Ratio (28.77), better than URAE and PixArt-Sigma-XL, suggesting a good balance between detail and redundancy. These quantitative results underscore the effectiveness of our LWD approach in enhancing the visual quality and detail of images generated at 4K resolution.

Qualitative Analysis

To complement the quantitative evaluation, Figures 9.6, 9.7, 9.9, and 9.8 provide a qualitative comparison of LWD against selected baselines, presenting side-by-side comparisons of the generated images.

The visual comparisons highlight the benefits of our LWD enhancement. Our method demonstrates the generation of images with finer and more intricate details, particularly noticeable in complex textures and object boundaries. These qualitative observations align with our quantitative findings, reinforcing the effectiveness of the proposed LWD for high-resolution image generation.



Figure 9.5. 4K images generated by LWD with different architectures.



Figure 9.6. 4K generation of URAE vs LWD + URAE.
Upper caption: *"Eiffel Tower was Made up of more than 2 million translucent straws to look like a cloud, with the bell tower at the top of the building, Michel installed huge foam-making machines in the forest to blow huge amounts of unpredictable wet clouds in the building's classic architecture."*. Lower caption: *"Barbarian woman riding a red dragon, holding a broadsword, in gold armour."*

Table 9.3. Evaluation on ultra resolution (4096 × 4096) image generation task with [8] prompts.

Method	MAN-IQA (↑)	QualiCLIP (↑)	GLCM Score ↑	Compression Ratio ↓
PixArt-Sigma-XL [265]	0.2935	0.2308	0.48	45.15
Sana-1.6B [277]	0.3288	**0.4979**	0.71	**25.89**
FLUX-1.dev [54]	0.3673	0.2564	-	-
URAE [276]	0.3850	0.3758	0.41	38.86
LWD + URAE	**0.4011**	0.3855	**0.74**	28.77

Table 9.4. Quantitative reconstruction results of VAEs with and without the Scale-Consistency Loss 9.2.1 on Aesthetic-4K [9] validation set.

Model	rFID ↓	NMSE ↓	PSNR ↑	SSIM ↑	LPIPS ↓
SD3-VAE [262]	1.05	0.01	**26.54**	**0.86**	**0.08**
SD3-VAE-F16 [9]	0.70	0.07	19.82	0.63	0.30
SD3-VAE-F16-SC	0.70	**0.04**	22.58	0.75	0.18
Flux-VAE [54]	0.73	0.01	27.18	0.89	0.07
Flux-VAE-SC	**0.50**	0.01	**28.14**	**0.90**	**0.06**
Sana-AE [277]	0.74	0.04	22.16	0.70	**0.159**
Sana-AE-SC	**0.55**	**0.02**	**23.64**	**0.73**	0.163

Ablation Study: Effect of Scale-Consistency on VAEs

Table 9.4 reports quantitative reconstruction metrics for various VAEs, with and without the proposed Scale-Consistency (SC) loss, evaluated on the Aesthetic-4K validation set. Across different architectures SD3-VAE, Flux-VAE, and Sana-AE, the addition of SC consistently improves performance, particularly in rFID and perceptual quality (LPIPS). For instance, Flux-VAE-SC outperforms its baseline with a significant reduction in rFID (0.50 vs. 0.73) and an increase in PSNR and SSIM, indicating sharper and more faithful reconstructions. Notably, SD3-VAE-F16-SC achieves a substantial LPIPS improvement (0.18 vs. 0.30), suggesting better perceptual fidelity despite using a more aggressive compression factor (F16). These results confirm that scale-consistent regularization enhances latent representations, making them more robust and structurally aligned, critical properties for downstream diffusion tasks.

Frequency-Aware Evaluation

To rigorously assess the frequency characteristics of generated images, we propose a suite of frequency-sensitive metrics that extend beyond standard perceptual scores. These metrics are designed to quantify the degree to which generated images preserve the natural frequency distribution observed in real images, with particular attention to the presence and quality of high-frequency details.

High/Low Frequency Ratio (HLFR): We decompose each image using a 2D discrete wavelet transform (DWT) and compute the total energy in the detail coefficients (high-frequency subbands: LH, HL, HH) and in the approximation coefficients (low-frequency LL subband). The HLFR is defined as the ratio of

high-frequency to low-frequency energy:

$$\text{HLFR} = \frac{E_{LH} + E_{HL} + E_{HH}}{E_{LL}}.$$

This ratio reflects the relative emphasis on fine-scale structures. A value similar to the reference (real images) indicates a natural distribution of frequency content. Large deviations can signal oversmoothing or hallucinated detail.

Ratio Difference from Real (RDR): To quantify deviation from the natural HLFR, we compute the absolute difference between the HLFR of the generated image and the real reference:

$$\text{RDR} = \left|\text{HLFR}_{\text{gen}} - \text{HLFR}_{\text{real}}\right|.$$

Lower values are better, indicating better alignment with the natural frequency distribution.

Wavelet Quality Score (WQS): This metric evaluates the structural similarity between generated and real images in the wavelet domain, where frequency components are explicitly separated by scale and orientation. Given a multi-level discrete wavelet transform (DWT) of both the generated image I_g and the reference image I_r, we compute the SSIM and MSE for each corresponding subband s across all decomposition levels l. The final WQS aggregates the per-subband scores using frequency-aware weights:

$$\text{WQS} = \sum_{l=1}^{L} \sum_{s \in \{LL,LH,HL,HH\}} w_{l,s} \cdot \text{SSIM}(I_r^{l,s}, I_g^{l,s}) - \lambda \cdot \text{MSE}(I_r^{l,s}, I_g^{l,s}),$$

where $w_{l,s}$ are weights that can prioritize perceptually important subbands (e.g., low-frequency LL or high-frequency HH), and λ is a scaling factor that penalizes distortion. The score is normalized to $[0, 1]$, where 1 indicates perfect structural alignment. Higher WQS values reflect better reconstruction fidelity across frequency scales, meaning the model preserves both coarse structure and fine texture.

High-Frequency Energy (HFE): This metric quantifies the total energy of the image's high-frequency components after wavelet decomposition. For a given decomposition level, we define:

$$\text{HFE} = \sum_{l=1}^{L} \left(\|I^{l,LH}\|^2 + \|I^{l,HL}\|^2 + \|I^{l,HH}\|^2 \right).$$

This value provides an absolute measure of fine-scale activity in the image. While real images have characteristic HFE ranges, excessive HFE may indicate artifacts or noise, and too little HFE suggests oversmoothing. Alignment with the real HFE is typically ideal.

High-Frequency Emphasis Index (HFEI): This metric evaluates how much the model over- or under-emphasizes high-frequency content relative to the real distribution. We define it as:

$$\text{HFEI} = \left(\frac{\text{HFE}_{\text{gen}}}{\text{TotalEnergy}_{\text{gen}}} \right) - \left(\frac{\text{HFE}_{\text{real}}}{\text{TotalEnergy}_{\text{real}}} \right),$$

where total energy is computed over all wavelet subbands. HFEI > 0 indicates the generated image places more emphasis on high frequencies than real images



Figure 9.7. 2K generation of PixArt-Sigma-XL vs LWD + PixArt-Sigma-XL.
Upper caption: *"An elderly man with a prominent, bushy beard and deep-set eyes wears a vibrant orange turban, his weathered face marked by lines of age and experience."*.
Lower caption: *"A lone figure on a horse stands in a misty forest, gazing up at a tall, multi-tiered temple surrounded by towering trees and soft, diffused light. Steam rises from the rocks near a stream, creating an atmospheric scene of tranquility and mystery."*

Table 9.5. Comparison of frequency-sensitive metrics across different methods on Aesthetic-4K [9] validation set.

Metric	HLFR	RDFR ↓	WQS ↑	HFE	HFEI ↓	FSIM ↑	MS-SSIM ↑
Real	0.0560	0.0000	1.0000	0.0140	0.0000	1.0000	1.0000
Sana-1.6B [277]	0.0784	0.0558	0.4673	0.0196	0.6108	0.6128	0.1324
LWD + Sana-1.6B	**0.0610**	**0.0537**	**0.4701**	**0.0144**	**0.5227**	**0.6217**	0.1324
SD3-Diff4k-F16 [9]	0.0691	0.0470	0.4624	0.0158	0.8064	0.6155	0.1296
LWD + SD3-F16	**0.0555**	**0.0437**	**0.4735**	**0.0144**	**0.4826**	**0.6245**	**0.1521**
PixArt-Sigma-XL	0.0550	**0.0409**	**0.4763**	0.0119	0.6296	0.1354	0.4255
LWD + PixArt-Sigma-XL	**0.0564**	0.0500	0.4730	**0.0150**	**0.6239**	**0.1478**	**0.5094**

(potentially hallucinated detail), while HFEI < 0 indicates a loss of fine detail. An HFEI near zero is ideal.

Perceptual Metrics: For completeness, we also report FSIM [280] and MS-SSIM [281], which capture visual similarity and structural coherence, respectively. Both metrics are bounded between 0 and 1, with higher values indicating better perceptual quality.

Results: Table 9.5 summarizes the performance of diffusion backbones, along with their respective versions enhanced with LWD. LWD consistently improves frequency alignment: for instance, the LWD+Diff4K variant achieves a high/low frequency ratio of 0.0556, nearly identical to the real reference (0.0560), while also minimizing the ratio difference (0.0438). Wavelet Quality Scores improve in both cases, and the HF energy is more tightly controlled, preventing oversharpening. Importantly, LWD preserves or improves FSIM and MS-SSIM, confirming that frequency fidelity does not come at the expense of perceptual quality. These results demonstrate that LWD enhances the frequency realism of generated images in a measurable and interpretable way, offering a principled approach to frequency-aware image synthesis.

9.3 Conclusion and Future Work

In this chapter, we introduced *Latent Wavelet Diffusion* (LWD), a general and modular framework that integrates frequency-based supervision into latent diffusion models for ultra-high-resolution image synthesis. By computing wavelet energy maps in the latent space and applying spatially and temporally adaptive masking, LWD selectively emphasizes high-detail regions during training. Our experiments demonstrate that without requiring architectural modifications or incurring additional inference cost, LWD consistently improves perceptual fidelity and semantic alignment across multiple state-of-the-art models like Flux and SD3. It preserves high-frequency detail and structural coherence more effectively than prior methods, demonstrating the value of signal-aware supervision in guiding the generative process.

This frequency-guided supervision introduces a form of spatial curriculum learning, where complex regions receive more focused updates. This principle opens avenues for future research into dynamic training strategies, such as frequency-aware learning rates or hybrid spatial-frequency schedules. These mechanisms may be especially valuable in domains where structural detail is critical but semantic guidance is weak, such as scientific visualization, material design, or multimodal generation.

While LWD improves generation quality, it inherits limitations from its underlying VAE, where compression can lead to the loss of fine-grained semantic detail. Future work could address this by exploring higher-fidelity latent spaces or hybrid approaches



Figure 9.8. 4K generation of Sana vs LWD + Sana.
Upper caption: *"A litter of golden retriever puppies playing in the snow. Their heads pop out of the snow, covered in."*. Lower caption: *"A curvy timber house near a sea, designed by Zaha Hadid, represent the image of a cold, modern architecture, at night, white lighting, highly detailed."*



Figure 9.9. 2K generation of SD3-Diff4k-F16 vs LWD + SD3-F16.
Upper caption: *"A serene landscape features a winding river, flanked by trees with autumn foliage, leading to a rustic wooden cabin with a corrugated roof, set against a softly blurred background."*. Lower caption: *"A grand interior featuring intricate stained glass windows, an elaborate rose window, ornate frescoes depicting biblical scenes, and elegant chandeliers illuminating the richly decorated walls and arches."*

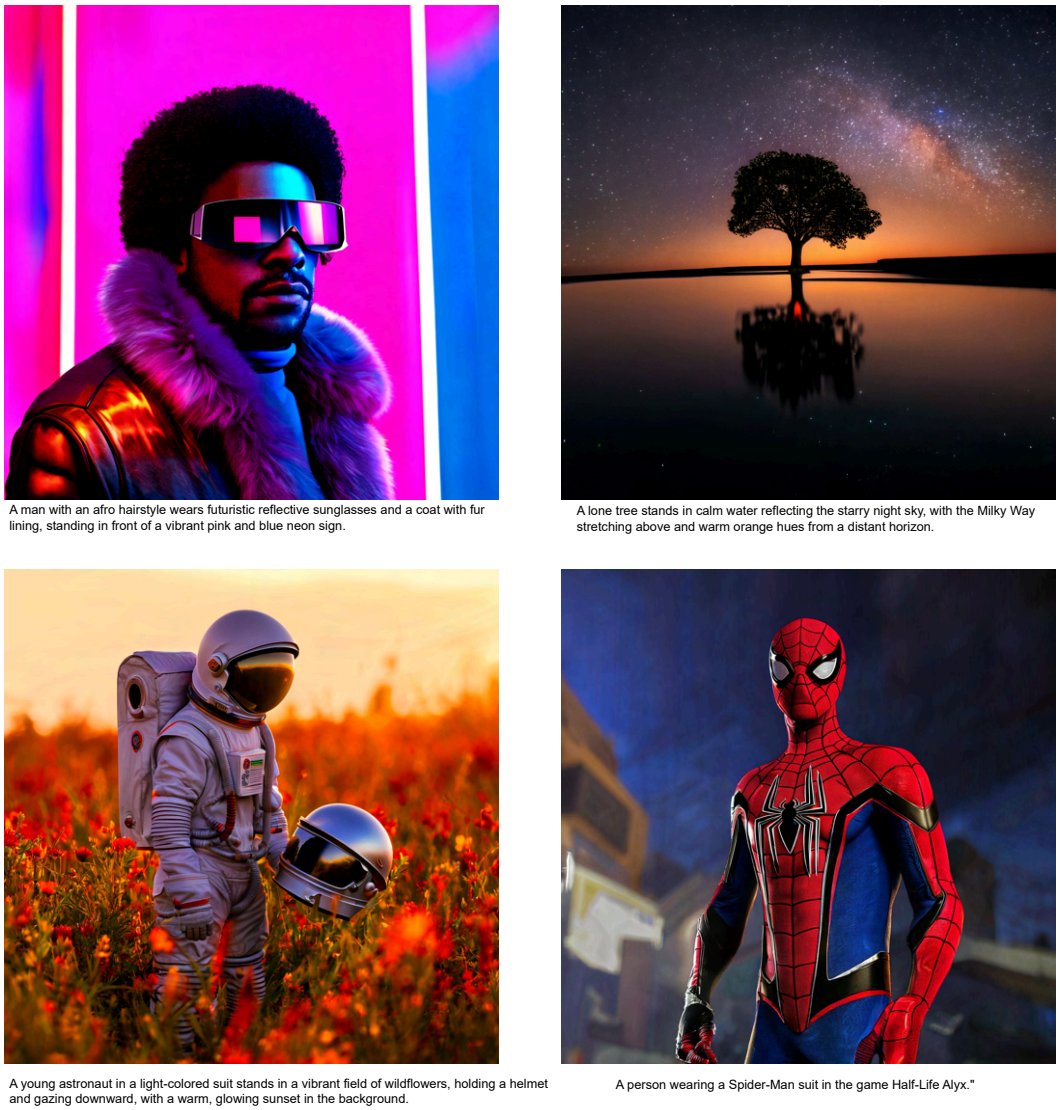


Figure 9.10. Images generated at 4K resolution with LWD+SANA.

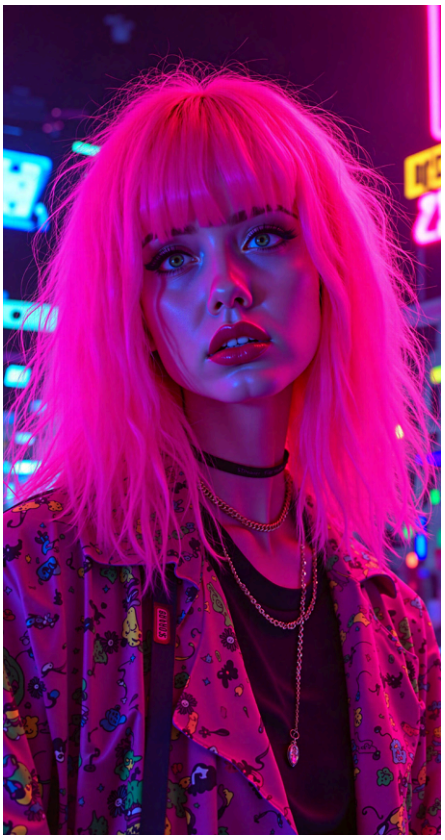


A surreal landscape depicting an ethereal fusion of natural beauty and fantastical architecture, reminiscent of Salvador Dali's dreamlike paintings. From above the clouds, one gazes upon a colossal tower emerging from the earth, its intricate gears visible as it merges seamlessly with a tranquil mountain lake. The scene is bathed in an otherworldly glow, casting lavender and gold hues across the sky, while delicate cherry blossoms flutter gently in the foreground, adding a sense of serenity to this breathtaking vision where time and nature intertwine.



Steampunk airship floating above a misty Victorian cityscape, intricate brass and copper mechanical details, golden hour lighting, billowing clouds, detailed architectural elements, rich warm color palette, cinematic composition.

Figure 9.11. Images generated at 4K resolution with LWD+URAE.



Girl with pink hair, vaporwave style, retro aesthetic, cyberpunk, vibrant, neon colors, vintage 80s and 90s style, highly detailed.



A sleek black luxury sedan parked on a rain-soaked city street at night, reflecting neon lights from nearby buildings. The wet pavement glistens, and the car's smooth curves are highlighted by the ambient glow of the urban environment.



A narrow and picturesque alley in the historic center of Naples, with laundry hanging out to dry between flower-filled balconies and the inviting aroma of freshly baked pizza in the air.

Figure 9.12. Images generated at 4K resolution with LWD+URAE.

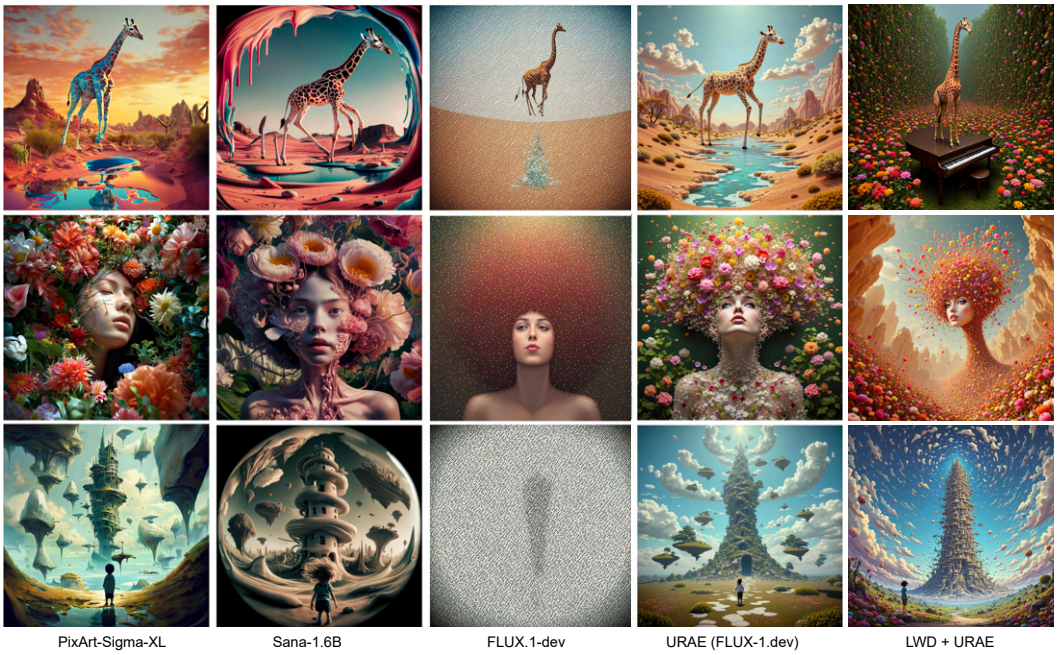


Figure 9.13. Visual comparison of 4K image generations from LWD and competing baselines.

that combine latent and pixel-space supervision. Extending LWD to video generation, depth-aware synthesis, or other conditional modalities also represents a promising direction.

Ultimately, by unifying principles from classical signal processing and modern diffusion modeling, LWD offers a scalable, interpretable, and efficient approach to high-fidelity image generation, pushing the boundaries of what is possible in a computationally constrained environment.

Broader Impact

LWD promotes efficient and interpretable generation by aligning supervision with signal-level detail. This may benefit applications requiring controllable high-resolution synthesis, while raising familiar concerns around synthetic media misuse. Incorporating safeguards and provenance tools remains an important direction for all powerful generative models.

Part IV

Synthesis and Future Directions

Chapter 10

Discussion and Future Endeavours

This thesis has journeyed through the exciting landscape of generative image synthesis, focusing on how wavelet and quaternion wavelet representations can bring about fundamental improvements. We have seen that by equipping generative models with a deeper understanding of multi-scale image structure, we can effectively tackle some of the most persistent challenges in creating high-fidelity, controllable, and broadly applicable image generation. This final chapter brings together the core contributions and insights from our research, reflecting on the path taken from initial explorations to novel wavelet-centric methods, and then looks ahead to the promising horizons for future work in this dynamic field.

10.1 Discussion

At the heart of this thesis, as we set out in Chapter 1, was the challenge posed by many generative models operating primarily in the pixel domain. This approach often falls short of optimally capturing the rich, multi-scale tapestry of visual information, making it difficult to simultaneously represent global context and intricate details. Our research path began with early explorations in Generative Adversarial Networks, leading to the **StawGAN** framework presented in Chapter 5, an initial step towards structural awareness in the demanding task of infrared-to-RGB translation. As the field of generative modeling evolved, so did our focus, shifting towards the stability and versatility of diffusion models (Section 5.3.3). This pivot enabled us to develop specialized solutions, such as **Ship-in-Sight** for advanced ship image super-resolution and **GWIT** for streamlined EEG-to-image synthesis, both detailed in Chapter 6.

However, the most significant leap forward, detailed in Part III, came from deeply integrating wavelet theory into the core of generative processes. In Chapter 7, we introduced versatile frameworks like **GROUSE** and **QUAVE**. These established model-agnostic ways to select and learn from vital quaternion wavelet features, greatly boosting performance by arming neural networks with both low-frequency context and critical high-frequency details. This even allowed us to extend the use of quaternion networks to single-channel data.

Building on this foundation, our dedicated work on wavelet-based super-resolution (Chapter 8), including **ResQu** using quaternion wavelet conditioning, **MWT-Diff** tailored for satellite imagery, and the efficient **WaDiGANSR** hybrid, delivered substantial gains in fidelity and speed. The culmination of this research was **Latent**

Wavelet Diffusion (LWD) in Chapter 9. LWD introduced a groundbreaking method for frequency-sensitive spatial supervision that unlocked state-of-the-art 4K image synthesis without altering existing model architectures or adding inference cost, marking a significant stride towards practical ultra-high-resolution generation.

Looking back at these varied contributions, several key insights about the powerful role of wavelets in generative modeling become clear.

Firstly, making multi-scale processing an explicit part of the model, as wavelets allow, is often crucial for capturing the complex interplay of spatial and frequency information in images. This principle was a consistent thread through all our models.

Secondly, wavelet representations can bring about notable efficiency gains. This is not just about computation as seen in Chapter 8.4 but also about how information is represented, leading to more compact yet powerful models like LWD (Chapter 9).

Thirdly, wavelets offer natural and effective pathways for enhanced conditioning and controllability. The ability to guide generation across different frequency bands, so clearly shown by the frequency-guided masking in LWD, gives us more intuitive control.

Lastly, our research consistently indicated that wavelet-based strategies lead to improved generalization across different datasets and tasks, as demonstrated by QUAVE in Chapter 7.3. The use of quaternion wavelets further enriched this by offering ways to capture complex inter-channel relationships and phase information. Our collective journey really highlights the immense value of bringing together classical signal processing wisdom with modern deep learning tools.

10.2 Future Works

While this thesis has significantly advanced the application of wavelets in generative modeling, the explored avenues also illuminate several limitations of current methods and exciting future research directions that promise to build upon these foundations.

The computational overhead associated with wavelet transforms, although often offset by downstream model efficiencies, warrants continued attention for achieving truly real-time deployment in resource-constrained settings. Future work should thus explore further optimized wavelet operations, dedicated hardware accelerations, and advanced model compression techniques specifically designed for wavelet-based neural networks. Furthermore, the selection of wavelet bases in our models, from **QUAVE** to **LWD**, has been largely guided by established practices (e.g., Daubechies, Haar). A promising direction is the development of methods for learning adaptive, task-specific wavelet-like decompositions or dynamically selecting from a broader range of wavelet families, potentially yielding representations even more attuned to specific data characteristics and generative objectives.

The exploration of quaternion algebra, while shown to be beneficial in **QUAVE**, also brings added complexity that deserves further investigation. Future research could focus on developing more sophisticated and potentially more efficient quaternion neural architectures, such as quaternion transformers or native quaternion diffusion models, alongside a deeper theoretical understanding of their learning dynamics and precise advantages in various generative contexts. In the realm of ultra-high-resolution synthesis, refining multi-scale fusion techniques to prevent subtle artifacts at frequency band interfaces remains an important goal. The integration of wavelet principles with emerging paradigms like flow matching, potentially extending the adaptive flow concepts from **LWD**, offers a rich avenue for creating even smoother and more robust generative processes across scales.

While this work has laid some theoretical groundwork, a more profound mathematical analysis of the convergence properties, generalization bounds, and information-theoretic characteristics of wavelet-based generative models would significantly strengthen the field. In parallel, the powerful synergy between transformer architectures and wavelet representations is another area ripe for exploration. Developing wavelet-based attention mechanisms or utilizing transformers to model long-range dependencies within the wavelet domain, as explored with WaveViT in our **MWT-Diff** framework, could lead to substantial breakthroughs, particularly in complex conditional synthesis tasks such as text-to-image generation.

Expanding the application of these wavelet-driven generative techniques to a wider array of modalities is a compelling direction that leverages the cross-modal insights gained from work like **GWIT**. Specific extensions include video synthesis (a noted extension for **LWD**), 3D volumetric data generation (a potential next step for **QUAVE**), and audio signal modeling. Moreover, while wavelets offer a degree of inherent interpretability through their multi-scale decomposition, enhancing the explainability of the deep networks that operate on these features is crucial, especially for safety-critical scientific and medical applications.

Finally, the representational efficiency and structural priors offered by wavelets make them highly suitable for scaling up to next-generation foundation models. Future work could investigate wavelet-native architectures for generative models comprising billions of parameters, which may also necessitate new solutions for existing bottlenecks in latent models, such as the VAE challenges we addressed with the scale-consistency loss in our **LWD** framework.

In conclusion, this thesis has robustly demonstrated that incorporating multi-scale perspectives through wavelet and quaternion wavelet transforms is not merely an incremental improvement but a foundational step towards more powerful, efficient, and controllable generative image synthesis. The principles and methodologies developed herein, evolving from enhancing GANs with frameworks like **QUAVE** to revolutionizing diffusion models with state-of-the-art architectures like **LWD** and **MWT-Diff**, offer a strong platform for a multitude of future innovations. These explorations promise to further bridge the gap between mathematical signal processing and data-driven deep learning, ultimately advancing our collective ability to generate visual content with unprecedented realism, utility, and scale.

Bibliography

[1] C.-Y. Wang, A. Bochkovskiy, and H.-Y. M. Liao, "YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors," in *Proceedings of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, June 2023, pp. 7464–7475.

[2] P. Tirupattur, Y. S. Rawat, C. Spampinato, and M. Shah, "ThoughtViz: Visualizing human thoughts using generative adversarial network," in *ACM Int. Conf. on Multimedia*, 2018, pp. 950–958.

[3] P. Wei, Z. Xie, H. Lu, Z. Zhan, Q. Ye, W. Zuo, and L. Lin, "Component divide-and-conquer for real-world image super-resolution," in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part VIII 16*. Springer, 2020, pp. 101–117.

[4] J. Cai, H. Zeng, H. Yong, Z. Cao, and L. Zhang, "Toward real-world single image super-resolution: A new benchmark and a new model," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 3086–3095.

[5] G. Christie, N. Fendley, J. Wilson, and R. Mukherjee, "Functional map of the world," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 6172–6180.

[6] Y. Cong, S. Khanna, C. Meng, P. Liu, E. Rozi, Y. He, M. Burke, D. Lobell, and S. Ermon, "Satmae: Pre-training transformers for temporal and multi-spectral satellite imagery," *Advances in Neural Information Processing Systems*, vol. 35, pp. 197–211, 2022.

[7] L. Sigillo, R. F. Gramaccioni, A. Nicolosi, and D. Comminiello, "Ship in sight: Diffusion models for ship-image super resolution," in *2024 International Joint Conference on Neural Networks (IJCNN)*, 2024, pp. 1–8.

[8] X. Wu, Y. Hao, K. Sun, Y. Chen, F. Zhu, R. Zhao, and H. Li, "Human preference score v2: A solid benchmark for evaluating human preferences of text-to-image synthesis," 2023. [Online]. Available: https://arxiv.org/abs/2306.09341

[9] J. Zhang, Q. Huang, J. Liu, X. Guo, and D. Huang, "Diffusion-4k: Ultra-high-resolution image synthesis with latent diffusion models," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025.

[10] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Advances in*

Neural Information Processing Systems, Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, and K. Weinberger, Eds., vol. 27. Curran Associates, Inc., 2014.

[11] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," *Advances in neural information processing systems*, vol. 33, pp. 6840–6851, 2020.

[12] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 10 684–10 695.

[13] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," 2022. [Online]. Available: https://arxiv.org/abs/1312.6114

[14] Y. Song, J. Sohl-Dickstein, D. P. Kingma, A. Kumar, S. Ermon, and B. Poole, "Score-based generative modeling through stochastic differential equations," in *International Conference on Learning Representations*, 2021. [Online]. Available: https://openreview.net/forum?id=PxTIG12RRHS

[15] Y. Lipman, R. T. Q. Chen, H. Ben-Hamu, M. Nickel, and M. Le, "Flow matching for generative modeling," in *The Eleventh International Conference on Learning Representations*, 2023. [Online]. Available: https://openreview.net/forum?id=PqvMRDCJT9t

[16] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 1125–1134.

[17] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2223–2232.

[18] M. Arjovsky, S. Chintala, and L. Bottou, "Wasserstein generative adversarial networks," in *International conference on machine learning*. PMLR, 2017, pp. 214–223.

[19] P. Dhariwal and A. Nichol, "Diffusion models beat GANs on image synthesis," in *Advances in Neural Information Processing Systems*, M. Ranzato, A. Beygelzimer, Y. Dauphin, P. Liang, and J. W. Vaughan, Eds., vol. 34. Curran Associates, Inc., 2021, pp. 8780–8794.

[20] C. Saharia, J. Ho, W. Chan, T. Salimans, D. J. Fleet, and M. Norouzi, "Image super-resolution via iterative refinement," *IEEE transactions on pattern analysis and machine intelligence*, vol. 45, no. 4, pp. 4713–4726, 2022.

[21] J. Wang, Z. Yue, S. Zhou, K. C. Chan, and C. C. Loy, "Exploiting diffusion prior for real-world image super-resolution," *International Journal of Computer Vision*, pp. 1–21, 2024.

[22] J. Ho, C. Saharia, W. Chan, D. J. Fleet, M. Norouzi, and T. Salimans, "Cascaded diffusion models for high fidelity image generation," *Journal of Machine Learning Research*, vol. 23, no. 47, pp. 1–33, 2022.

[23] W. Peebles and S. Xie, "Scalable diffusion models with transformers," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2023, pp. 4195–4205.

[24] S. Mallat, "A theory for multiresolution signal decomposition: the wavelet representation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 11, no. 7, pp. 674–693, 1989.

[25] H. Phung, Q. Dao, and A. Tran, "Wavelet diffusion models are fast and scalable image generators," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2023, pp. 10 199–10 208.

[26] Y. Huang, J. Huang, J. Liu, M. Yan, Y. Dong, J. Lyu, C. Chen, and S. Chen, "Wavedm: Wavelet-based diffusion models for image restoration," *IEEE Transactions on Multimedia*, 2024.

[27] L. Zhang, A. Rao, and M. Agrawala, "Adding conditional control to text-to-image diffusion models," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2023, pp. 3836–3847.

[28] L. Sigillo, E. Grassucci, and D. Comminiello, "StawGAN: Structural-aware generative adversarial networks for infrared image translation," in *IEEE Int. Symposium on Circuits and Systems (ISCAS)*, 2023.

[29] E. Lopez, L. Sigillo, F. Colonnese, M. Panella, and D. Comminiello, "Guess what i think: Streamlined eeg-to-image generation with latent diffusion models," in *ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2025, pp. 1–5.

[30] E. Grassucci, L. Sigillo, A. Uncini, and D. Comminiello, "GROUSE: A task and model agnostic wavelet-driven framework for medical imaging," *IEEE Signal Processing Letters*, pp. 1–5, 2023.

[31] L. Sigillo, E. Grassucci, A. Uncini, and D. Comminiello, "Generalizing medical image representations via quaternion wavelet networks," *Neurocomputing*, vol. 638, p. 130195, 2025. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0925231225008677

[32] L. Sigillo, C. Bianchi, A. Uncini, and D. Comminiello, "Quaternion wavelet-conditioned diffusion models for image super-resolution," in *2025 International Joint Conference on Neural Networks (IJCNN)*, 2025, pp. 1–8.

[33] L. Sigillo, R. Giamba, and D. Comminiello, "Metadata, wavelet, and time aware diffusion models for satellite image super resolution," *Pattern Recognition Letters*, vol. 202, pp. 102–107, 2026. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0167865526000589

[34] L. Aloisi, L. Sigillo, A. Uncini, and D. Comminiello, *A Wavelet Diffusion GAN for Image Super-Resolution*. Springer Nature Singapore, 2026, https://arxiv.org/abs/2410.17966. [Online]. Available: https://doi.org/10.1007/978-981-95-4072-3_36

[35] L. Sigillo, S. He, and D. Comminiello, "Latent wavelet diffusion for ultra high-resolution image synthesis," in *The Fourteenth International Conference on Learning Representations*, 2026. [Online]. Available: https://openreview.net/forum?id=5og80LMVxG

[36] D. P. Kingma, M. Welling *et al.*, "An introduction to variational autoencoders," *Foundations and Trends® in Machine Learning*, vol. 12, no. 4, pp. 307–392, 2019.

[37] Y. Zhi-Han, "Training latent variable models with auto-encoding variational bayes: A tutorial," 2022. [Online]. Available: https://arxiv.org/abs/2208.07818

[38] I. Goodfellow, "Nips 2016 tutorial: Generative adversarial networks," 2017. [Online]. Available: https://arxiv.org/abs/1701.00160

[39] P. Salehi, A. Chalechale, and M. Taghizadeh, "Generative adversarial networks (gans): An overview of theoretical model, evaluation metrics, and recent developments," 2020. [Online]. Available: https://arxiv.org/abs/2005.13178

[40] K. Xu, C. Li, J. Zhu, and B. Zhang, "Understanding and stabilizing gans' training dynamics using control theory," in *International conference on machine learning*. PMLR, 2020, pp. 10 566–10 575.

[41] T. Salimans, I. J. Goodfellow, W. Zaremba, V. Cheung, A. Radford, and X. Chen, "Improved techniques for training GANs," in *Neural Information Processing Systems (NIPS)*, 2016, pp. 2234–2242.

[42] I. Gulrajani, F. Ahmed, M. Arjovsky, V. Dumoulin, and A. C. Courville, "Improved training of wasserstein gans," *Advances in neural information processing systems*, vol. 30, 2017.

[43] T. Karras, T. Aila, S. Laine, and J. Lehtinen, "Progressive growing of gans for improved quality, stability, and variation," 2018. [Online]. Available: https://arxiv.org/abs/1710.10196

[44] R. E. Turner, C.-D. Diaconu, S. Markou, A. Shysheya, A. Y. K. Foong, and B. Mlodozeniec, "Denoising diffusion probabilistic models in six simple steps," 2024. [Online]. Available: https://arxiv.org/abs/2402.04384

[45] M. Mirza and S. Osindero, "Conditional generative adversarial nets," *arXiv preprint arXiv:1411.1784*, 2014.

[46] X. Chen, Y. Duan, R. Houthooft, J. Schulman, I. Sutskever, and P. Abbeel, "Infogan: Interpretable representation learning by information maximizing generative adversarial nets," *Advances in neural information processing systems*, vol. 29, 2016.

[47] H. Zhang, I. J. Goodfellow, D. N. Metaxas, and A. Odena, "Self-attention generative adversarial networks," in *Int. Conf. on Machine Learning (ICML)*, ser. Proceedings of Machine Learning Research (PMLR), vol. 97, 2019, pp. 7354–7363.

[48] T. Karras, S. Laine, and T. Aila, "A style-based generator architecture for generative adversarial networks," in *IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*. Computer Vision Foundation / IEEE, 2019, pp. 4401–4410.

[49] J. Sohl-Dickstein, E. Weiss, N. Maheswaranathan, and S. Ganguli, "Deep unsupervised learning using nonequilibrium thermodynamics," in *Proceedings of the 32nd Int. Conf. on Machine Learning*, ser. Proceedings of Machine Learning Research, F. Bach and D. Blei, Eds., vol. 37. Lille, France: PMLR, 07–09 Jul 2015, pp. 2256–2265.

[50] D. Kingma, T. Salimans, B. Poole, and J. Ho, "Variational diffusion models," *Advances in neural information processing systems*, vol. 34, pp. 21 696–21 707, 2021.

[51] J. Song, C. Meng, and S. Ermon, "Denoising diffusion implicit models," in *Int. Conf. on Learning Representations*, 2021.

[52] Y. Shi, V. De Bortoli, A. Campbell, and A. Doucet, "Diffusion schr\" odinger bridge matching," *arXiv preprint arXiv:2303.16852*, 2023.

[53] C. Saharia, W. Chan, S. Saxena, L. Li, J. Whang, E. L. Denton, S. K. S. Ghasemipour, R. G. Lopes, B. Karagol-Ayan, T. Salimans, J. Ho, D. J. Fleet, and M. N. 0002, "Photorealistic text-to-image diffusion models with deep language understanding," in *Advances in Neural Information Processing Systems 35*, S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, Eds., 2022.

[54] B. F. Labs, "Flux," https://github.com/black-forest-labs/flux, 2024.

[55] A. Creswell, T. White, V. Dumoulin, K. Arulkumaran, B. Sengupta, and A. A. Bharath, "Generative adversarial networks: An overview," *IEEE signal processing magazine*, vol. 35, no. 1, pp. 53–65, 2018.

[56] N. Huang, A. Gokaslan, V. Kuleshov, and J. Tompkin, "The gan is dead; long live the gan! a modern gan baseline," *Advances in Neural Information Processing Systems*, vol. 37, pp. 44 177–44 215, 2024.

[57] E. Reinhard, P. Shirley, and T. Troscianko, "Natural image statistics for computer graphics," *Univ. Utah Tech Report UUCS-01-002 (Mar. 2001)*, 2001.

[58] D. Saxena and J. Cao, "Generative adversarial networks (gans) challenges, solutions, and future directions," *ACM Computing Surveys (CSUR)*, vol. 54, no. 3, pp. 1–42, 2021.

[59] S. Chakraborty and P. L. Bartlett, "On the statistical properties of generative adversarial models for low intrinsic data dimension," *arXiv preprint arXiv:2401.15801*, 2024.

[60] C. Doersch, "Tutorial on variational autoencoders," *arXiv preprint arXiv:1606.05908*, 2016.

[61] S. Ma, X. Zhang, C. Jia, Z. Zhao, S. Wang, and S. Wang, "Image and video compression with neural networks: A review," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 30, no. 6, pp. 1683–1698, 2019.

[62] M. Yang, Z. Wang, Z. Chi, and W. Feng, "Wavegan: Frequency-aware gan for high-fidelity few-shot image generation," in *European conference on computer vision*. Springer, 2022, pp. 1–17.

[63] S. Ren, Q. Yu, J. He, A. Yuille, and L.-C. Chen, "Grouping first, attending smartly: Training-free acceleration for diffusion transformers," *arXiv preprint arXiv:2505.14687*, 2025.

[64] O. Bar-Tal, L. Yariv, Y. Lipman, and T. Dekel, "MultiDiffusion: Fusing diffusion paths for controlled image generation," in *Proceedings of the 40th International Conference on Machine Learning*, ser. Proceedings of Machine

Learning Research, A. Krause, E. Brunskill, K. Cho, B. Engelhardt, S. Sabato, and J. Scarlett, Eds., vol. 202. PMLR, 23–29 Jul 2023, pp. 1737–1752. [Online]. Available: https://proceedings.mlr.press/v202/bar-tal23a.html

[65] Y. He, S. Yang, H. Chen, X. Cun, M. Xia, Y. Zhang, X. Wang, R. He, Q. Chen, and Y. Shan, "Scalecrafter: Tuning-free higher-resolution visual generation with diffusion models," in *The Twelfth International Conference on Learning Representations*, 2023.

[66] C. Ledig, L. Theis, F. Huszár, J. Caballero, A. Cunningham, A. Acosta, A. Aitken, A. Tejani, J. Totz, Z. Wang *et al.*, "Photo-realistic single image super-resolution using a generative adversarial network," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 4681–4690.

[67] Y. Pang, J. Lin, T. Qin, and Z. Chen, "Image-to-image translation: Methods and applications," *IEEE Transactions on Multimedia*, vol. 24, pp. 3859–3881, 2021.

[68] R. Zhang, P. Isola, and A. A. Efros, "Colorful image colorization," in *Proc. Eur. Conf. on Computer Vision (ECCV)*, 2016.

[69] R. Mechrez, I. Talmi, F. Shama, and L. Zelnik-Manor, "Maintaining natural image statistics with the contextual loss," in *Computer Vision–ACCV 2018: 14th Asian Conference on Computer Vision, Perth, Australia, December 2–6, 2018, Revised Selected Papers, Part III 14*. Springer, 2019, pp. 427–443.

[70] M. J. Chong and D. Forsyth, "Effectively unbiased fid and inception score and where to find them," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 6070–6079.

[71] A. Lucas, S. Lopez-Tapia, R. Molina, and A. K. Katsaggelos, "Generative adversarial networks and perceptual losses for video super-resolution," *IEEE Transactions on Image Processing*, vol. 28, no. 7, pp. 3312–3327, 2019.

[72] J. Yu, Z. Lin, J. Yang, X. Shen, X. Lu, and T. S. Huang, "Generative image inpainting with contextual attention," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 5505–5514.

[73] H. Huang, R. He, Z. Sun, T. Tan *et al.*, "Introvae: Introspective variational autoencoders for photographic image synthesis," *Advances in neural information processing systems*, vol. 31, 2018.

[74] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, "The unreasonable effectiveness of deep features as a perceptual metric," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 586–595.

[75] R. Suvorov, E. Logacheva, A. Mashikhin, A. Remizova, A. Ashukha, A. Silvestrov, N. Kong, H. Goka, K. Park, and V. Lempitsky, "Resolution-robust large mask inpainting with fourier convolutions," in *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, 2022, pp. 2149–2159.

[76] A. Q. Nichol and P. Dhariwal, "Improved denoising diffusion probabilistic models," in *International conference on machine learning*. PMLR, 2021, pp. 8162–8171.

[77] J. Chen, H. Cai, J. Chen, E. Xie, S. Yang, H. Tang, M. Li, Y. Lu, and S. Han, "Deep compression autoencoder for efficient high-resolution diffusion models," *arXiv preprint arXiv:2410.10733*, 2024.

[78] T. Dai, J. Cai, Y. Zhang, S.-T. Xia, and L. Zhang, "Second-order attention network for single image super-resolution," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 11 065–11 074.

[79] C. Dong, C. C. Loy, K. He, and X. Tang, "Learning a deep convolutional network for image super-resolution," in *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part IV 13*. Springer, 2014, pp. 184–199.

[80] ——, "Image super-resolution using deep convolutional networks," *IEEE transactions on pattern analysis and machine intelligence*, vol. 38, no. 2, pp. 295–307, 2015.

[81] C. Dong, C. C. Loy, and X. Tang, "Accelerating the super-resolution convolutional neural network," in *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part II 14*. Springer, 2016, pp. 391–407.

[82] B. Lim, S. Son, H. Kim, S. Nah, and K. Mu Lee, "Enhanced deep residual networks for single image super-resolution," in *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, 2017, pp. 136–144.

[83] Y. Zhang, K. Li, K. Li, L. Wang, B. Zhong, and Y. R. Fu, "Image super-resolution using very deep residual channel attention networks," in *European Conf. on Computer Vision*, 2018.

[84] Y. Zhang, Y. Tian, Y. Kong, B. Zhong, and Y. Fu, "Residual dense network for image super-resolution," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 2472–2481.

[85] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16x16 words: Transformers for image recognition at scale," in *Int. Conf. on Learning Representations*, 2021.

[86] J. Liang, J. Cao, G. Sun, K. Zhang, L. Van Gool, and R. Timofte, "SwinIR: Image restoration using swin transformer," in *2021 IEEE/CVF Int. Conf. on Computer Vision Workshops (ICCVW)*, 2021.

[87] X. Wang, K. Yu, S. Wu, J. Gu, Y. Liu, C. Dong, Y. Qiao, and C. Change Loy, "ESRGAN: Enhanced super-resolution generative adversarial networks," in *Proceedings of the European Conf. on computer vision (ECCV) workshops*, 2018, pp. 0–0.

[88] X. Wang, L. Xie, C. Dong, and Y. Shan, "Real-esrgan: Training real-world blind super-resolution with pure synthetic data," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 1905–1914.

[89] S. Menon, A. Damian, S. Hu, N. Ravi, and C. Rudin, "Pulse: Self-supervised photo upsampling via latent space exploration of generative models," *IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, pp. 2434–2442, 2020.

[90] C. Saharia, J. Ho, W. Chan, T. Salimans, D. J. Fleet, and M. Norouzi, "Image super-resolution via iterative refinement," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 45, no. 4, 2023.

[91] A. Lugmayr, M. Danelljan, A. Romero, F. Yu, R. Timofte, and L. Van Gool, "Repaint: Inpainting using denoising diffusion probabilistic models," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 11 461–11 471.

[92] F. A. Fardo, V. H. Conforto, F. C. de Oliveira, and P. S. Rodrigues, "A formal evaluation of psnr as quality measurement parameter for image segmentation algorithms," *arXiv preprint arXiv:1605.07116*, 2016.

[93] O. Keleş, M. A. Yılmaz, A. M. Tekalp, C. Korkmaz, and Z. Doğan, "On the computation of psnr for a set of images or video," in *2021 Picture Coding Symposium (PCS)*. IEEE, 2021, pp. 1–5.

[94] D. Tao, S. Di, X. Liang, Z. Chen, and F. Cappello, "Fixed-psnr lossy compression for scientific data," in *2018 IEEE International Conference on Cluster Computing (CLUSTER)*. IEEE, 2018, pp. 314–318.

[95] A. K. Venkataramanan, C. Wu, A. C. Bovik, I. Katsavounidis, and Z. Shahid, "A hitchhiker's guide to structural similarity," *IEEE Access*, vol. 9, pp. 28 872–28 896, 2021.

[96] B. Ghojogh, F. Karray, and M. Crowley, "Theoretical insights into the use of structural similarity index in generative models and inferential autoencoders," in *International Conference on Image Analysis and Recognition*. Springer, 2020, pp. 112–117.

[97] S. Barratt and R. Sharma, "A note on the inception score," *arXiv preprint arXiv:1801.01973*, 2018.

[98] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter, "Gans trained by a two time-scale update rule converge to a local nash equilibrium," *Advances in neural Inf. Proc. systems, (NeurIPS)*, vol. 30, 2017.

[99] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 2818–2826.

[100] J. Hessel, A. Holtzman, M. Forbes, R. Le Bras, and Y. Choi, "Clipscore: A reference-free evaluation metric for image captioning," in *EMNLP (1)*, 2021.

[101] L. Agnolucci, L. Galteri, and M. Bertini, "Quality-aware image-text alignment for opinion-unaware image quality assessment," *arXiv preprint arXiv:2403.11176*, 2024.

[102] S. Yang, T. Wu, S. Shi, S. Lao, Y. Gong, M. Cao, J. Wang, and Y. Yang, "Maniqa: Multi-dimension attention network for no-reference image quality assessment," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 1191–1200.

[103] Y. Kirstain, A. Polyak, U. Singer, S. Matiana, J. Penna, and O. Levy, "Pick-a-pic: An open dataset of user preferences for text-to-image generation," *Advances in Neural Information Processing Systems*, vol. 36, pp. 36 652–36 663, 2023.

[104] J. Gilles, "Empirical wavelet transform," *IEEE transactions on signal processing*, vol. 61, no. 16, pp. 3999–4010, 2013.

[105] J. Morlet, G. Arens, E. Fourgeau, and D. Giard, "Wave propagation and sampling theory; part ii, sampling theory and complex waves," *Geophysics*, vol. 47, no. 2, pp. 222–236, 02 1982. [Online]. Available: https://doi.org/10.1190/1.1441329

[106] A. Grossmann and J. Morlet, "Decomposition of functions into wavelets of constant shape, and related transforms," in *Mathematics+ Physics: Lectures on Recent Results (Volume 1)*. World Scientific, 1985, pp. 135–165.

[107] ——, "Decomposition of hardy functions into square integrable wavelets of constant shape," *SIAM Journal on Mathematical Analysis*, vol. 15, no. 4, pp. 723–736, 1984. [Online]. Available: https://doi.org/10.1137/0515056

[108] I. W. Selesnick, "Wavelets, a modern tool for signal processing," *Physics Today*, vol. 60, no. 10, pp. 78–79, 2007.

[109] J. Gilles, G. Tran, and S. Osher, "2d empirical transforms. wavelets, ridgelets, and curvelets revisited," *SIAM Journal on Imaging Sciences*, vol. 7, no. 1, pp. 157–186, 2014.

[110] W. Mattar, I. Levy, N. Sharon, and S. Dekel, "Wavelets are all you need for autoregressive image generation," *arXiv preprint arXiv:2406.19997*, 2024.

[111] M. W. Marcellin, M. J. Gormish, A. Bilgin, and M. P. Boliek, "An overview of jpeg-2000," in *Proceedings DCC 2000. Data compression conference*. IEEE, 2000, pp. 523–541.

[112] I. Selesnick, R. Baraniuk, and N. Kingsbury, "The dual-tree complex wavelet transform," *IEEE Signal Process. Mag.*, vol. 22, no. 6, pp. 123–151, nov 2005.

[113] W. Xiang, C. Liu, H. Yu, and X. Chen, "Wavelet-driven masked image modeling: A path to efficient visual representation," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 39, no. 8, 2025, pp. 8611–8619.

[114] J. P. Ward, *Quaternions and Caley Numbers. Algebra and Applications*, ser. Mathematics and Its Applications. Kluwer Academic Publishers, 1997, vol. 403.

[115] T. Parcollet, M. Morchid, and G. Linarès, "A survey of quaternion neural networks," *Artif. Intell. Rev.*, Aug. 2019.

[116] P. Arena, L. Fortuna, L. Occhipinti, and M. G. Xibilia, "Neural networks for quaternion-valued function approximation," in *IEEE Int. Symp. on Circuits and Syst. (ISCAS)*, London, UK, May 1994, pp. 307–310.

[117] T. Parcollet, M. Morchid, and G. Linarès, "Quaternion convolutional neural networks for heterogeneous image processing," in *IEEE Int. Conf. on Acoust., Speech and Signal Process. (ICASSP)*, Brighton, UK, May 2019, pp. 8514–8518.

[118] E. Grassucci, D. Comminiello, and A. Uncini, "An information-theoretic perspective on proper quaternion variational autoencoders," *Entropy*, vol. 23, no. 7, 2021.

[119] W. L. Chan, H. Choi, and R. Baraniuk, "Quaternion wavelets for image analysis and processing," in *2004 International Conference on Image Processing, 2004. ICIP '04.*, vol. 5, 2004, pp. 3057–3060 Vol. 5.

[120] L. H. Ngo, M. Luong, N. M. Sirakov, E. Viennet, and T. Le-Tien, "Skin lesion image classification using sparse representation in quaternion wavelet domain," *Signal, Image and Video Processing*, vol. 16, no. 7, p. 1721–1729, Jan. 2022.

[121] P. Carré and P. Denis, "Quaternionic wavelet transform for colour images," in *Wavelet Applications in Industrial Processing IV*, vol. 6383. SPIE, 2006, p. 638301.

[122] E. Bayro-Corrochano, "The theory and use of the quaternion wavelet transform," *Journal of Mathematical Imaging and Vision*, vol. 24, no. 1, pp. 19–35, 2006.

[123] M. Vetterli and C. Herley, "Wavelets and filter banks: theory and design," *IEEE Trans. on Signal Process.*, vol. 40, no. 9, pp. 2207–2232, 1992.

[124] W. L. Chan, H. Choi, and R. G. Baraniuk, "Coherent multiscale image processing using dual-tree quaternion wavelets," *IEEE Transactions on Image Processing*, vol. 17, no. 7, pp. 1069–1082, 2008.

[125] L. H. Ngo, M. Luong, N. M. Sirakov, T. Le-Tien, S. Guerif, and E. Viennet, "Sparse representation wavelet based classification," in *Int. Conf. on Image Process. (ICIP)*, vol. 5, 2018, p. 2974–2978.

[126] W. Zou and Y. Li, "Image classification using wavelet coefficients in lowpass bands," in *IEEE Int. Joint Conf. on Neural Networks (IJCNN)*, 2007, pp. 114–118.

[127] W. Liu, Q. Yan, and Y. Zhao, "Densely self-guided wavelet network for image denoising," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, 2020, pp. 432–433.

[128] G. Michau, G. Frusque, and O. Fink, "Fully learnable deep wavelet transform for unsupervised monitoring of high-frequency time series," *Proceedings of the National Academy of Sciences*, vol. 119, no. 8, p. e2106598119, 2022.

[129] Y. Zhuang, Z. Wang, F. Tao, and J. Shang, "Wavspa: Wavelet space attention for boosting transformers' long sequence learning ability," in *Proceedings of UniReps: the First Workshop on Unifying Representations in Neural Models*, ser. Proceedings of Machine Learning Research, M. Fumero, E. Rodolá, C. Domine, F. Locatello, K. Dziugaite, and C. Mathilde, Eds., vol. 243. PMLR, 15 Dec 2024, pp. 27–46. [Online]. Available: https://proceedings.mlr.press/v243/zhuang24a.html

[130] T. Yao, Y. Pan, Y. Li, C.-W. Ngo, and T. Mei, "Wave-vit: Unifying wavelet and transformers for visual representation learning," in *Proceedings of the European conference on computer vision (ECCV)*, 2022.

[131] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Advances in Neural Inf. Proc. Systems (NeurIPS)*, Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, and K. Weinberger, Eds., vol. 27. Curran Associates, Inc., 2014.

[132] E. Grassucci, L. Sigillo, A. Uncini, and D. Comminiello, "Hypercomplex image-to- image translation," in *Int. Joint Conf. on Neural Networks (IJCNN)*, 2022, pp. 1–8.

[133] A. Nyberg, A. Eldesokey, D. Bergstrom, and D. Gustafsson, "Unpaired thermal to visible spectrum transfer using adversarial training," in *Proc. Eur. Conf. on Computer Vision (ECCV)*, 2018, pp. 0–0.

[134] T. Le-Tien, T. H. D. Quang, H. Y. Vy, T. Nguyen-Thanh, and H. Phan-Xuan, "Gan-based thermal infrared image colorization for enhancing object identification," in *Int. Symp. Elect. and Electron. Eng. (ISEE)*. IEEE, 2021, pp. 90–94.

[135] A. Toet and M. Hogervorst, "Progress in color night vision," *Optical Engineering*, vol. 51, 02 2012.

[136] S. Park, D. H. Choi, J. U. Kim, and Y. M. Ro, "Robust thermal infrared pedestrian detection by associating visible pedestrian knowledge," in *IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*. IEEE, 2022, pp. 4468–4472.

[137] M. Zhang, R. Zhang, Y. Yang, H. Bai, J. Zhang, and J. Guo, "Isnet: Shape matters for infrared small target detection," in *IEEE/CVF Conf. on Computer Vision and Pattern Recog. (CVPR)*, 2022, pp. 867–876.

[138] G.-A. Wang, T. Zhang, Y. Yang, J. Cheng, J. Chang, X. Liang, and Z.-G. Hou, "Cross-modality paired-images generation for rgb-infrared person re-identification," *Proc. of the AAAI Conf. on Artif. Intell.*, vol. 34, pp. 12 144–12 151, 04 2020.

[139] J. Kim, H. Kim, T. Kim, N. Kim, and Y. Choi, "Mlpd: Multi-label pedestrian detector in multispectral domain," *IEEE Robot. Autom.*, vol. 6, no. 4, pp. 7846–7853, 2021.

[140] Y. Wang, M. Xia, L. Qi, J. Shao, and Y. Qiao, "Palgan: Image colorization with palette generative adversarial networks," in *Proc. Eur. Conf. on Computer Vision (ECCV)*, 2022.

[141] M. Kumar, D. Weissenborn, and N. Kalchbrenner, "Colorization transformer," in *Int. Conf. on Learning Repr., (ICLR)*, 2021. [Online]. Available: https://openreview.net/forum?id=5NA1PinlGFu

[142] R. Zhang, J.-Y. Zhu, P. Isola, X. Geng, A. S. Lin, T. Yu, and A. A. Efros, "Real-time user-guided image colorization with learned deep priors," *ACM Trans. on Graphics (TOG)*, vol. 9, no. 4, 2017.

[143] P. L. Suárez, A. D. Sappa, and B. X. Vintimilla, "Infrared image colorization based on a triplet dcgan architecture," *2017 IEEE Conf. on Computer Vision and Pattern Recog. Workshops (CVPRW)*, pp. 212–217, 2017.

[144] F. Luo, Y. Li, G. Zeng, P. Peng, G. Wang, and Y. Li, "Thermal infrared image colorization for nighttime driving scenes with top-down guided attention," *IEEE Trans. on Intell. Transportation Systems*, vol. 23, no. 9, pp. 15 808–15 823, 2022.

[145] K.-F. Yang, C. Li, and Y. Li, "Multifeature-based surround inhibition improves contour detection in natural images," *IEEE Trans. on Image Proc.*, vol. 23, pp. 5020–5032, 2014.

[146] C. Junxiao, W. Jia, and L. Rui, "TarGAN: Target-aware generative adversarial networks for multi-modality medical image translation," in *Int. Conf. on Medical Image Computing and Computer Assisted Intervention (MICCAI)*, Oct. 2021.

[147] Y. Sun, B. Cao, P. Zhu, and Q. Hu, "Drone-based rgb-infrared cross-modality vehicle detection via uncertainty-aware learning," *IEEE Trans. on Circuits and Systems for Video Technology*, pp. 1–1, 2022.

[148] D. Zhu, S. Liu, W. Jiang, C. Gao, T. Wu, and G. Guo, "Ugan: Untraceable gan for multi-domain face translation," *ArXiv*, vol. abs/1907.11418, 2019.

[149] Z. Wang, A. Bovik, H. Sheikh, and E. Simoncelli, "Image quality assessment: from error visibility to structural similarity," *IEEE Trans. on Image Proc.*, vol. 13, no. 4, pp. 600–612, 2004.

[150] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *2017 IEEE International Conference on Computer Vision (ICCV)*, 2017, pp. 2242–2251.

[151] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," *IEEE Conf. on Computer Vision and Pattern Recog. (CVPR)*, pp. 5967–5976, 2017.

[152] T. Miyato, T. Kataoka, M. Koyama, and Y. Yoshida, "Spectral normalization for generative adversarial networks," in *International Conference on Learning Representations*, 2018. [Online]. Available: https://openreview.net/forum?id=B1QRgziT-

[153] L. R. Dice, "Measures of the amount of ecologic association between species," *Ecology*, vol. 26, pp. 297–302, 1945.

[154] Z. Zhang, L. Yang, and Y. Zheng, "Translating and segmenting multimodal medical volumes with cycle-and shape-consistency generative adversarial network," in *IEEE/CVF Conf. on Computer Vision and Pattern Recog. (CVPR)*, 2018, pp. 9242–9251.

[155] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Int. Conf. on Learning Repr., (ICLR)*, Y. Bengio and Y. LeCun, Eds., 2015. [Online]. Available: http://arxiv.org/abs/1412.6980

[156] T.-C. Wang, M.-Y. Liu, J.-Y. Zhu, A. Tao, J. Kautz, and B. Catanzaro, "High-resolution image synthesis and semantic manipulation with conditional gans," *IEEE/CVF Conf. on Computer Vision and Pattern Recog. (CVPR)*, pp. 8798–8807, 2018.

[157] Y. Choi, Y. Uh, J. Yoo, and J.-W. Ha, "StarGAN v2: Diverse image synthesis for multiple domains," *IEEE/CVF Conf. on Computer Vision and Pattern Recog. (CVPR)*, pp. 8185–8194, 2020.

[158] C. Ledig, L. Theis, F. Huszár, J. Caballero, A. Aitken, A. Tejani, J. Totz, Z. Wang, and W. Shi, "Photo-realistic single image super-resolution using a generative adversarial network," *IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, pp. 105–114, 2017.

[159] O. Karakus, I. G. Rizaev, and A. M. Achim, "Ship wake detection in sar images via sparse regularization," *IEEE Trans. on Geoscience and Remote Sensing*, vol. 58, pp. 1665–1677, 2019.

[160] X. Yang, H. Sun, X. Sun, M. Yan, Z. Guo, and K. Fu, "Position detection and direction prediction for arbitrary-oriented ships via multitask rotation region convolutional neural network," *IEEE Access*, vol. 6, pp. 50 839–50 849, 2018.

[161] C. Dao-Duc, H. Xiaohui, and O. Morère, "Maritime vessel images classification using deep convolutional neural networks," in *Proceedings of the 6th Int. Symposium on Information and Communication Technology*, ser. SoICT '15. New York, NY, USA: Association for Computing Machinery, 2015, p. 276–281.

[162] L. Sigillo, A. Marzilli, D. Moretti, E. Grassucci, C. Greco, and D. Comminiello, "Sailing the Seaformer: A transformer-based model for vessel route forecasting," in *IEEE 33rd Int. Workshop on Machine Learning for Signal Processing (MLSP)*, 2023, pp. 1–6.

[163] Z. Shao, W. Wu, Z. Wang, W. Du, and C. Li, "Seaships: A large-scale precisely annotated dataset for ship detection," *IEEE Trans. on Multimedia*, vol. 20, no. 10, pp. 2593–2604, 2018.

[164] M. Leclerc, R. Tharmarasa, M. C. Florea, A.-C. Boury-Brisset, T. Kirubarajan, and N. Duclos-Hindié, "Ship classification using deep learning techniques for maritime target tracking," in *2018 21st Int. Conf. on Information Fusion (FUSION)*, 2018, pp. 737–744.

[165] Y. Shan, X. Zhou, S. Liu, Y. Zhang, and K. Huang, "SiamFPN: A deep learning method for accurate and real-time maritime ship tracking," *IEEE Trans. on Circuits and Systems for Video Technology*, vol. 31, no. 1, pp. 315–325, 2021.

[166] Z. Wang, J. Chen, and S. C. H. Hoi, "Deep learning for image super-resolution: A survey," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 43, pp. 3365–3387, 2019.

[167] Z. Wan, B. Zhang, D. Chen, P. Zhang, F. Wen, and J. Liao, "Old photo restoration via deep latent space translation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 2, pp. 2071–2087, 2022.

[168] R. A. Yeh, C. Chen, T.-Y. Lim, A. G. Schwing, M. A. Hasegawa-Johnson, and M. N. Do, "Semantic image inpainting with deep generative models," *2017 IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pp. 6882–6890, 2016.

[169] D. Ulyanov, A. Vedaldi, and V. S. Lempitsky, "Deep image prior," *Int. Journal of Computer Vision*, vol. 128, pp. 1867–1888, 2017.

[170] S. W. Zamir, A. Arora, S. H. Khan, M. Hayat, F. S. Khan, and M.-H. Yang, "Restormer: Efficient transformer for high-resolution image restoration," *2022 IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, pp. 5718–5729, 2021.

[171] Y. Xiao, Q. Yuan, K. Jiang, J. He, X. Jin, and L. Zhang, "EDiffSR: An efficient diffusion probabilistic model for remote sensing image super-resolution," *IEEE Trans. on Geoscience and Remote Sensing*, 2024.

[172] S. K. Mudgal, S. K. Sharma, J. Chaturvedi, and A. Sharma, "Brain computer interface advancement in neurosciences: Applications and issues," *Interdisciplinary Neurosurgery*, vol. 20, p. 100694, 2020.

[173] Z. Kourtzi and N. Kanwisher, "Cortical regions involved in perceiving object shape," *Journal of Neuroscience*, vol. 20, no. 9, 2000.

[174] Z. Chen, J. Qing, T. Xiang, W. L. Yue, and J. H. Zhou, "Seeing beyond the brain: Conditional diffusion model with sparse masked modeling for vision decoding," in *IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2023, pp. 22 710–22 720.

[175] B. Zeng *et al.*, "Controllable mind visual diffusion model," *AAAI Conf. Artif. Intell.*, vol. 38, no. 7, pp. 6935–6943, 2024.

[176] Y. Takagi and S. Nishimoto, "High-resolution image reconstruction with latent diffusion models from human brain activity," in *IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2023, pp. 14 453–14 463.

[177] Y.-T. Lan, K. Ren, Y. Wang, W.-L. Zheng, D. Li, B.-L. Lu, and L. Qiu, "Seeing through the brain: Image reconstruction of visual perception from human brain signals," *arXiv preprint arXiv:2308.02510*, 2023.

[178] W. Mai, J. Zhang, P. Fang, and Z. Zhang, "Brain-conditional multimodal synthesis: A survey and taxonomy," *arXiv preprint arXiv:2401.00430*, 2023.

[179] M. Ferrante, T. Boccato, S. Bargione, and N. Toschi, "Decoding visual brain representations from electroencephalography through knowledge distillation and latent diffusion models," *Comput. Biol. and Med.*, p. 108701, 2024.

[180] P. Singh, P. Pandey, K. Miyapuram, and S. Raman, "EEG2IMAGE: image reconstruction from EEG brain signals," in *IEEE Int. Conf. Acoust., Speech and Signal Process. (ICASSP)*. IEEE, 2023.

[181] P. Singh, D. Dalal, G. Vashishtha, K. Miyapuram, and S. Raman, "Learning robust deep visual representations from EEG brain recordings," in *IEEE/CVF Winter Conf. on App. of Comput. Vis. (WACV)*, 2024, pp. 7553–7562.

[182] H. Fu, H. Wang, J. J. Chin, and Z. Shen, "Brainvis: Exploring the bridge between brain and visual signals via image reconstruction," in *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2025, pp. 1–5.

[183] Y. Bai, X. Wang, Y.-p. Cao, Y. Ge, C. Yuan, and Y. Shan, "DreamDiffusion: Generating high-quality images from brain EEG signals," *arXiv preprint arXiv:2306.16934*, 2023.

[184] J. Wang, Z. Yue, S. Zhou, K. C. Chan, and C. C. Loy, "Exploiting diffusion prior for real-world image super-resolution," *International Journal of Computer Vision*, pp. 1–21, 2024.

[185] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever, "Learning transferable visual models from natural language supervision," in *Int. Conf. on Machine Learning*, 2021.

[186] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *2016 IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 770–778.

[187] X. Wang, K. Yu, C. Dong, and C. Change Loy, "Recovering realistic texture in image super-resolution by deep spatial feature transform," in *IEEE/CVF Conf. on Computer Vision and Pattern Recognition*, 2018.

[188] J. Choi, J. Lee, C. Shin, S. Kim, H. Kim, and S. Yoon, "Perception prioritized training of diffusion models," in *IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2022.

[189] E. Gundogdu, B. Solmaz, V. Yücesoy, and A. Koç, "MARVEL: A large-scale image dataset for maritime vessels," in *ACCV*, S.-H. Lai, V. Lepetit, K. Nishino, and Y. Sato, Eds. Springer Int. Publishing, 2017.

[190] X. Wang, L. Xie, C. Dong, and Y. Shan, "Real-ESRGAN: Training real-world blind super-resolution with pure synthetic data," in *IEEE/CVF Int. Conf. on Computer Vision Workshops (ICCVW)*, 2021.

[191] E. Postolache, N. Polouliakh, H. Kitano, A. Connelly, E. Rodolà, and T. Akama, "Naturalistic music decoding from EEG data via latent diffusion models," *arXiv preprint arXiv:2405.09062*, 2024.

[192] A. Défossez, C. Caucheteux, J. Rapin, O. Kabeli, and J.-R. King, "Decoding speech perception from non-invasive brain recordings," *Nature Mach. Intell.*, vol. 5, no. 10, pp. 1097–1107, 2023.

[193] I. Kavasidis, S. Palazzo, C. Spampinato, D. Giordano, and M. Shah, "Brain2Image: Converting brain signals into images," in *ACM Int. Conf. Multimedia*, 2017, pp. 1809–1817.

[194] S. Khare, R. N. Choubey, L. Amar, and V. Udutalapalli, "NeuroVision: perceived image regeneration using cProGAN," *Neural Comput. and App.*, vol. 34, no. 8, pp. 5979–5991, 2022.

[195] X. Zheng, W. Chen, M. Li, T. Zhang, Y. You, and Y. Jiang, "Decoding human brain activity with deep learning," *Biomed. Signal Process. and Control*, vol. 56, p. 101730, 2020.

[196] A. Fares, S.-h. Zhong, and J. Jiang, "Brain-media: A dual conditioned and lateralization supported GAN (DCLS-GAN) towards visualization of image-evoked brain activities," in *ACM Int. Conf. Multimedia*, 2020, pp. 1764–1772.

[197] A. Odena, C. Olah, and J. Shlens, "Conditional image synthesis with auxiliary classifier gans," in *International conference on machine learning*. PMLR, 2017, pp. 2642–2651.

[198] R. Mishra, K. Sharma, R. R. Jha, and A. Bhavsar, "NeuroGAN: image reconstruction from EEG signals via an attention-based GAN," *Neural Computing and Applications*, vol. 35, no. 12, pp. 9181–9192, 2023.

[199] M. Mustafa, L. Lindemann, and M. Magnor, "EEG analysis of implicit human visual perception," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 2012, pp. 513–516.

[200] C. Spampinato, S. Palazzo, I. Kavasidis, D. Giordano, N. Souly, and M. Shah, "Deep learning human mind for automated visual classification," in *IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2017, pp. 6809–6817.

[201] S. Palazzo, C. Spampinato, I. Kavasidis, D. Giordano, J. Schmidt, and M. Shah, "Decoding brain representations by multimodal learning of neural activity and visual features," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 11, pp. 3833–3849, 2020.

[202] P. Kumar, R. Saini, P. P. Roy, P. K. Sahu, and D. P. Dogra, "Envisioned speech recognition using EEG sensors," *Pers. and Ubiquitous Comput.*, vol. 22, pp. 185–199, 2018.

[203] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*. IEEE, 2009, pp. 248–255.

[204] Z. Zhou, L. Qi, and Y. Shi, "Generalizable medical image segmentation via random amplitude mixup and domain-specific image restoration," in *European Conference on Computer Vision*. Springer, 2022, pp. 420–436.

[205] G. Zhao, Q. Feng, C. Chen, Z. Zhou, and Y. Yu, "Diagnose like a radiologist: Hybrid neuro-probabilistic reasoning for attribute-based medical image diagnosis," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, pp. 1–18, 2021.

[206] T. Zhou, Q. Cheng, H. Lu, Q. Li, X. Zhang, and S. Qiu, "Deep learning methods for medical image fusion: A review," *Computers in Biology and Medicine*, vol. 160, p. 106959, 2023.

[207] N. K. Tomar, D. Jha, M. A. Riegler, H. D. Johansen, D. Johansen, J. Rittscher, P. Halvorsen, and S. Ali, "Fanet: A feedback attention network for improved biomedical image segmentation," *IEEE Transactions on Neural Networks and Learning Systems*, pp. 1–14, 2022.

[208] Y. Song, L. Shen, L. Xing, and S. Ermon, "Solving inverse problems in medical imaging with score-based generative models," in *Int. Conf. on Learning Representations (ICLR)*, 2022.

[209] S. Masoudi, S. Harmon, S. Mehralivand, S. Walker, H. Raviprakash, U. Bagci, P. Choyke, and T. B., "Quick guide on radiology image pre-processing for deep learning applications in prostate cancer research," *Journal of Medical Imaging*, vol. 8, no. 1, pp. 893–912, 2021.

[210] Y. Yang, "Medical image fusion through a new DWT based technique," in *Int. Conf. on Bioinformatics and Biomedical Eng.*, 2010, pp. 1–4.

[211] Y. Yan, L. Li, and Y. Yang, "Image fusion based on principal component analysis in dual-tree complex wavelet transform domain," in *Int. Conf. on Wavelet Active Media Technology and Information Process. (ICWAMTIP)*, 2012, p. 70.73.

[212] Z. Zhancheng, L. Xiaoqing, X. Mengyu, W. Zhiwen, and L. Kai, "Medical image fusion based on quaternion wavelet transform and visibility feature," *Journal of Algorithms and Computational Technology*, vol. 14, p. 1–12, 2020.

[213] P. Geng, X. Sun, and J. Liu, "Adopting quaternion wavelet transform to fuse multi-modal medical images," *Journal of Medical and Biological Eng.*, vol. 37, pp. 230 – 239, 2017.

[214] A. Muppidi and M. Radfar, "Speech emotion recognition using quaternion convolutional neural networks," in *IEEE Int. Conf. on Acoustics, Speech and Signal Process. (ICASSP)*, 2021, pp. 6309–6313.

[215] E. Grassucci, D. Comminiello, and A. Uncini, "A quaternion-valued variational autoencoder," in *IEEE Int. Conf. on Acoust., Speech and Signal Process. (ICASSP)*, Jun. 2021.

[216] C. Brignone, G. Mancini, E. Grassucci, A. Uncini, and D. Comminiello, "Efficient sound event localization and detection in the quaternion domain," *IEEE Trans. on Circuits and Systems II: Express Brief*, vol. 69, no. 5, pp. 2453–2457, 2022.

[217] T. Chang and C. Kuo, "Texture analysis and classification with treestructured wavelet transform," *IEEE Trans. Img. Proc.*, vol. 2, no. 4, p. 429–441, oct 1993.

[218] K. Huang and S. Aviyente, "Wavelet feature selection for image classification," *IEEE Trans. Img. Proc.*, vol. 17, no. 9, pp. 1709–1720, sept 2008.

[219] T. Cover and J. Thomas, *Elements of Information Theory*. Wiley, 1991.

[220] J. Liu and P. Moulin, "Information-theoretic analysis of interscale and intrascale dependencies between image wavelet coefficients," *IEEE Trans. Img. Proc.*, vol. 10, no. 11, p. 1647–1658, nov 2001.

[221] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, 2015.

[222] Z. Zhou, M. M. R. Siddiquee, N. Tajbakhsh, and J. Liang, "U-Net++: Redesigning skip connections to exploit multiscale features in image segmentation," *IEEE Trans. on Medical Imaging*, vol. 39, pp. 1856–1867, 2020.

[223] P. Chai, X.-Q. Luo, and Z. Zhang, "Image fusion using quaternion wavelet transform and multiple features," *IEEE Access*, vol. 5, pp. 6724–6734, 2017.

[224] H. Dastmalchi and H. Aghaeinia, "Super-resolution of very low-resolution face images with a wavelet integrated, identity preserving, adversarial network," *Signal Processing: Image Communication*, vol. 107, p. 116755, 2022.

[225] W. Zou, L. Chen, Y. Wu, Y. Zhang, Y. Xu, and J. Shao, "Joint wavelet subbands guided network for single image super-resolution," *IEEE Transactions on Multimedia*, 2022.

[226] D. Jha, P. H. Smedsrud, M. A. Riegler, P. Halvorsen, T. de Lange, D. Johansen, and H. D. Johansen, "Kvasir-seg: A segmented polyp dataset," in *MultiMedia Modeling: 26th International Conference*, 2020, p. 451–462.

[227] A. E. Kavur, M. A. Selver, O. Dicle, M. Barış, and N. S. Gezer, "CHAOS - Combined (CT-MR) Healthy Abdominal Organ Segmentation Challenge Data," *Medical Image Analysis*, Apr 2019.

[228] Z. Zhou, V. Sodha, M. M. Rahman S., R. Feng, N. Tajbakhsh, M. B. Gotway, and J. Liang, "Models genesis: Generic autodidactic models for 3d medical image analysis," in *Int. Conf. on Medical Image Computing and Computer Assisted Intervention (MICCAI)*, 2019, pp. 384–393.

[229] A. I. Hazmy, A. Hawbani, X. Wang, A. Al-Dubai, A. Ghannami, A. A. Yahya, L. Zhao, and S. H. Alsamhi, "Potential of satellite-airborne sensing technologies for agriculture 4.0 and climate-resilient: A review," *IEEE Sensors Journal*, vol. 24, no. 4, pp. 4161–4180, 2024.

[230] J. Cornebise, I. Oršolić, and F. Kalaitzis, "Open high-resolution satellite imagery: The worldstrat dataset–with application to super-resolution," *Advances in Neural Information Processing Systems*, vol. 35, pp. 25 979–25 991, 2022.

[231] Z. Xiao, K. Kreis, and A. Vahdat, "Tackling the generative learning trilemma with denoising diffusion GANs," in *International Conference on Learning Representations*, 2022.

[232] D. C. Lepcha, B. Goyal, A. Dogra, and V. Goyal, "Image super-resolution: A comprehensive review, recent trends, challenges and applications," *Information Fusion*, vol. 91, pp. 230–260, 2023.

[233] J. Kim, J. Oh, and K. M. Lee, "Beyond image super-resolution for image recognition with task-driven perceptual loss," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 2651–2661.

[234] W.-Y. Hsu and P.-W. Jian, "Wavelet pyramid recurrent structure-preserving attention network for single image super-resolution," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 35, no. 11, pp. 15 772–15 786, 2024.

[235] N. Ruiz, Y. Li, V. Jampani, Y. Pritch, M. Rubinstein, and K. Aberman, "Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2023, pp. 22 500–22 510.

[236] Y. Takagi and S. Nishimoto, "High-resolution image reconstruction with latent diffusion models from human brain activity," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2023, pp. 14 453–14 463.

[237] S. Shang, Z. Shan, G. Liu, L. Wang, X. Wang, Z. Zhang, and J. Zhang, "Resdiff: Combining cnn and diffusion model for image super-resolution," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 38, no. 8, 2024, pp. 8975–8983.

[238] R. Wu, T. Yang, L. Sun, Z. Zhang, S. Li, and L. Zhang, "Seesr: Towards semantics-aware real-world image super-resolution," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2024, pp. 25 456–25 467.

[239] B. B. Moser, S. Frolov, F. Raue, S. Palacio, and A. Dengel, "Waving goodbye to low-res: A diffusion-wavelet approach for image super-resolution," in *2024 International Joint Conference on Neural Networks (IJCNN)*, 2024, pp. 1–8.

[240] X. Ji, Y. Cao, Y. Tai, C. Wang, J. Li, and F. Huang, "Real-world super-resolution via kernel estimation and noise injection," in *proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, 2020, pp. 466–467.

[241] K. Zhang, J. Liang, L. Van Gool, and R. Timofte, "Designing a practical degradation model for deep blind image super-resolution," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 4791–4800.

[242] C. Chen, X. Shi, Y. Qin, X. Li, X. Han, T. Yang, and S. Guo, "Real-world blind super-resolution via feature matching with implicit high-resolution priors," in *Proceedings of the 30th ACM International Conference on Multimedia*, 2022, pp. 1329–1338.

[243] Z. Yue, J. Wang, and C. C. Loy, "Resshift: Efficient diffusion model for image super-resolution by residual shifting," *Advances in Neural Information Processing Systems*, vol. 36, 2024.

[244] E. Agustsson and R. Timofte, "Ntire challenge on single image super-resolution: Dataset and study," in *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, 2017.

[245] R. Timofte, E. Agustsson, L. Van Gool, M.-H. Yang, and L. Zhang, "Ntire challenge on single image super-resolution: Methods and results," in *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, 2017.

[246] X. Wang, K. Yu, C. Dong, and C. C. Loy, "Recovering realistic texture in image super-resolution by deep spatial feature transform," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 606–615.

[247] S. Khanna, P. Liu, L. Zhou, C. Meng, R. Rombach, M. Burke, D. B. Lobell, and S. Ermon, "Diffusionsat: A generative foundation model for satellite imagery," in *The Twelfth International Conference on Learning Representations*, 2024. [Online]. Available: https://openreview.net/forum?id=I5webNFDgQ

[248] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever, "Learning transferable visual models from natural language supervision," in *Int. Conf. on Machine Learning*, 2021.

[249] C. Zhao, W. Cai, C. Dong, and C. Hu, "Wavelet-based fourier information interaction with frequency diffusion adjustment for underwater image restoration," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 8281–8291.

[250] X. Wang, K. Yu, S. Wu, J. Gu, Y. Liu, C. Dong, Y. Qiao, and C. Change Loy, "Esrgan: Enhanced super-resolution generative adversarial networks," in *Proceedings of the European conference on computer vision (ECCV) workshops*, 2018, pp. 0–0.

[251] M. Haris, G. Shakhnarovich, and N. Ukita, "Deep back-projection networks for super-resolution," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 1664–1673.

[252] Z. Luo, B. Song, and L. Shen, "Satdiffmoe: A mixture of estimation method for satellite image super-resolution with latent diffusion models," in *2nd Workshop on Advancing Neural Network Training: Computational Efficiency, Scalability, and Resource Optimization (WANT@ICML 2024)*, 2024. [Online]. Available: https://openreview.net/forum?id=dpK5a5xnel

[253] B. B. Moser, S. Frolov, F. Raue, S. Palacio, and A. Dengel, "Dwa: Differential wavelet amplifier for image super-resolution," in *Artificial Neural Networks and Machine Learning – ICANN*. Cham: Springer Nature Switzerland, 2023, pp. 232–243.

[254] J. Choi, J. Lee, C. Shin, S. Kim, H. Kim, and S. Yoon, "Perception prioritized training of diffusion models," in *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022, pp. 11 462–11 471.

[255] D. Podell, Z. English, K. Lacey, A. Blattmann, T. Dockhorn, J. Müller, J. Penna, and R. Rombach, "SDXL: Improving latent diffusion models for high-resolution image synthesis," in *The Twelfth International Conference on Learning Representations*, 2024. [Online]. Available: https://openreview.net/forum?id=di52zR8xgf

[256] A. Sauer, D. Lorenz, A. Blattmann, and R. Rombach, "Adversarial diffusion distillation," in *European Conference on Computer Vision*. Springer, 2024, pp. 87–103.

[257] Z. Wang, J. Bao, S. Gu, D. Chen, W. Zhou, and H. Li, "Designdiffusion: High-quality text-to-design image generation with diffusion models," 2025. [Online]. Available: https://arxiv.org/abs/2503.01645

[258] Y. Wang, W. Yang, X. Chen, Y. Wang, L. Guo, L.-P. Chau, Z. Liu, Y. Qiao, A. C. Kot, and B. Wen, "Sinsr: diffusion-based image super-resolution in a single step," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 25 796–25 805.

[259] T. Brooks, A. Holynski, and A. A. Efros, "Instructpix2pix: Learning to follow image editing instructions," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2023, pp. 18 392–18 402.

[260] C. Meng, Y. He, Y. Song, J. Song, J. Wu, J.-Y. Zhu, and S. Ermon, "SDEdit: Guided image synthesis and editing with stochastic differential equations," in *International Conference on Learning Representations*, 2022. [Online]. Available: https://openreview.net/forum?id=aBsCjcPu_tE

[261] J. Song, C. Meng, and S. Ermon, "Denoising diffusion implicit models," in *International Conference on Learning Representations*, 2021. [Online]. Available: https://openreview.net/forum?id=St1giarCHLP

[262] P. Esser, S. Kulal, A. Blattmann, R. Entezari, J. Müller, H. Saini, Y. Levi, D. Lorenz, A. Sauer, F. Boesel *et al.*, "Scaling rectified flow transformers for high-resolution image synthesis," in *Forty-first international conference on machine learning*, 2024.

[263] R. Du, D. Chang, T. Hospedales, Y.-Z. Song, and Z. Ma, "Demofusion: Democratising high-resolution image generation with no $$$," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 6159–6168.

[264] J. Ren, W. Li, H. Chen, R. Pei, B. Shao, Y. Guo, L. Peng, F. Song, and L. Zhu, "Ultrapixel: Advancing ultra high-resolution image synthesis to new peaks," in *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. [Online]. Available: https://openreview.net/forum?id=voJCpdlw53

[265] J. Chen, C. Ge, E. Xie, Y. Wu, L. Yao, X. Ren, Z. Wang, P. Luo, H. Lu, and Z. Li, "Pixart-σ: Weak-to-strong training of diffusion transformer for 4k text-to-image generation," in *European Conference on Computer Vision*. Springer, 2024, pp. 74–91.

[266] Y. Liu, K. Zhang, Y. Li, Z. Yan, C. Gao, R. Chen, Z. Yuan, Y. Huang, H. Sun, J. Gao, L. He, and L. Sun, "Sora: A review on background, technology, limitations, and opportunities of large vision models," 2024.

[267] G. Liu, Y. Wang, Z. Feng, Q. Wu, L. Tang, Y. Gao, Z. Li, S. Cui, J. McAuley, Z. Yang, E. P. Xing, and Z. Hu, "Unified generation, reconstruction, and representation: Generalized diffusion with adaptive latent encoding-decoding," in *Forty-first International Conference on Machine Learning*, 2024. [Online]. Available: https://openreview.net/forum?id=igRjCCAz2a

[268] J. Hahm, J. Lee, S. Kim, and J. Lee, "Isometric representation learning for disentangled latent space of diffusion models," in *Forty-first International Conference on Machine Learning*, 2024. [Online]. Available: https://openreview.net/forum?id=ufCptn28vG

[269] I. Skorokhodov, S. Girish, B. Hu, W. Menapace, Y. Li, R. Abdal, S. Tulyakov, and A. Siarohin, "Improving the diffusability of autoencoders," 2025. [Online]. Available: https://arxiv.org/abs/2502.14831

[270] T. Kouzelis, I. Kakogeorgiou, S. Gidaris, and N. Komodakis, "Eq-vae: Equivariance regularized latent space for improved generative image modeling," *arXiv preprint arXiv:2502.09509*, 2025.

[271] M. Caron, H. Touvron, I. Misra, H. Jégou, J. Mairal, P. Bojanowski, and A. Joulin, "Emerging properties in self-supervised vision transformers," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 9650–9660.

[272] S. Mallat and W. L. Hwang, "Singularity detection and processing with wavelets," *IEEE transactions on information theory*, vol. 38, no. 2, pp. 617–643, 1992.

[273] B. B. Moser, S. Frolov, F. Raue, S. Palacio, and A. Dengel, " Dynamic Attention-Guided Diffusion for Image Super-Resolution ," in *2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*. Los Alamitos, CA, USA: IEEE Computer Society, Mar. 2025, pp. 451–460. [Online]. Available: https://doi.ieeecomputersociety.org/10.1109/WACV61041.2025.00054

[274] C. Schuhmann, R. Beaumont, R. Vencu, C. Gordon, R. Wightman, M. Cherti, T. Coombes, A. Katta, C. Mullis, M. Wortsman, P. Schramowski, S. Kundurthy, K. Crowson, L. Schmidt, R. Kaczmarczyk, and J. Jitsev, "Laion-5b: an open large-scale dataset for training next generation image-text models," in *Proceedings of the 36th International Conference on Neural Information Processing Systems*, ser. NIPS '22. Red Hook, NY, USA: Curran Associates Inc., 2022.

[275] F. Cotter, "Uses of complex wavelets in deep convolutional neural networks," Ph.D. dissertation, Apollo - University of Cambridge Repository, 2019. [Online]. Available: https://www.repository.cam.ac.uk/handle/1810/306661

[276] R. Yu, S. Liu, Z. Tan, and X. Wang, "Ultra-resolution adaptation with ease," *International Conference on Machine Learning*, 2025.

[277] E. Xie, J. Chen, J. Chen, H. Cai, H. Tang, Y. Lin, Z. Zhang, M. Li, L. Zhu, Y. Lu, and S. Han, "SANA: Efficient high-resolution text-to-image synthesis with linear diffusion transformers," in *The Thirteenth International Conference on Learning Representations*, 2025. [Online]. Available: https://openreview.net/forum?id=N8Oj1XhtYZ

[278] R. Du, D. Liu, L. Zhuo, Q. Qi, H. Li, Z. Ma, and P. Gao, "I-max: Maximize the resolution potential of pre-trained rectified flow transformers with projected flow," 2024. [Online]. Available: https://arxiv.org/abs/2410.07536

[279] Q. Qin, L. Zhuo, Y. Xin, R. Du, Z. Li, B. Fu, Y. Lu, J. Yuan, X. Li, D. Liu *et al.*, "Lumina-image 2.0: A unified and efficient image generative framework," *arXiv preprint arXiv:2503.21758*, 2025.

[280] L. Zhang, L. Zhang, X. Mou, and D. Zhang, "Fsim: A feature similarity index for image quality assessment," *IEEE Transactions on Image Processing*, vol. 20, no. 8, pp. 2378–2386, 2011.

[281] Z. Wang, E. Simoncelli, and A. Bovik, "Multiscale structural similarity for image quality assessment," in *The Thrity-Seventh Asilomar Conference on Signals, Systems & Computers, 2003*, vol. 2, 2003, pp. 1398–1402 Vol.2.