Neutralizing Proactive Defense using Diffusion-based Upsampling

Giuseppe Daidone*
giuseppe.daidone@uniroma1.it
Sapienza University of Rome
Rome, Italy

Filippo Bartolucci
filippo.bartolucci3@unibo.it
University of Bologna
Bologna, Italy

Maria Rosaria Briglia
briglia@di.uniroma1.it
Sapienza University of Rome
Rome, Italy

Mujtaba Hussain Mirza
mirza@di.uniroma1.it
Sapienza University of Rome
Rome, Italy

Giuseppe Lisanti
giuseppe.lisanti@unibo.it
University of Bologna
Bologna, Italy

Iacopo Masi
masi@di.uniroma1.it
Sapienza University of Rome
Rome, Italy

Abstract

The rapid spread of open-source generative models has made it easy to create highly realistic manipulated media, posing a critical threat to content authenticity and provenance. Proactive image protection mechanisms have recently emerged as a promising defense, embedding imperceptible or characteristic signals into images to enable reliable manipulation detection. However, their robustness under realistic and adversarial post-release conditions remains largely unexplored. In this work, we present a systematic evaluation of the robustness of recent state-of-the-art proactive image protection schemes in a *black-box* setting. We analyze the resilience of PADL [6] and DiffVax [20] protections against a broad range of attacks, including classical image transformations and diffusion-based reconstruction attacks that implicitly re-synthesize image content while preserving perceptual quality. Our results reveal that, despite strong performance under limited perturbations, current proactive defenses are vulnerable to unseen image manipulations and generative reconstruction attacks. Considering PADL, we empirically demonstrate that adding diffusion-based upsampling attacks in the training does not improve robustness, without increasing protection intensity. These findings expose critical gaps between assumed and real-world threat models, highlighting the need for more robust proactive protection designs and standardized evaluation protocols for trustworthy digital media.

CCS Concepts

• **Computing methodologies** → **Machine learning**; *Computer vision.*

Keywords

watermarking, robustness, proactive defense

ACM Reference Format:
Giuseppe Daidone, Filippo Bartolucci, Maria Rosaria Briglia, Mujtaba Hussain Mirza, Giuseppe Lisanti, and Iacopo Masi. 2026. Neutralizing Proactive Defense using Diffusion-based Upsampling. In *ACM Workshop on Information Hiding and Multimedia Security (IH&MMSec '26), June 17–19, 2026, Firenze, Italy*. ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3785353.3815091

*Corresponding author.

IH&MMSec '26, Firenze, Italy

ACM ISBN 979-8-4007-2376-6/2026/06
https://doi.org/10.1145/3785353.3815091

1 Introduction

Media manipulation via open-source generative models is rapidly transitioning from a localized technical challenge into a pervasive systemic issue. To understand the current situation, we must recognize the two recent distinct technological waves that have redefined the limits of synthetic media. The initial wave was marked by the emergence of deepfakes [13], a class of AI-driven media manipulation techniques primarily focused on face swapping and facial reenactment. These early systems relied largely on Generative Adversarial Networks (GANs) [8], enabling high levels of local visual realism. While GAN-based approaches achieved convincing facial detail, they often struggled with training instability and visual glitches. The second wave is more impactful and is dominated by text-to-image diffusion models (DMs) and large-scale multimodal systems. Unlike their predecessors, diffusion models leverage iterative denoising to achieve a level of structural coherence and semantic detail that was previously impossible. This shift has effectively closed the "distinction gap". Today, the output from open-source DMs, Large Language Models (LLMs), and high-fidelity video generators is becoming virtually indistinguishable from authentic media. As these tools become increasingly decentralized and accessible, our collective ability to verify digital truth faces an unprecedented inflection point. The resulting challenge of content authenticity and provenance has emerged as a critical point in this scenario where the main focus is on maintaining trust in digital ecosystems [17]. The current defense mechanisms based on content watermarking have become crucial for detecting and establishing the provenance of AI-manipulated content [37]. On the other hand, this technology, unfortunately, demonstrated low robustness against trivial circumvention techniques such as removal attacks, manipulation, and forgery attacks [14]. The presence of adversarial settings adds further requirements for the defensive mechanisms, which must not only embed robust signals but also withstand sophisticated adversarial attacks [7]. In this setting, proactive defense mechanisms emerge as the most suitable option to immunize digital content against manipulation through imperceptible perturbations [5]. Although these pipelines natively add engineered noise to the input

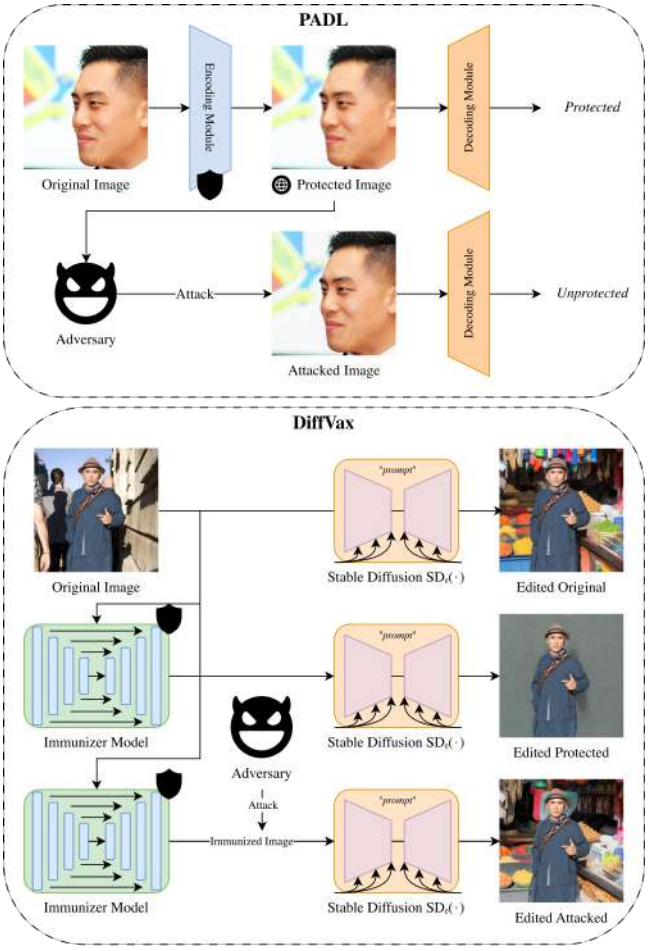


Figure 1: Robustness analysis overview of PADL [6] protection and DiffVax [20] immunization mechanisms under real-world black-box adversarial perturbations, where PADL encodes a proactive watermark detected through a decoder that localizes image forgery, and DiffVax injects an immunization noise that deviates from any attempt of Stable Diffusion editing. For each method, we first encode the protection or the immunization, respectively. Then, for PADL, we attack the protected image and test if the decoder no longer detects the presence of the watermark. While for DiffVax, we attack the immunized image and compare the Stable Diffusion editing model capability between the immunized images and the attacked images.

to proactively discourage edits, their robustness to unseen and common real-world image modifications remains unclear.

In this paper, we analyze the robustness of the state-of-the-art proactive mechanisms for detecting manipulated images against removal attacks and image perturbations. Figure 1 shows the setting of our evaluation: in the most general setting, we take a proactive defense mechanism: first, we encode the input, i.e., add the watermark, then we attack the protected images by perturbing them and verifying their persistence and detectability. We draw inspiration from [11]: while Honig et al. tested the persistence of proactive defense against style mimicry and artistic content, in this paper, we extend this study to generic proactive defenses like [6, 20] and natural images. For evaluation, we measure attack effectiveness, image degradation metrics, and the logit distribution. Our main contributions are summarized as follows:

- We provide a comprehensive robustness evaluation of recent state-of-the-art proactive image protection mechanisms, specifically PADL [6] and DiffVax [20] under a realistic black-box threat model. We assess the resilience of these frameworks against both classical image transformations (such as noise injection and geometric shifts) and, taking inspiration from [11], against newer generative reconstruction attacks.
- We demonstrate that these proactive defenses are highly susceptible to our proposed black-box diffusion-based upsampling attacks. By evaluating compressed, noisy, and standard upscaling variants, we show that these methods effectively neutralize embedded protective signals. These attacks achieve near-perfect success rates across varying protection depths while strictly preserving the semantic content and visual fidelity of the original media.
- We investigate whether this vulnerability can be mitigated through targeted data augmentation. In the case of PADL [6], by integrating upscaling augmentations directly into the defense model's training pipeline, we show that exposing the model to these operations during training is insufficient to restore robustness. This reveals an inherent limitation in current proactive defense designs that rely on low-magnitude embedded signatures, which remain vulnerable to generative upscaling operations.

2 Related Work

To proactively protect images against text-to-image diffusion models, Salman et al. [24] proposed an encoder attack that maps the original image to a target image, and a diffusion attack that deviates the diffusion model's output using projected gradient descent (PGD). More effectively, Lo et al. [16] proposed an analogous proactive approach by semantically attacking the cross-attention mechanism, resulting in corrupted results. Similarly, Dual Attention Guided Noise Perturbation (DANP) [34] encodes perturbations to attack the attention mechanism and noise prediction processes, leveraging a given text prompt. Conversely, in [28], the authors perturbed the cross-attention by using an automatically generated caption of the source image as a proxy, without requiring knowledge of the editing prompt. To address the limitation of computational overhead in the immunization process, DiffVax [20] first generates a mask on the subject to protect, then injects noise into the mask to minimize perturbation to preserve image quality and maximize error in the edited image. VoidFace [29] proactively prevents images from face-swapping manipulations on stable diffusion models, leveraging an adaptive step-by-step destruction of the face-swapping pipeline, by generating adversarial images for deviating the face localization and bounding box selection, erasing the facial identity recognition in the embedding space, decoupling the attention mechanism, and corrupting the image features, though showing highly visible artifacts on the facial structure. PADL [6] leverages cross-attention for encoding an invisible watermark to protect images proactively.

When a generative model uses a watermarked image for editing, a decoder can detect and localize the manipulations on the modified image. Comparably, Sander et al. [25] proposed a deep learning model that embeds a 32-bit watermark into localized regions of an image via segmentation, making it resilient to editing. Glaze [26] introduces a style-based protection by shifting the image representation in the model's feature space toward a different target style, effectively misleading the model's ability to learn a specific artistic aesthetic. Notably, noising a watermarked image and then upscaling it was practically shown to be effective for vanishing anti-plagiarism watermarks [11].

To enhance robustness and reduce complexity, in the literature, alternative solutions to pixel space protections have been proposed, leveraging the model's latent space for watermark addition. Wen et al. [32] embedded a Fourier-space pattern into the initial noise vector of the diffusion model, resulting in a model fingerprint on the generated image. The watermark is structured as a ring to be resistant to classical image transformations (i.e., rotation, translation, compression, and color jitter), taking advantage of Fourier transform properties for periodic signals. However, the watermark can be verified only by the model owner and it is limited to diffusion models, hindering large-scale adoption. Similarly, in [10], a binary watermark is first mapped into a unit sphere and then embedded into the noise input of the diffusion model. The spherical mapping process first projects the bits onto a unit sphere, then applies an orthogonal rotation, and scales it by a chi-square distributed radius. This framework shows improvements in terms of robustness and computational efficiency, while still suffering from the model's owner-only verification process. To address the issue of distribution preserving, Yang et al. [33] proposed Gaussian Shading, a framework that embeds a latent-space watermark following a Gaussian distribution through a stream key. For extraction, the authors employed Denoising Diffusion Implicit Model (DDIM) inversion, achieving robustness to noise perturbations. However, the watermarking detection is limited to the key owner, and the scheme can be vulnerable to forgery attacks. Therefore, Aremu et al. [2] proposed a generalized multiple-key scheme in which a single randomly sampled key, from a set of pre-generated secret keys, is used for embedding the watermark in the latent space. During the detection pipeline, every key is tested and scored; a high score from no keys indicates non-watermarked content, from a single key indicates a valid watermark, and from multiple keys indicates forgery. While watermark generation requires slight overhead, the complexity of the detection mechanisms increases linearly with respect to the number of keys. Similarly, SEAL [1] employs multiple key-based watermarking, but the diffusion initial noise is conditioned on the keys derived from the image semantic embedding rather than based on pre-generated keys, i.e., without requiring storing them. While this method is still robust against forgery attacks, it showed high vulnerability to removal attacks. PSyDUCK [18] leverages a model-independent steganography approach by using pre-shared keys to embed a secret message in the latent space of a diffusion model for images and videos. This scheme does not require re-training and exploits denoising trajectories to embed messages of arbitrary length. However, PSyDUCK utilizes the same set of keys for both encoding and decoding, attaining confidentiality rather than non-repudiation, unlike what is desirable in watermarking for content provenance.

3 Method

In this section, we will present the threat model, specifying the possibilities of the attacker and the available tools. Furthermore, we present the two state-of-the-art proactive protection techniques under consideration, highlighting their potential vulnerabilities and the current lack of thorough analysis, which may result in improper operation of these techniques.

3.1 Threat Model

We study the robustness of proactive image protection mechanisms: we consider a defender that applies a proactive protection to an image prior to public dissemination, embedding an imperceptible signal intended to enable the detection of manipulations or provenance verification at a later stage; on the other side, an attacker that manipulates and perturbs the data to remove the protection. An overview of our analysis is shown in Figure 1.

Defender The defender has full control over the protection pipeline, including the encoding procedure and the corresponding verification or decoding mechanism. Given a clean image $\boldsymbol{x}$, the defender produces a protected image $\tilde{\boldsymbol{x}} = \mathcal{E}_\theta(\boldsymbol{x})$, where $\mathcal{E}$ denotes the encoder of the protecting network. At verification time, a decoder $\mathcal{D}_\theta(\cdot)$ can be used to assess whether the protection signal is present in a given image and returns 1 if protected and 0 otherwise. We assume that the defender can reliably verify the presence of the protection signal on $\tilde{\boldsymbol{x}}$ in the absence of attacks.

Adversary The adversary operates after content release and has access only to the protected image $\tilde{\boldsymbol{x}}$. The adversary aims to produce a manipulated image $\boldsymbol{x}'$ such that either 1) the protection signal is no longer detectable, i.e., $\mathcal{D}_\theta(\boldsymbol{x}') = 0$ (PADL case [6]) or 2) the image is not immunized anymore, and the malicious editing of the image correctly occurs (DiffVax case [20]). The adversary seeks to preserve the perceptual quality and semantic content of the original image, that is $\boldsymbol{x}' \approx \tilde{\boldsymbol{x}}$.

We consider adaptive adversaries that may apply a sequence of transformations, including both standard image manipulations and editing based on neural models. We analyze a black-box setting with respect to the defender's protection mechanism.

Attack capabilities The adversary is allowed to apply any image-space transformation $\mathcal{T}$ satisfying a bounded distortion constraint, producing $\boldsymbol{x}' = \mathcal{T}(\tilde{\boldsymbol{x}})$. These transformations include common real-world post-processing operations such as geometric transformations, downscaling, and upscaling via classic interpolations, as well as more advanced reconstruction-based attacks using stable diffusion models [11]. All attacks are constrained to maintain visual fidelity, ensuring that $\boldsymbol{x}'$ remains perceptually indistinguishable from $\tilde{\boldsymbol{x}}$ to a human observer.

A defense is considered successful if the protection signal remains detectable after the attack, and broken otherwise. This protocol allows us to systematically evaluate robustness under both known and previously unseen image manipulations.

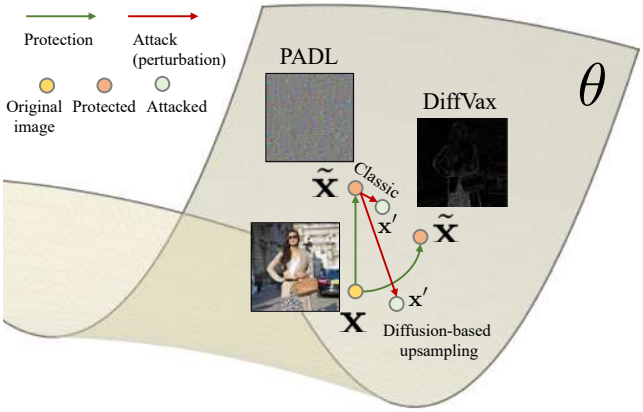


Figure 2: We conceptualize the upsampling based on Diffusion models as projecting an image x into the "manifold" learned by the generative model θ, resulting in x'. Protected images $\tilde{x}$ push it outside of it with different strengths: PADL [6] creates more off-manifold perturbations, while DiffVax [20] produces more on-the-manifold perturbations by creating a border around the mask. Classic attacks sometimes nudge the immunization off, but only with diffusion-based upsampling, which completely reconstructs the image, does the watermark get removed more often.

3.2 Proactive Defenses

In this section, we'll discuss PADL [6] and DiffVax [20], two state-of-the-art proactive protection frameworks. We'll analyse their application in real-life scenarios and highlight potential limitations.

PADL Architecture PADL [6] is a proactive framework for manipulated image detection. In contrast to passive detectors, which are applied only after an image has potentially been altered, PADL operates before any manipulation occurs. Specifically, it enables images to be protected beforehand and later verified, allowing the integrity of the image to be assessed and any subsequent manipulation to be detected. Protection is achieved by embedding an almost imperceptible signal into the image, whose alteration enables manipulation detection. Moreover, unlike previous proactive approaches [3, 4], which relied on a limited set of protections (at most three) and therefore exposed a larger attack surface due to protection reuse, PADL generates image-specific protections. Each image is assigned a unique, tailored protection, significantly reducing the risk of reuse-based attacks. This is achieved through the Encoding module responsible for image protection. The module is based on a transformer architecture, composed of N attention layers, in which a learnable sequence of tokens is customized to the specific input image via a cross-attention mechanism. Once protected, the image can be shared "in the wild". At any later time, its integrity can be verified using the Decoding module, which is responsible for manipulation detection. The latter, also based on a transformer architecture with N repeated attention layers, extracts the image-specific protection and verifies whether it has been altered, thereby assessing the image's integrity. The PADL framework is trained using four loss terms, among which Equations (1) and (2) are the primary objectives. The reconstruction loss in Equation (1) enforces consistency between the protection Δ_e, generated by the Encoding module, and the corresponding protection Δ_d, reconstructed by the Decoding module. The diversity loss in Equation (2) promotes diversity across protections by encouraging those produced from different images within a batch to be mutually orthogonal.

$$\mathcal{L}_{\text{rec}} = 1 - \frac{\Delta_e^\top \Delta_d}{\|\Delta_e\|\|\Delta_d\|}, \tag{1}$$

$$\mathcal{L}_{\text{div}} = \sum_{\substack{i,j=1 \\ i \neq j}}^{B} \max\left(\frac{\Delta_e[i]^\top \Delta_e[j]}{\|\Delta_e[i]\|\|\Delta_e[j]\|}, 0\right). \tag{2}$$

DiffVax Architecture DiffVax [20] is an optimization-free framework for image and video immunization against diffusion-based editing. Different from prior defense techniques, which require solving an optimization problem separately for each target image, DiffVax operates through a learned, single-forward approach. Specifically, it enables images to be protected in the order of milliseconds, allowing for real-time and scalable immunization that effectively generalizes to unseen content. The immunization process is achieved through the Immunizer model $\mathcal{I}_\theta$ that, given an image $\boldsymbol{x}$ and a protection mask $\mathcal{M}$, embeds an imperceptible noise applied to the masked region of the input image, whose presence causes any subsequent diffusion-based editing attempt to fail, resulting in an immunized image $\tilde{\boldsymbol{x}}$. During training, this immunized image is passed to a diffusion-based editing model $\text{SD}_e(\cdot)$, guided by a text prompt $\mathcal{P}$ and the complement of the mask $\tilde{\mathcal{M}}$. The framework is optimized end-to-end by minimizing $\mathcal{L}_{noise}$, defined in Equation (3):

$$\mathcal{L}_{noise} = \frac{1}{\text{sum}(\mathcal{M})}\|(\tilde{\boldsymbol{x}} - \boldsymbol{x}) \odot \mathcal{M}\|, \tag{3}$$

which penalizes the magnitude of the applied noise to preserve visual fidelity, and $\mathcal{L}_{edit}$, given in Equation (4):

$$\mathcal{L}_{edit} = \frac{1}{\text{sum}(\tilde{\mathcal{M}})}\left\|\text{SD}_e\left(\tilde{\boldsymbol{x}}, \tilde{\mathcal{M}}, \mathcal{P}\right) \odot \tilde{\mathcal{M}}\right\|, \tag{4}$$

which penalizes successful edits to ensure disruption. Once trained, the immunizer can be deployed to instantly protect unseen images and videos, neutralizing malicious diffusion-based modifications.

3.3 Attacks

The application of proactive protections, as described in the previous section, is based on the use of additional protective noise applied to the starting image. We test the robustness of the protection algorithm against an attacker that applies different transformations to the protected image to reduce its protection effectiveness. The proposed set of attacks consists solely of black-box attacks and spans from simple noise addition operations to reconstruction attacks. The reason for this choice is to evaluate protection behavior for both common image transformations and diffusion-based image operations, which are nowadays commonly adopted for tasks such as image quality enhancement or upscaling. We conceptualize the difference between these two attacks in Figure 2.

Classical Image Transformations We first evaluate the robustness of protected images against bicubic interpolation upsampling, salt-and-pepper noise, and shear transformations. These operations

were selected for two reasons. First, they represent straightforward, non-optimization-based corruptions that attackers can readily apply without knowledge of the underlying protection mechanism. Second, despite posing genuine threats to widely adopted protection schemes, these transformations remain underexplored in current literature, which predominantly focuses on JPEG compression and Gaussian blur [6, 9]. This gap is notable given that existing protection methods are often vulnerable to these simpler attack vectors.

Diffusion-based Upsampling The second category comprises diffusion-based reconstruction attacks, in which an attacker employs a learned model that performs some sort of image super-resolution, reconstruction or upsampling [19]. We evaluate three upscaling-based variants: *compressed upscaling*, *noisy upscaling*, and *standard upscaling*. All three variants follow the same pipeline: the protected image is passed through a preprocessing step before being fed to a Stable Diffusion ×4 upscaler [23] with generic enhancing quality and negative prompts. Compressed upscaling applies JPEG compression before upscaling and has previously been explored as an adversarial training technique (e.g., in Glaze [26]). Noisy upscaling instead adds Gaussian noise before upscaling, and was proposed by Honig et al. [11] to facilitate watermark removal. Standard upscaling applies the diffusion upscaler directly without any preprocessing. This simpler approach, despite being absent from prior literature, proves equally effective at removing protective signals. We include this straightforward reconstruction attack to demonstrate that current proactive protection paradigms fail even against the most basic diffusion-based threats.

4 Experiments

4.1 Experimental Setting

We evaluate the robustness of PADL- and DiffVax-protected images under the two attack categories mentioned in Section 3.3. For each attack, we measure the PADL decoder's ability to correctly recover the embedded information after the protected image has been corrupted. For DiffVax, we measure the immunization capability against image editing by stable diffusion on attacked protected images.

Classical Image Transformations Classical transformations directly manipulate the protected image without employing learned models. The attack pipeline for PADL follows the Equation (5):

$$\boldsymbol{x} \xrightarrow{\text{PADL enc. } \mathcal{E}_\theta(\boldsymbol{x})} \tilde{\boldsymbol{x}} \xrightarrow{\text{Transformation } \mathcal{T}} \boldsymbol{x'} \xrightarrow{\text{PADL decoder}} \mathcal{D}_\theta(\boldsymbol{x'}), \quad (5)$$

while, for DiffVax follows the Equation (6):

$$\boldsymbol{x} \xrightarrow{\text{DiffVax imm. } \mathcal{I}_\theta(\boldsymbol{x'},\mathcal{M})} \tilde{\boldsymbol{x}} \xrightarrow{\text{Transformation } \mathcal{T}} \boldsymbol{x'} \xrightarrow{\text{SD}_e\left(\boldsymbol{x'},\tilde{\mathcal{M}},\mathcal{P}\right)} \boldsymbol{x'}_{\text{edit}}, \quad (6)$$

where $\text{SD}_e(\cdot)$ denotes the Stable Diffusion image editing model on mask $\tilde{\mathcal{M}}$ with prompt $\mathcal{P}$, and $\boldsymbol{x'}_{\text{edit}}$ is the edited attacked image. We evaluate three transformations with fixed intensity parameters. For *bicubic interpolation upsampling*, we employ weighted averaging to resample the image as ×4 bigger than the original, serving as a non-diffusion upscaling baseline, then we resize it to match the input size of the model. For *salt-and-pepper noise*, we randomly corrupt 5% of pixels by setting them to either minimum or maximum intensity values with equal probability, while for *shear transformation*, we apply a horizontal shear with magnitude 0.1, which displaces pixels horizontally proportional to their vertical position.

Diffusion-based Upsampling Reconstruction attacks leverage diffusion models to reconstruct the image while potentially removing the protective signal. We employ a pre-trained Stable Diffusion ×4 upscaler [23] as the reconstruction mechanism. The attack pipeline for PADL is given in Equation (7):

$$\boldsymbol{x} \xrightarrow{\text{PADL encoder } \mathcal{E}_\theta(\boldsymbol{x})} \tilde{\boldsymbol{x}} \xrightarrow{\mathscr{P}} \boldsymbol{x'} \xrightarrow{\text{SD } \times 4 \text{ upscaler}} \boldsymbol{x'}_{\uparrow} \xrightarrow{\text{PADL decoder}} \mathcal{D}_\theta(\boldsymbol{x'}_{\uparrow}), \quad (7)$$

where $\mathscr{P}$ denotes the optional preprocessing step, $\boldsymbol{x'}_{\uparrow}$ the upscaled image, and $\mathcal{D}_\theta(\cdot)$ represents the PADL decoder output. Similarly, the attack pipeline for DiffVax is given in Equation (8):

$$\boldsymbol{x} \xrightarrow{\text{DiffVax imm. } \mathcal{I}_\theta(\boldsymbol{x},\mathcal{M})} \tilde{\boldsymbol{x}} \xrightarrow{\mathscr{P}} \boldsymbol{x'} \xrightarrow{\text{SD } \times 4 \text{ upscaler}} \boldsymbol{x'}_{\uparrow} \xrightarrow{\text{SD}_e\left(\boldsymbol{x'}_{\uparrow},\tilde{\mathcal{M}},\mathcal{P}\right)} \boldsymbol{x'}_{\uparrow,\text{edit}}. \quad (8)$$

The upscaler performs image reconstruction by first encoding the image into a latent representation, then decoding it at ×4 the original resolution, and finally, we employ image resize to match the input size of the model. We evaluate three variants that differ in their preprocessing step: (i) Compressed upscaling applies JPEG compression at 80% quality before upscaling. The lossy compression removes high-frequency information, potentially weakening the protective signal before reconstruction; (ii) Noisy upscaling adds Gaussian noise with standard deviation $\sigma = 0.01$ before upscaling. The noise injection aims to disrupt the protective signal since it is known from [27] that this noise perturbs off-manifold components where the adversarial noise may reside; (iii) Standard upscaling applies the upscaler directly without preprocessing, so in that case, $\mathscr{P}$ is the identity function. This represents the simplest reconstruction attack, relying solely on the diffusion model's inherent denoising and reconstruction capabilities.

4.2 Robustness Evaluation

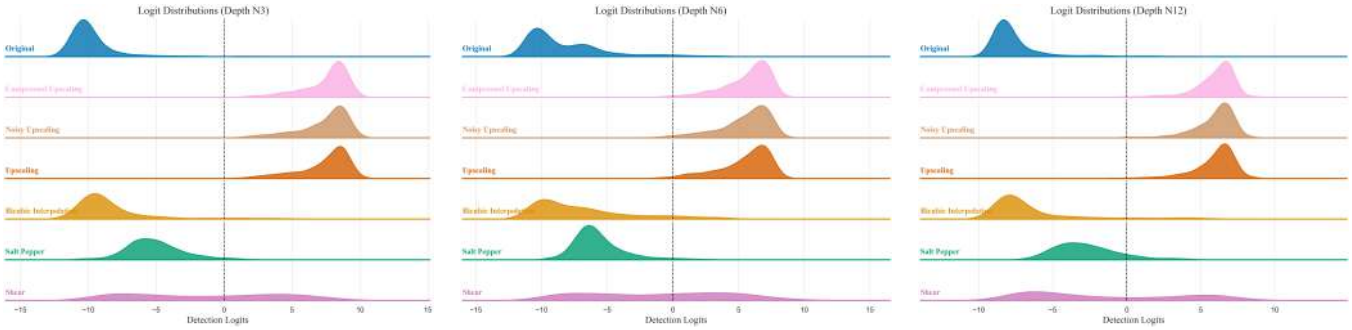
We evaluate the robustness of the PADL model against both classical image transformations and diffusion-based reconstruction attacks. In Table 1, we provide a qualitative comparison of these techniques across datasets. We observe that images reconstructed via Stable Diffusion (SD) upscaler exhibit minimal visible artifacts, thereby maintaining high functional utility for an adversary in a real-world scenario, unlike classical image transformations which often introduce significant visual degradation. We report in Table 2 the Attack Success Rate (ASR) and the False Negative Rate (FNR) of the PADL decoder. Reconstruction attacks yield the highest watermark removal rates, averaging 99.50% for compressed upscaling, 98.94% for noisy upscaling, and 99.22% for pure upscaling. In contrast, geometric shear achieves a moderate ASR of 43.22%, while salt-and-pepper and bicubic interpolation obtain only 5.22% and 6.89% ASR, respectively, indicating the robustness of the PADL watermark against high-frequency noise and classical image upsampling. To objectively assess the trade-off between attack success and image quality, we report the Learned Perceptual Image Patch Similarity [36] (LPIPS), Mean Squared Error (MSE), Peak Signal-to-Noise Ratio (PSNR), and Structural Similarity Index Measure [31] (SSIM) in Table 3. Corroborating the qualitative analysis in Table 1, reconstruction attacks consistently preserve perceptual fidelity, yielding an average LPIPS of 0.06 and SSIM of 0.84. To gain deeper

Table 1: Qualitative comparison of original, PADL [6] encoded, and attacked images across different datasets. The visual examples compare image quality after applying classical image transformations (i.e., Bicubic Interpolation, Salt & Pepper, Shear) and diffusion-based reconstruction attacks (i.e., Compressed Upscaling, Noisy Upscaling, and Upscaling). Images subject to reconstruction attacks exhibit minimal visual degradation and fewer artifacts, effectively maintaining the functional utility and perceptual quality of the original content compared to classical methods.

Dataset	Original	PADL	Compressed Upsc. (Ours)	Noisy Upsc. (Ours)	Upscaling (Ours)	Bicubic Interpolation	Salt & Pepper	Shear
CelebA-HQ [15]								
FFHQ [12]								
GTA2CITY [22]								
Horse2Zebra [38]								
SKETCH2PHOTO [30]								
Summer2Winter [38]								

insight into the decoder's decision process, we analyze the distribution of detection logits across all datasets and attention layers, as illustrated in Figure 3. The decision threshold is set at 0, where negative values indicate a *protected* classification and positive values indicate *unprotected*. The original protected images exhibit a sharp distribution centered well into the negative domain, confirming the decoder's high confidence in detecting the watermark on clean data. The diffusion-based reconstruction attacks induce a drastic shift in the logit distributions toward positive values. This clean separation from the decision boundary corroborates the high values of ASR reported in Table 2, indicating that the reconstruction process effectively neutralizes the watermark signal perceived by the decoder. Conversely, the bicubic upsampling and the salt-and-pepper noise result in a distribution that largely overlaps with the original protected signals, remaining mostly negative, revealing the model's resilience to non-diffusion upsampling and high-frequency noise. The shear attack displays a high-variance distribution that spans the decision boundary, reflecting its partial success rate. Notably, these distributional patterns remain consistent across all protection levels ($N = 3, 6, 12$), suggesting that the susceptibility to reconstruction attacks remains high among all encoding depths.

As described in Section 3.2, the DiffVax model does not detect the presence of a protection watermark through a decoder as in PADL. Instead, it encodes a perturbation into images or video frames through an immunizer model to make them useless on the diffusion-based editing task. Therefore, the evaluation of the DiffVax model is conducted by comparing the editing capability of the SD model on the protected images and the attacked images against classical image transformations and diffusion-based reconstruction attacks. In Table 4, we provide a qualitative comparison of



(a) $N = 3$ **(b)** $N = 6$ **(c)** $N = 12$

Figure 3: Distributions of PADL [6] detection logits for each attack across attention protection layers. Decision threshold is set at 0, negative values represent *protected*, and positive values represent *unprotected*. Reconstruction attacks completely shift the distribution towards positive values for all the depths, indicating watermark removal.

Table 2: Attack Success Rate (ASR) and False Negative Rate (FNR) for each dataset and attack type on PADL [6], across protection levels N3, N6, and N12. The arrows ↑ and ↓ indicate the direction toward a better metric. Bold values indicate the best attack success. Compressed Upscaling, Noisy Upscaling, and Upscaling reveal high watermarking removal efficacy.

		N3		N6		N12	
Dataset	**Attack**	**ASR ↑**	**FNR ↓**	**ASR ↑**	**FNR ↓**	**ASR ↑**	**FNR ↓**
CelebA-HQ [15]	Comp. Upscaling	**1.0000**	0.0000	**1.0000**	0.0050	0.9800	0.0050
	Noisy Upscaling	**1.0000**	0.0000	**1.0000**	0.0050	0.9800	0.0050
	Upscaling	**1.0000**	0.0000	**1.0000**	0.0050	0.9700	0.0050
	Bicubic	0.0000	0.0000	0.0200	0.0050	0.0100	0.0050
	Salt & Pepper	0.0050	0.0000	0.0850	0.0050	0.1400	0.0050
	Shear	0.3600	0.0000	0.3950	0.0050	0.1950	0.0050
FFHQ [12]	Comp. Upscaling	**1.0000**	0.0000	**1.0000**	0.0100	0.9850	0.0000
	Noisy Upscaling	**1.0000**	0.0000	0.9950	0.0100	0.9700	0.0000
	Upscaling	**1.0000**	0.0000	0.9950	0.0100	0.9900	0.0000
	Bicubic	0.0050	0.0000	0.0300	0.0100	0.0200	0.0000
	Salt & Pepper	0.0250	0.0000	0.0550	0.0100	0.1500	0.0000
	Shear	0.4750	0.0000	0.5750	0.0100	0.3950	0.0000
GTA2CITY [22]	Comp. Upscaling	**1.0000**	0.0000	**1.0000**	0.0000	**1.0000**	0.0000
	Noisy Upscaling	**1.0000**	0.0000	**1.0000**	0.0000	**1.0000**	0.0000
	Upscaling	**1.0000**	0.0000	**1.0000**	0.0000	**1.0000**	0.0000
	Bicubic	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
	Salt & Pepper	0.0200	0.0000	0.0000	0.0000	0.0500	0.0000
	Shear	0.1700	0.0000	0.2400	0.0000	0.1100	0.0000
Horse2Zebra [38]	Comp. Upscaling	**1.0000**	0.0500	0.9950	0.0700	**1.0000**	0.0700
	Noisy Upscaling	**1.0000**	0.0500	0.9700	0.0700	**1.0000**	0.0700
	Upscaling	**1.0000**	0.0500	0.9750	0.0700	**1.0000**	0.0700
	Bicubic	0.1700	0.0500	0.1850	0.0700	0.2100	0.0700
	Salt & Pepper	0.0550	0.0500	0.0000	0.0700	0.0700	0.0700
	Shear	0.6150	0.0500	0.6850	0.0700	0.5900	0.0700
SKETCH2PHOTO [30]	Comp. Upscaling	**1.0000**	0.0000	**1.0000**	0.0000	**1.0000**	0.0000
	Noisy Upscaling	**1.0000**	0.0000	**1.0000**	0.0000	**1.0000**	0.0000
	Upscaling	**1.0000**	0.0000	**1.0000**	0.0000	**1.0000**	0.0000
	Bicubic	0.0000	0.0000	0.0050	0.0000	0.0000	0.0000
	Salt & Pepper	0.0150	0.0000	0.0000	0.0000	0.1450	0.0000
	Shear	0.3700	0.0000	0.1950	0.0000	0.3800	0.0000
Summer2Winter [38]	Comp. Upscaling	**1.0000**	0.0350	0.9500	0.1050	**1.0000**	0.0700
	Noisy Upscaling	0.9950	0.0350	0.9050	0.1050	0.9950	0.0700
	Upscaling	0.9950	0.0350	0.9350	0.1050	**1.0000**	0.0700
	Bicubic	0.1250	0.0350	0.2300	0.1050	0.2300	0.0700
	Salt & Pepper	0.0450	0.0350	0.0100	0.1050	0.0900	0.0700
	Shear	0.6950	0.0350	0.7000	0.1050	0.6350	0.0700

the original images, edited original images, edited protected images, and edited attacked images. Similar to what we observed in the previous method, diffusion-based upsampling techniques yield the best results in providing functional yet adversarially effective images, producing meaningful editing results similar to unprotected image editing. Unlike PADL, the edited attacked images of DiffVax reveal less degradation since the attack is computed on protected images, while Table 4 shows the output of the SD editing process to compare it with the SD editing output on non-protected images. In Table 5, we report a quantitative evaluation on the distance between edited unprotected images and attacked images by measuring SSIM, PSNR, Feature Similarity Index Method [35] (FSIM), and the distance between the attacked image and the editing text prompt through CLIP-T [21], where the metric is computed by following Equation (9):

$$\text{CLIP-T} = \max\left[(\text{clip}_{\text{score}} \cdot 100), 0\right]. \tag{9}$$

We also measure similarity metrics on the clean test set for comparison with the DiffVax baseline. From the attacker's point of view, a higher metric of similarity indicates greater attack success, and therefore lower robustness of the DiffVax immunization method. In Table 5, we observe that SSIM, PSNR, and FSIM metrics do not provide meaningful insights into attack efficacy, as most of the image is composed of the masked subject, yet diffusion-based reconstruction attacks achieve higher PSNR and FSIM values, while bicubic interpolation achieves a higher SSIM value. Conversely, the CLIP-T reflects the qualitative evaluation of Table 4, where attacked images reveal significantly higher adherence to the editing text prompt than baseline. In particular, salt-and-pepper noise achieves the highest CLIP-T metric because it suppresses the immunization noise of DiffVax. Indeed, compressed and noisy upscaling achieves a higher CLIP-T value than pure diffusion upscaling and bicubic interpolation upscaling because the preprocessing noising step adds to immunization noise, suppressing it more effectively. It is worth noting that some immunized images exhibit resistance to reconstruction attacks, thereby preventing image editing; therefore, the DiffVax method shows greater robustness than PADL.

4.3 Training PADL with Upscaling Augmentation

As shown in Table 2, PADL is vulnerable to upscaling attacks. To further investigate this limitation, we trained a new PADL model

Table 3: Average LPIPS, MSE, PSNR (dB), and SSIM metrics for each dataset and attack type, among PADL [6] protection levels, between protected images and attacked images. The arrows ↑ and ↓ indicate the direction toward a better metric. Bold and underlined values indicate the best similarity score for diffusion-based attacks and classical perturbation, respectively, where a high similarity score means the adversary does not get a corrupted image. We observe that reconstruction attacks (i.e., Compressed Upscaling, Noisy Upscaling, and Upscaling) show less image degradation than classical image transformation attacks, while degrading more than bicubic upsampling.

Dataset	Attack	LPIPS ↓	MSE ↓	PSNR ↑	SSIM ↑
CelebA-HQ [15]	Comp. Upscaling	**0.0614**	**0.0014**	28.6749	0.8600
	Noisy Upscaling	0.0629	**0.0014**	28.7361	0.8601
	Upscaling	0.0629	**0.0014**	**28.7802**	**0.8633**
	Bicubic	0.0003	0.0000	54.3680	0.9996
	Salt & Pepper	0.5666	0.0167	17.7849	0.4553
	Shear	0.1152	0.0290	15.7119	0.3887
FFHQ [12]	Comp. Upscaling	0.0527	0.0015	28.4037	0.8686
	Noisy Upscaling	0.0465	0.0013	29.3414	0.8882
	Upscaling	**0.0449**	**0.0012**	**29.4774**	**0.8915**
	Bicubic	0.0004	0.0000	53.5882	0.9995
	Salt & Pepper	0.5300	0.0164	17.8610	0.4696
	Shear	0.1211	0.0310	15.3635	0.3890
GTA2CITY [22]	Comp. Upscaling	0.0822	0.0027	26.5588	0.8005
	Noisy Upscaling	**0.0771**	**0.0019**	**27.4603**	0.8317
	Upscaling	0.0786	0.0024	27.4114	**0.8321**
	Bicubic	0.0006	0.0000	51.6065	0.9993
	Salt & Pepper	0.5451	0.0158	18.0168	0.4599
	Shear	0.1079	0.0173	17.9593	0.3863
Horse2Zebra [38]	Comp. Upscaling	0.0755	0.0026	26.4022	0.7886
	Noisy Upscaling	**0.0657**	0.0022	27.4706	0.8301
	Upscaling	0.0663	**0.0021**	**27.5474**	**0.8337**
	Bicubic	0.0007	0.0000	51.5281	0.9993
	Salt & Pepper	0.4382	0.0155	18.1121	0.4672
	Shear	0.1119	0.0217	17.4008	0.3912
SKETCH2PHOTO [30]	Comp. Upscaling	**0.0440**	**0.0022**	26.9771	0.8632
	Noisy Upscaling	0.0497	**0.0022**	27.2387	0.8671
	Upscaling	0.0482	**0.0022**	**27.4484**	**0.8718**
	Bicubic	0.0005	0.0000	52.0701	0.9995
	Salt & Pepper	0.5109	0.0168	17.7660	0.5344
	Shear	0.1126	0.0304	15.2627	0.3170
Summer2Winter [38]	Comp. Upscaling	0.0732	0.0032	25.7956	0.8001
	Noisy Upscaling	0.0671	0.0027	26.5543	0.8321
	Upscaling	**0.0660**	**0.0026**	**26.6874**	**0.8376**
	Bicubic	0.0008	0.0000	50.3230	0.9992
	Salt & Pepper	0.4532	0.0166	17.8137	0.4907
	Shear	0.1261	0.0239	16.9993	0.3675

from scratch with $N = 6$. During training, images were subjected to an upscaling attack with a probability of 10%, while all other training settings were kept identical to those of the original model. In particular, the parameter α, which controls the strength of the perturbation, was left unchanged. This exposure was intended to improve the model's robustness to upscaling-based tampering; however, it proved insufficient. The newly trained model still exhibits a 100% ASR across all datasets, indicating that it remains susceptible to upscaling attacks. This limitation is inherent to proactive approaches in general, as these methods rely entirely on the presence of an embedded signature. The protection is implemented as low-magnitude, near-imperceptible noise that can be effectively removed by simple operations, such as upscaling followed by downsampling, allowing an attacker to strip the embedded protection from images. Fine-tuning the PADL training parameters may improve robustness to upsampling attacks, but doing so would require increasing α, which in turn would amplify the perturbation and lead to greater visual degradation of the protected images, or increasing the number of times we apply the upscaling perturbation during training, thereby increasing training complexity.

In addition, training the model with a larger α and deploying it at inference with a lower one could appear as an alternative strategy to preserve visual quality. However, this approach is not effective in this scenario. Increasing α during training biases the model toward detecting stronger perturbations. As a consequence, when evaluated on samples protected with a smaller α, the protection becomes too weak relative to what the model has learned to detect, leading to a significant drop in detection performance. This mismatch between training and inference yields suboptimal performance without addressing the vulnerability to upscaling attacks.

5 Conclusions

In this work, we evaluate the robustness of state-of-the-art models for proactive watermarking in a black-box model-agnostic setting via diffusion-based upsampling and classical image transformations. The experimental results show that while the models exhibit robustness against unseen image transformations, diffusion-based attacks achieve nearly perfect success in watermark removal, effectively shifting the detection logits from protected to unprotected across all datasets and for any attention layer of PADL and compromising the image editing immunization of DiffVax. Moreover, we evaluate the degradation of images both qualitatively and quantitatively; however, reconstruction attacks via SD achieve the minimum image corruption, posing a serious threat in real-world case scenarios. Furthermore, computing adversarial training in PADL with reconstruction attacks does not improve the robustness of the model for the same threshold of image perturbation. Future studies should focus on developing new approaches for watermarking multimedia content, considering hybrid approaches to mitigate diffusion-based upsampling attacks.

Acknowledgments

Giuseppe Daidone was supported by an ACN (Italian National Cybersecurity Agency) Ph.D. fellowship.

References

[1] Kasra Arabi, R. Teal Witter, Chinmay Hegde, and Niv Cohen. 2025. SEAL: Semantic Aware Image Watermarking. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. 16196–16205.

[2] Toluwani Aremu, Noor Hussein, Munachiso Nwadike, Samuele Poppi, Jie Zhang, Karthik Nandakumar, Neil Gong, and Nils Lukas. 2025. Mitigating Watermark Forgery in Generative Models via Randomized Key Selection. arXiv:2507.07871 [cs.CR] https://arxiv.org/abs/2507.07871

Table 4: Qualitative comparison of original, edited original, edited protected with DiffVax [20], and edited attacked images. Diffusion-based reconstruction attacks (i.e., Compressed Upscaling, Noisy Upscaling, Upscaling), Bicubic Interpolation, and Salt & Pepper noise effectively remove the immunization, allowing meaningful image editing with SD, while Shear degrades the quality of edited images significantly.

Original	Edited Original	Edited Protected	Compressed Upsc. (Ours)	Noisy Upsc. (Ours)	Upscaling (Ours)	Bicubic Interpolation	Salt & Pepper	Shear

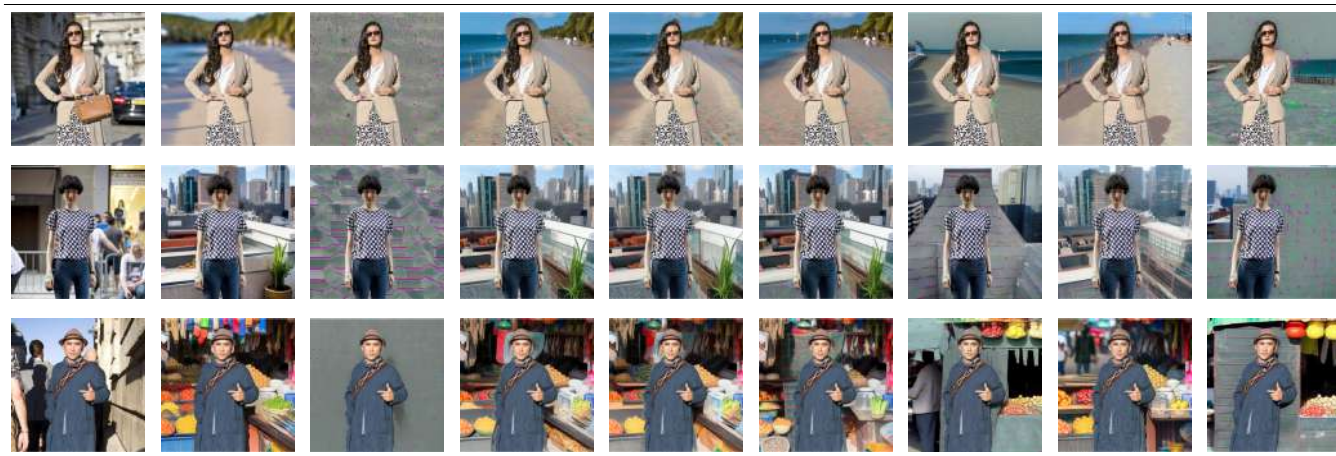


Table 5: Average SSIM, PSNR (dB), FSIM, and CLIP-T metrics evaluated on DiffVax [20] between edited images and edited protected images. The arrows ↑ and ↓ indicate the direction toward a better metric. Bold values indicate the highest similarity scores, where higher similarity indicates less effective protection by DiffVax. All attack methods significantly increase the CLIP-T score, indicating greater adherence to the editing prompt and thereby decreasing the effectiveness of the DiffVax immunization.

Attack	SSIM ↑	PSNR ↑	FSIM ↑	CLIP-T ↑
DiffVax [20] (baseline)	0.5218	14.2432	0.7134	17.9231
Compressed Upscaling	0.4933	14.8247	0.7220	22.0568
Noisy Upscaling	0.4940	**15.0524**	0.7290	22.4584
Upscaling	0.4896	14.7792	**0.7295**	21.6338
Bicubic	**0.5345**	13.9038	0.7067	21.6118
Salt & Pepper	0.2973	14.1111	0.6756	**25.8850**
Shear	0.2096	10.4120	0.5117	23.0509

[3] Vishal Asnani, Xi Yin, Tal Hassner, Sijia Liu, and Xiaoming Liu. 2022. Proactive Image Manipulation Detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR).* 15386–15395.

[4] Vishal Asnani, Xi Yin, Tal Hassner, and Xiaoming Liu. 2023. Malp: Manipulation localization using a proactive scheme. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR).* 12343–12352.

[5] Vishal Asnani, Xi Yin, and Xiaoming Liu. 2024. Proactive Schemes: A Survey of Adversarial Attacks for Social Good. arXiv:2409.16491 [cs.CV] https://arxiv.org/abs/2409.16491

[6] Filippo Bartolucci, Iacopo Masi, and Giuseppe Lisanti. 2025. Perturb, Attend, Detect, and Localize (PADL): Robust Proactive Image Defense. *IEEE Access* 13 (2025), 81755–81768. doi:10.1109/ACCESS.2025.3565824

[7] Jingyi Deng, Chenhao Lin, Zhengyu Zhao, Shuai Liu, Zhe Peng, Qian Wang, and Chao Shen. 2025. A Survey of Defenses Against AI-Generated Visual Media: Detection, Disruption, and Authentication. *ACM Comput. Surv.* 58, 5, Article 123 (Nov. 2025), 35 pages. doi:10.1145/3770916

[8] Ian J Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative adversarial nets. In *NeurIPS.*

[9] Xiao Guo, Xiaohong Liu, Zhiyuan Ren, Steven Grosz, Iacopo Masi, and Xiaoming Liu. 2023. Hierarchical fine-grained image forgery detection and localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR).*

[10] Xiaoxiao Hu, Jiaqi Jin, Sheng Li, Wanli Peng, Xinpeng Zhang, and Zhenxing Qian. 2026. Spherical Watermark: Encryption-Free, Lossless Watermarking for Diffusion Models. In *The Fourteenth International Conference on Learning Representations.*

[11] Robert Hönig, Javier Rando, Nicholas Carlini, and Florian Tramèr. 2025. Adversarial Perturbations Cannot Reliably Protect Artists From Generative AI. In *The Thirteenth International Conference on Learning Representations (ICLR).*

[12] Tero Karras, Samuli Laine, and Timo Aila. 2019. A Style-Based Generator Architecture for Generative Adversarial Networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR).*

[13] Pavel Korshunov and Sébastien Marcel. 2018. Deepfakes: a new threat to face recognition? assessment and detection. *arXiv preprint arXiv:1812.08685* (2018).

[14] Yihao Liu, Jinhe Huang, Yanjie Li, Dong Wang, and Bin Xiao. 2024. Generative AI model privacy: a survey. *Artificial Intelligence Review* 58, 1 (04 Dec 2024), 33. doi:10.1007/s10462-024-11024-6

[15] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. 2015. Deep Learning Face Attributes in the Wild. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV).*

[16] Ling Lo, Cheng Yu Yeo, Hong-Han Shuai, and Wen-Huang Cheng. 2024. Distraction is All You Need: Memory-Efficient Image Immunization against Diffusion-Based Image Editing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR).* 24462–24471. doi:10.1109/CVPR52733.2024.02309

[17] Shayne Longpre, Nikhil Singh, Manuel Cherep, Kushagra Tiwary, Joanna Materzynska, William Brannon, Robert Mahari, Naana Obeng-Marnu, Manan Dey, Mohammed Hamdy, Nayan Saxena, Ahmad Mustafa Anis, Emad A. Alghamdi, Vu Minh Chien, Da Yin, Kun Qian, Yizhi LI, Minnie Liang, An Dinh, Shrestha Mohanty, Deividas Mataciunas, Tobin South, Jianguo Zhang, Ariel N. Lee, Campbell S. Lund, Christopher Klamm, Damien Sileo, Diganta Misra, Enrico Shippole, Kevin Klyman, Lester James Validad Miranda, Niklas Muennighoff, Seonghyeon Ye, Seungone Kim, Vipul Gupta, Vivek Sharma, Xuhui Zhou, Caiming Xiong, Luis Villa, Stella Biderman, Alex Pentland, Sara Hooker, and Jad Kabbara. 2025. Bridging the Data Provenance Gap Across Text, Speech, and Video. In *The Thirteenth International Conference on Learning Representations (ICLR).* https://openreview.net/forum?id=G5DziesYxL

[18] Aqib Mahfuz, Georgia Channing, Mark van der Wilk, Philip Torr, Fabio Pizzati, and Christian Schroeder de Witt. 2025. PSyDUCK: Training-Free Steganography for Latent Diffusion. arXiv:2501.19172 [cs.LG] https://arxiv.org/abs/2501.19172

[19] Aamir Mustafa, Salman H Khan, Munawar Hayat, Jianbing Shen, and Ling Shao. 2019. Image super-resolution as a defense against adversarial attacks. *IEEE Transactions on Image Processing* 29 (2019), 1711–1724.

[20] Tarik Can Ozden, Ozgur Kara, Oguzhan Akcin, Kerem Zaman, Shashank Srivastava, Sandeep P Chinchali, and James M Rehg. 2026. DiffVax: Optimization-Free Image Immunization Against Diffusion-Based Editing. In *The Fourteenth International Conference on Learning Representations (ICLR)*.

[21] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning Transferable Visual Models From Natural Language Supervision. In *Proceedings of the 38th International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 139)*, Marina Meila and Tong Zhang (Eds.). PMLR, 8748–8763. https://proceedings.mlr.press/v139/radford21a.html

[22] Stephan R. Richter, Vibhav Vineet, Stefan Roth, and Vladlen Koltun. 2016. Playing for Data: Ground Truth from Computer Games. In *Computer Vision – ECCV 2016*, Bastian Leibe, Jiri Matas, Nicu Sebe, and Max Welling (Eds.). Springer International Publishing, Cham, 102–118.

[23] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2022. High-Resolution Image Synthesis With Latent Diffusion Models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 10684–10695.

[24] Hadi Salman, Alaa Khaddaj, Guillaume Leclerc, Andrew Ilyas, and Aleksander Madry. 2023. Raising the Cost of Malicious AI-Powered Image Editing. In *Proceedings of the 40th International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 202)*. PMLR, 29894–29918. https://proceedings.mlr.press/v202/salman23a.html

[25] Tom Sander, Pierre Fernandez, Alain Oliviero Durmus, Teddy Furon, and Matthijs Douze. 2025. Watermark Anything With Localized Messages. In *The Thirteenth International Conference on Learning Representations (ICLR)*. https://openreview.net/forum?id=IkZVDzdC8M

[26] Shawn Shan, Jenna Cryan, Emily Wenger, Haitao Zheng, Rana Hanocka, and Ben Y Zhao. 2023. Glaze: Protecting artists from style mimicry by {Text-to-Image} models. In *32nd USENIX Security Symposium (USENIX Security 23)*. 2187–2204.

[27] Yang Song and Stefano Ermon. 2019. Generative modeling by estimating gradients of the data distribution. In *NeurIPS*.

[28] Matteo Trippodo, Federico Becattini, and Lorenzo Seidenari. 2025. Immunizing Images from Text to Image Editing via Adversarial Cross-Attention. In *ACMMM* (Dublin, Ireland) *(MM '25)*. Association for Computing Machinery, New York, NY, USA, 10535–10543. doi:10.1145/3746027.3755755

[29] Liqin Wang, Qianyue Hu, Wei Lu, and Xiangyang Luo. 2026. Safeguarding Facial Identity against Diffusion-based Face Swapping via Cascading Pathway Disruption. arXiv:2601.14738 [cs.CV] https://arxiv.org/abs/2601.14738

[30] Xiaogang Wang and Xiaoou Tang. 2009. Face Photo-Sketch Synthesis and Recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 31, 11 (2009), 1955–1967. doi:10.1109/TPAMI.2008.222

[31] Zhou Wang, A.C. Bovik, H.R. Sheikh, and E.P. Simoncelli. 2004. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing* 13, 4 (2004), 600–612. doi:10.1109/TIP.2003.819861

[32] Yuxin Wen, John Kirchenbauer, Jonas Geiping, and Tom Goldstein. 2023. Tree-Ring Watermarks: Fingerprints for Diffusion Images that are Invisible and Robust. In *NeurIPS*. https://par.nsf.gov/biblio/10522361

[33] Zijin Yang, Kai Zeng, Kejiang Chen, Han Fang, Weiming Zhang, and Nenghai Yu. 2024. Gaussian Shading: Provable Performance-Lossless Image Watermarking for Diffusion Models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 12162–12171.

[34] Jie Zhang, Shuai Dong, Shiguang Shan, and Xilin Chen. 2025. Dual Attention Guided Defense Against Malicious Edits. arXiv:2512.14333 [cs.CV] https://arxiv.org/abs/2512.14333

[35] Lin Zhang, Lei Zhang, Xuanqin Mou, and David Zhang. 2011. FSIM: A Feature Similarity Index for Image Quality Assessment. *IEEE Transactions on Image Processing* 20, 8 (2011), 2378–2386. doi:10.1109/TIP.2011.2109730

[36] Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. 2018. The Unreasonable Effectiveness of Deep Features as a Perceptual Metric. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

[37] Xuandong Zhao, Sam Gunn, Miranda Christ, Jaiden Fairoze, Andres Fabrega, Nicholas Carlini, Sanjam Garg, Sanghyun Hong, Milad Nasr, Florian Tramer, Somesh Jha, Lei Li, Yu-Xiang Wang, and Dawn Song. 2025. SoK: Watermarking for AI-Generated Content. In *IEEE Symposium on Security and Privacy (SP)*. 2621–2639. doi:10.1109/SP61157.2025.00178

[38] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A. Efros. 2017. Unpaired Image-To-Image Translation Using Cycle-Consistent Adversarial Networks. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*.