Enhancing Scene Realism through Neural Radiance Fields and Monocular Depth Estimation

Giorgio De Magistris[1], Juan David Rodriguez[1] and Christian Napoli[1]

[1]Department of Computer, Control and Management Engineering, Sapienza University of Rome

Abstract

This paper addresses the challenge of differentiable rendering, focusing on a novel implementation designed to integrate 3D objects seamlessly into reconstructed 3D environments, thereby creating entirely new perspectives of the scene. Our methodology leverages Neural Radiance Field (NeRF) models to reconstruct the 3D environments with high fidelity, alongside monocular depth estimation algorithms for deriving the 3D characteristics of objects from single images. The main goal of our approach lies in harmonizing the depth map output from the NeRF model with the depth data of the inserted object. This synergy enables the accurate and space-coherent placement of the object within the scene, ensuring a natural integration that enhances the overall realism of the virtual environment.

Keywords

Differentiable rendering, NeRF model, 3D reconstruction, depth map

1. Introduction

This work bridges two pivotal areas of computer vision: view synthesis and 3D reconstruction, each targeting distinct goals yet sharing a fundamental connection. View synthesis, an image-based rendering technique, creates new scenes from various images and perspectives. Conversely, 3D reconstruction aspires to model real-world scenes in three dimensions, crafting geometric representations from visual data.

In our approach, we amalgamate elements from both view synthesis and 3D reconstruction to fabricate entirely new visual perspectives of environments, incorporating 3D objects previously absent from these scenes. Achieving this level of realism and visual coherence necessitates the use of 3D reconstruction methods to ascertain spatial details such as the distance and depth of scene elements. This integration allows for synthesized images that authentically mirror the spatial dynamics and geometry of the newly added objects within their respective environments.

Recent years have underscored the significance of 3D reconstruction from 2D imagery within computer vision, propelled by its vast application potential and foundational ties to 3D perception—endeavoring to endow systems with a nuanced understanding of scene compositions. Amidst various methodologies aimed at enhancing efficiency, structure-from-motion and multi-view stereo techniques have gained prominence. These methods excel in spatial object organization and camera positioning through the exploitation of visual and motion cues, offering a cost-effective yet accurate 3D modeling approach by producing sparse point clouds from image correspondences.

Despite these advantages, the advent of deep learning has pivoted the focus towards neural network-based solutions for 3D reconstruction, exemplified by the application of convolutional neural networks (CNNs) in stereo reconstruction. These learning-based strategies integrate global semantic insights for improved matching accuracy [1, 2, 3, 4, 5, 6, 7].

View synthesis techniques, ranging from depth-based rendering to texture mapping, utilize existing data like depth or disparity maps to forge new visual perspectives through pixel reconfiguration and amalgamation.

This paper specifically delves into a cutting-edge view synthesis method—Neural Radiance Fields (NeRFs)—building on prior advancements to train a neural network in mapping a 5D vector (comprising position and orientation) to the emitted radiance at that location. This technique eschews convolutions for a deep, fully-connected network learning a regression from the 5D input to RGB color and volume density, facilitating rendering via traditional volume rendering techniques.

A notable aspect of NeRFs is their ability to deduce 3D geometric details, such as depth, from the model's learned representation, positioning NeRF at the core of our proposed solution that adeptly leverages both view synthesis and 3D reconstruction.

However, the NeRF model's tendency to simply memorize scene radiance poses challenges for scene editing or manipulation—core objectives of this paper. We propose a novel solution integrating monocular depth estima-

ICYRIME 2023: 8th International Conference of Yearly Reports on Informatics, Mathematics, and Engineering. Naples, July 28-31, 2023

demagistris@diag.uniroma1.it (G. D. Magistris);
rodriguezgomez@diag.uniroma1.it (J. D. Rodriguez);
cnapoli@diag.uniroma1.it (C. Napoli)

tion to render a 3D object representation from a single image, enabling precise object integration into the NeRF-derived 3D scene map. This synthesis not only enhances the model's editing capabilities but also marks a significant advancement in combining view synthesis with 3D reconstruction for dynamic scene generation.

2. Related work

The convergence of view synthesis and 3D reconstruction in computer vision has been the subject of extensive research, driven by their potential to revolutionize how machines perceive and interact with their environments. This section reviews the foundational works and recent advancements in these fields, setting the stage for our proposed methodology.

Differentiable rendering has emerged as a critical bridge between 3D models and their 2D projections, allowing gradients of the image loss to be backpropagated through the rendering process to the model parameters. [8] and [9] have laid the groundwork in this domain, proposing frameworks that enable optimization over mesh vertices and textures. However, the combinatorial nature of meshes makes it difficult to directly optimize the geometry from multi-view images.

Neural Radiance Fields (NeRFs), introduced in [10], have significantly advanced view synthesis by learning a continuous volumetric scene function from a sparse set of images. Subsequent research has expanded on NeRFs to improve training efficiency [11], inference speed [12] [13] and reducing constrains [14]. We extend the NeRF methodology to incorporate 3D objects into the scenes, leveraging its depth inference capabilities for realistic scene reconstruction.

The field of 3D reconstruction has evolved from geometry-based techniques to deep learning approaches. Early works on structure-from-motion [15] and Multi-View Stereo [16] laid the foundation for understanding scene geometry from image sequences. More recently, CNN-based methods have demonstrated superior performance in extracting 3D information from 2D images, with notable contributions from [17] and [18] in applying deep learning for spatial understanding. Our approach synergizes with these advancements, utilizing deep learning for enhanced depth estimation and scene reconstruction. Monocular depth estimation has seen rapid progress, transitioning from traditional methods reliant on hand-crafted features to learning-based approaches that utilize neural networks for depth prediction from a single image. The authors of [19] initially explored the potential of using supervised learning for this task, while more recent efforts by [20] and [21] have introduced self-supervised and semi-supervised techniques, achieving remarkable accuracy. Our model integrates the more advanced monocular depth estimation method introduced in [22] to facilitate the seamless embedding of 3D objects into NeRF-generated scenes.

Combining View Synthesis and 3D Reconstruction: While several studies have independently explored view synthesis and 3D reconstruction, few have investigated their integration for enhanced scene rendering and object insertion. Our work is inspired by the pioneering efforts in both fields, aiming to create a cohesive framework that leverages the strengths of each to produce photorealistic and spatially coherent scene augmentations.

In summary, our proposed method stands at the intersection of multiple research domains, drawing from and contributing to a rich body of knowledge on differentiable rendering, NeRFs, 3D reconstruction, and monocular depth estimation. By synthesizing these technologies, we aspire to advance the capabilities of computer vision systems in understanding and manipulating complex 3D environments.

3. Implementation

The proposed approach was tested with two scenarios with different levels of illumination, to study how this aspect affects the final representation of the environment. The pictures for the first scenario were taken manually with a cellphone camera. In total, we collected 40 pictures with size 652x367 following the protocol described in [23], where the maximum disparity between views was no more than about 64 pixels. The images for the second scenario instead are taken from [24], which provides 41 images with size 504x378 pixels showcasing an office. A sample from both scenarios is illustrated in Figure 1.

While NeRF can be initialized with the poses extracted from COLMAP [15] we opted for a fully data-driven pipeline, removing the requirement of both poses and camera parameters as shown in [25]. The images of the object to embed in the two scenarios are taken from [24], which provided multiples views of the object as well as segmentation masks. However the monocular depth estimation model [26] inputs a single image and returns the estimated depth map, hence the multi-view dataset is used only to choose among different poses of the embedding object. Figure 2 shows one of the views of the object.

3.1. Model Pipeline

The model pipeline for this task is divided into two main components: NeRF model training for scenario representation and image embedding for adding 3D objects into these scenarios. Initially, depth maps are generated to describe the scenes in 3D, a crucial step for seamlessly integrating new objects in a spatially coherent manner.

Figure 1: The two test scenarios. Scenario A (on top): a room with low illumination (pictures taken by author), and Scenario B (bottom): an office with better light (pictures taken from [24]).

Detailed explanations of each component in the pipeline follow, covering the specifics of both the NeRF training process and the methodology for embedding objects into the scenes.

3.1.1. NeRF model and training

The model basically optimizes a NeRF architecture network, and the camera parameters (intrinsic and extrinsic) of the images of a scenario, by minimizing photometric reconstruction errors [25].

In general terms, a 3D sampling process is done for each pixel, where camera rays are traced through the scene to collect a set of samples at (x, d) locations (position and view direction). These tuples are then used at each sample as input to the NeRF model to generate a continuous function that outputs an RGB color and its corresponding density.

During the training process, a random selection of pixels is rendered per input image. The purpose is to minimize the reconstruction loss by comparing the rendered colors of these pixels with the corresponding ground-truth colors. It is important to note that the complete

Figure 2: Object to embed in the two scenarios.

pipeline is fully differentiable, allowing for the simultaneous optimization of NeRF and camera parameters.

In order to obtain fast training and comply with computational power limitations, it was decided to implement a tinyNeRF model which simplifies the original NeRF structure presented in [24], using just 4 layers within the multi-layer perceptron (MLP) architecture (instead of 8), also discarding the skip recurrent connection, due to the shallowness of the structure. The input and output of the model remained the same, a single continuous 5D coordinate (the spatial location (x, y, z) and viewing direction (theta, phi) as inputs, and the volume density and view-dependent emitted radiance at that spatial location as outputs.

The hyper-parameters used in the training process are presented in the table 1.

Table 1
NeRF parameters

Parameter	Value	Comments
`# epochs`	5000	
`Optimizer`	Adam	
`Learning rate`	0.001	with scheduler
`Loss`	MSE	

Initially, the number of epochs was set to a few hundred, but the synthesized results indicated the need for a larger number of epochs. This was necessary to achieve not only improved output image quality, avoiding blurriness, but also a more accurate and well-defined depth map. Consequently, the final number of epochs used for the training process reached approximately 5k.

3.1.2. Image embedding

The embedding process started with the characterization of the object to integrate into the scenario. For this purpose, the image of the object (with no background) was used as input to the monocular depth estimation algorithm [22]. This technique estimated the distance to the camera for each pixel in the RGB image, giving as a result a depth map represented as a matrix. On the other hand the NeRF model does not provide directly a depth map, but a density field, i.e. a scalar density for each point in space. This information however can be used to approximate depth as follows: 1) First the surface is extracted as a level set of the density function for a fixed threshold, i.e. the surface is represented implicitly as $\sigma(x) = \tau$ where τ is found empirically. 2) Then the Depth Map is computed through Ray Casting. For the given camera pose a ray $\mathbf{r}(t) = \mathbf{o} + t\mathbf{d}$ is shot from each pixel and sampled. The first value t such that $\mathbf{r}(t) \geq \tau$ is considered the depth value for the current pixel.

Once we have the two images with the corresponding depth maps we simply fuse the images according to the depth values, such that each pixel in the final image takes the value of the image with the smallest depth. To adjust the scale of the embedding object relative to the scene we can simply scale the width and height of the corresponding image and depth map. Moreover we can place the object at different distances simply scaling the depth map.

4. Results

The results of the NeRF training are shown in Figure 3 while the depth maps extracted with different values for the threshold τ are shown in Figure 4.

Figure 3: Novel views obtained after a few hundred epochs (left column) and after 5k epochs (right column).

The results show that with just 40 front-faced pictures of the selected scenarios, the reconstruction and the depth map achieve a realistic outcome for each of the novel view.

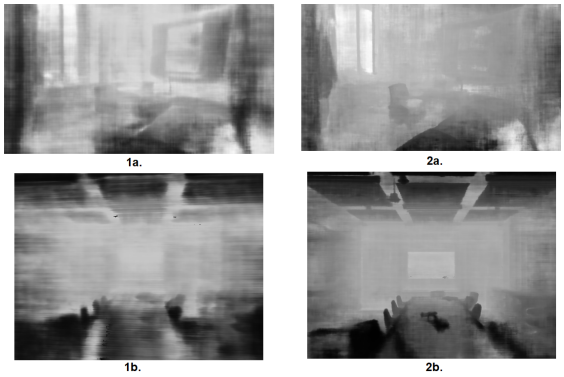

Figure 4: Depth maps obtained using different thresholds τ

Figure 5 shows the 3D depth map of the embedding object, obtained from the monocular depth estimation algorithm.

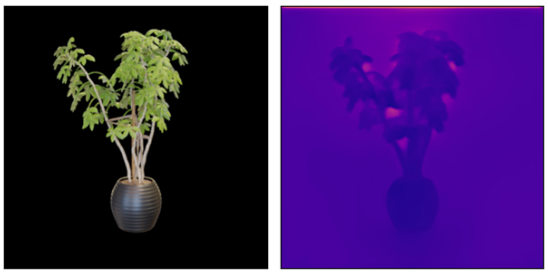

Figure 5: Object depth map.

As described in Section 3, a crucial step in the image embedding process was the normalization procedure. Its purpose was to establish consistency between the depth measurements of the object and the surrounding environment. In particular, the depth values of the object were scaled in the interval [0, 20]. This adjustment ensures that the embedding object is ten times smaller than the maximum length of the rooms (whose depth values range from 0 to 200), aligning appropriately with the nature of the item selected. This choice maintains coherence within the scenarios and ensures space-consistent proportions.

The result of the merging is shown in figures 6 and 9.

Some statistics of the images of the scene and the embedding object are shown in Table 2.

Upon initial observation, it became evident that the monocular depth estimation model faced challenges in generating high-quality predictions near the edges of the object image showing some noise around the edges of the

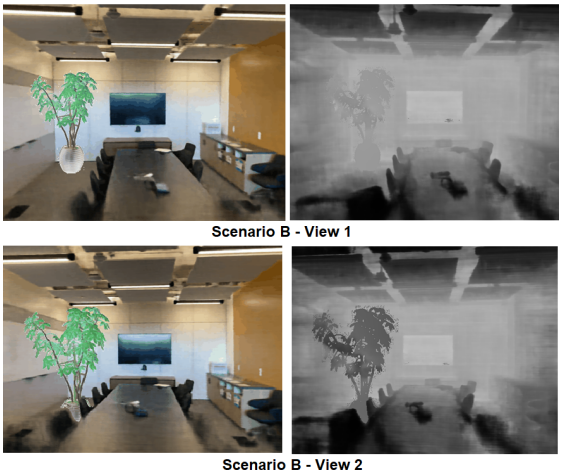

Figure 6: Scenario B novel views examples, with 3D object embedded.

Table 2
Parameters in novel views scenario B figure 6.

Parameter	Value
View 1	
Room img size	504x378 px
Object img size	225x225 px
Object x shift	50 px
Object y shift	120 px
Object closer depth	100
View 2	
Room img size	504x378 px
Object img size	300x300 px
Object x shift	0 px
Object y shift	100 px
Object closer depth	0.3

embedding object(see Figure 7 A). However, to mitigate this phenomenon, a simple Nearest Neighbour filter was applied to the final image (see Figure 7B).

Figure 7: Resulting image before (A) and after (B) filtering

Figure 8 shows in more detail the positive result of the merging process between the reconstructed space and the 3D object. In particular, our approaches correctly represent occlusion.

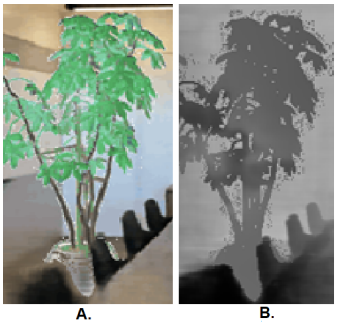

Figure 8: Some details of the synthetic image

Table 2 shows the parameters use to generate the final merged novel image.

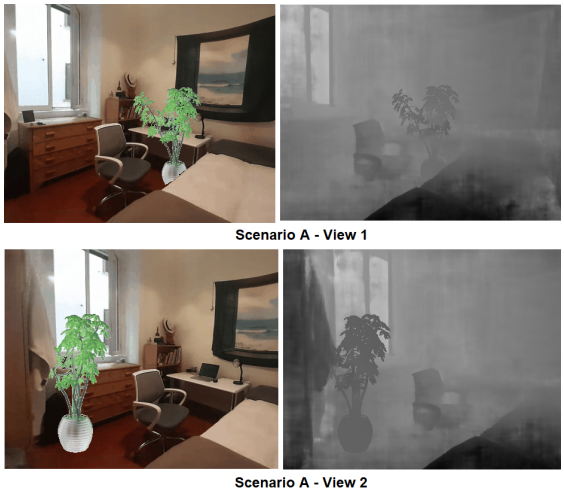

Figure 9: Scenario B novel views examples, with 3D object embedded.

5. Conclusions

This study presents a novel approach to scene manipulation using neural radiance fields, demonstrating promising results in integrating objects into scenes with realistic depth and color. Key to success is the precise depth characterization by the NeRF model and the streamlined optimization of camera parameters within the NeRF architecture, eliminating the need for external processing and enhancing novel view rendering. However, achieving truly realistic embeddings demands careful adjustment of object positioning and meticulous pre-processing to

Table 3
Parameters in novel views scenario A figure 9.

Parameter	Value
View 1	
Room img size	652x367 px
Image size	220x220 px
Image size	220x220 px
Object x shift	240 px
Object y shift	120 px
Object closer depth	100
View 2	
Room img size	652x367 px
Object img size	300x300 px
Object x shift	10 px
Object y shift	110 px
Object closer depth	90

ensure seamless integration, highlighting the importance of coherent spatial logic and image adjustments for authentic synthesis.

Acknowledgments

This work has been developed at is.Lab() Intelligent Systems Laboratory at the Department of Computer, Control, and Management Engineering, Sapienza University of Rome (https:// islab.diag.uniroma1.it). The work has also been partially supported from Italian Ministerial grant PRIN 2022 "ISIDE: Intelligent Systems for Infrastructural Diagnosis in smart-concretE", n. 2022S88WAY - CUP B53D2301318, and by the Age-It: Ageing Well in an ageing society project, task 9.4.1 work package 4 spoke 9, within topic 8 extended partnership 8, under the National Recovery and Resilience Plan (PNRR), Mission 4 Component 2 Investment 1.3—Call for tender No. 1557 of 11/10/2022 of Italian Ministry of University and Research funded by the European Union—NextGenerationEU, CUP B53C22004090006.

References

[1] G. De Magistris, E. Iacobelli, R. Brociek, C. Napoli, An automatic cnn-based face mask detection algorithm tested during the covid-19 pandemics, volume 3398, 2022, pp. 36 – 41.

[2] D. Połap, M. Woźniak, C. Napoli, E. Tramontana, R. Damaševičius, Is the colony of ants able to recognize graphic objects?, Communications in Computer and Information Science 538 (2015) 376 – 387. doi:10.1007/978-3-319-24770-0_33.

[3] N. Brandizzi, S. Russo, G. Galati, C. Napoli, Addressing vehicle sharing through behavioral analysis: A solution to user clustering using recency-frequency-monetary and vehicle relocation based on neighborhood splits, Information (Switzerland) 13 (2022). doi:10.3390/info13110511.

[4] S. Pepe, S. Tedeschi, N. Brandizzi, S. Russo, L. Iocchi, C. Napoli, Human attention assessment using a machine learning approach with gan-based data augmentation technique trained using a custom dataset, OBM Neurobiology 6 (2022). doi:10.21926/obm.neurobiol.2204139.

[5] A. Alfarano, G. De Magistris, L. Mongelli, S. Russo, J. Starczewski, C. Napoli, A novel convmixer transformer based architecture for violent behavior detection 14126 LNAI (2023) 3 – 16. doi:10.1007/978-3-031-42508-0_1.

[6] G. De Magistris, M. Romano, J. Starczewski, C. Napoli, A novel dwt-based encoder for human pose estimation, volume 3360, 2022, pp. 33 – 40.

[7] M. Woźniak, D. Połap, M. Gabryel, R. K. Nowicki, C. Napoli, E. Tramontana, Can we process 2d images using artificial bee colony?, volume 9119, 2015, pp. 660 – 671. doi:10.1007/978-3-319-19324-3_59.

[8] M. M. Loper, M. J. Black, Opendr: An approximate differentiable renderer, in: Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part VII 13, Springer, 2014, pp. 154–169.

[9] S. Liu, T. Li, W. Chen, H. Li, Soft rasterizer: A differentiable renderer for image-based 3d reasoning, in: Proceedings of the IEEE/CVF international conference on computer vision, 2019, pp. 7708–7717.

[10] B. Mildenhall, P. P. Srinivasan, M. Tancik, J. T. Barron, R. Ramamoorthi, R. Ng, Nerf: Representing scenes as neural radiance fields for view synthesis, Communications of the ACM 65 (2021) 99–106.

[11] T. Müller, A. Evans, C. Schied, A. Keller, Instant neural graphics primitives with a multiresolution hash encoding, ACM transactions on graphics (TOG) 41 (2022) 1–15.

[12] S. J. Garbin, M. Kowalski, M. Johnson, J. Shotton, J. Valentin, Fastnerf: High-fidelity neural rendering at 200fps, in: Proceedings of the IEEE/CVF international conference on computer vision, 2021, pp. 14346–14355.

[13] C. Reiser, S. Peng, Y. Liao, A. Geiger, Kilonerf: Speeding up neural radiance fields with thousands of tiny mlps, in: Proceedings of the IEEE/CVF international conference on computer vision, 2021, pp. 14335–14345.

[14] R. Martin-Brualla, N. Radwan, M. S. Sajjadi, J. T. Barron, A. Dosovitskiy, D. Duckworth, Nerf in the wild: Neural radiance fields for unconstrained photo collections, in: Proceedings of the IEEE/CVF

Conference on Computer Vision and Pattern Recognition, 2021, pp. 7210–7219.
[15] J. L. Schönberger, J.-M. Frahm, Structure-from-motion revisited, in: Conference on Computer Vision and Pattern Recognition (CVPR), 2016.
[16] J. L. Schönberger, E. Zheng, M. Pollefeys, J.-M. Frahm, Pixelwise view selection for unstructured multi-view stereo, in: European Conference on Computer Vision (ECCV), 2016.
[17] A. Newell, K. Yang, J. Deng, Stacked hourglass networks for human pose estimation, in: Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part VIII 14, Springer, 2016, pp. 483–499.
[18] X.-F. Han, H. Laga, M. Bennamoun, Image-based 3d object reconstruction: State-of-the-art and trends in the deep learning era, IEEE transactions on pattern analysis and machine intelligence 43 (2019) 1578–1604.
[19] A. Saxena, S. H. Chung, A. Y. Ng, 3-d depth reconstruction from a single still image, International journal of computer vision 76 (2008) 53–69.
[20] C. Godard, O. Mac Aodha, G. J. Brostow, Unsupervised monocular depth estimation with left-right consistency, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, pp. 270–279.
[21] N. Wang, Y. Zhang, Z. Li, Y. Fu, W. Liu, Y.-G. Jiang, Pixel2mesh: Generating 3d mesh models from single rgb images, in: Proceedings of the European conference on computer vision (ECCV), 2018, pp. 52–67.
[22] D. Kim, W. Ka, P. Ahn, D. Joo, S. Chun, J. Kim, Global-local path networks for monocular depth estimation with vertical cutdepth, arXiv preprint arXiv:2201.07436 (2022).
[23] B. Mildenhall, P. P. Srinivasan, R. Ortiz-Cayon, N. K. Kalantari, R. Ramamoorthi, R. Ng, A. Kar, Local light field fusion: Practical view synthesis with prescriptive sampling guidelines, 2019. arXiv:1905.00889.
[24] B. Mildenhall, P. P. Srinivasan, M. Tancik, J. T. Barron, R. Ramamoorthi, R. Ng, Nerf: Representing scenes as neural radiance fields for view synthesis, 2020. arXiv:2003.08934.
[25] Z. Wang, S. Wu, W. Xie, M. Chen, V. A. Prisacariu, Nerf–: Neural radiance fields without known camera parameters, 2022. arXiv:2102.07064.
[26] D. Kim, W. Ka, P. Ahn, D. Joo, S. Chun, J. Kim, Global-local path networks for monocular depth estimation with vertical cutdepth, 2022. arXiv:2201.07436.