This ICCV workshop paper is the Open Access version, provided by the Computer Vision Foundation. Except for this watermark, it is identical to the accepted version; the final published version of the proceedings is available on IEEE Xplore.

Targeted Adversarial Attacks on Generalizable Neural Radiance Fields

András Horváth
Pázmány Péter Catholic University
Faculty of Information Technology and Bionics
horvath.andras@itk.ppke.hu

Csaba M. Józsa
Nokia Bell Labs
csaba.jozsa@nokia-bell-labs.com

Abstract

Neural Radiance Fields (NeRFs) have recently emerged as a powerful tool for 3D scene representation and rendering. These data-driven models can learn to synthesize high-quality images from sparse 2D observations, enabling realistic and interactive scene reconstructions. However, the growing usage of NeRFs in critical applications such as augmented reality, robotics, and virtual environments could be threatened by adversarial attacks.

In this paper we present how generalizable NeRFs can be attacked by both low-intensity adversarial attacks and adversarial patches, where the later could be robust enough to be used in real world applications. We also demonstrate targeted attacks, where a specific, predefined output scene is generated by these attack with success.

1. Introduction

Neural Radiance Fields (NeRFs) [1] have emerged as a groundbreaking paradigm in the domain of 3D scene representation and rendering, revolutionizing the way we perceive and interact with virtual environments. NeRFs leverage the power of deep learning to capture intricate scene details [2], enabling the synthesis of photorealistic images from sparse 2D observations [3]. The ability to reconstruct high-quality scenes from limited input data has propelled NeRFs into the forefront of computer vision, computer graphics, augmented reality [4], robotics [5], and other related fields.

NeRFs represent 3D scenes as continuous functions, mapping 3D coordinates to their corresponding scene appearance properties, such as color and opacity. This continuous representation distinguishes them from most traditional 3D models, which often rely on discrete voxels or point clouds. In essence, NeRFs can be seen as implicit functions that define the scene's surface, depth and appearance properties, making them particularly suited for complex and detailed scene reconstruction. They can generate depth maps [6] and can be used in navigation [7], [8], localization [5] and six degrees of freedom orientation estimation [9].

The significance of Neural Radiance Fields (NeRFs) lies in their widespread applicability and apart from image rendering, in generating 3D scenes, depth maps, and aiding navigation. However, it is essential to acknowledge that the susceptibility of NeRFs to adversarial attacks can introduce complications and challenges. These attacks could have the potential to produce unrealistic maps and representations, leading to the hallucination of non-existent objects within the scene or the omission of existing objects. As a result, in various applications employing NeRFs, these adversarial perturbations may give rise to erroneous outcomes and hinder accurate scene reconstruction and navigation.

The training process of NeRFs involves capturing multi-view image observations of the scene and optimizing the model to predict accurate color and opacity values for any novel viewing angle within the scene's spatial extent. This approach enables NeRFs to not only render novel viewpoints but also handle dynamic scenes and incorporate additional observations over time. Consequently, NeRFs have opened up exciting possibilities for applications like real-time virtual reality experiences, interactive architectural visualizations [10], and advanced autonomous robotic systems [7].

As NeRFs find increasing adoption in real-world applications, concerns surrounding their vulnerability to adversarial attacks have surfaced. Adversarial attacks aim to exploit vulnerabilities in machine learning models by introducing carefully crafted perturbations to the input data. These perturbations are imperceptible to the human eye but can lead to drastic misclassifications or erroneous predictions.

In their conventional configurations, NeRFs are trained in a scene-specific and object-specific manner, involving the training of a dedicated neural network for each scene. The neural network's weights store the scene-specific representations and knowledge of views and camera angles. While these networks could potentially be vulnerable to attacks during the training process, exploiting data poisoning [11]

or backdoor attacks [12], resulting in the production of invalid three-dimensional representations, their lack of generality limits the potential issues in real-world applications. As a consequence, the specialized nature of NeRF training offers a degree of protection against such adversarial perturbations in practical scenarios.

As research on NeRFs has progressed, recent advancements have led to the development of Generalizable Neural Radiance Fields (GeNeRFs) [13]. These extensions go beyond the original NeRF formulation where scene specific models had to be trained. The capabilities of these models can encompass both the generation of novel views and the creation of implicit three-dimensional representations using known previous views and camera poses. Due to the general nature of these methods, there arises a suspicion that they might be susceptible to attacks through perturbations of input pixels in the images. Such attacks could potentially enable the creation of scenes with arbitrary objects.

In this paper, we aim to substantiate this hypothesis by providing a demonstration of the vulnerability of these models to adversarial perturbations on one of the most commonly used GeNeRF variant: IBRNet [14], showcasing the potential for generating scenes with arbitrary objects through these attacks.

Adversarial attacks can take various forms within the context of NeRFs, including attacks on the embedded 3D representation, the weights of the trained models or the input pixels. Attacking input pixels is relatively easy, and this method remains the most significant form of attack as it does not require access to the image processing pipeline, making it a potential real-world threat. Consequently, this study focuses on this form of attack by employing targeted attack strategies involving both low-intensity attacks [15] covering all input pixels and patch based attacks [16] being limited to only a certain region of the image.

Attack strategies can also be distinguished based on the expected output of the attacks. In case of untargeted attacks our aim is to modify the output of the network as much as possible, without any restrictions on the output scene of the model. Meanwhile in case of targeted attack a predefined output scene has to be generated by the model as the result of the attack.

In [17] untargeted attacks have been introduced using GeNeRFs. The attack methodology and results are interesting, but untargeted attacks do not pose a substantial real-world threat, as the resulting outputs are often easily detected as non-realistic images, hence their unrestrictedness.

In contrast, this research delves into targeted attacks, wherein the objective is to create realistic scenes featuring unreal objects on the rendered images, while the generated depth map were not investigated in the current work. Given the importance of rendered images as the most commonly investigated element, the study specifically focuses on attacking this aspect. By exploring the vulnerability of NeRFs to targeted attacks on the rendered image, our research aims to shed light on potential security risks and the extent of their impact on NeRF-based systems. This investigation is expected to provide valuable insights into safeguarding NeRFs against adversarial threats and further enhancing their reliability and practicality in various real-world applications.

In this paper, we embark on a comprehensive exploration of adversarial attacks on NeRFs. We investigate the efficacy of different attack strategies and evaluate their impact on the rendering quality, scene reconstruction accuracy, and generalization capabilities of NeRFs.

Our paper is structured the following way: in section 2 we briefly describe Generalized Neural Radiance Fields, in section 3 we introduce the most commonly applied adversarial attack methodologies and algorithms, in section 4 we describe our experiments and results and in section 5 we draw conclusion from them.

2. Generalizable NeRFs

NeRFs present a cutting-edge approach in leveraging deep neural networks to generate 3D representations of objects or scenes from 2D images. This innovative technique involves encoding the complete object or scene within an artificial neural network, which then predicts the light intensity, also known as radiance, at any specific point in the 2D image. As a result, NeRFs enable the creation of novel 3D views from various angles, revolutionizing the generation of highly realistic 3D objects automatically.

The exceptional potential of NeRFs lies in their capacity to represent 3D data more efficiently compared to other existing methods. This efficiency opens new avenues for generating highly realistic 3D objects with remarkable promise. Moreover, when combined with complementary techniques, NeRFs offer the exciting prospect of significantly compressing 3D representations of the world, reducing data sizes from gigabytes to mere tens of megabytes [18]. Such advancements hold significant implications for various fields, enabling streamlined and versatile 3D data generation and manipulation.

GeNeRF variants like [19],[20], [21], [14] enable cross-scene generalization via two modifications on top of traditional NeRFs: Firstly, these variants condition NeRFs on the source views of new scenes. This involves utilizing a limited number of observed source views from a new scene to extract features via a Convolutional Neural Network (CNN) encoder. These features are then used as scene priors and fed into mostly feed-forward neural networks combined with transformer architectures. Secondly, the variants incorporate a ray transformer, which operates on all points along the same ray, enhancing the density prediction.

The most common steps implemented by the GeNeRF models can be summarized as follows: 2D feature maps $\{W_i\}_{i=1}^S$ are inferred for a total of S source views $\{I_i\}_{i=1}^S$ using a pretrained CNN encoder E, where $W_i = E(I_i)$ represents a 3D tensor. (Notably, this process requires only a one-time effort for each new scene.) A ray $r(t) = o + td$ is emitted from the origin o of the virtual camera along the view direction d to pass through the pixel to be rendered. 3D points x_k are sampled along the ray based on an ordered depth sequence t_k drawn from a certain distribution. Each sampled 3D point x_k is projected onto the image planes of source views using a project transformation π, obtaining the corresponding scene features $W_i(\pi(x_k))$ for all S source views. The scene features acquired in the previous step are applied to an MLP model f to derive the color c_k and volumetric density σ_k for each point. Compared to scene specific NeRF models, instead of directly predicting volumetric densities σ_k some architectures implement a two-step process where density features f_k^σ are predicted, and the final volumetric density prediction is determined by a transformer architecture T having as input all the f_k^σ vectors of every sample. Occlusion aware volume rendering is performed in the final step by taking into account the relative viewing directions or predicting visibility probabilities. During training, the networks E, f, and T are updated using the Mean Squared Error (MSE) loss or other pixel-based distance metrics, ensuring effective learning of the rendering process.

From the various variants of GeNeRFs we have selected IBRNet [14] for our investigations, which is commonly applied and highly cited variant, capable of rendering state of the art images from new views on novel scenes. Since the whole rendering pipeline is differentiable, pixels or parts of the source images can be modified according to the planned adversarial attacks. We have used a pretrained model, which was trained on multiple datasets simultaneously (LLF [22], RealEstate 10k [23], Google Scanned Objects [24], etc.) to be able to cope with generic scenes. For the sake of reproducibility, the same pretrained model and data for training and evaluation are available at the following link [1]

There are more recent implementations and variants of GeNeRFs, such as [25], which apply geometric constraints to be more efficient, or [13] where even hardware constraints were considered, but these approaches do not differ significantly from the model of our selection, therefor we believe that the attacks presented here can be generalized for these variants as well.

GeNeRFs represent a highly promising real-world solution for novel view synthesis, owing to their remarkable ability to generalize across different scenes, facilitating instant rendering on previously unseen environments. Despite the critical significance of adversarial robustness in practical applications, limited attention has been given to exploring its implications specifically for GeNeRF. We postulate that GeNeRF's conditioning on source views from new scenes, often sourced from the Internet or third-party providers, may introduce novel security concerns in real-world scenarios. Additionally, the conventional understanding and solutions for achieving adversarial robustness in neural networks may not directly apply to GeNeRFs, given its distinctive 3D nature and diverse operations.

3. Adversarial attacks

The concept of adversarial attacks originated from the pioneering work of [26]. It brought to light a crucial revelation about deep neural networks. Despite their ability to generalize effectively and perform well on conventional input data and even on similar inputs, they possess a vulnerability to exploitation by malicious agents. This vulnerability stems from the high-dimensional nature of inputs, enabling the generation of non-realistic input samples that generate outputs, which deviate drastically from human judgment and the expected outcomes.

The initial adversarial attacks proposed by Goodfellow et al. [15] involved calculating the sign of the gradient of the cost function (J) with respect to the input ($\boldsymbol{x}$) and expected output (y), which was then scaled by a constant (ϵ) to control the intensity of the noise. This method, known as the Fast Gradient Sign Method (FGSM), allowed for rapid generation of attacks.

Rozsa et al. [27] extended FGSM by utilizing not just the sign of the raw gradient but also a scaled version of the gradient's magnitude, termed the Fast Gradient Value method.

Furthermore, Dong et al. [28] proposed an iterative version of FGSM that incorporated momentum into the equation. The inclusion of momentum was inspired by the concept of optimization during model training, with the goal of avoiding poor local minima and non-convex patterns in the objective function's landscape.

Moosavi et al. [29] approached adversarial attacks from the perspective of binary classifier robustness. They formulated the idea that a binary classifier's robustness at a given point $\boldsymbol{x}_0$ is determined by its distance from the separating hyperplane $\Delta(\boldsymbol{x}_0; f)$. They derived a closed-form formula to calculate the smallest perturbation required to change the classifier's output and applied these perturbations iteratively to the image until the classifier's decision changed. This approach was later extended to address multiclass classification problems as well.

While these methods were crucial for theoretical understanding, their application to neural networks in practical, real-world applications has limited significance due to their low-intensity, constrained noise application. In real-world

[1]. https://drive.google.com/drive/folders/1qfcPffMy8-rmZjbapLAtdrKwg3AV-NJe

scenarios, even the smallest perturbations, such as those arising from environmental factors like perspective, illumination changes, or lens distortion, can completely disrupt the desired results. Therefore, the utilization of these attacks in practical applications is not feasible [30].

In [16], [31] robust and real-world attacks were presented against various classification networks. These methods create an adversarial patch, where instead of the global, but low-intensity approaches, distortions appear in a region with limited area, but intensity values are not bounded[2]. Successful attacks with adversarial patches were also demonstrated using black and white patches only [32], where not the intensities of the patch, but the locations and sizes of the stickers are optimized. These attacks, where the gradients of the networks are not necessarily used during optimization open space towards black-box attacks [33], [34], where the attacker needs access only to the final responses, confidence values to generate attacks using evolutionary algorithms.

A general overview of adversarial attacks, containing a more detailed description of most of the previously mentioned methods can be found in the following survey paper [35]. The resilience of segmentation networks against adversarial attacks was investigated heavily in the past years [36], [37], [38], [39].

Subsequent years witnessed extensive investigations into the potential of exploiting adversarial attacks. Researchers developed novel attack strategies to enhance the robustness of generated attacks [16], [31], even enabling black-box attacks, which do not require access to the network gradients [32], [33], [34].

Moreover, advancements were made in extending adversarial attacks to more complex tasks beyond classification, such as detection and localization problems [40]. These innovative techniques were applied to diverse network architectures, including Faster-RCNN [41].

According to our best knowledge adversarial attacks has not be presented and investigated in GeNeRF models apart from [17], which is restricted to low-intensity and untargeted attacks.

4. Method and Results

Our proposed method for the adversarial attack on GeNeRFs is shown in Fig. 1. We have selected a pretrained model of the IBRNet as a GeNeRF. For a certain pose and source images we created a new rendered image. We manually placed a hallucinated object on the rendered image. The resulting image serves as the adversarial ground truth image. The residual loss is always computed between the adversarial ground truth image and the currently rendered image at the same pose. This is an important regularizer because ignoring the gradients coming from the non-attacked regions might significantly deteriorate the image quality in these parts.

Since these modifications were manual we have to admit the they can be biased in two ways. On one hand they might disturb the real structure of the images (artificial insertion and deletion might cause extremely strong edges in the image), on the other hand the modifications are subjective and other people might desire different modifications. We would argue that this subjectivity is unavoidable and we were carefully generating three different types of modifications:

- types where the shape of existing objects are modified.
- types where existing images were deleted from scenes and substituted by background pixels
- types where new objects were added to the scenes

A few samples of these modifications and the result of attacks using these image as desired outputs can be seen in Fig. 2.

4.1. Low-intensity Attacks

For our investigation into low-intensity attacks, we opted for the iterative version of FGSM with momentum [28] as the attack mechanism. Our setup involved 1000 iterations, with parameter ϵ set to 0.01.

In a typical low-intensity attack on classification problems, a single input image is used, allowing modifications to all its pixels until a predefined threshold is reached. However, since GeNeRFs utilize multiple input images, referred to as source images or source views, attackers can simultaneously modify all or a subset of these images. To explore the impact of different attack scenarios, we devised five setups with varying numbers of source images: 10, 8, 6, 5, and 4. The quality of the generated image depends on the number of source images, generally improving with an increase in this number. Our investigation covered cases where one, two, three, and so on, up to all source images were subject to modification.

This investigation holds significance as it addresses real-world scenarios where images from events are uploaded to a common dataset by users or multiple autonomous robots. In such cases, understanding the necessary number and percentage of images to be attacked for successful modifications in the rendered output image becomes crucial. This way Our research aims could provide valuable insights into enhancing the security and reliability of GeNeRFs in various practical applications.

The quality of the attack was measured as the average ℓ_2 distance between the generated image and our hand-modified ground truth image. We have executed this experiment on ten different scenes, repeating each attack ten

[2] apart from the global bounds of image values

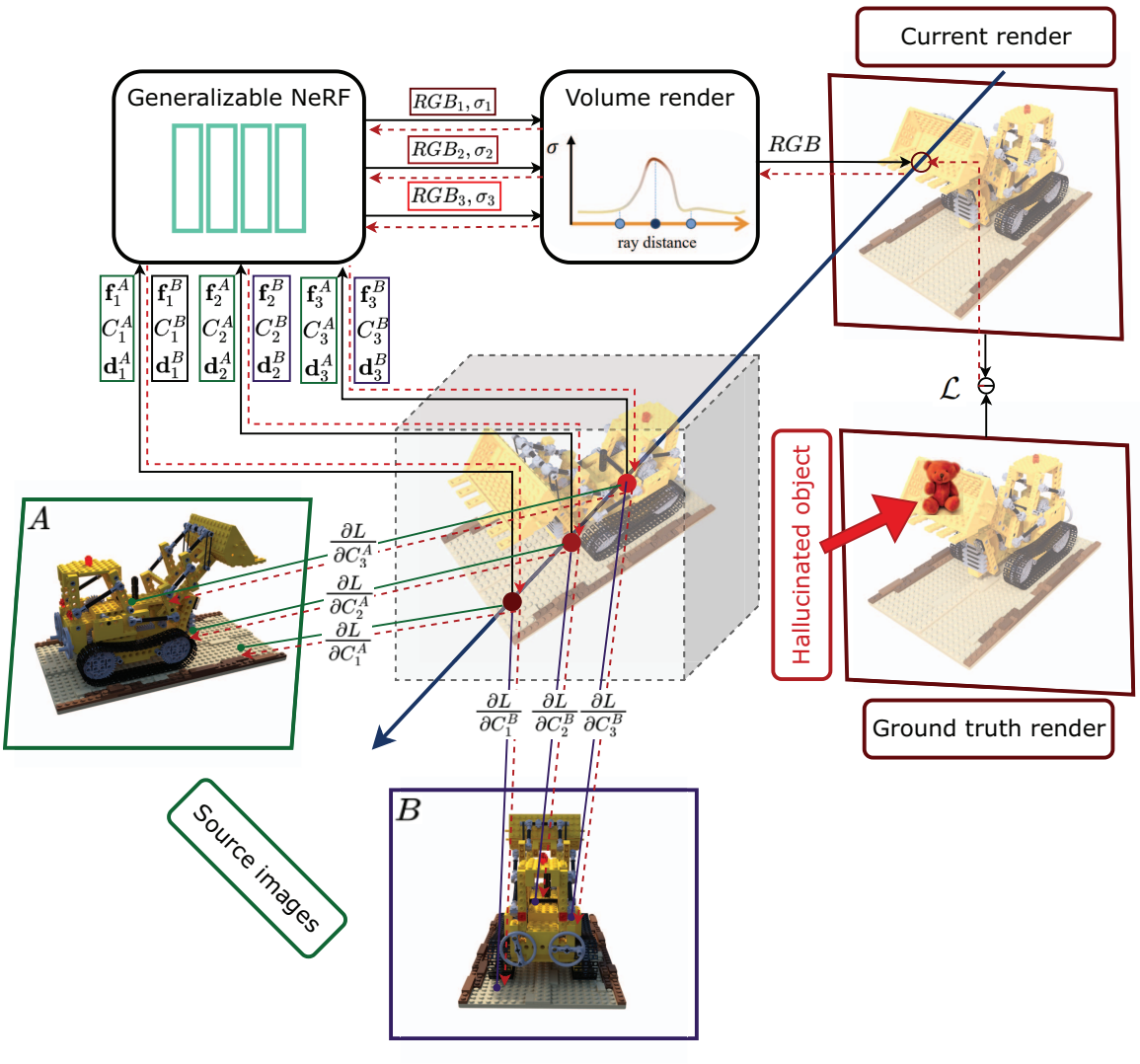

Figure 1: Adversarial attack on generalizable NeRFs. A ground truth render is created for the adversarial pose and a hallucinated object is randomly placed on the rendered image. Since all the components of the pipeline are fully differentiable, the goal is to modify the source images such that the resulting render will be close to the attacked ground truth render. The back-propagated gradients of the residual image are used to alter the source images within certain bounds.

times (to average out the stochastic nature of the attack algorithm) and the quantitative summary of the results can be seen in Fig. 3.

These results clearly indicate that attacks were successful in most cases when a significant majority of the source views were targeted. In our setup, an attack can be considered successful when the average pixel distance dropped below 0.015, while unsuccessful attacks resulted in values above 0.020. It is important to note that these threshold values may vary depending on the scene, but as observed in Fig. 2, scenes with only a small region altered in the image can be used as a rule of thumb. Additional results containing PSNR, SSIM and LPIPS measures can be found in table 1.

These findings highlight the overall robustness of GeNeRFs, as the generated images remained reliable in cases where the majority of the source images were left untouched. However, the study also underscores the vulnerability of the system when an attacker gains access to most of the source images, enabling arbitrary modifications to the output. Understanding and addressing these security implications are crucial as GeNeRFs and similar technologies advance, ensuring their safe and reliable application in various practical scenarios.

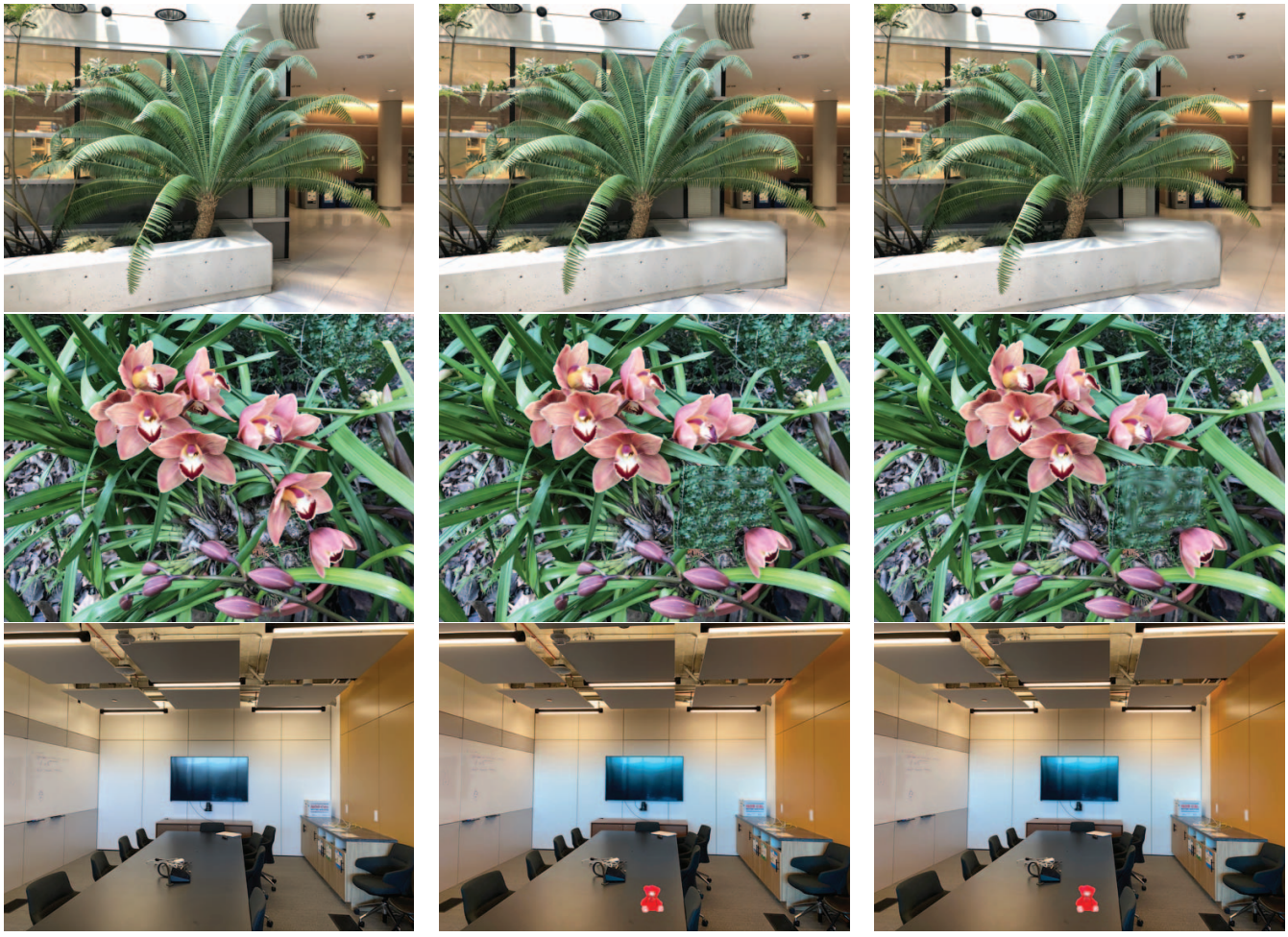

Figure 2: Samples cases from evaluation part of the LLFF dataset. Here we display three different samples, one in each row. The first column contains the original output images of the network without any attacks, the second column contains the modified images which were used as ground truth during the attacks. These modifications were done manually. The third column contains the output images of the network after the attack. In this setup the images were generated using ten different views and attacks were applied on all input images. The attacks were generated using FGSM for 1000 iterations with and ϵ value of 0.01. As these images demonstrate adversarial attacks were successful and we were able to modify objects in the scene (fern), delete objects from the scene (orchids) and render non-existing objects in the scene (room).

4.2. Patch-based attacked

Low-intensity attacks may hold academic interest, but their significance diminishes when considering real-world applications, primarily due to the limited access attackers have to the image processing pipeline. However, the most straightforward and practical way to target neural networks is by modifying the real environment itself. In such scenarios, attackers can manipulate small regions within the image while freely altering the pixel values in this designated region. To effectively simulate and study these real-world threats, we have focused our investigation on patch-based attacks.

Patch-based attacks provide a suitable framework to understand the vulnerabilities of neural networks in the face of real-world adversarial manipulations. By restricting our attention to specific regions in the image, we emulate the scenario where an attacker can locally modify the environment while leaving the rest of the scene intact. The arbitrary nature of pixel values within these patches allows us to evaluate the robustness of the neural networks against unpredictable and potentially damaging alterations.

For low-intensity attacks, the algorithm's crucial parameter is the ϵ value, intended to ensure the challenging detectability of these modifications. Similarly the size of the patch applied is the most crucial parameter in patch-based attacks, akin to the significance of the amount of maximal change in low-intensity attacks. To examine the impact of patch size on these attacks, we employed the same set of

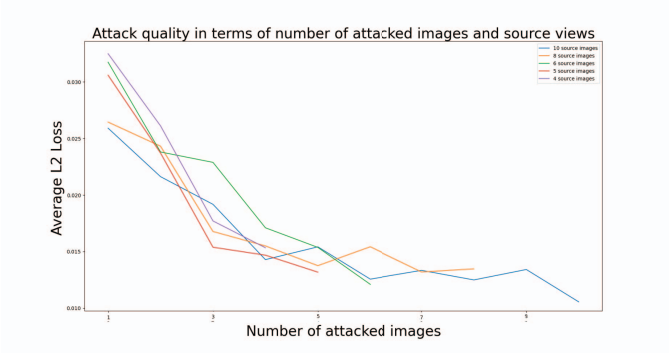

Figure 3: This plot depicts the dependence of attack quality on the number of source views and the number of attacked samples in case of GeNeRFs. The Y axes plots the average ℓ_2 distance between the pixels of the ground truth image and the image generated by the network after the attack. Lower values mean the attack was more successful, since this case the network output was closer to our desired output. The X axes contains the number of attacked images, meanwhile the different colored plots depicts outputs generated from different number of source views. As it can be seen from these results attacks are not successful (they generate a larger distance) until the number of attacked views will not reach the majority of the source views. Each point in these measurements were generated as the average of 10 independent runs and on ten different scenes.

10 scenes previously generated. For each scene, desired attack outputs were manually specified, and patches were automatically placed at the center of the images. This approach ensured that the patches were not closely positioned to the regions already modified.

Clearly, a patch covering the modified region could influence the outcome, especially when applied near or at the boundary of the effect. However, the most critical scenario to consider is when patches have far-reaching effects, altering pixels that are not in close proximity to them and keeping the original output value of other regions.

Our experimental investigations involved generating patches of sizes 2×2, 5×5, 10×10, and 20×20, and then assessing their respective effects on the scenes. The results of these experiments are illustrated in Fig. 4, providing valuable insights into the relationship between patch size and the success of patch-based attacks. Additional results containing PSNR, SSIM and LPIPS measures can be found in table 1.

The results clearly demonstrate the feasibility of patch-based attacks when the patch size is sufficiently large (typically 10×10 patches in our experiments) and when the patches are prevalent in the majority of images. In our investigations, utilizing ten source views, attacks were generally successful if at least four of them contained a patch large enough to cause significant impact.

It is essential to highlight that in this experiment, the

Attack	L2 ↓	PSNR ↑	SSIM ↑	LPIPS ↓
Low (2/10)	0.022	19.75	0.537	0.242
Low (4/10)	0.016	21.83	0.841	0.168
Low (10/10)	0.011	24.72	0.910	0.163
Patch (2/10)	0.022	18.43	0.588	0.256
Patch (4/10)	0.017	21.60	0.792	0.173
Patch (10/10)	0.011	24.33	0.903	0.114

Table 1: Within this table, one can find evaluations of attack methodologies showcased across diverse attributes. The rows labeled as "Low" exhibit outcomes of low-intensity adversarial attacks, whereas those designated as "Patch" reveal findings from attacks rooted in patch patch based attack (with size 20x20). Each entry corresponds to L2, PSNR, SSIM, and LPIPS metrics (in different columns accordingly). These analyses involve 10 source views (indicated by the second value in brackets following the attack methods) while the quantity of attacked images is denoted by the first figure (2, 4, or 10). These measurements were calculated on the average of 10 independent runs and conducted on ten different scenes.

patches were independently optimized for each source image. Consequently, the pixel values at the same location could differ across different images, enabling the attacker to tailor their patches specifically to exploit the vulnerabilities in each individual source view.

These findings underscore the potential threat posed by patch-based attacks and emphasize the importance of developing robust defenses against such manipulation techniques. Understanding the adaptability of these attacks to various scenarios is crucial for strengthening the security of neural network systems in real-world applications.

These preliminary results demonstrate the feasibility of patch-based attacks on GeNeRFs. However, our simulations deviate from real-world setups in the following aspects:

- The patches are consistently positioned at the center of images, and their locations remain unchanged regardless of the viewpoint.
- The pixel values within the patches are optimized independently from each other and can vary across different input images.

Addressing these differences in the future is essential to simulate scenarios where an attacker introduces a real object into a scene. Despite these limitations, we are optimistic that this research paves the way for real-world adversarial applications, such as stickers on GeNeRFs.

5. Conclusion

We have demonstrated targeted adversarial attacks on GeNeRF, revealing important insights into the security vul-

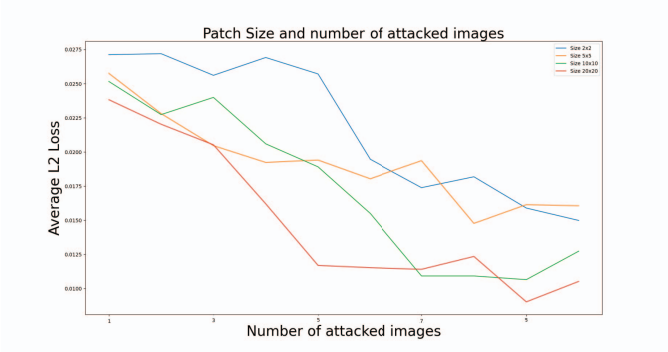

Figure 4: This plot illustrates how the attack quality is influenced by the number of source views and the size of the patch applied during the attack. The Y-axis represents the average ℓ_2 distance between the pixels of the ground truth image and the image generated by the network after the attack. Lower values indicate more successful attacks, as they result in the network output being closer to our desired output. On the other hand, the X-axis represents the number of attacked images, while the various colored plots depict outputs generated from different patch sizes. The results demonstrate that attacks are not successful (they generate a larger distance) until the number of attacked views encompasses the majority of the source views or when the patch size is too small. These measurements are based on the average of 10 independent runs and conducted on ten different scenes.

nerabilities of these networks. The success of the attacks, utilizing methods commonly employed in classification tasks, emphasizes the ease with which malevolent attackers can manipulate the generated images. However, our findings also demonstrate the relative robustness of NeRFs when multiple views are utilized and not all source images are accessible to the attacker. In such cases, the effectiveness of the attack diminishes, indicating the importance of safeguarding access to critical source images. In cases where the attacker has access to the majority of the views the quality of the attacks increases significantly.

Additionally, we explored patch-based attacks, where limited regions of the image are targeted, but arbitrary values can be introduced. Remarkably, these attacks are not restricted to local neighborhoods, as even distant regions can be manipulated with such modifications. The position and view angle of these patches proved to have little impact on their efficacy, further accentuating the potential threat posed by these attacks.

While our results indicate that these attacks have the potential to be robust enough for real-world applications, it is essential to acknowledge that further investigations are necessary to fully comprehend their implications and develop effective countermeasures. As the field of NeRFs continues to advance, addressing security concerns and improving defenses against adversarial attacks becomes imperative to ensure the trustworthy deployment of these technologies in various domains.

Acknowledgement

This research has been partially supported by the Hungarian Government by the following grants: 2018-1.2.1-NKP00008: Exploring the Mathematical Foundations of Artificial Intelligence and TKP2021_02-NVA-27 – Thematic Excellence Program. The support of the Alfréd Rényi Institute of Mathematics if also gratefully acknowledged.

References

[1] B. Mildenhall, P. P. Srinivasan, M. Tancik, J. T. Barron, R. Ramamoorthi, and R. Ng, "Nerf: Representing scenes as neural radiance fields for view synthesis," *Communications of the ACM*, vol. 65, no. 1, pp. 99–106, 2021. 1

[2] V. Rudnev, M. Elgharib, W. Smith, L. Liu, V. Golyanik, and C. Theobalt, "Nerf for outdoor scene relighting," in *European Conference on Computer Vision*, pp. 615–631, Springer, 2022. 1

[3] Z. Zhang, Y. Liu, C. Han, Y. Pan, T. Guo, and T. Yao, "Transforming radiance field with lipschitz network for photorealistic 3d scene stylization," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 20712–20721, 2023. 1

[4] N. Deng, Z. He, J. Ye, B. Duinkharjav, P. Chakravarthula, X. Yang, and Q. Sun, "Fov-nerf: Foveated neural radiance fields for virtual reality," *IEEE Transactions on Visualization and Computer Graphics*, vol. 28, no. 11, pp. 3854–3864, 2022. 1

[5] D. Maggio, M. Abate, J. Shi, C. Mario, and L. Carlone, "Loc-nerf: Monte carlo localization using neural radiance fields," in *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 4018–4025, IEEE, 2023. 1

[6] K. Deng, A. Liu, J.-Y. Zhu, and D. Ramanan, "Depth-supervised nerf: Fewer views and faster training for free," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12882–12891, 2022. 1

[7] M. Adamkiewicz, T. Chen, A. Caccavale, R. Gardner, P. Culbertson, J. Bohg, and M. Schwager, "Vision-only robot navigation in a neural radiance world," *IEEE Robotics and Automation Letters*, vol. 7, no. 2, pp. 4606–4613, 2022. 1

[8] B. Xie, B. Li, Z. Zhang, J. Dong, X. Jin, J. Yang, and W. Zeng, "Navinerf: Nerf-based 3d representation disentanglement by latent semantic navigation," *arXiv preprint arXiv:2304.11342*, 2023. 1

[9] F. Li, H. Yu, I. Shugurov, B. Busam, S. Yang, and S. Ilic, "Nerf-pose: A first-reconstruct-then-regress approach for weakly-supervised 6d object pose estimation," *arXiv preprint arXiv:2203.04802*, 2022. 1

[10] D. Zimny, T. Trzciński, and P. Spurek, "Points2nerf: Generating neural radiance fields from 3d point cloud," *arXiv preprint arXiv:2206.01290*, 2022. 1

[11] X. Chen, C. Liu, B. Li, K. Lu, and D. Song, "Targeted backdoor attacks on deep learning systems using data poisoning," *arXiv preprint arXiv:1712.05526*, 2017. 1

[12] P. Kiourti, K. Wardega, S. Jha, and W. Li, "Trojdrl: evaluation of backdoor attacks on deep reinforcement learning," in *2020 57th ACM/IEEE Design Automation Conference (DAC)*, pp. 1–6, IEEE, 2020. 2

[13] Y. Fu, Z. Ye, J. Yuan, S. Zhang, S. Li, H. You, and Y. Lin, "Gen-nerf: Efficient and generalizable neural radiance fields via algorithm-hardware co-design," in *Proceedings of the 50th Annual International Symposium on Computer Architecture*, pp. 1–12, 2023. 2, 3

[14] Q. Wang, Z. Wang, K. Genova, P. Srinivasan, H. Zhou, J. T. Barron, R. Martin-Brualla, N. Snavely, and T. Funkhouser, "Ibrnet: Learning multi-view image-based rendering," in *CVPR*, 2021. 2, 3

[15] I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," *arXiv preprint arXiv:1412.6572*, 2014. 2, 3

[16] T. B. Brown, D. Mané, A. Roy, M. Abadi, and J. Gilmer, "Adversarial patch," *arXiv preprint arXiv:1712.09665*, 2017. 2, 4

[17] Y. Fu, Y. Yuan, S. Kundu, S. Wu, S. Zhang, and Y. Lin, "Nerfool: Uncovering the vulnerability of generalizable neural radiance fields against adversarial perturbations," *arXiv preprint arXiv:2306.06359*, 2023. 2, 4

[18] C. L. Deng and E. Tartaglione, "Compressing explicit voxel grid representations: fast nerfs become also small," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 1236–1245, 2023. 2

[19] A. Chen, Z. Xu, F. Zhao, X. Zhang, F. Xiang, J. Yu, and H. Su, "Mvsnerf: Fast generalizable radiance field reconstruction from multi-view stereo," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 14124–14133, 2021. 2

[20] Y. Liu, S. Peng, L. Liu, Q. Wang, P. Wang, C. Theobalt, X. Zhou, and W. Wang, "Neural rays for occlusion-aware image-based rendering. arxiv cs," *CV*, vol. 2107, p. 1, 2021. 2

[21] J. Reizenstein, R. Shapovalov, P. Henzler, L. Sbordone, P. Labatut, and D. Novotny, "Common objects in 3d: Large-scale learning and evaluation of real-life 3d category reconstruction," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 10901–10911, 2021. 2

[22] B. Mildenhall, P. P. Srinivasan, R. Ortiz-Cayon, N. K. Kalantari, R. Ramamoorthi, R. Ng, and A. Kar, "Local light field fusion: Practical view synthesis with prescriptive sampling guidelines," *ACM Transactions on Graphics (TOG)*, 2019. 3

[23] T. Zhou, R. Tucker, J. Flynn, G. Fyffe, and N. Snavely, "Stereo magnification: Learning view synthesis using multiplane images," *arXiv preprint arXiv:1805.09817*, 2018. 3

[24] L. Downs, A. Francis, N. Koenig, B. Kinman, R. Hickman, K. Reymann, T. B. McHugh, and V. Vanhoucke, "Google scanned objects: A high-quality dataset of 3d scanned household items," in *2022 International Conference on Robotics and Automation (ICRA)*, pp. 2553–2560, IEEE, 2022. 3

[25] M. M. Johari, Y. Lepoittevin, and F. Fleuret, "Geonerf: Generalizing nerf with geometry priors," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18365–18375, 2022. 3

[26] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, and R. Fergus, "Intriguing properties of neural networks," *arXiv preprint arXiv:1312.6199*, 2013. 3

[27] A. Rozsa, E. M. Rudd, and T. E. Boult, "Adversarial diversity and hard positive generation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pp. 25–32, 2016. 3

[28] Y. Dong, F. Liao, T. Pang, H. Su, J. Zhu, X. Hu, and J. Li, "Boosting adversarial attacks with momentum," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 9185–9193, 2018. 3, 4

[29] S.-M. Moosavi-Dezfooli, A. Fawzi, and P. Frossard, "Deepfool: a simple and accurate method to fool deep neural networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2574–2582, 2016. 3

[30] J. Lu, H. Sibai, E. Fabry, and D. Forsyth, "No need to worry about adversarial examples in object detection in autonomous vehicles," *arXiv preprint arXiv:1707.03501*, 2017. 4

[31] A. Athalye, L. Engstrom, A. Ilyas, and K. Kwok, "Synthesizing robust adversarial examples," *arXiv preprint arXiv:1707.07397*, 2017. 4

[32] K. Eykholt, I. Evtimov, E. Fernandes, B. Li, A. Rahmati, C. Xiao, A. Prakash, T. Kohno, and D. Song, "Robust physical-world attacks on deep learning models," *arXiv preprint arXiv:1707.08945*, 2017. 4

[33] M. Alzantot, Y. Sharma, S. Chakraborty, and M. Srivastava, "Genattack: Practical black-box attacks with gradient-free optimization," *arXiv preprint arXiv:1805.11090*, 2018. 4

[34] N. Papernot, P. McDaniel, I. Goodfellow, S. Jha, Z. B. Celik, and A. Swami, "Practical black-box attacks against machine learning," in *Proceedings of the 2017 ACM on Asia Conference on Computer and Communications Security*, pp. 506–519, ACM, 2017. 4

[35] N. Akhtar and A. Mian, "Threat of adversarial attacks on deep learning in computer vision: A survey," *IEEE Access*, vol. 6, pp. 14410–14430, 2018. 4

[36] C. Xie, J. Wang, Z. Zhang, Y. Zhou, L. Xie, and A. Yuille, "Adversarial examples for semantic segmentation and object detection," in *Proceedings of the IEEE International Conference on Computer Vision*, pp. 1369–1378, 2017. 4

[37] J. H. Metzen, M. C. Kumar, T. Brox, and V. Fischer, "Universal adversarial perturbations against semantic image segmentation," in *2017 IEEE International Conference on Computer Vision (ICCV)*, pp. 2774–2783, IEEE, 2017. 4

[38] A. Arnab, O. Miksik, and P. H. Torr, "On the robustness of semantic segmentation models to adversarial attacks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 888–897, 2018. 4

[39] J. Al-afandi and H. András, "Class retrieval of detected adversarial attacks," *Applied Sciences*, vol. 11, no. 14, p. 6438, 2021. 4

[40] S. Thys, W. Van Ranst, and T. Goedemé, "Fooling automated surveillance cameras: adversarial patches to attack person detection," *arXiv preprint arXiv:1904.08653*, 2019. 4

[41] S.-T. Chen, C. Cornelius, J. Martin, and D. H. P. Chau, "Shapeshifter: Robust physical adversarial attack on faster r-cnn object detector," in *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pp. 52–68, Springer, 2018. 4