Adversarial Attacks on MedNeRF

Attacking Neural Radiance Field on Medical Care Scenarios

Rui Chu
Electrical and Computer Engineering, Tufts University

School of Engineering
rui.chu@tufts.edu

***Abstract*—*As generative AI technologies continue to develop, their significant influence on the field of large predictive models has become clear, indicating a mature phase in this area. Along with this, the progression of computer vision is likely to be greatly advanced by generative AI methods. Even Though there have been some initial successes, such as the system known as Sora, the focus on this area still needs to be stronger. At the same time, the security aspects of generative AI have attracted a lot of attention, particularly concerning language models where the risk of adversarial attacks is high. This increased concern highlights the importance of having strong protections against possible weaknesses. Yet, the specific security issues related to generative AI in visual contexts have not been addressed as much, even though they could lead to serious problems. This paper argues that the security of visual generative AI needs urgent and thorough attention. We use Neural Radiance Fields (NeRF) as an example of visual generative AI and apply it in the medical field. We also carry out targeted adversarial attacks to reveal its vulnerabilities, showing the urgent need for better security measures in visual AI applications. This research aims to fill the gaps in the discussion about security by focusing on the unique challenges and risks related to visual generative AI and promoting a proactive approach to protect these systems.*

Keywords*—*Generative AI, Adversarial Attacks, Neural Radiance Fields (NeRF), Visual AI, Medical Applications

I. Introduction

The Author is using two main concepts to prove the Hypothesis: NeRF and Adversarial Attacks in Artificial intelligence.

A. Neural Radiance Fields (NeRF)

Neural Radiance Fields (NeRF) represent an innovative approach in the field of three-dimensional scene reconstruction, utilizing deep learning to produce complex scenes with great detail and realism. This technique models a scene as a continuous volume where a neural network determines the color and density at any point in 3D space based on coordinates. The essential idea of NeRF is to train a deep neural network with a collection of 2D images from different perspectives, accompanied by camera parameters. This network calculates the density and radiance dependent on the viewpoint, facilitating high-quality rendering by combining these predictions along the paths of camera rays using volumetric rendering methods.

One key benefit of NeRF over traditional 3D reconstruction approaches is its precision in capturing detailed features and managing hidden parts of scenes with great accuracy, as the way shown in Figure 1. NeRF works within a continuous space, unlike methods that use discrete voxel grids or mesh-based structures, allowing for smoother transitions and more precise interpolations between different viewpoints. This feature is especially valuable for applications that require lifelike renderings, such as virtual reality, augmented reality, and film visual effects.

Moreover, the NeRF framework has sparked various improvements that overcome some of its initial drawbacks, like lengthy training periods and the need for precise camera positioning. These developments have broadened NeRF's use to include moving scenes, vast environments, and real-time applications, establishing it as a key advancement in the areas of computational photography and computer vision. This introduction section intends to explore the basic concepts of NeRF, explain how it works, and discuss its significant effects on transforming 3D scene modeling.

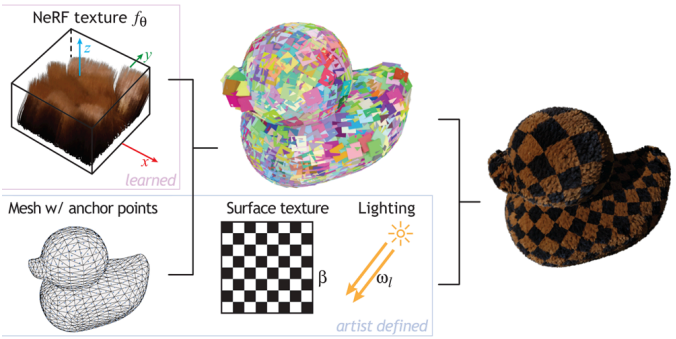

Fig. 1. NeRF Theory

B. Adversarial Attack in AI Security

Adversarial attacks are a significant and growing concern in the field of artificial intelligence (AI) security, as they challenge the stability and reliability of AI systems used in many important areas such as finance, healthcare, and autonomous vehicles. These attacks involve creating and modifying inputs to AI models to make them perform incorrectly or unpredictably. The risks associated with these security issues are particularly high due to the increasing dependence on AI technologies in critical domains, where failures could lead to serious consequences.

Adversarial attacks exploit the weaknesses of AI models that focus primarily on recognizing patterns in large data sets. These models can be deceived by small, often barely noticeable changes to input data, which lead them to make wrong decisions. This happens because the models do not fully understand the data they are processing, focusing only on statistical patterns.

The study of these attacks not only points out the vulnerabilities of current AI systems but also motivates the development of more secure and resilient AI technologies. Strategies to reduce these risks include adversarial training (training models with manipulated data) and implementing robustness checks. Additionally, the creation of rules and ethical guidelines for AI deployment highlights the importance of tackling these security issues effectively.

II. BACKGROUND

There are several current works on NeRF and Adversarial attacks. As an example, medical AI is a heat area in vision fields. Thus, the author is using Medical AI as an example to set the experiments.

A. NeRF in Medical Area

Medical imaging, particularly computed tomography (CT), plays a pivotal role in clinical diagnostics. However, the process typically involves exposing patients to significant levels of ionizing radiation, which can be harmful. Recent developments in deep learning have introduced innovative methods to reduce this exposure by enhancing image reconstruction techniques. A notable contribution in this field is the "MedNeRF" model, which leverages Neural Radiance Fields (NeRF) for reconstructing CT projections from minimal X-ray views, thus significantly reducing the required radiation dose.

The "MedNeRF" model, proposed by Corona-Figueroa et al., is designed to generate 3D-aware CT projections using a novel architecture that builds upon generative radiance fields. This model adeptly learns a continuous representation of the 3D structure of anatomical features directly from one or a few 2D X-ray images [1]. By utilizing the principles of Neural Radiance Fields, "MedNeRF" disentangles the volumetric depth and surface details of internal structures without the need for extensive 3D supervision. This is achieved through a sophisticated deep learning framework that integrates the Generative Radiance Fields (GRAF) approach, tailored specifically for medical applications.

"MedNeRF" adapts the GRAF methodology to medical imaging by introducing a discriminator architecture that is robust to the unique challenges presented by medical datasets, which are often limited in size and complexity compared to standard image datasets. This adaptation allows for high-fidelity synthesis of CT projections, capturing the intricate variations in anatomical structures as viewed from different angles. The model demonstrates its effectiveness on diverse medical instances such as chest and knee datasets, producing results that are not only realistic but also clinically relevant due to their reduced demand for radiation.

Furthermore, the approach taken by "MedNeRF" showcases the potential for future applications in medical imaging where reduced radiation exposure is crucial. By synthesizing CT projections from limited X-ray views, the model opens up new possibilities for diagnostic imaging in scenarios where traditional multi-view CT scans may not be feasible.

B. Adversarial Attack on NeRF

Recent advancements in generative neural networks have expanded the scope of adversarial attacks to encompass 3D multiview data, enhancing the complexity and effectiveness of these threats against deep learning models. A pivotal study by Jiang et al. introduced "NeRFail," an innovative method that utilizes Neural Radiance Fields (NeRF) to craft multiview adversarial perturbations and do classification as it was shown in Figure 2. This technique constructs a 3D scene from various 2D views and generates photorealistic renderings from novel vantages to deceive recognition systems systematically [2]. Specifically, their approach involves a transformation mapping from 2D-pixel representations to 3D spatial configurations. By manipulating a subset of these views, the adversarial perturbations are designed to mislead deep neural networks about the scene's geometry, affecting both the trained and novel viewpoints. This manipulation highlights the vulnerabilities in applications relying on synthesized 3D data from multiple 2D images, raising significant security concerns.

In our research, we adapt the "NeRFail" method to the domain of medical imaging, a field where diagnostic accuracy is crucial, and data sensitivity is paramount. Our adaptation involves refining the algorithm to work effectively with medical datasets, exploring potential risks and the necessary defensive mechanisms for AI applications in healthcare diagnostics. By leveraging the foundational insights from Jiang et al., we aim to assess the robustness of medical imaging systems against advanced adversarial attacks that exploit the 3D rendering capabilities of NeRF, thereby contributing to safer AI implementations in sensitive environments.

Fig. 2. Classification of NeRFail

III. Methodology

In this study, we enhance the adversarial attack approach originally developed in the NeRFail model, applying it to the medical imaging dataset from MedNeRF. Our methodology consists of several key steps, starting with the design of an improved adversarial attack algorithm. We then replicate the NeRF framework to accommodate our modifications, ensuring that the foundational model is compatible with our advanced attack strategies.

After establishing a robust NeRF base, we implement our refined adversarial attack algorithm. The MedNeRF dataset, primarily consisting of medical images, is then fine-tuned to align with the adversarial model requirements. This adaptation is critical for ensuring that the attack effectively induces misclassifications within the MedNeRF framework.

Our objective is to evaluate the impact of the adversarial attacks on the MedNeRF model, specifically targeting its diagnostic accuracy. We assess the effectiveness of these attacks by measuring changes in the Mean Squared Error (MSE) and Peak Signal-to-Noise Ratio (PSNR) of the medical images processed under adversarial conditions. By doing so, we aim to quantify the vulnerability of the MedNeRF system to adversarial perturbations, providing insights into the potential risks and necessary enhancements needed for medical imaging technologies.

Finally, complete content and organizational editing before formatting. Please take note of the following items when proofreading spelling and grammar:

A. Adversarial Attacking Algorithm

The Adversarial Attacking Algorithm is mostly inherited from the NeRFail section.

Algorithm 1: NeRFail-S Attack

Input: X: 3D points constructed with p camera views according to (9)
Input: $\mathcal{I}$: images from a view set $\mathcal{V}$
Input: l_g: true label, K: maximum iterations
Input: α, ϵ: step size
Output: $\mathcal{Y}_K$

$\mathbf{z} \leftarrow \mathbf{0}$
while $i < K$ **do**
  while $v \in \mathcal{V}$ **do**
    $X_v \leftarrow \cup_{\mathbf{r} \in \mathcal{R}_v}(M(\hat{C}(\mathbf{r}, F)))$
    $I'_v \leftarrow \mathbf{clip}(I(M'(X, \mathbf{z}), v) + I_v, 0, 1)$
    $\Delta z \leftarrow \mathbf{sign}(\nabla_{\mathbf{z}} \mathbf{CrossEntropy}(f(I'_v)))$
    $\mathbf{z} \leftarrow \mathbf{z} + \alpha \Delta z$
    $\mathbf{z} \leftarrow \mathbf{clip}(\mathbf{z}, -\epsilon, \epsilon)$
  end while
end while

B. Updating Attacking Algorithms

- The Adjustable α in auto finding value
- Updated Algorithm: α ← AdaptStepSize(α, Δz, lg)

Function AdaptStepSize(α, Δz, lg):

  if Success(Δz, lg):

    return Increase(α)

  else:

    return Decrease(α)

In refining the NeRFail-S attack mechanism, we introduced an adaptive step-size adjustment to enhance the perturbation efficiency. Unlike the conventional method with a fixed step size, our algorithm dynamically adjusts the magnitude of alterations to the 3D points at each iteration. This adaptive process commences with an initial step size, which is subsequently modulated based on the success of the adversarial perturbations in inducing misclassification.

The core of our improvement lies in the AdaptStepSize function. This function evaluates whether the current gradient direction, indicated by the sign of the cross-entropy loss, successfully deteriorates the model's performance. If a successful perturbation is detected, the function escalates the step size to accelerate convergence towards an effective adversarial example. Conversely, suppose the perturbation fails to mislead the model. In that case, the function decreases the step size to fine-tune the perturbations, thereby preventing overshooting which could lead to detectable or ineffective adversarial noise.

The adaptive step-size search not only aligns with the objective of generating more potent adversarial examples but also contributes to a more nuanced and controlled search process. Such a tailored approach is particularly crucial in medical applications, where the interpretability and subtlety of perturbations can significantly impact the diagnostic outcomes. Through iterative adjustments and evaluations, our methodology aspires to yield adversarial examples that subtly yet effectively compromise the integrity of the medical imaging model, ensuring that the induced perturbations remain imperceptible to human observers while still deceiving the machine learning model.

The efficacy of this adaptive method is gauged through repeated experiments, assessing the impact on Mean Squared Error (MSE) and Peak Signal-to-Noise Ratio (PSNR) metrics under adversarial conditions. By doing so, we aim to comprehensively understand the resilience of the MedNeRF model against such sophisticated adversarial tactics and to establish a robust measure of the adversarial attack's potency.

IV. Evaluation

The entire process is designed to run on an L40s GPU and use Jupyter Notebook to run the code.

A. Experiments Design

In our study, we carefully designed experiments to test our improved adversarial attack on medical images. We adapted the setup to ensure it's suitable for medical data, which can be quite different from other types of data used in NeRF models.

1) Theoretical Implementation rules

a) Dataset Usage: We began with the NeRF Lego dataset to enhance our model's ability to understand images from many camera angles. Then, we applied what the model learned to the MedNeRF dataset. This step is called the

transfer learning and is important for helping our model get used to medical images.

b) Model Structure: Our experiments continue to use the NeRF model with its Multilayer Perceptron (MLP) setup. We didn't change this structure because it's essential for the type of work we're doing, which is about creating 3D images from 2D pictures.

c) Setting the Parameters: For our experiment, we used the Adam optimizer to adjust the α value. This approach is different from NeRFail, which uses a fixed α. Our choice aims to let the model decide the best α during the learning process. The initial learning rate was set to 0.001, which is a common starting point for many models. We also set the parameters (beta1 and beta2) for Adam to standard values (0.9 and 0.999), which often work well.

d) Training Method: We trained the NeRF model with new camera angles from the Lego dataset for six hours on a computer called L40S. This step was crucial for making sure the model could create new directions accurately.

e) Evaluation Measures: We used PSNR to see how well the model was doing after the attack. We also looked at MSE and how often the model was correct in classifying images. Training speed was also important to us because we wanted to know if our new method made the training slower.

f) Goals and Expectations: Our main goal was to see if the adversarial attack would change the PSNR on medical images. We believed that our attack could make it harder for the model to classify images correctly without making the training too slow.

2) Implementation Core Snaps

The main difficulties we were facing was the speed in both training and attacking. In this case, the author tried to improve the attacking method to improve the general running time.

a) Update of Adversarial Attack: Since the main update of the attacking algorithm is $\alpha \leftarrow \text{AdaptStepSize}(\alpha, \Delta z, lg)$, thus we made the update in Jupyter as shown in Figure 3:

```
def adam_gradient_descent(X, y, theta, iterations):
    m = len(y)
    t = 0
    m_t = np.zeros_like(theta)
    v_t = np.zeros_like(theta)

    for i in range(iterations):
        t += 1
        grad = compute_gradient(X, y, theta)
        grad = perturb(grad)  # Adjust gradients for perturbation

        m_t = beta1 * m_t + (1 - beta1) * grad
        v_t = beta2 * v_t + (1 - beta2) * np.square(grad)

        m_t_hat = m_t / (1 - np.power(beta1, t))
        v_t_hat = v_t / (1 - np.power(beta2, t))

        theta = theta - alpha * m_t_hat / (np.sqrt(v_t_hat) + epsilon)
```

Fig. 3. Snap of Code Implementation in Algorithm Update

b) Transfer Learning and Fine-Tunning on Medical X-Ray Dataset: Because the Fine-Tuning Dataset of MedNeRF needs to be set up as a proper dataset folder based on the set photoing direction as it was set in NeRFail, we re-arranged the dataset provided by MedNeRF to get ready for the proper training.

c) *Setting Required Photoing Direction on NeRF:* Since we need to have a co-related output photo, we need to set the output photo direction as what was used in NeRFail to help us do the classification.

3) Difficulties Faced

Because of the time difference and diversity of the research goal, there are a lot of steps that need to be adjusted between those two projects,

a) *The dataset provided by MedNeRF is not sufficient enough*: Since we were trying to do Transfer-Learning, there are only 1 dataset of Knee provided by the MedNeRF team, which is not sufficient enough for us to Fine-Tunning the NeRFail into MedNeRFail.

b) *Difference in used packages*: Because of the year of publication, the package used by the two projects has huge differences. We adjusted a lot of packages and re-illustrated some pieces to make the entire project run successfully.

In order to realize the experiments, we adjusted the research goal from classification to grading PSNR to see the performance of applying X-ray to NeRFail directly. To compare the difference, we can also prove if the NeRFail attack works towards the Medical data.

B. Result Evaluation

The following parts are required to be evaluated.

1) The speed improvements of attacking training progress: As discussed before, we have updated the attacking progress. To measure the performance, the following Table 1 shows how the performance was improved. As was shown, there are some efficiency improvements based on the update of the attacking during training progress with the attacking method.

TABLE I. COMPARISON BETWEEN TRAINING SPEED

Goal	Table Column Head		
	Iteration Amount	***Original***	***Updated***
NeRF Camera Direction	20,000 by NeRF Training	6 Hour	6 Hour
NeRFail Attack	5,000	1 Hour	0.9 Hour
MedNeRFail	100	140 Sec	120 Sec

2) Training NeRF to have proper camera direction: We want the MedNeRF to generate the same camera direction as NeRFail, so we re-trained the NeRF with a proper camera direction as it was shown in Figure 4, which represents the training set and targeted training outcome of a pre-set camera direction.

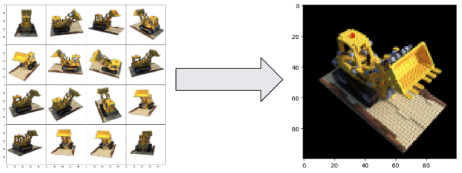

Fig. 4. Output of Retraining NeRF to have a pre-set camera direction

3) Tunning the adversarial attacking method: After having a pre-trained NeRF with the camera direction we need, we applied the NeRFail attacking algorithm to the model and Fine-tuned it. Figure 5 shows how the perturbation works during the progress. It can be noticed that comparing iterations 600 and 700, the PSNR will drop to a stable level, and thus the attacking progress with the updated NeRFail to be working properly.

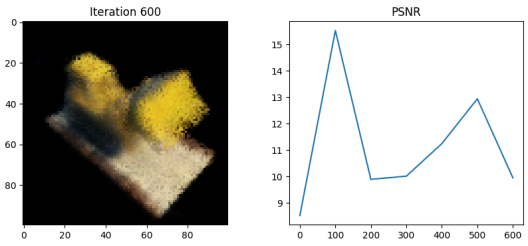

Fig. 5. Success rate of NeRfail when Iteration 600

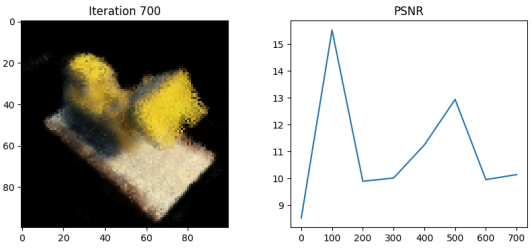

Fig. 6. Success rate of NeRfail when Iteration 700

4) Applying the attacked model with the MedNeRF dataset: The MedNeRF dataset contains the X-Ray Images of Knee from different camera directions. Since we had trained a NeRF with and without a NeRFail attack, we can apply the image to both of the models and see the performance. As shown in Figure 7, there is a difference before and after the NeRFail towards the X-ray image of the Knee.

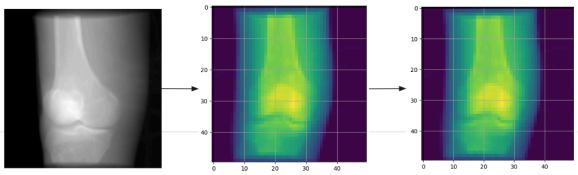

Fig. 7. X-ray image of Knee after NeRF and NeRFail

TABLE II. PERFORMANCE OF NERFAIL ON X-RAY MEDICAL IMAGES

Comparing NeRFail on the original set and Medical Set		
Iteration amount	***PSNR on MedNeRFail***	***PSNR on NeRFail***
25% of the entire Iteration	25.16	10

Table II shows that the PSNR on MedNeRFail has an acceptable drop compared to NeRFail. Since we are attacking, we hope the lower PSNR will be better.

V. CONCLUSION

After the experiments, we have supported the hypothesis that the adversarial attack can be improved through auto-adjustment progress.

Meanwhile, there is transfer-learning can be made for MedNeRF towards the attacked NeRFail model.

Because of the limitation of the dataset, we did not to Fine-tune the NeRF on the MedNeRF. We believe that the MedNeRF will be improved into a better MedNeRFail as long as the dataset is sufficient.

It might cause a huge impact in the area of misdiagnosing in the following research. Thus, further protection methods should be proposed.

ACKNOWLEDGMENT

Appreciation towards Professor Yingjie Lao at the Electrical and Computer Engineering department at Tufts University for providing instruction and lectures on Artificial Intelligence Security to guide the paper.

REFERENCES

[1] A. Corona-Figueroa, J. Frawley, S. Bond-Taylor, S. Bethapudi, H. P. H. Shum, C. G. Willcocks, "MedNeRF: Medical Neural Radiance Fields for Reconstructing 3D-aware CT-Projections from a Single X-ray," 2022 IEEE Engineering in Medicine & Biology Society (EMBC), Glasgow, UK, July 11-15, 2022.

[2] W. Jiang et al., "NeRFail: Neural Radiance Fields-Based Multiview Adversarial Attack," in Proceedings of the Thirty-Eighth AAAI Conference on Artificial Intelligence, 2024.