DefenseSplat: Enhancing the Robustness of 3D Gaussian Splatting via Frequency-Aware Filtering

Yiran Qiao, Yiren Lu, Yunlai Zhou, Rui Yang, Linlin Hou, Yu Yin, and Jing Ma

Case Western Reserve University, Cleveland OH 44106, USA
{yxq350, yxl3538, yxz3057, rxy337, lxh633, yxy1421, jxm1384}@case.edu
Code: https://github.com/yrqiao/DefenseSplat

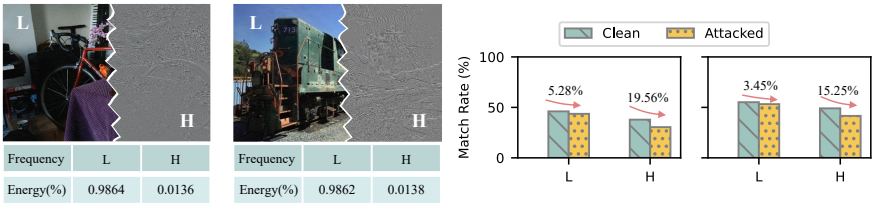

Fig. 1: Effects of Attacks for 3DGS. **Left**: Two examples showing the comparison between low- and high-frequency components on the same image after attack, with energy ratios of different frequency components in clean images at the bottom. The left side shows results on the *Mip-NeRF 360* dataset, while the right side shows *Tanks-and-Temples*. We take the *bonsai* and *Train* as examples, respectively. **Right**: Matching Rate (a signal for multi-view consistency) in low-frequency (L) and high-frequency components (H) of the clean/attacked images.

Abstract. 3D Gaussian Splatting (3DGS) has emerged as a powerful paradigm for real-time and high-fidelity 3D reconstruction from posed images. However, recent studies reveal its vulnerability to adversarial corruptions in input views, where imperceptible yet consistent perturbations can drastically degrade rendering quality, increase training and rendering time, and inflate memory usage, even leading to server denial-of-service. In our work, to mitigate this issue, we begin by analyzing the distinct behaviors of adversarial perturbations in the low- and high-frequency components of input images using wavelet transforms. Based on this observation, we design a simple yet effective frequency-aware defense strategy that reconstructs training views by filtering high-frequency noise while preserving low-frequency content. This approach effectively suppresses adversarial artifacts while maintaining the authenticity of the original scene. Notably, it does not significantly impair training on clean data, achieving a desirable trade-off between robustness and performance on clean inputs. Through extensive experiments under a wide range of attack intensities on multiple benchmarks, we demonstrate that

our method substantially enhances the robustness of 3DGS without access to clean ground-truth supervision. By highlighting and addressing the overlooked vulnerabilities of 3D Gaussian Splatting, our work paves the way for more robust and secure 3D reconstructions.

Keywords: Adversarial Attack & Defense · Gaussian Splatting · Discrete Wavelet Transform

1 Introduction

3D reconstruction, exemplified by Neural Radiance Fields (NeRF) [30] and the more recent 3D Gaussian Splatting (3DGS) [14], refers to the process of recovering a scene's spatial geometry and structure from multi-view 2D images. It is a fundamental task in computer vision with wide-ranging applications in robotics [13,23,28,36,51], autonomous driving [8,15,38,50], and healthcare [4,42]. Especially, 3DGS has recently gained prominence as a dominant approach in 3D vision [2]. Unlike NeRF, which encodes scenes as continuous volumetric radiance fields parameterized by a multi-layer perceptron, 3DGS explicitly represents the scene using a collection of 3D Gaussian ellipsoids. This explicit modeling confers several advantages over NeRF, including faster rendering speeds, higher visual fidelity, greater interpretability, and improved scalability. These benefits make 3DGS well-suited for deployment on remote servers, where it can generate high-quality scene representations from user-provided images.

1.1 Vulnerabilities of 3DGS

While 3DGS offers strong performance, its strengths also reveal critical vulnerabilities. Because it explicitly fits object textures and edges, its rendering quality is highly sensitive to input image quality [25]. Moreover, 3DGS does not use a fixed network architecture; instead, its adaptive density control dynamically adjusts the number of Gaussian primitives, creating a flexible, data-dependent parameter space [14]. This explicitness and flexibility make 3DGS particularly vulnerable to adversarial perturbations, which can degrade rendering and increase computational cost.

Although adversarial threats to 3DGS are only beginning to be explored, initial work has emerged. Zeybey *et al.* [47] introduces a two-stage image-space attack that transfers adversarial noise into 3DGS representations, but it operates solely in 2D and lacks true 3D-level manipulation. In contrast, Poison-Splat [24] performs a genuine 3D-level attack by training a surrogate model and applying bi-level optimization, achieving stronger disruption at the cost of significant computational overhead.

1.2 Why is Defense for 3DGS Underexplored?

Given the emergence of 3DGS-specific adversarial attacks and the increasing awareness of its inherent vulnerabilities, developing effective defense mechanisms

for 3DGS is of growing importance. However, to the best of our knowledge, **defense strategies against adversarial attacks on 3DGS remain largely underexplored.** This gap stems from several unique obstacles that arise from the intersection of 3DGS's design and the nature of adversarial threats:

Challenge 1: Incompatibility with Standard Defense Frameworks. Most adversarial defenses are tailored for supervised classification models with fixed neural architectures, where bi-level optimization is used (an inner loop crafts adversarial perturbations to maximize loss, while the outer loop minimizes this loss via adversarial training [27]). In contrast, 3DGS is a *self-supervised* method *without a fixed network structure*, leading to fundamentally different attack and defense formulations. As a result, standard defense strategies do not apply effectively in this setting.

Challenge 2: Non-Invertible Attack Objectives. Even when using a quick proxy workaround, such as a pre-trained 3DGS model, reversing the bi-level optimization objective for defense remains ineffective. This is because adversarial objectives in 3DGS often target *non-invertible* image statistics, such as increasing the total variation of rendered images to inject subtle pixel-level noise. Attempting to reverse such objectives introduces unintended artifacts like image blurring or texture distortion, which ultimately degrade reconstruction quality instead of mitigating the attack.

Challenge 3: Absence of Ground-Truth Clean Data. Unlike supervised settings where clean labels are available, 3DGS lacks a clear notion of "ground-truth" clean input. This makes it *impossible to distinguish clean from adversarially perturbed images* during training or inference. As a result, designing defenses that restore adversarial inputs without overcorrecting and harming genuinely clean images becomes highly challenging, especially when performance degradation cannot be tolerated.

Challenge 4: Limited Effectiveness of Traditional 2D Image-Space Defenses. Common 2D image-level or frequency-based defense techniques often prove insufficient or harmful in the 3DGS setting. For instance, Gaussian smoothing and bilinear filtering may reduce adversarial perturbations, but they also blur legitimate details critical to accurate scene reconstruction. Similarly, reducing the number of Gaussian primitives can limit adversarial effects but comes at the cost of reconstructive fidelity. Even frequency-domain filtering (e.g., via Fourier transforms) falls short, as it neglects the spatial structure essential to 3D rendering, weakening its ability of meaningful defense for 3DGS.

1.3 A Spark of Opportunity: Frequency

In this work, we present the first comprehensive study of defense mechanisms against adversarial attacks on 3DGS. To address the fundamental differences from conventional attack and defense (Challenges 1 and 2), we forgo adversarial training and instead propose a defense strategy that operates directly on the input multi-view images. We begin by analyzing different frequency components of the input images and how adversarial perturbations affect them. Specifically, we employ wavelet transforms to decompose the input into low- and high-frequency

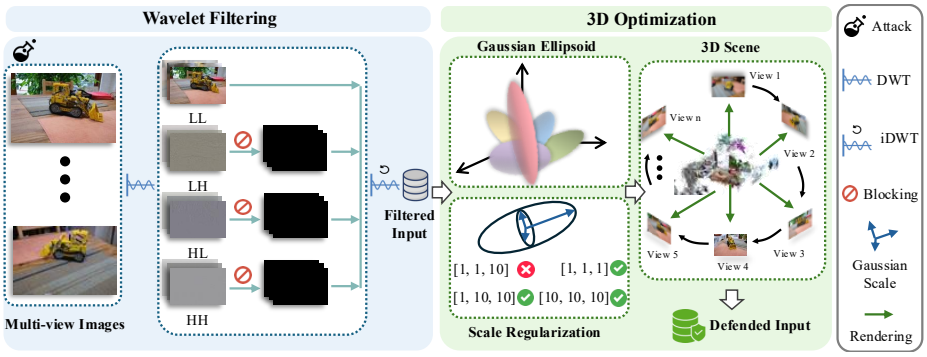

Fig. 2: The overview of our proposed method.

components, leveraging their ability to provide both spatial and frequency representations. This enables a spatially-aware, frequency-sensitive consistency analysis across multiple views. Using deep image matching techniques [32], we evaluate the consistency of these frequency bands and have two important observations, revealing a new opportunity for defense:

- **Observation 1:** The adversarial perturbations on 3DGS inputs often manifest as **high-frequency noises** (as shown in the right of Fig. 1).
- **Observation 2:** The **low-frequency component retains the main content** of the original scene, as the ratios of energy (a signal for image structure and content) are shown in the bottom left of Fig. 1.

Building on this observation, we introduce *DefenseSplat* (as shown in Fig. 2), a frequency-aware defense method designed to suppress high-frequency adversarial noises while preserving informative low-frequency content. By doing so, the 3DGS optimization process naturally blurs out inconsistent perturbations during reconstruction, enhancing robustness without requiring access to clean ground-truth data (addressing Challenge 3). Importantly, this approach maintains high reconstruction quality on clean inputs, addressing Challenge 4 by balancing robustness with fidelity, especially in scenarios where significant performance degradation is not acceptable. Our main contribution can be summarized as:

- We investigate an important yet previously unexplored problem of defending 3DGS against adversarial attacks **(in this paper, we consider PoisonSplat as the attack to defend against).** We analyze the significance of this problem and identify the unique challenges it presents.
- We introduce a novel frequency-aware defense strategy DefenseSplat for 3DGS by leveraging wavelet analysis to characterize the behavior of adversarial perturbations across frequency bands. To the best of our knowledge, this is the *first* work specifically targeting adversarial defense for 3DGS.

– Our comprehensive experiments show that the proposed method enhances the robustness of 3DGS across diverse datasets and a range of attack strengths, without substantially compromising clean-data performance.

2 Related Works

2.1 3D Gaussian Splatting.

3DGS [14] is a recently introduced framework for 3D scene representation, which employs an explicit set of anisotropic 3D Gaussian primitives parameterized by position, scale, orientation, and color to model the underlying geometry and appearance. Unlike implicit methods such as NeRF [30], 3DGS supports efficient, real-time rendering through a differentiable rasterization process. Owing to its high rendering quality, interpretability, and scalability, 3DGS has been rapidly adopted across a wide spectrum of vision and graphics applications, including dynamic scene reconstruction [40, 45], open-vocabulary 3D semantic segmentation [26,37,43,46], and text-to-3D generation. However, as 3DGS is increasingly deployed on cloud-based servers and used in commercial applications, its potential vulnerabilities and susceptibility to adversarial manipulation have become critical concerns.

Adversarial Attack. Adversarial attacks are typically formulated as a bi-level optimization problem, where imperceptible perturbations are crafted to maximize the model's prediction error while remaining constrained in norm. In 2D image classification, such attacks have been extensively studied under white-box and black-box settings. White-box methods [7, 27, 31] utilize gradient access to generate perturbations, whereas black-box attacks [1, 21] rely on transferability or query-based optimization. Due to the inherent vulnerabilities of 3D Gaussian Splatting, several recent works have emerged aiming to exploit these weaknesses through adversarial attacks at the 3D level. Poison-Splat [24] is a representative work in this direction, which performs a genuine 3D adversarial attack by pretraining a surrogate 3DGS model and adopting the total variance score of the rendered images as the attack objective. However, adversarial attacks targeting the domain of 3D Gaussian Splatting remain largely underexplored.

2.2 Defense on Images.

The rise of adversarial attacks has spurred advances in defense strategies in images [34]. Among them, adversarial training [27, 49] is the most straightforward and widely used approach. It involves augmenting the training process with adversarial samples, allowing the model to learn parameters that are robust to specific perturbation patterns. However, as mentioned in the challenges discussed in Sec. 1, adversarial training cannot be directly applied to 3D adversarial attacks. Another line of defense leverages diffusion models to remove artifacts and distortions from images [41]. However, such methods typically rely on large-scale pretraining and are primarily designed for scenarios where the degradation in rendering quality arises from sparse-view supervision during 3D reconstruction. As

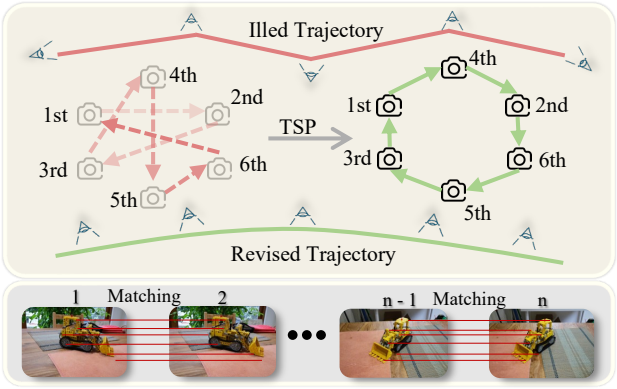

Fig. 3: To accurately measure image matching, we order the camera poses using a Traveling Salesman Problem formulation, which yields an optimized traversal path.

far as we are aware, neither prior work has addressed defense against 3D adversarial attacks, nor are existing defense approaches directly applicable to this task.

2.3 Frequency Transform.

The Discrete Fourier Transform (DFT) plays a pivotal role in image processing. Previous works have leveraged DFT on images to enhance synthesis quality [12] and improve image classification accuracy [35]. More recent studies have extended the application of DFT to the domain of 3D reconstruction. FreGS [48] leverages additional frequency spectrum supervision to mitigate the issue of over-reconstruction in Gaussian Splatting. 3D-GSW [11] employs DFT to guide the splitting of 3D Gaussians in high-frequency regions, thereby enhancing reconstruction quality. DFT only captures global frequency information, whereas the Discrete Wavelet Transform (DWT) provides a joint representation of both spatial and frequency domains. DWT can be seamlessly integrated with radiance field representations [10, 22, 44], significantly enhancing the quality of 3D reconstruction. Building upon the advantageous properties of DWT and its demonstrated success in 3D reconstruction, we leverage DWT to analyze adversarial attacks on 3DGS from both spatial and frequency perspectives and further design a wavelet-based defense strategy.

3 Methodology

3.1 Preliminaries

3D Gaussian Splatting. 3DGS represents a 3D scene using an explicit set of anisotropic 3D Gaussian primitives. Each primitive is defined as:

$$G(\mathbf{x}; \boldsymbol{\Sigma}) = \exp\left(-\frac{1}{2}(\mathbf{x})^\top \boldsymbol{\Sigma}^{-1}(\mathbf{x})\right), \tag{1}$$

where $\mathbf{x}$ is the center of each primitive. $\boldsymbol{\Sigma}$ is the covariance matrix, which, together with $\mathbf{x}$ determines the spatial distribution of the Gaussian. To enable differentiable optimization, $\boldsymbol{\Sigma}$ is expressed as a combination of a scaling matrix R and a rotation matrix S, shown as:

$$\boldsymbol{\Sigma} = RSS^{\top}R^{\top}. \tag{2}$$

Given a specific camera pose π, each pixel (x, y) color $I_{\pi}(x, y)$ rendered onto the 2D image is calculated via alpha-blending [14], taking into account a set of Gaussians ($\mathcal{N}$) that are simultaneously visible and contribute to the pixel through spatial overlap:

$$I_{\pi}(x, y) = \sum_{i \in \mathcal{N}} c_i \alpha_i \prod_{j=1}^{i-1} (1 - \alpha_j), \tag{3}$$

where c_i is each projected Gaussian given by spherical harmonics, and α_i is the opacity. During the reconstruction training process, the parameters of the 3D Gaussian primitives are optimized by jointly minimizing the L_1 loss and the SSIM loss [39] between the rendered images and the ground-truth images.

Discrete Wavelet Transform (DWT). DWT decomposes an image into four frequency subbands: LL (low-low), LH (low-high), HL (high-low), and HH (high-high). The LL subband captures the *low-frequency* components in both horizontal and vertical directions and typically contains the majority of the image's energy, representing its overall structure and smooth content. This subband can be further recursively decomposed across multiple levels to extract hierarchical low-frequency representations. In contrast, the LH, HL, and HH subbands capture *high-frequency* information along different orientations: LH highlights horizontal edge-like features by encoding low-frequency horizontal and high-frequency vertical changes; HL emphasizes vertical edges by capturing high-frequency horizontal and low-frequency vertical variations; and HH contains diagonal details with high-frequency variations in both directions. Together, these subbands provide a joint spatial–frequency representation of the image. The DWT [6] is defined as:

$$\begin{aligned} W_{\varphi}(j_0, m, n) &= \frac{1}{\sqrt{MN}} \sum_{x=0}^{M-1} \sum_{y=0}^{N-1} I(x, y)\, \varphi_{j_0, m, n}(x, y), \\ W_{\psi}^{i}(j, m, n) &= \frac{1}{\sqrt{MN}} \sum_{x=0}^{M-1} \sum_{y=0}^{N-1} I(x, y)\, \psi_{j, m, n}^{i}(x, y), \end{aligned} \tag{4}$$

where W_{φ} is the LL subband, and W_{ψ}^{i} represents the LH, HL, and HH subbands. j_0 is the initial approximation scale, j is the scale index for detail subbands. M and N are width and height of the RGB image, while m and n are the spatial location indices in the wavelet coefficient domain. $i = \{H, V, D\}$ denotes

the horizontal, vertical, and diagonal coefficients, respectively. $\varphi(x, y)$ is a scaling function and $\psi(x, y)$ is a wavelet function. $I(x, y)$ denotes the pixel intensity of a given position.

3.2 Frequency-Aware Analysis of Adversarial Perturbations

Frequency Decomposition. To design more targeted defense strategies, it is essential to first conduct a comprehensive analysis of the adversarial attacks. Prior works [11, 48] have demonstrated the power of frequency-domain analysis, establishing it as a powerful tool for enhancing scene reconstruction in 3DGS. For example, it has proven effective in mitigating issues such as over-reconstruction [48] and supporting applications like watermark embedding [11]. Building on these successful precedents, we adopt DWT to decompose adversarial images. Owing to its joint representation of both spatial and frequency information, DWT facilitates comprehensive analysis across 2D and 3D domains. Moreover, its frequency decomposition enables a fine-grained dissection of adversarial behavior across different frequency subbands, laying the foundation for targeted defense strategies. Specifically, given an adversarial image $I' \in \mathbb{R}^{3\times H\times W}$, we apply wavelet decomposition to obtain a set of subbands $\{LL_l, LH_l, HL_l, HH_l\}$, where l denotes the level of DWT. For simplicity and without loss of effectiveness, we set $l = 1$. To constrain each subband within the RGB domain, we perform a wavelet transform on I' channel-wise, normalize each channel to [0, 1], rescale it to [0, 255], and finally fuse the three channels.

Vulnerability Analysis. Next, we use deep image matching [9,32,33] to assess the 3D consistency of corresponding subbands between clean and adversarial images, identifying *the most vulnerable regions* where perturbations concentrate. A key challenge before this step is ensuring efficient and accurate multi-view matching. Common strategies include brute-force, retrieval-based, and sequential matching. Brute-force has quadratic complexity, making it inefficient at scale. Retrieval-based methods rely on global features, introducing error without optimal speed. Sequential matching is most efficient but assumes a smooth view sequence, which is often violated in practice due to *randomly ordered* camera poses, leading to failures or reduced accuracy from large viewpoint differences and occlusions.

To address this issue, we formulate it as a Traveling Salesman Problem (TSP). We first convert each camera pose into the form of the Special Euclidean group in 3D, and then compute the pose loss between every pair of cameras, defined as:

$$\mathcal{L}_{\text{pose}} = w_g \, \mathcal{L}_{\text{geo}} + w_t \, \mathcal{L}_{\text{trans}}, \tag{5}$$

Table 1: Quantitative comparison on Mip-NeRF 360 and Tanks-and-Temples datasets. Colors indicate the best and second best results. #G/M indicates the number of Gaussians (Million).

Method	Training Time ↓	#G/M ↓	GPU Mem (MB) ↓	PSNR ↑	SSIM ↑	LPIPS ↓	FPS ↑
Mip-NeRF 360							
3DGS	1:01:01	5.91	20965	25.18	0.6356	0.4481	70
Compact GS	1:22:26	3.00	21072	24.95	0.6374	0.4590	128
Difix	0:45:44	3.65	15894	24.52	0.7326	0.3575	137
Ours	0:34:26	2.24	12567	27.32	0.7968	0.3469	243
Ours+ReLU	0:33:49	2.16	11400	27.47	0.8025	0.3417	258
Tanks-and-Temples							
3DGS	0:31:22	2.76	10089	24.88	0.6425	0.4190	165
Compact GS	0:43:29	1.73	10930	24.37	0.6540	0.4278	244
Difix	0:24:12	1.71	7749	24.75	0.8092	0.3031	290
Ours	0:18:56	1.06	6296	26.29	0.8281	0.3199	468
Ours+ReLU	0:19:19	1.01	5858	26.42	0.8373	0.3134	498

where $\mathcal{L}_{\text{geo}}$ is the geodesic loss, $\mathcal{L}_{\text{trans}}$ is the translation loss, w_g and w_t are their corresponding weights. Specifically, these two losses are defined as:

$$\mathcal{L}_{\text{geo}} = \arccos\left(\frac{\text{trace}\big(R_{\text{pred}}^{\top} R_{\text{gt}}\big) - 1}{2}\right), \tag{6}$$

$$\mathcal{L}_{\text{trans}} = \|t_{\text{pred}} - t_{\text{gt}}\|_2^2, \tag{7}$$

where $R_{(\cdot)}$ and $t_{(\cdot)}$ represent the rotation matrix and the translation vector, respectively. The subscripts indicate prediction and ground truth, respectively. We treat each camera as a city and the pose loss between cameras as the distance between cities in TSP. By solving the resulting instance of the TSP using the Lin–Kernighan [19] algorithm, we obtain a smooth and optimized camera trajectory (as shown in Fig. 3). Based on the optimized trajectory, we adopt the state-of-the-art SuperPoint [5] as the feature extractor and LightGlue [20] as the matcher to perform pairwise matching across all wavelet subbands of clean and adversarial images. Each image is only matched with subsequent images along the trajectory. For each image pair, the *matching rate* is defined as the ratio of the number of matched keypoints to the number of extracted keypoints, indicating the *consistency of the image across multiple viewpoints*. The overall multi-view matching rate is computed by averaging the pairwise matching rates. Since the HH subband contains the least energy (only $< 0.08\%$ in almost all datasets, with details in Appendix B) and diagonal details are relatively rare in natural images, we omit HH and use LH and HL to represent the high-frequency information. As shown in Fig. 1, both the bar charts and subband visualizations indicate that **the vulnerability is concentrated in the high-frequency regions**. The consistency degradation in high-frequency subbands

is significantly more pronounced than in low-frequency ones, suggesting that **3D adversarial attacks primarily target high-frequency components by injecting random noise, while introducing relatively smooth and consistent pseudo-textures in the low-frequency regions.**

3.3 Defense via Frequency-Based Filtering

Based on the above analysis and insights, we propose a simple yet effective defense strategy, DefenseSplat. Given adversarial images, we first apply DWT to obtain four subbands. The high-frequency subbands (*i.e.*, LH, HL, and HH) are then set to *zero*, and the inverse wavelet transform is performed using the LL subband together with the zeroed high-frequency subbands to reconstruct the filtered image in the spatial domain. This process can be defined as:

$$I_f = \texttt{iDWT}(\texttt{DWT}(I')_{LL}, 0, 0, 0), \tag{8}$$

where I_f represents the filtered images, $\texttt{iDWT}(\cdot)$ is the inverse Discrete Wavelet Transform, I' denotes the adversarial images and $\texttt{DWT}(I')_{LL}$ is the LL subband extracted from I'. Merely relying on the I_f component is far from sufficient, as it still contains artificial textures with relatively low consistency in the low-frequency domain. To mitigate this problem, we exploit an intrinsic property of 3DGS: when fitting regions with low cross-view consistency, the Gaussians tend to average the colors across views to minimize the L_1 loss, thereby producing a blurred reconstruction over these regions, which naturally further suppresses adversarial artifacts in I_f that lack consistent support across views.

Another common observation is: For certain artificially injected textures that exhibit high multi-view consistency, the reconstruction process tends to generate *elongated* Gaussians to fit these patterns, which increases the number of Gaussians and memory consumption, thereby undermining the effectiveness of the defense. To tackle this issue, we design a ReLU-based scale regularization loss on the normalized variance of the scales along the three principal axes of each Gaussian, defined as follows:

$$\mathcal{L}_{scale} = \text{ReLU}(\nu - \tau), \tag{9}$$

where ν denotes the normalized variance, τ is a predefined threshold that limits the range of scale regularization. For Gaussians whose normalized variance falls below τ, the loss does not propagate gradients. This loss *constrains elongated Gaussians without affecting the other two types of Gaussians*: small spherical Gaussians and large flat Gaussians (because their normalized variance of scales does not exceed that of elongated Gaussians). Both of them are crucial: the former contributes to reconstructing fine details and alleviates over-reconstruction, while the latter enhances the reconstruction quality of large smooth regions and mitigates under-reconstruction. Therefore, this loss can effectively suppress adversarial textures while preserving the reconstruction of genuine scene details. Finally, after training, the 3D Gaussians are rendered on all camera views to generate the final defended images for downstream tasks.

Table 2: Quantitative comparison under different attack levels on Tanks-and-Temples.

ϵ	Method	Training Time ↓	#G/M ↓	GPU Mem (MB) ↓	PSNR ↑	SSIM ↑	LPIPS ↓	FPS ↑
16/255	3DGS	0:27:38	2.17	8558	24.03	0.662	0.375	215
	Compact GS	0:36:58	1.47	9466	23.59	0.669	0.386	306
	Difix	0:21:39	1.48	7238	23.59	0.787	0.298	295
	Ours	0:17:44	0.99	5642	24.89	0.815	0.294	449
	Ours + ReLU	0:18:17	0.96	5418	25.03	0.824	0.291	473
32/255	3DGS	0:36:14	3.24	11541	20.36	0.449	0.485	140
	Compact GS	0:50:13	2.20	12997	20.26	0.458	0.491	199
	Difix	0:23:16	1.69	7417	22.22	0.707	0.348	263
	Ours	0:20:02	1.25	6182	23.47	0.690	0.377	362
	Ours + ReLU	0:19:43	1.19	5806	23.61	0.703	0.374	388
64/255	3DGS	0:43:00	4.07	12418	15.55	0.272	0.586	119
	Compact GS	0:59:48	2.69	15807	15.65	0.285	0.591	178
	Difix	0:25:46	1.99	8067	19.18	0.504	0.458	243
	Ours	0:23:42	1.79	7651	19.27	0.443	0.509	281
	Ours + ReLU	0:23:16	1.71	6958	19.38	0.454	0.505	301

3.4 Fine-Grained Cross-View Consistency Filtering (Optional)

The coarse filter in Sec. 3.3 removes all high-frequency coefficients, effectively suppressing adversarial perturbations but also discarding genuine scene details and causing over-smoothed reconstructions on structured regions. To avoid this all-or-nothing suppression, we introduce a fine-grained cross-view-consistency filter. We first order all camera poses using the TSP tour from Sec. 3.2. Each adversarial image is then compared with its left neighbour along this pose-ordered cycle, with the first pose matched back to the last. For each high-frequency wavelet subband, we retain coefficients around locations that match the corresponding subband in the neighbouring view, treating them as multi-view consistent scene details, and suppress the unmatched coefficients as view-inconsistent perturbations. The defended image is finally reconstructed by keeping the low-frequency subband unchanged and inverting the wavelet transform with only the matched high-frequency coefficients. This preserves authentic details erased by the coarse filter while removing adversarial components that do not recur across nearby views.

4 Experiments

In this section, we conduct extensive experiments on multiple datasets. Specifically, we address the following research questions based on the experimental results. **RQ1:** How does DefenseSplat perform compared to other baselines? Is it an effective defense strategy? **RQ2:** Can DefenseSplat consistently demonstrate effective defense capability under varying levels of attack intensity? **RQ3:** Given the diversity of user-submitted inputs, can DefenseSplat maintain a balance between robustness and performance on clean data? **RQ4:** Can the Gaussian scale loss (Eq. (9)) further enhance the overall performance of DefenseSplat (ablation study)? **RQ5:** Can our method remain resilient to stronger attacks, such

Table 3: Quantitative results of our method on clean inputs.

Method	Training Time ↓	#G/M ↓	GPU Mem (MB) ↓	PSNR ↑	SSIM ↑	LPIPS ↓	FPS ↑
Mip-NeRF 360							
Clean	0:38:20	2.62	20965	29.06	0.868	0.248	223
Clean + Ours	0:32:11	1.83	11012	28.58	0.844	0.280	314
Clean + Ours + ReLU	0:32:23	1.79	10945	28.79	0.847	0.271	321
Tanks-and-Temples							
Clean	0:22:48	1.56	7328	28.06	0.898	0.212	327
Clean + Ours	0:18:24	0.97	6161	27.18	0.868	0.254	517
Clean + Ours + ReLU	0:18:34	0.94	5857	27.33	0.875	0.245	526

as the recently proposed Adaptive Attack [18]? **RQ6:** Can the fine-grained filter improve reconstruction performance?

Fig. 4: Comparison of reconstruction quality of all methods and ground truth (GT) on Mip-NeRF 360 and Tanks-and-Temples datasets.

4.1 Experiment Settings

Datasets. We use Mip-NeRF 360 [3], Tanks-and-Temples [16] and LLFF [29], which are commonly used benchmarks in the field of 3D reconstruction. For the Tanks and Temples dataset, we exclude scenes containing dynamic objects and

retain 8 static scenes for evaluation. For Mip-NerF 360 and LLFF datasets, we include all available scenes.

Baselines. Due to the absence of prior research dedicated to the defense of 3D adversarial attacks, no existing method serves as a directly comparable baseline. Therefore, in addition to the original **3DGS** as a baseline, we select two other methods whose research directions are most closely related to defense as comparative baselines. The first one is **Difix3D+** [41], which leverages diffusion models to enhance the quality of reconstructed images while mitigating 3D inconsistency. Moreover, it shares a similar setting with ours, as both methods operate directly on the input images. The second is **CompactGS** [17], which achieves reduced storage and faster rendering speed while maintaining high-quality reconstruction by effectively removing redundant Gaussians that do not significantly contribute to the overall performance. This objective closely aligns with the essence of defense.

Implementation Details. Our method and all baselines are trained on the same server equipped with four A6000 GPUs. To ensure fairness, all methods are trained for 30,000 iterations. The remaining hyperparameters of the baselines are kept unchanged from their original settings. Unless otherwise specified, we use adversarial images with an attack strength of $\epsilon = 16/255$ as the training input. The weight for the scale loss λ_{scale} is set to 1×10^5.

Evaluation Metrics We evaluate the effectiveness of our defense strategy along two dimensions: **robustness** and **reconstruction fidelity**. For robustness, we evaluate training time, the number of Gaussians, peak GPU memory usage, and rendering speed, which together reflect the server's capacity to sustain training under the load imposed by the input images. For reconstruction fidelity, we adopt widely used metrics including peak signal-to-noise ratio (PSNR), structural similarity (SSIM), and learned perceptual image patch similarity (LPIPS).

4.2 Experiment Results

Defense Effectiveness of DefenseSplat (RQ1). We conduct a comprehensive comparison between our method and the baselines. As shown in Tab. 1, our method **achieves the best performance in both robustness and rendering quality** on the Mip-NeRF 360 and the Tanks-and-Temples datasets with the sole exception of the LPIPS on the Tanks-and-Temples dataset, where our score is only slightly higher than the best-performing baseline. However, this does not indicate a performance drawback: our method still achieves the lowest peak GPU memory usage—an essential metric, as exceeding hardware limits can lead to denial-of-service issues and prevent rendering. Fig. 4 presents visualizations of the rendered images. Since the perturbations introduced by the 3D adversarial attacks are nearly imperceptible to humans, we zoom in to better highlight the details. As observed in both the room and truck scenes, the rendered images from the original 3DGS contain numerous adversarial artifacts, which are more clearly revealed in the corresponding LH subband. Although CompactGS reduces the number of Gaussians, it still fails to effectively constrain the Gaussians from

Table 4: Results of Adaptive Attack on two scenes of NeRF Synthetic Dataset (TT=Training Time, #G=Number of Gaussians, GPU=GPU Memory)

Scene	Settings	TT	#G	GPU	PSNR	SSIM	LPIPS	FPS
$\beta = 1$								
chair	Attack	0:16:00	889700	6052	30.14	0.3483	0.3971	436
	Ours	0:08:06	200798	2338	30.31	0.4123	0.3241	1653
ficus	Attack	0:09:17	280991	5982	30.44	0.3015	0.4365	1276
	Ours	0:06:07	121204	2118	31.68	0.3405	0.3589	2381
$\beta = 5$								
chair	Attack	0:16:43	897341	6186	30.04	0.3428	0.4051	437
	Ours	0:08:10	202392	2182	30.32	0.4085	0.3337	1654
ficus	Attack	0:09:17	280642	6194	30.41	0.3012	0.4396	1288
	Ours	0:06:04	121065	2120	31.69	0.3438	0.3627	2358
Clean								
chair		0:09:17	439996	2846	39.08	0.9948	0.0083	892
ficus		0:06:07	170861	1818	38.00	0.9954	0.0080	1930

overfitting to adversarial perturbations. The rendered images produced by Difix3D+ appear smoother; however, they tend to introduce new spurious textures or remove authentic ones. In the room scene, wavy artifacts emerge on the carpet, while in the Truck scene, genuine details such as the rust on the hub and the tire tread patterns are mistakenly eliminated. Our method effectively removes adversarial artifacts while preserving the original textures to the greatest extent, resulting in reconstructions that are closest to the ground truth.

Defense Effectiveness Under Various Attack Strengths (RQ2). As shown in Tab. 2, our method consistently **shows effective defense under varying attack strengths**, due to the fact that the attack primarily targets high-frequency information, which our method blocks. Even when the attack strength increases and low-frequency components are possibly affected, our defense still remains effective, since our method consistently filters out high-frequency content.

Robustness and Fidelity Trade-Off (RQ3). As shown in Tab. 3, even when the input images are clean, our method **does not exhibit any significant performance degradation**. The PSNR and SSIM drop by only 1.75% and 2.5%, respectively, and LPIPS increases by just 11%. Although our method removes high-frequency information, the LL subband retains over 95% of the total energy across all subbands. As a result, even for clean images, the majority of information is preserved, leading to no significant degradation in reconstruction quality.

Effectiveness of Scale Loss (RQ4). As evidenced by Tab. 1 and Tab. 3, introducing the scale loss leads to a noticeable improvement in our method's performance, yielding the best results across most evaluation metrics. We provide additional experimental results in Appendix B.

Defense Effectiveness against Adaptive Attack (RQ5). Adaptive Attack is constructed by adding a weighted term, the mutual information between the clean and the poisoned images, to the PoisonSplat objective. This modification makes the poisoned images more imperceptible and harder to defend against.

Table 5: Quantitative results comparison of basic filter and fine-grained filter.

Method	Training Time ↓	#G/M ↓	GPU Mem (MB) ↓	PSNR ↑	SSIM ↑	LPIPS ↓	FPS ↑
basic	0:33:49	2.16	11400	27.47	0.8025	0.3417	258
fine-grained	0:34:15	2.19	11373	27.49	0.8027	0.3388	265

The formulation is as follows:

$$\max_{\mathcal{V}_{\text{poi}}} \mathcal{L} = \max_{\mathcal{V}_{\text{poi}}} \{\mathcal{S}_{TV}(\mathcal{V}_{\text{poi}}) + \beta \cdot I(\mathcal{V}_{\text{cln}}; \mathcal{V}_{\text{poi}})\}, \tag{10}$$

where $\mathcal{S}_{TV}$ denotes the TV score, $I(\mathcal{V}_{\text{cln}}; \mathcal{V}_{\text{poi}})$ denotes the mutual information between the clean images and poisoned images, β is the weight of mutual information. Since Adaptive Attack does not report the exact number of attack iterations, a fair comparison is not possible; therefore, it is not the primary attack method we focus on. We conduct defense evaluations on two scenes from the NeRF Synthetic dataset. As shown in Tab. 4, our method provides effective defense against Adaptive Attack across different strength levels (β). In our experiments, we adopt an extremely large number of attack steps (36,000), which leads to relatively worse post-attack SSIM and LPIPS values. Under this severe setting, our method still provides effective defense and does not require access to the clean images, which is more reasonable.

Effectiveness of the fine-grained filter (RQ6). We evaluate both filters on the Mip-NeRF 360 dataset and report the average results. As shown in Tab. 5, the fine-grained filter slightly improves reconstruction quality, validating our previous analysis and the effectiveness of the proposed method. However, since the improvement brought by the fine-grained filter is limited and its data preprocessing is more complex, this component can be optional when the overall efficiency is preferred.

5 Discussion and Conclusion

In this work, we propose a novel and effective defense strategy DefenseSplat against 3DGS adversarial attacks. By applying filtering on high-frequency wavelet subbands, our method achieves both robust 3D reconstruction training on the server side and faithful image rendering quality. Since only one main attack and its variant on 3DGS have been proposed so far, our evaluation is necessarily limited to this specific method. However, given our observation that adversarial attacks primarily target vulnerable high-frequency components in the input, we believe our defense strategy is broadly generalizable. We remain confident and look forward to testing its effectiveness against future, unseen attacks. It is worth noting that our defense strategy can also be flexibly employed as a plug-and-play module, allowing seamless integration with existing methods to achieve better defense performance.

References

1. Andriushchenko, M., Croce, F., Flammarion, N., Hein, M.: Square attack: a query-efficient black-box adversarial attack via random search. In: European conference on computer vision. pp. 484–501. Springer (2020)
2. Bao, Y., Ding, T., Huo, J., Liu, Y., Li, Y., Li, W., Gao, Y., Luo, J.: 3d gaussian splatting: Survey, technologies, challenges, and opportunities. IEEE Transactions on Circuits and Systems for Video Technology (2025)
3. Barron, J.T., Mildenhall, B., Verbin, D., Srinivasan, P.P., Hedman, P.: Mip-nerf 360: Unbounded anti-aliased neural radiance fields. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 5470–5479 (2022)
4. Cai, Y., Wang, J., Yuille, A., Zhou, Z., Wang, A.: Structure-aware sparse-view x-ray 3d reconstruction. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 11174–11183 (2024)
5. DeTone, D., Malisiewicz, T., Rabinovich, A.: Superpoint: Self-supervised interest point detection and description. In: Proceedings of the IEEE conference on computer vision and pattern recognition workshops. pp. 224–236 (2018)
6. Gonzalez, R.C.: Digital image processing. Pearson education india (2009)
7. Goodfellow, I.J., Shlens, J., Szegedy, C.: Explaining and harnessing adversarial examples. arXiv preprint arXiv:1412.6572 (2014)
8. Hess, G., Lindström, C., Fatemi, M., Petersson, C., Svensson, L.: Splatad: Real-time lidar and camera rendering with 3d gaussian splatting for autonomous driving. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 11982–11992 (2025)
9. Ioli, F., Dematteis, N., Giordan, D., Nex, F., Livio, P.: Deep learning low-cost photogrammetry for 4d short-term glacier dynamics monitoring. PFG – Journal of Photogrammetry, Remote Sensing and Geoinformation Science (2024). `https://doi.org/10.1007/s41064-023-00272-w`
10. Jang, Y., Lee, D.I., Jang, M., Kim, J.W., Yang, F., Kim, S.: Waterf: Robust watermarks in radiance fields for protection of copyrights. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 12087–12097 (2024)
11. Jang, Y., Park, H., Yang, F., Ko, H., Choo, E., Kim, S.: 3d-gsw: 3d gaussian splatting for robust watermarking. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 5938–5948 (2025)
12. Jiang, L., Dai, B., Wu, W., Loy, C.C.: Focal frequency loss for image reconstruction and synthesis. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 13919–13929 (2021)
13. Keetha, N., Karhade, J., Jatavallabhula, K.M., Yang, G., Scherer, S., Ramanan, D., Luiten, J.: Splatam: Splat track & map 3d gaussians for dense rgb-d slam. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 21357–21366 (2024)
14. Kerbl, B., Kopanas, G., Leimkühler, T., Drettakis, G.: 3d gaussian splatting for real-time radiance field rendering. ACM Trans. Graph. **42**(4), 139–1 (2023)
15. Khan, M., Fazlali, H., Sharma, D., Cao, T., Bai, D., Ren, Y., Liu, B.: Autosplat: Constrained gaussian splatting for autonomous driving scene reconstruction. In: 2025 IEEE International Conference on Robotics and Automation (ICRA). pp. 8315–8321. IEEE (2025)

16. Knapitsch, A., Park, J., Zhou, Q.Y., Koltun, V.: Tanks and temples: Benchmarking large-scale scene reconstruction. ACM Transactions on Graphics **36**(4) (2017)
17. Lee, J.C., Rho, D., Sun, X., Ko, J.H., Park, E.: Compact 3d gaussian representation for radiance field. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 21719–21728 (2024)
18. Li, Y., Liu, Z., Li, Z., Lin, Z., Zhang, J.: Remedygs: Defend 3d gaussian splatting against computation cost attacks. arXiv preprint arXiv:2511.22147 (2025)
19. Lin, S., Kernighan, B.W.: An effective heuristic algorithm for the traveling-salesman problem. Operations research **21**(2), 498–516 (1973)
20. Lindenberger, P., Sarlin, P.E., Pollefeys, M.: Lightglue: Local feature matching at light speed. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 17627–17638 (2023)
21. Liu, Y., Chen, X., Liu, C., Song, D.: Delving into transferable adversarial examples and black-box attacks. arXiv preprint arXiv:1611.02770 (2016)
22. Lou, A., Planche, B., Gao, Z., Li, Y., Luan, T., Ding, H., Chen, T., Noble, J., Wu, Z.: Darenerf: Direction-aware representation for dynamic scenes. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5031–5042 (2024)
23. Lu, G., Zhang, S., Wang, Z., Liu, C., Lu, J., Tang, Y.: Manigaussian: Dynamic gaussian splatting for multi-task robotic manipulation. In: European Conference on Computer Vision. pp. 349–366. Springer (2024)
24. Lu, J., Zhang, Y., Shen, Q., Wang, X., Yan, S.: Poison-splat: Computation cost attack on 3d gaussian splatting. arXiv preprint arXiv:2410.08190 (2024)
25. Lu, Y., Zhou, Y., Liu, D., Liang, T., Yin, Y.: Bard-gs: Blur-aware reconstruction of dynamic scenes via gaussian splatting. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 16532–16542 (2025)
26. Lu, Y., Zhou, Y., Qiao, Y., Song, C., Liang, T., Ma, J., Yin, Y.: Segment then splat: A unified approach for 3d open-vocabulary segmentation based on gaussian splatting. arXiv preprint arXiv:2503.22204 (2025)
27. Madry, A., Makelov, A., Schmidt, L., Tsipras, D., Vladu, A.: Towards deep learning models resistant to adversarial attacks. arXiv preprint arXiv:1706.06083 (2017)
28. Matsuki, H., Murai, R., Kelly, P.H., Davison, A.J.: Gaussian splatting slam. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 18039–18048 (2024)
29. Mildenhall, B., Srinivasan, P.P., Ortiz-Cayon, R., Kalantari, N.K., Ramamoorthi, R., Ng, R., Kar, A.: Local light field fusion: Practical view synthesis with prescriptive sampling guidelines. ACM Transactions on Graphics (ToG) **38**(4), 1–14 (2019)
30. Mildenhall, B., Srinivasan, P.P., Tancik, M., Barron, J.T., Ramamoorthi, R., Ng, R.: Nerf: Representing scenes as neural radiance fields for view synthesis. Communications of the ACM **65**(1), 99–106 (2021)
31. Moosavi-Dezfooli, S.M., Fawzi, A., Frossard, P.: 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)
32. Morelli, L., Ioli, F., Maiwald, F., Mazzacca, G., Menna, F., Remondino, F.: Deep-image-matching: A toolbox for multiview image matching of complex scenarios. The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences **XLVIII-2/W4-2024**, 309–316 (2024). https://doi.org/10.5194/isprs-archives-XLVIII-2-W4-2024-309-2024

33. Morelli, L., Bellavia, F., Menna, F., Remondino, F.: Photogrammetry now and then–from hand-crafted to deep-learning tie points–. The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences **48**, 163–170 (2022)
34. Qiao, Y., Liu, D., Lu, Y., Yin, Y., Du, M., Ma, J.: Counterfactual visual explanation via causally-guided adversarial steering. arXiv preprint arXiv:2507.09881 (2025)
35. Rao, Y., Zhao, W., Zhu, Z., Lu, J., Zhou, J.: Global filter networks for image classification. Advances in neural information processing systems **34**, 980–993 (2021)
36. Rosinol, A., Leonard, J.J., Carlone, L.: Nerf-slam: Real-time dense monocular slam with neural radiance fields. In: 2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). pp. 3437–3444. IEEE (2023)
37. Shi, J.C., Wang, M., Duan, H.B., Guan, S.H.: Language embedded 3d gaussians for open-vocabulary scene understanding. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5333–5343 (2024)
38. Tonderski, A., Lindström, C., Hess, G., Ljungbergh, W., Svensson, L., Petersson, C.: Neurad: Neural rendering for autonomous driving. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 14895–14904 (2024)
39. Wang, Z., Bovik, A.C., Sheikh, H.R., Simoncelli, E.P.: Image quality assessment: from error visibility to structural similarity. IEEE transactions on image processing **13**(4), 600–612 (2004)
40. Wu, G., Yi, T., Fang, J., Xie, L., Zhang, X., Wei, W., Liu, W., Tian, Q., Wang, X.: 4d gaussian splatting for real-time dynamic scene rendering. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 20310–20320 (2024)
41. Wu, J.Z., Zhang, Y., Turki, H., Ren, X., Gao, J., Shou, M.Z., Fidler, S., Gojcic, Z., Ling, H.: Difix3d+: Improving 3d reconstructions with single-step diffusion models. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 26024–26035 (2025)
42. Wu, S., Lu, Y., Guo, Y., Ji, W., Huang, S., Yang, F., Sirejiding, S., He, Q., Tong, J., Ji, Y., et al.: Discretized gaussian representation for tomographic reconstruction. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 25073–25082 (2025)
43. Wu, Y., Meng, J., Li, H., Wu, C., Shi, Y., Cheng, X., Zhao, C., Feng, H., Ding, E., Wang, J., et al.: Opengaussian: Towards point-level 3d gaussian-based open vocabulary understanding. Advances in Neural Information Processing Systems **37**, 19114–19138 (2024)
44. Xu, M., Zhan, F., Zhang, J., Yu, Y., Zhang, X., Theobalt, C., Shao, L., Lu, S.: Wavenerf: Wavelet-based generalizable neural radiance fields. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 18195–18204 (2023)
45. Yang, Z., Gao, X., Zhou, W., Jiao, S., Zhang, Y., Jin, X.: Deformable 3d gaussians for high-fidelity monocular dynamic scene reconstruction. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 20331–20341 (2024)
46. Ye, M., Danelljan, M., Yu, F., Ke, L.: Gaussian grouping: Segment and edit anything in 3d scenes. In: European conference on computer vision. pp. 162–179. Springer (2024)
47. Zeybey, A., Ergezer, M., Nguyen, T.: Gaussian splatting under attack: Investigating adversarial noise in 3d objects. arXiv preprint arXiv:2412.02803 (2024)

48. Zhang, J., Zhan, F., Xu, M., Lu, S., Xing, E.: Fregs: 3d gaussian splatting with progressive frequency regularization. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 21424–21433 (2024)
49. Zhang, Y., Gong, M., Liu, T., Niu, G., Tian, X., Han, B., Schölkopf, B., Zhang, K.: Causaladv: Adversarial robustness through the lens of causality. arXiv preprint arXiv:2106.06196 (2021)
50. Zhou, X., Lin, Z., Shan, X., Wang, Y., Sun, D., Yang, M.H.: Drivinggaussian: Composite gaussian splatting for surrounding dynamic autonomous driving scenes. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 21634–21643 (2024)
51. Zou, X., Song, Y., Qiu, R.Z., Peng, X., Ye, J., Liu, S., Wang, X.: 3d-spatial multimodal memory. In: The Thirteenth International Conference on Learning Representations (2025)