

Generating Multi-view Adversarial Examples for Visual Geometry Grounded Transformer

Qi Song, Ziyuan Luo*, Haoliang Han, and Renjie Wan

Department of Computer Science,
Hong Kong Baptist University, Hong Kong, China
{qisong, ziyuanluo, haolianghan}@life.hkbu.edu.hk,
renjiewan@hkbu.edu.hk

Abstract. The Visual Geometry Grounded Transformer (VGGT) enables unified feed-forward 3D reconstruction from multi-view images. However, deploying such a high-performance model may expose critical security vulnerabilities. Traditional adversarial perturbations require costly per-scene optimization, while Universal Adversarial Perturbations (UAPs) rely on a single static pattern and fail to effectively attack VGGT. To address these limitations, we propose **MVAP-G**, a multi-view adversarial perturbation generator that produces imperceptible consistent perturbations across multiple views in a single feed-forward pass. To ensure perturbation consistency across diverse scenes, we design a cross-view adversarial alignment mechanism to process multi-view images. Experiments demonstrate that MVAP-G significantly degrades VGGT performance without iterative optimization during inference. This work pioneers multi-view adversarial attacks on 3D foundation models, uncovering severe vulnerabilities and underscoring the urgent need for robust 3D vision systems. The code is available at <https://github.com/qsong2001/mvap-g>.

Keywords: Adversarial Examples · VGGT · 3D Safety

1 Introduction

Recent advances in 3D vision have led to new 3D foundation models like Visual Geometry Grounded Transformer (VGGT) [54]. It unifies diverse 3D tasks like 3D reconstruction, SLAM, tracking [9, 47, 53, 56] into a single feed-forward pass. However, as these models are increasingly deployed in safety-critical applications [38, 58], their vulnerability to invisible adversarial perturbations emerges as a fundamental security concern [36, 52].

Adversarial attacks are a major threat to the security of most vision applications [2, 37, 43, 59, 67]. However, large foundation models like VGGT [54], with their massive parameter scales, exhibit robustness to naive adversarial perturbations [37] in *real-world applications*. Traditional adversarial attacks [14, 37]

* Corresponding author. This work was carried out at the Renjie Group, Hong Kong Baptist University.

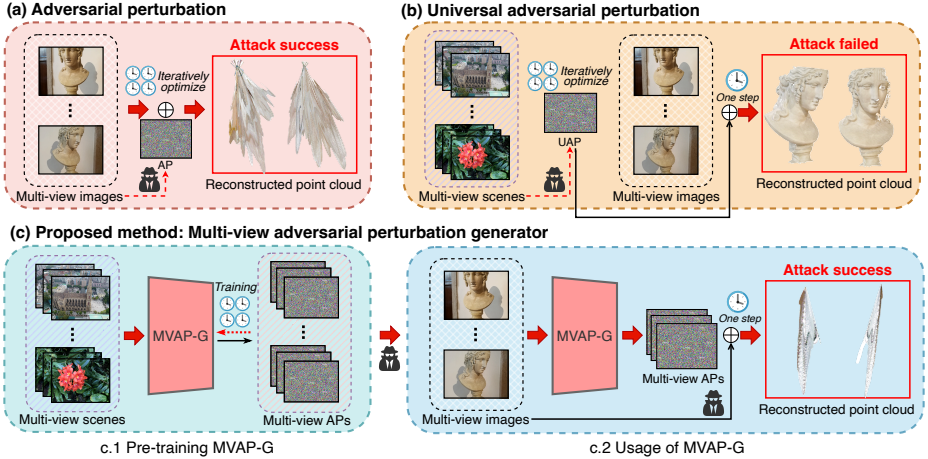


Fig. 1: Different adversarial examples against VGGT. Although traditional (a) adversarial perturbation (AP) [37] can achieve effective attack performance, it must be optimized for each scene with significant computational cost. (b) Universal adversarial perturbation [41] (UAP) is optimized on multi-view scenes data, yet it lacks attack effectiveness. (c) Our proposed MVAP-G is trained on multi-view data and can generate multi-view perturbations (Multi-view APs) against VGGT in real-time.

require continuous per-scene optimization for every new multi-view input (as shown in Fig. 1). This computational overhead makes them impractical in dynamic environments where scenes change continuously in the real world. To eliminate per-scene optimization, Universal Adversarial Perturbations (UAPs) [41, 67] employ data-agnostic, universal perturbations applicable to all inputs. This “one-for-all” strategy eliminates iterative computation, offering high efficiency. However, UAPs fail against VGGT due to their limited expressiveness: a static perturbation cannot adapt to the rich geometric variations in multi-view scenes [19, 21, 65]. As a result, UAPs exhibit poor capability against VGGT across diverse scenes.

A promising direction is to leverage generative models that can dynamically synthesize adversarial perturbations in real time. Prior generative adversarial example methods [2, 16, 43] have demonstrated success in this paradigm, efficiently creating perturbations through learned generators without requiring per-instance optimization. However, these methods are fundamentally designed for single-view architectures and cannot effectively handle multi-view inputs. They fail to model cross-view consistency and geometric relationships, which are critical factors for attacking multi-view foundation models like VGGT that explicitly exploit geometric correspondences across views.

To address this limitation, we propose **Multi-View Adversarial Perturbations Generator (MVAP-G)**, the first generative approach specifically designed for attacking multi-view 3D foundation models. MVAP-G directly processes raw multi-view images and produces imperceptible yet geometrically consistent ad-

versarial perturbations across all views in real time. By learning to generate view-coherent perturbations, MVAP-G combines the efficiency of generative methods with the geometric awareness required for multi-view scenarios, requiring no test-time optimization while adapting seamlessly to diverse scene configurations.

We achieve this through three core designs. **First**, a cross-view adversarial alignment (CAA) module explicitly models inter-view correspondences, forcing the generated perturbations to remain visually imperceptible while being destructive to VGGT’s inference. **Second**, dynamic perturbation regularization is applied during training to stabilise optimization under strong attack objectives, balancing attack performance and the invisibility of generated perturbations. **Third**, we demonstrate that pre-training the generator on a single-view dataset [45] can reduce total training costs. This simple yet effective strategy learns rich prior knowledge about adversarial vulnerability against VGGT. With these designs, MVAP-G consistently reduces VGGT’s performance in real time and across unseen scenes. Our contributions can be summarized:

- The first study on adversarial examples of the visual geometry grounded transformer, exploiting and revealing the potential vulnerabilities in the visual geometry grounded transformer.
- A real-time multi-view perturbation generation method that dynamically injects adversarial perturbations onto multi-view images, markedly surpassing prior adversarial perturbation methods.
- This work establishes a new paradigm for adversarial robustness in 3D vision, emphasizing the urgent need for secure deployment strategies of 3D foundation models.

Extensive experiments demonstrate the effectiveness and generalizability of our method across various input scenes.

2 Related work

2.1 Universal adversarial perturbations

Adversarial examples challenge the security of deep learning systems. These carefully crafted perturbations cause models to misclassify without being perceptible to humans. Early methods such as the Fast Gradient Sign Method (FGSM) [14] and Projected Gradient Descent (PGD) [37] use gradient-based optimization to generate instance-specific perturbations. They expose the vulnerability of neural networks [13, 17, 60] to manipulated inputs. However, these attacks require iterative optimization per input, which proves computationally expensive at scale. This limitation motivates the study of Universal Adversarial Perturbations (UAPs) [41]. UAPs enable one-shot attacks via a single, precomputed perturbation. Unlike sample-specific attacks, a UAP deceives the model across diverse inputs, revealing systemic risks in deployment. The original UAP [41] aggregates instance-specific directions iteratively and fools ImageNet-trained networks [8]. Later work branches into two directions: generative methods use GANs

to improve transferability across architectures [16, 43, 66], while gradient-based approaches optimize in feature subspaces [28, 57, 67]. Recent studies highlight limitations of CNN-focused UAPs on vision transformers [20, 34, 63]. New paradigms emerge, including geometric transformations [25] and attacks on attention mechanisms [30]. Nonetheless, conventional scene-agnostic UAPs falter in multi-view, geometry-aware tasks like 3D reconstruction. They lack sufficient capacity and explicit cross-view modeling.

2.2 Adversarial attack in 3D

3D foundation models transform multi-view reconstruction tasks. However, their complex architectures introduce unique adversarial vulnerabilities. Traditional ℓ_p -bounded perturbations effectively disrupt 2D CNNs with imperceptible noise that alters classification outputs [21]. Similarly, attacks like NeRFool [11] target generalizable NeRFs [5, 55]. These rely on iterative optimization, which is incompatible with feed-forward 3D models due to pipeline differences. Universal Adversarial Perturbations (UAPs) [41] improve efficiency via input-agnostic noise. However, they fail to generalize across diverse multi-view inputs in 3D tasks. They lack mechanisms to handle geometric and temporal dependencies [34]. For instance, NerFail [24] applies multi-view attacks on NeRFs but struggles with varying viewpoints. This highlights the limitations of static perturbations in dynamic 3D settings. More broadly, INR-based representations have also been studied for security-oriented information hiding and steganography, reflecting growing interest in the security of neural scene representations [48, 51]. Recent work explores more targeted 3D attacks. Poison-splat [32] poisons input images to increase training costs in 3D Gaussian Splatting (3DGS) [26]. MPAM-3DGS [23] manipulates Gaussian parameters to degrade reconstruction quality. Targeted attacks [18] on generalizable NeRFs mislead downstream tasks like object detection [6, 15, 33, 59]. Recent studies have investigated copyright protection for 3D representations through reconstruction prevention and ownership verification [35, 49, 50]. Despite their effectiveness, most methods require iterative optimization [37]. This incurs a high computational cost with multi-view inputs and large models such as VGGT [54]. Consequently, they remain impractical for real-world deployment.

3 Background

Visual geometry grounded transformer. VGGT [54] leverages a vision foundation transformer model [42] combined with DPT [44] to achieve 3D scene reconstruction inference in seconds [9]. VGGT integrates tasks such as camera pose estimation, depth prediction, and point cloud reconstruction in a single feed-forward pass, using cross-view feature consistency to generate viewpoint-invariant 3D representations. Formally, VGGT can be regarded as a function $\mathcal{F}$ that processes a multi-view image sequence $\mathcal{S} = \{I_t \in \mathbb{R}^{H \times W \times 3}\}_{t=1}^T$ to produce geometric outputs as:

$$\mathbf{Y} = \mathcal{F}(\mathcal{S}; \phi), \quad (1)$$

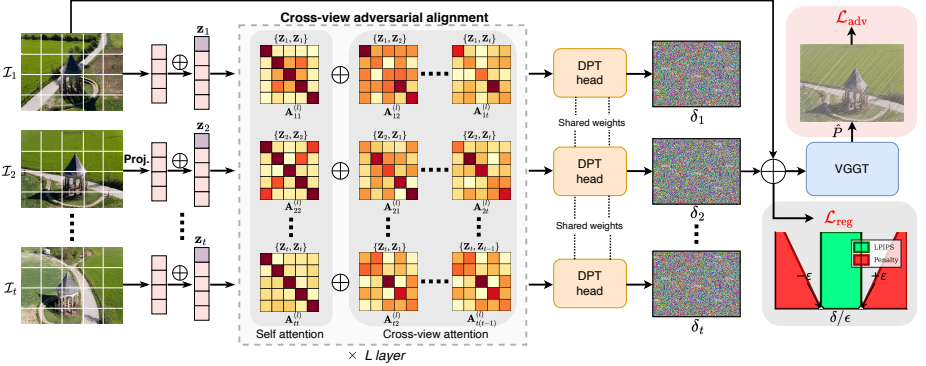


Fig. 2: Overview of the Multi-view Adversarial Perturbation Generator (MVAP-G). The generator $\mathcal{G}$ takes a multi-view image sequence $\mathcal{S}$ and produces perturbations δ_t for each frame in a single feed-forward pass. The perturbed images $\hat{I}_t = I_t + \delta_t$ are then fed into the pre-trained VGGT model $\mathcal{F}$ to produce adversarial 3D point clouds $\hat{P}$. The generator is built on a transformer architecture with cross-view adversarial alignment (CAA) and a DPT head for generating cross-view pixel-level perturbations.

where ϕ denotes the parameters of the pretrained vision transformer [4] and DPT backbone [44]. $\mathbf{Y} = (P, \{D_t\}_{t=1}^T, \{\theta_t\}_{t=1}^T)$ contains the 3D attributes, including point clouds $P \in \mathbb{R}^{N \times 3}$, depth maps $\{D_t \in \mathbb{R}^{H \times W}\}_{t=1}^T$, and camera poses $\{\theta_t \in \mathbb{R}^6\}_{t=1}^T$. VGGT effectively processes multi-view images and produces high-quality 3D content in a single forward pass, establishing a new stage in robust 3D foundation modeling.

Problem setup. VGGT enables unified, feed-forward reconstruction of 3D attributes from arbitrary multi-view image sequences, powering downstream applications such as SLAM [38], spatial tracking [58], and real-time 3D inference [9, 47, 53]. However, adversarial attacks pose a critical threat in real-world deployments, where subtle perturbations can distort depth maps or camera poses, undermining the safety of safety-critical systems.

Formally, given a clean multi-view sequence $\mathcal{S} = \{I_t \in \mathbb{R}^{H \times W \times 3}\}_{t=1}^T$, an adversarial perturbation δ is added to induce significant errors in VGGT’s output $\mathcal{F}(\mathcal{S})$. The attack objective is defined as:

$$\delta^* = \arg \max_{\delta} \mathcal{L}_{\text{adv}}(\mathcal{F}(\mathcal{S} + \delta), \mathcal{F}(\mathcal{S})), \quad \text{s.t.} \quad \|\delta\|_p \leq \epsilon, \quad (2)$$

where ϵ enforces imperceptibility under an ℓ_p -norm constraint, and $\mathcal{L}_{\text{adv}}$ quantifies the deviation in 3D outputs.

Traditional methods like FGSM [14] and PGD [37] require per-scene iterative optimization, rendering them impractical for real-time, dynamic environments. Universal Adversarial Perturbations (UAPs) [41] eliminate iteration by applying a fixed pattern across inputs, but fail to model multi-view geometry, resulting in poor transferability against transformer-based 3D models like VGGT.

These shortcomings highlight the need for single-pass perturbations that preserve cross-view consistency and scale to real-world conditions. Our MVAP-G

framework meets this demand by generating input-conditioned, aligned cross-view adversarial noise in a single feed-forward pass. This design enables efficient, transferable attacks on VGGT in real-time deployment scenarios.

4 Proposed method

The key challenge in generating adversarial examples for VGGT [54] is creating perturbations that are both effective and computationally efficient. We address this with the Multi-view Adversarial Perturbation Generator (MVAP-G), which learns an input-conditioned mapping to generate multi-view adversarial examples against VGGT in a single feed-forward pass (as shown in Fig. 2).

4.1 Multi-view perturbation generator

MVAP-G takes a multi-view image sequence $\mathcal{S} = \{I_t \in \mathbb{R}^{H \times W \times 3}\}_{t=1}^T$ as input and generates adversarial examples in a single feed-forward pass. These perturbations are optimized to disrupt the 3D attribute point clouds $P \in \mathbb{R}^{N \times 3}$ generated by the pretrained VGGT model $\mathcal{F}$. The adversarial objective is formulated to maximize the loss function $\mathcal{L}_{\text{adv}}$, thereby degrading the performance of VGGT $\mathcal{F}$:

$$\delta^* = \arg \max_{\{\delta_t\} \in \Delta} \mathcal{L}_{\text{adv}}(\mathcal{F}(\{I_t + \delta_t\}), \mathcal{F}(\mathcal{S})), \quad (3)$$

where the perturbations are generated via the MVAP-G generator $\mathcal{G}$ as:

$$\{\delta_t\} = \mathcal{G}(\mathcal{S}; \mathbf{w}), \quad (4)$$

with $\mathbf{w}$ representing the learnable parameters of the generator. The constraint set Δ ensures imperceptibility by enforcing $\|\delta_t\|_{\infty} \leq \epsilon$.

The MVAP-G is built on a transformer architecture [10, 42, 44] with introduced cross-view adversarial alignment (CAA) and a DPT head for generating cross-view pixel-level perturbations. For an input sequence $\mathcal{S}$, each frame I_t is divided into N_p 16×16 patches, projected into latent vectors $\mathbf{e} \in \mathbb{R}^D$ ($D = 384$). Spatio-temporal relationships are encoded via *learnable* camera embeddings $\mathbf{p}_t$, yielding:

$$\mathbf{z}_t^{(0)} = [\mathbf{e}_{t,1}, \mathbf{e}_{t,2}, \dots, \mathbf{e}_{t,N_p}] + \mathbf{p}_t, \quad (5)$$

where $\mathbf{e}_{t,i}$ denotes the embedding vector of the i -th patch in the t -th view, and $[\cdot]$ represents the concatenation operation. These initial embeddings, $\mathbf{z}_t^{(0)}$, are then processed through multiple transformer layers, each incorporating the proposed cross-view adversarial alignment mechanism to enable information exchange across views.

Cross-view adversarial alignment. A cross-view adversarial alignment (CAA) module is proposed to facilitate understanding across multiple views. The attention mechanism combines both self-attention within each view and cross-attention across different views:

$$\mathbf{z}_t^{(l)} = \text{LayerNorm}(\mathbf{z}_t^{(l-1)} + \mathbf{A}_t^{(l)}), \quad (6)$$

where the attention output $\mathbf{A}_t^{(l)}$ for view t is the sum of attention contributions from all views:

$$\mathbf{A}_t^{(l)} = \sum_{k=1}^T \mathbf{A}_{tk}^{(l)}, \quad (7)$$

with each component computed as:

$$\mathbf{A}_{tk}^{(l)} = \text{softmax}\left(\frac{\mathbf{z}_t^{(l-1)}(\mathbf{z}_k^{(l-1)})^\top}{\sqrt{D}}\right)\mathbf{z}_k^{(l-1)}. \quad (8)$$

When $t = k$, $\mathbf{A}_{tt}^{(l)}$ represents self-attention within the same view. When $t \neq k$, $\mathbf{A}_{tk}^{(l)}$ represents cross-view attention between different views. This formulation enables comprehensive information exchange across the multi-view sequence. After processing the multi-view features through the transformer layers with CAA, the resulting feature representations are passed to the Dense Prediction Transformer (DPT) head to generate pixel-level perturbations.

Dense prediction transformer head. To obtain pixel-level perturbations from patch tokens, we adopt a Dense Prediction Transformer (DPT) head [44]. Specifically, we extract features from the 3rd, 6th, 9th, and 12th transformer layers using four learned readout projections:

$$\mathbf{h}_t^{(l)} = \mathbf{z}_t^{(l)} W_{\text{readout}}^{(l)} \in \mathbb{R}^{N_p \times D}, \quad l \in \{3, 6, 9, 12\}, \quad (9)$$

where $N_p = (H/16) \times (W/16)$ is the number of patches, and $D=384$ is the embedding dimension. Following DPT [44, 61], these multi-scale features are progressively fused through four Reassemble stages ($s=1\sim 4$). Each stage up-samples to resolution $H/2^s \times W/2^s$ by concatenating the input feature with $K_s \in \{1, 4, 16, 64\}$ learnable fusion tokens and applying a linear projection.

After the final stage, a lightweight refinement module (3×3 depth-wise + 1×1 convolution) outputs the unbounded perturbation:

$$\tilde{\delta}_t = \text{Refine}(\mathbf{r}_t^{(4)}) \in \mathbb{R}^{H \times W \times 3}, \quad (10)$$

where $\mathbf{r}_t^{(4)} \in \mathbb{R}^{H \times W \times D}$ denotes the full-resolution feature map. The perturbation is constrained *only during the inference* within the ℓ_∞ -ball via:

$$\delta_t = \text{clip}(\tilde{\delta}_t, -\epsilon, \epsilon), \quad (11)$$

which guarantees $\|\delta_t\|_\infty \leq \epsilon$ by clipping the unbounded perturbation to the predefined budget. MVAP-G adopts a multi-scale feature-decoding [44] strategy similar to VGGT to process hierarchical features. This enables stable pixel-level perturbation generation, producing visually imperceptible noise that remains perfectly consistent across all views.

4.2 Training objective

Given a pre-trained VGGT model $\mathcal{F}$ and our MVAP-G generator $\mathcal{G}$, the training objective follows a step-by-step adversarial process to optimize the generator $\mathcal{G}$

Algorithm 1: MVAP-G training procedure

Input : VGGT model $\mathcal{F}$, multi-view image dataset $\mathcal{D}_{\text{train}}$, Learning rate α ,
 Perturbation bound ϵ , regularization coefficient λ , number of
 iterations N_{iter}

Output: MVAP-G ($\mathcal{G}$) parameters $\mathbf{w}$

- 1 Initialize MVAP-G generator: $\mathcal{G}_{\mathbf{w}}$;
- 2 **for** $iter$ in 0 to N_{iter} **do**
 - // Batch multi-view sequences
 - 3 $\mathcal{S}_{\text{batched}} \leftarrow \text{Batch}(\mathcal{D}_{\text{train}})$;
 - // Generate perturbations
 - 4 $\Delta = \{\delta_t\}_{t=1}^T \leftarrow \mathcal{G}_{\mathbf{w}}(\mathcal{S})$;
 - // Create perturbed sequence
 - 5 $\hat{\mathcal{S}} \leftarrow \{I_t + \delta_t\}_{t=1}^T$;
 - // Compute adv. point cloud
 - 6 $\hat{P} \leftarrow \mathcal{F}(\hat{\mathcal{S}})$;
 - // Compute adv. loss
 - 7 $\mathcal{L}_{\text{adv}} \leftarrow \frac{1}{|\hat{P}|} \sum_{p \in \hat{P}} \|p\|_2$;
 - // Compute regularization loss
 - 8 $\mathcal{L}_{\text{reg}} \leftarrow \text{ReLU}(\cdot) + \text{LPIPS}(\cdot)$;
 - // Compute total loss
 - 9 $\mathcal{L}_{\text{total}} \leftarrow \mathcal{L}_{\text{adv}} + \lambda \mathcal{L}_{\text{reg}}$;
 - // Update parameters
 - 10 $\mathbf{w} \leftarrow \mathbf{w} - \alpha \cdot \nabla_{\mathbf{w}}(\mathcal{L}_{\text{total}})$;
- 11 **return** $\mathbf{w}$;

for disrupting the 3D reconstruction model $\mathcal{F}$. First, each input image I_t from the multi-view sequence $\mathcal{S} = \{I_t \in \mathbb{R}^{H \times W \times 3}\}_{t=1}^T$ is fed into the generator $\mathcal{G}$, parameterized by weights $\mathbf{w}$, to produce per-frame perturbations $\{\delta_t\}_{t=1}^T$ via $\{\delta_t\} = \mathcal{G}(\mathcal{S}; \mathbf{w})$. The perturbed sequence $\hat{\mathcal{S}} = \{I_t + \delta_t\}_{t=1}^T$ is processed by VGGT to generate the adversarial output:

$$\hat{\mathbf{Y}} = \mathcal{F}(\hat{\mathcal{S}}). \quad (12)$$

where $\hat{\mathbf{Y}} = (\hat{P}, \{\hat{D}_t\}_{t=1}^T, \{\hat{\theta}_t\}_{t=1}^T)$. We target the output point cloud $\hat{P}$ as the primary attack objective, given its central role in VGGT's 3D reconstruction pipeline [54] (We also discuss the impacts on other outputs of VGGT in Sec. 5.2). To effectively degrade the output, the adversarial loss $\mathcal{L}_{\text{adv}}$ is formulated as the mean L2 norm of the adversarial point cloud coordinates, designed to shrink the point cloud to the origin $(0, 0, 0)$ [49]:

$$\mathcal{L}_{\text{adv}} = \mathbb{E}_{\mathcal{S} \sim \mathcal{D}_{\text{train}}} \left[\frac{1}{|\hat{P}|} \sum_{p \in \hat{P}} \|p\|_2 \right]. \quad (13)$$

This simple yet effective metric efficiently disrupts the 3D structure (further discussed in Sec. 5.2).

To constrain the generated perturbations δ , we apply a multi-component regularization loss *during the training*:

$$\mathcal{L}_{\text{reg}} = \sum_{t=1}^T \left[\text{LPIPS}(\hat{I}_t, I_t) + \text{ReLU}(\|\delta_t\|_{\infty} - \epsilon) \right], \quad (14)$$

where $\text{LPIPS}(\cdot, \cdot)$ [64] measures perceptual similarity in deep feature space. The ReLU term enforces the hard ℓ_{∞} constraint $\|\delta_t\|_{\infty} \leq \epsilon$.

The total objective combines both terms with a balancing coefficient λ :

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{adv}} + \lambda \mathcal{L}_{\text{reg}}. \quad (15)$$

The generator $\mathcal{G}$ parameters $\mathbf{w}$ are optimized using AdamW [31] to minimize $\mathcal{L}_{\text{total}}$ over the training dataset $\mathcal{D}_{\text{train}}$, which contains diverse multi-view sequences for robust perturbation learning.

4.3 Implementation details

MVAP-G undergoes a two-stage training procedure. First, the model is pre-trained on single-view images from the COCO dataset [29] for 4 million iterations using AdamW [31] optimizer with a learning rate of 1×10^{-5} . The regularization loss is introduced after 2.4 million iterations to allow stable feature learning. This single-view pre-training enables the generator to learn robust image features for effective perturbation generation before multi-view adaptation. Subsequently, the model is fine-tuned on multi-view imagery. Each training batch samples 2-10 frames randomly from a scene. The multi-view training corpus comprises seven diverse 3D datasets: CO3Dv2 [45], BlendMVS [62], ScanNet [7], Virtual KITTI [12], DTU [22], ETH3D [46], and FlyingThings3D [39]. These datasets encompass both indoor/outdoor environments and synthetic/real-world data, providing diversity similar to MAST3R [27] and VGGT [54]. Following VGGT’s preprocessing procedure, input frames are resized to a maximum dimension of 518 pixels. The complete training requires approximately 20 days on a single NVIDIA RTX 4090 GPU. More implementation details can be found in the *supplemental material*.

5 Experiments

Setup. To evaluate the effectiveness of our proposed MVAP-G framework, we conduct extensive experiments targeting the Visual Geometry Grounded Transformer (VGGT) [54]. We assess the attack performance on point cloud reconstruction and related tasks, measuring degradation using the Chamfer Distance (CD) [3] for point clouds and the Mean Squared Error (MSE) for the reconstructed point cloud. The evaluation spans four datasets: COCO [29], a large-scale dataset with single-view RGB images to test generalization to real-world scenes; ImageNet [8], a diverse single-view dataset to evaluate perturbation transferability across object categories; Co3D [45], a real-world multi-view

Table 1: Attack impacts on VGGT via different perturbation methods. We report Shifted CD distances between the reconstructed 3D point clouds from the clear and perturbed samples. Higher CD values denote a greater shifted distance and indicate better attack performance. MVAP-G achieves superior attack performance with **only a single feed-forward iteration**.

Method	COCO [29]	ImageNet [8]	LLFF [40]					CO3D [45]				
	$v = 1$	$v = 1$	$v = 1$	$v = 4$	$v = 8$	$v = 16$	$v = 25$	$v = 1$	$v = 4$	$v = 8$	$v = 16$	$v = 25$
Gaus. noise	0.534	0.505	0.912	0.900	0.941	0.949	0.982	0.678	0.666	0.671	0.632	0.672
UAP	0.565	0.556	0.940	0.964	0.980	1.010	1.043	0.673	0.675	0.671	0.673	0.659
AP (iter=10)	1.361	1.349	1.841	1.769	1.803	1.870	1.834	1.442	1.718	1.890	1.643	1.913
AP (iter=20)	1.535	1.584	1.842	1.811	1.817	1.882	1.925	1.889	1.892	1.958	1.875	2.00
MVAP-G (iter=1)	1.566	1.611	1.975	1.885	1.857	1.922	1.910	1.995	2.016	2.011	1.999	2.047

dataset contains common 3D object; and LLFF [40], a real-world multi-view dataset capturing dynamic scenes with complex lighting and geometry.

Benchmarks. MVAP-G is compared against several baseline methods: Random Noise, which adds Gaussian noise to input images as a naive baseline; Universal Adversarial Perturbations (UAP) [41], a scene-agnostic attack method; and naive adversarial attacks, *i.e.*, AP (iter=10 and iter=20) [37], optimized with steps of 10 and 20 iterations, respectively. We evaluate performance using Chamfer Distance (CD) [3]. We report the shifted CD distance between outputs from perturbed and clear multi-view images. For multi-view datasets, performance is assessed across varying numbers of views $v \in \{1, 4, 8, 16, 25\}$. Computational efficiency is measured by the inference time and peak GPU memory per view.

5.1 Main results

Attack performance. Table 1 compares MVAP-G against baselines on point cloud reconstruction. Random noise yields a negligible Chamfer Distance (CD), confirming its ineffectiveness. UAP [41] improves over noise but degrades sharply in multi-view settings due to the absence of cross-view modeling. Naive adversarial perturbation, AP (iter=10, iter=20) [37] achieves stronger disruption, with marginal gains from additional iterations. In contrast, MVAP-G consistently outperforms all baselines across datasets and view counts, achieving the highest CD in every configuration using only a single feed-forward pass. Notably, on multi-view CO3D and LLFF (Fig. 6), MVAP-G maintains robust attack strength, far surpassing UAP and closely competing with or exceeding iterative AP, demonstrating effective exploitation of VGGT vulnerabilities. Additional results are available in the *supplemental material*.

Efficiency. MVAP-G generates universal multi-view adversarial perturbations in a single feed-forward pass, achieving real-time inference and low memory footprint across increasing view counts. As shown in Fig. 4 (a), inference time scales sub-linearly with the number of views, remaining in the millisecond range even for large inputs. In contrast, Naive AP [37] requires per-scene iterative optimization [37], leading to exponential runtime growth and reaching minutes per

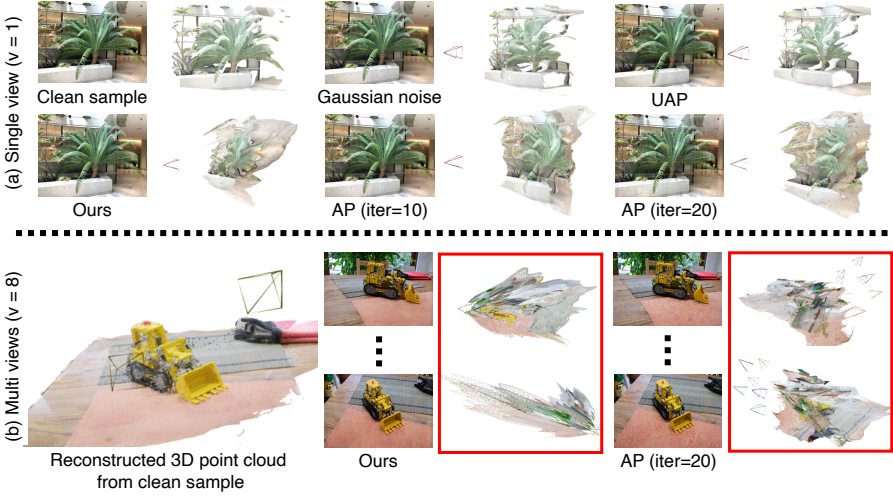


Fig. 3: 3D point cloud reconstruction under adversarial attacks for VGGT. (a) **Single-view** ($v=1$): clean input and its ground truth; reconstructions from five adversarial perturbations, including Gaussian, UAP [41], AP with iter=10 and iter=20 [37], and our MVAP-G. (b) **Multi-views** ($v=8$): comparison between reconstructions from AP (iter=20) and our MVAP-G, demonstrating our method’s superior effectiveness in a practical setting. Our MVAP-G causes severe degradation with a single pass. More results can be found in the supplemental material.

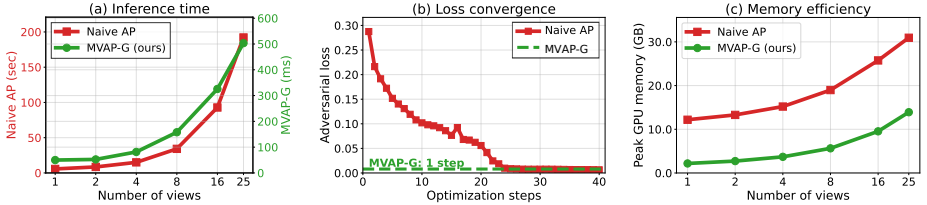


Fig. 4: MVAP-G achieves real-time, memory-efficient, and high-quality multi-view attacks with a single feed-forward pass. (a) **Inference time** demonstrates that Naive AP requires over 25–200s per scene, while MVAP-G achieves millisecond-level attack in 50–500ms across multi-views input in a single forward pass. (b) **Loss convergence** shows that Naive AP needs more than 25 steps to reach a loss of 0.008, whereas MVAP-G matches this performance in one step. (c) **Memory efficiency** demonstrates that MVAP-G uses less than 14 GB of GPU memory, compared to 30 GB for Naive AP.

scene at high view counts. Similarly, Fig. 4 (c) reveals that MVAP-G maintains significantly lower peak GPU memory than Naive AP, which exceeds 30 GB under heavy multi-view loads. These results highlight MVAP-G’s superior scalability, enabling efficient, real-time attacks on VGGT without the computational overhead of scene-specific AP [37].

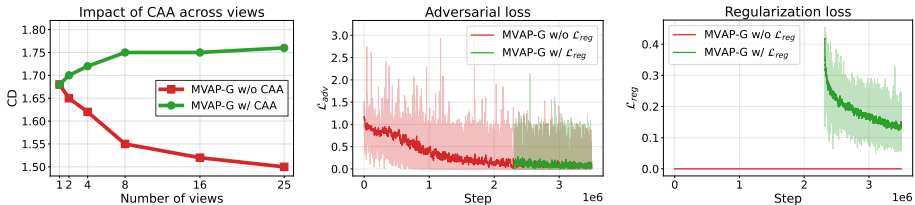


Fig. 5: (a) **Cross-view adversarial alignment (CAA).** Attack strength remains steady even with more input views. Yet, w/o CAA faces inadequate attack performance against VGGT, especially in multi-view inputs. (b) **Adversarial loss.** The adversarial loss $\mathcal{L}_{adv}$ converges slowly with oscillations, aligning with the training process of generative adversarial networks [1]. (c) **Regularization loss.** The convergence of $\mathcal{L}_{reg}$ indicates effective control over perturbation magnitude, ensuring visual imperceptibility while the adversarial loss remains low, thus achieving a balance between stealth and attack efficacy.

Table 2: Ablation study on key components of MVAP-G. Pre-training provides the largest boost in attack strength. The DPT outperforms the naive decoder, as it captures more features for pixel-level adversarial perturbations [44]. CAA enables multi-view performance (more in Fig. 5 (a)). $\mathcal{L}_{reg}$ minimize the size of perturbation and ensures perturbation invisibility.

Configuration	Attack performance		Invisibility	
	CD $\uparrow$	$\mathcal{L}_{adv}$ $\downarrow$	PSNR $\uparrow$	LPIPS $\downarrow$
Baseline	0.58	0.400	21.1	0.400
+ Pre-training	1.40	0.250	20.8	0.435
+ DPT head	1.68	0.120	20.5	0.478
+ CAA	1.75	0.008	20.3	0.512
+ $\mathcal{L}_{reg}$	1.78	0.008	29.8	0.107

5.2 Ablation studies

Component ablation. To understand the contributions of MVAP-G’s components, we conduct ablation studies to systematically evaluate the impact of each key module. Table 2 reports incremental ablation results averaged over 1, 2, and 4 views. The baseline configuration uses a Conv+Linear decoder without pre-training, CAA, or $\mathcal{L}_{reg}$. Pre-training on COCO [29] provides a boost in attack strength, as it can learn adversarial knowledge against VGGT. The DPT head outperforms the naive decoder by capturing richer features for pixel-level ad-



Fig. 6: Adversarial examples and visualized $\times 20$ perturbation.

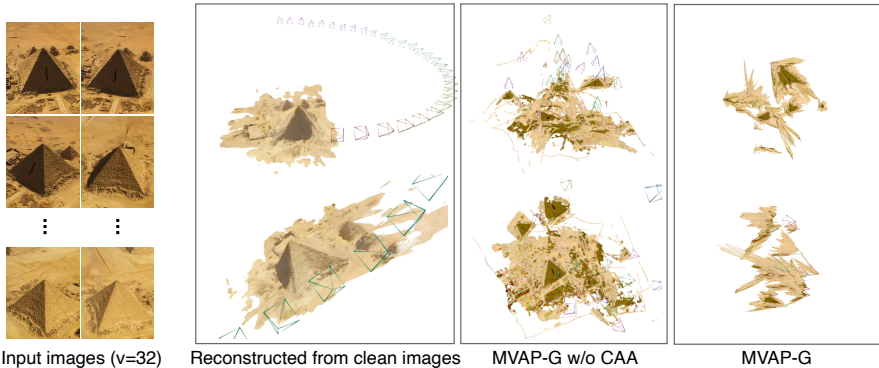


Fig. 7: Impacts of CAA when input 32 view images. *Without CAA*, perturbations allow VGGT to partially recover geometry results, yielding fragmented but recognizable structures.

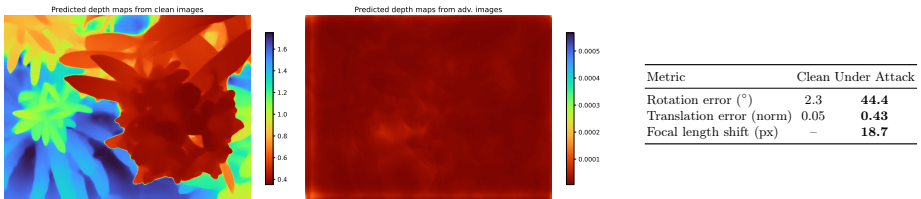


Fig. 8: (left) Impacts on depth. (right) Impacts on camera pose. Camera pose under attack (CO3D $v=8$, avg. 100 scenes) ([please zoom in for best review](#)).

versarial perturbations [44]. CAA plays a limited role in increasing performance under limited review. However, it could enable robust performance in multi-view scenarios, with its benefits growing significantly as the number of views increases (as shown in Fig. 5 (a)). Finally, $\mathcal{L}_{\text{reg}}$ effectively minimizes perturbation size and ensures invisibility without compromising attack efficacy.

Cross-view adversarial alignment (CAA). Fig. 5 (a) shows attack performance (CD) versus input view count. Without CAA, performance degrades as view counts increase due to misalignment perturbations across views. In contrast, our full model with CAA leverages additional views to steadily enhance attack strength by exploiting multi-view geometry. These results demonstrate that CAA achieves competitive performance across multi-view scenarios. To further illustrate the effectiveness of CAA, we visualize results at view number $v=32$. *Without CAA*, perturbations allow VGGT to partially recover geometry through multi-view attention fusion, yielding fragmented but recognizable structures.

Impact on Depth/Pose. Using the **same MVAP-G model** (trained *only* on point cloud loss $\mathcal{L}_{\text{adv}}$), we observe significant degradation on VGGT’s other geometric outputs (*i.e.*, depth maps and camera poses) as shown in Fig. 8. This occurs because VGGT’s outputs share a common backbone that captures key geometric information.

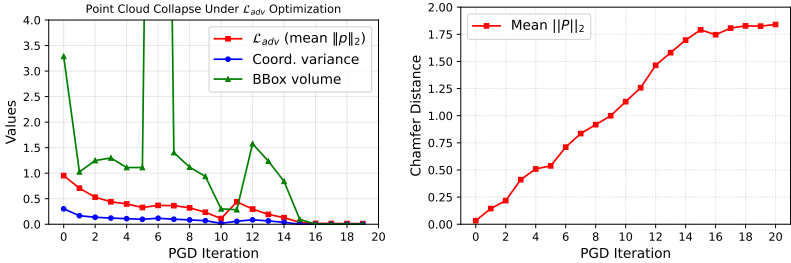


Fig. 9: Statistics under $\mathcal{L}_{adv}$. (Left) The point cloud collapses to the origin (0, 0, 0) via supervision of adversarial loss $\mathcal{L}_{adv}$. (Right) The CD distance between the reconstructed point cloud and the original increases in steps.

Regularization loss $\mathcal{L}_{reg}$. To ensure the invisibility of the generated adversarial perturbation, the regularization loss $\mathcal{L}_{reg}$ was introduced after 2.4×10^6 training steps. Fig. 5 (b) compares training dynamics with and without $\mathcal{L}_{reg}$. Regularized training yields stable optimization, reliably reaching low $\mathcal{L}_{adv}$ with improving perturbation invisibility. By limiting perturbation magnitude, $\mathcal{L}_{reg}$ encourages precise, minimal noise that retains full attack strength, demonstrating that effective disruption of VGGT stems from targeted precision, not large-scale distortion. Fig. 5 (c) and visualized perturbation in Fig. 6 further show that $\mathcal{L}_{reg}$ produces visually imperceptible, texture-preserving perturbations in input images, while effectively compromising VGGT’s inference. More examples and analysis are provided in the *supplemental material*.

Effectiveness of adversarial loss $\mathcal{L}_{adv}$. We conduct a case study on a PGD optimization of $\mathcal{L}_{adv}$ (*i.e.*, min mean $\|P\|_2$). As shown in Fig. 9, PGD optimization of $\mathcal{L}_{adv}$ increases the shifted CD from 0.02 to 1.82 within 20 iterations, while significantly reducing point cloud variance. This confirms that $\mathcal{L}_{adv}$ can effectively collapse the point cloud.

6 Conclusion

In this work, we present the first systematic study of adversarial vulnerabilities in 3D foundation models, focusing on the Visual Geometry Grounded Transformer (VGGT). To address the limitations of traditional adversarial attack methods, we propose MVAP-G, a novel multi-view adversarial perturbation generator that produces multi-view-consistent adversarial perturbations in real time. Our method eliminates the need for per-scene optimization through a feed-forward generation pipeline, while explicitly maintaining adversarial performance across multi-view images. This work exposes critical security risks in 3D foundation models for adversarial robustness research.

Limitations and future work. Our method poses a serious threat to real-world VGGT deployment. This powerful attack could be misused to compromise safety-critical systems, such as autonomous driving or robotic navigation, causing severe real-world harm. Future work should prioritize the development of robust defenses and detection mechanisms, as well as responsible policy guidelines to prevent misuse.

Acknowledgements

Renjie Group is supported by the National Natural Science Foundation of China under Grant No. 62302415, Guangdong Basic and Applied Basic Research Foundation under Grant No. 2024A1515012822, and the Research Grants Council (RGC) of Hong Kong SAR, China, under a GRF Grant 12203124 and an ECS Grant 22201125. Ziyuan Luo is supported by a fellowship award from the RGC of Hong Kong SAR, China (Project No. HKBU JRFS2627-2S03).

References

1. Arjovsky, M., Chintala, S., Bottou, L.: Wasserstein generative adversarial networks. In: ICML (2017)
2. Baluja, S., Fischer, I.: Adversarial transformation networks: Learning to generate adversarial examples. In: AAAI (2018)
3. Barrow, H., Tenenbaum, J., Bolles, R., Wolf, H.: Parametric correspondence and chamfer matching: two new techniques for image matching. In: IJCAI (1977)
4. Caron, M., Touvron, H., Misra, I., Jégou, H., Mairal, J., Bojanowski, P., Joulin, A.: Emerging properties in self-supervised vision transformers. In: CVPR (2021)
5. Chen, A., Xu, Z., Zhao, F., Zhang, X., Xiang, F., Yu, J., Su, H.: Mvsnerf: Fast generalizable radiance field reconstruction from multi-view stereo. In: ICCV (2021)
6. Chen, S.T., Cornelius, C., Martin, J., Chau, D.H.: Shapeshifter: Robust physical adversarial attack on faster r-cnn object detector. In: ECML-PKDD (2018)
7. Dai, A., Chang, A.X., Savva, M., Halber, M., Funkhouser, T., Nießner, M.: Scannet: Richly-annotated 3D reconstructions of indoor scenes. In: CVPR (2017)
8. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: CVPR (2009)
9. Deng, K., Ti, Z., Xu, J., Yang, J., Xie, J.: Vggt-long: Chunk it, loop it, align it—pushing vggt’s limits on kilometer-scale long rgb sequences. arXiv preprint arXiv:2507.16443 (2025)
10. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. In: ICLR (2021)
11. Fu, Y., Yuan, Y., Kundu, S., Wu, S., Zhang, S., Lin, Y.: Nerfool: uncovering the vulnerability of generalizable neural radiance fields against adversarial perturbations. In: ICML (2023)
12. Gaidon, A., Wang, Q., Cabon, Y., Vig, E.: Virtual worlds as proxy for multi-object tracking analysis. In: CVPR (2016)
13. Goodfellow, I., Warde-Farley, D., Mirza, M., Courville, A., Bengio, Y.: Maxout networks. In: ICML (2013)
14. Goodfellow, I.J., Shlens, J., Szegedy, C.: Explaining and harnessing adversarial examples. In: ICLR (2018)
15. Han, H., Song, Q., Qi, J., Wan, R.: Quantum-inspired frequency attenuation for enhanced targeted fabrication attacks in object detection. In: ICASSP (2026)
16. Hayes, J., Danezis, G.: Learning universal adversarial perturbations with generative models. In: IEEE SPW (2018)
17. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: CVPR (2016)

18. Horváth, A., Józsa, C.M.: Targeted adversarial attacks on generalizable neural radiance fields. In: ICCVW (2023)
19. Hu, Y., Zou, Y., Li, R., Li, Y.: Generate universal adversarial perturbations for few-shot learning. In: NeurIPS (2024)
20. Huang, H., Erfani, S.M., Li, Y., Ma, X., Bailey, J.: X-transfer attacks: Towards super transferable adversarial attacks on clip. In: ICML (2025)
21. Huang, Y., Guo, Q., Juefei-Xu, F., Hu, M., Jia, X., Cao, X., Pu, G., Liu, Y.: Texture re-scalable universal adversarial perturbation. IEEE TIFS (2024)
22. Jensen, R., Dahl, A., Vogiatzis, G., Tola, E., Aanaes, H.: Large scale multi-view stereopsis evaluation. In: CVPR (2014)
23. Jiang, W., Zhang, H., Wang, W., Guo, Z., Zhang, T., Wang, H.: Mpam-3dgs: Multi-parametric adversarial manipulation for 3D gaussian splatting. In: ICASSP (2025)
24. Jiang, W., Zhang, H., Wang, X., Guo, Z., Wang, H.: Nerfail: Neural radiance fields-based multiview adversarial attack. In: AAAI (2024)
25. Kanbak, C., Moosavi-Dezfooli, S.M., Frossard, P.: Geometric robustness of deep networks: analysis and improvement. In: CVPR (2018)
26. Kerbl, B., Kopanas, G., Leimkühler, T., Drettakis, G.: 3d gaussian splatting for real-time radiance field rendering. ACM TOG (2023)
27. Leroy, V., Cabon, Y., Revaud, J.: Grounding image matching in 3D with mast3r. In: ECCV (2024)
28. Lin, J., Song, C., He, K., Wang, L., Hopcroft, J.E.: Nesterov accelerated gradient and scale invariance for adversarial attacks. In: ICLR (2020)
29. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: ECCV (2014)
30. Liu, Y., An, J., Zhang, W., Wu, D., Gu, J., Lin, Z., Wang, W.: Disrupting diffusion: Token-level attention erasure attack against diffusion-based customization. In: ACM MM (2024)
31. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. In: ICLR (2019)
32. Lu, J., Zhang, Y., Shen, Q., Wang, X., YAN, S.: Poison-splat: Computation cost attack on 3D gaussian splatting. In: ICLR (2025)
33. Lu, J., Sibai, H., Fabry, E., Forsyth, D.: No need to worry about adversarial examples in object detection in autonomous vehicles. In: CVPRW (2017)
34. Lu, R., He, Z., Wen, J., Kang, X., Wang, Z.J.: Ca-uap: Content-agnostic universal adversarial perturbation for enhanced generalization. In: ICASSP (2025)
35. Luo, Z., Guo, Q., Cheung, K.C., See, S., Wan, R.: CopyRNeRF: Protecting the copyright of neural radiance fields. In: ICCV (2023)
36. Luo, Z., Song, Q., Li, H., Rocha, A., Wan, R.: Adversarially robust multimedia watermarking via data-centric optimization. PR (2026)
37. Madry, A., Makelov, A., Schmidt, L., Tsipras, D., Vladu, A.: Towards deep learning models resistant to adversarial attacks. In: ICLR (2018)
38. Maggio, D., Lim, H., Carlone, L.: Vggt-slam: Dense rgb slam optimized on the sl (4) manifold. In: NeurIPS (2025)
39. Mayer, N., Ilg, E., Hausser, P., Fischer, P., Cremers, D., Dosovitskiy, A., Brox, T.: A large dataset to train convolutional networks for disparity, optical flow, and scene flow estimation. In: CVPR (2016)
40. 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 TOG (2019)
41. Moosavi-Dezfooli, S.M., Fawzi, A., Fawzi, O., Frossard, P.: Universal adversarial perturbations. In: CVPR (2017)

42. Oquab, M., Darcet, T., Moutakanni, T., Vo, H.V., Szafraniec, M., Khalidov, V., Fernandez, P., HAZIZA, D., Massa, F., El-Nouby, A., et al.: Dinov2: Learning robust visual features without supervision. TMLR (2025)
43. Poursaeed, O., Katsman, I., Gao, B., Belongie, S.: Generative adversarial perturbations. In: CVPR (2018)
44. Ranftl, R., Bochkovskiy, A., Koltun, V.: Vision transformers for dense prediction. In: ICCV (2021)
45. Reizenstein, J., Shapovalov, R., Henzler, P., Sbordone, L., Labatut, P., Novotny, D.: Common objects in 3D: Large-scale learning and evaluation of real-life 3D category reconstruction. In: ICCV (2021)
46. Schops, T., Schonberger, J.L., Galliani, S., Sattler, T., Schindler, K., Pollefeys, M., Geiger, A.: A multi-view stereo benchmark with high-resolution images and multi-camera videos. In: CVPR (2017)
47. Shen, Y., Zhang, Z., Qu, Y., Cao, L.: Fastvggt: Training-free acceleration of visual geometry transformer. arXiv preprint arXiv:2509.02560 (2025)
48. Song, Q., Han, H., Luo, Z., Qi, J., Wan, R.: Securing inr-based steganography with quantum circuit-driven weight initialization. In: ICASSP (2026)
49. Song, Q., Luo, Z., Cheung, K.C., See, S., Wan, R.: Geometry cloak: Preventing tgs-based 3D reconstruction from copyrighted images. In: NeurIPS (2024)
50. Song, Q., Luo, Z., Cheung, K.C., See, S., Wan, R.: Protecting NeRFs' copyright via plug-and-play watermarking base model. In: ECCV (2024)
51. Song, Q., Luo, Z., Huang, X., Li, S., Wan, R.: Unified steganography via implicit neural representation. arXiv preprint arXiv:2505.01749 (2025)
52. Song, Q., Luo, Z., Wan, R.: Creating blank canvas against ai-enabled image forgery. In: AAAI (2026)
53. Wang, C.S.B., Schmidt, C., Piekenbrinck, J., Leibe, B.: Faster vggt with block-sparse global attention. arXiv preprint arXiv:2509.07120 (2025)
54. Wang, J., Chen, M., Karaev, N., Vedaldi, A., Rupprecht, C., Novotny, D.: Vggt: Visual geometry grounded transformer. In: CVPR (2025)
55. Wang, Q., Wang, Z., Genova, K., Srinivasan, P.P., Zhou, H., Barron, J.T., Martin-Brualla, R., Snavely, N., Funkhouser, T.: Ibrnet: Learning multi-view image-based rendering. In: CVPR (2021)
56. Wang, Y., Zhou, J., Zhu, H., Chang, W., Zhou, Y., Li, Z., Chen, J., Pang, J., Shen, C., He, T.: π^3 : Scalable permutation-equivariant visual geometry learning. arXiv preprint arXiv:2507.13347 (2025)
57. Wu, T., Wang, X., Qiao, S., Xian, X., Liu, Y., Zhang, L.: Small perturbations are enough: Adversarial attacks on time series prediction. Inf. Sci. (2022)
58. Xiao, Y., Wang, J., Xue, N., Karaev, N., Makarov, Y., Kang, B., Zhu, X., Bao, H., Shen, Y., Zhou, X.: Spatialtrackerv2: 3D point tracking made easy. In: ICCV (2025)
59. Xie, C., Wang, J., Zhang, Z., Zhou, Y., Xie, L., Yuille, A.: Adversarial examples for semantic segmentation and object detection. In: ICCV (2017)
60. Xie, Y., Zheng, W., Huang, H., Ye, G., Ma, X.: Towards million-scale adversarial robustness evaluation with stronger individual attacks. In: CVPR (2025)
61. Yang, L., Kang, B., Huang, Z., Zhao, Z., Xu, X., Feng, J., Zhao, H.: Depth anything v2. In: NeurIPS (2024)
62. Yao, Y., Luo, Z., Li, S., Zhang, J., Ren, Y., Zhou, L., Fang, T., Quan, L.: Blend-edmvs: A large-scale dataset for generalized multi-view stereo networks. In: CVPR (2020)
63. Zhang, J., Li, J., Huang, H., Erfani, S.M., Rubinstein, B.I.P., Liu, F.: Semantic-aware adversarial fine-tuning for CLIP. TMLR (2026)

64. Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O.: The unreasonable effectiveness of deep features as a perceptual metric. In: CVPR (2018)
65. Zhang, Y., Xu, Y., Shi, J., Zhang, L.Y., Hu, S., Li, M., Zhang, Y.: Improving generalization of universal adversarial perturbation via dynamic maximin optimization. In: AAAI (2025)
66. Zhao, P., Ram, P., Lu, S., Yao, Y., Bouneffouf, D., Lin, X., Liu, S.: Learning to generate image source-agnostic universal adversarial perturbations. In: IJCAI (2022)
67. Zhou, Z., Song, Y., Li, M., Hu, S., Wang, X., Zhang, L.Y., Yao, D., Jin, H.: Darksam: Fooling segment anything model to segment nothing. In: NeurIPS (2024)