# Fading the Digital Ink: A Universal Black-Box Attack Framework for 3DGS Watermarking Systems

**Qingyuan Zeng[1,4,5], Shu Jiang[1,4,5], Jiajing Lin[2,4,5], Zhenzhong Wang[2,4,5], Kay Chen Tan[3], Min Jiang[1,2,4,5*]**

[1] Institute of Artificial Intelligence, Xiamen University
[2] School of Informatics, Xiamen University
[3] Department of Data Science and Artificial Intelligence, The Hong Kong Polytechnic University
[4] Key Laboratory of Multimedia Trusted Perception and Efficient Computing, Ministry of Education of China, Xiamen University
[5] Key Laboratory of Digital Protection and Intelligent Processing of Intangible Cultural Heritage of Fujian and Taiwan, Ministry of Culture and Tourism, Xiamen University

## Abstract

With the rise of 3D Gaussian Splatting (3DGS), a variety of digital watermarking techniques, embedding either 1D bitstreams or 2D images, are used for copyright protection. However, the robustness of these watermarking techniques against potential attacks remains underexplored. This paper introduces the first universal black-box attack framework, the Group-based Multi-objective Evolutionary Attack (GMEA), designed to challenge these watermarking systems. We formulate the attack as a large-scale multi-objective optimization problem, balancing watermark removal with visual quality. In a black-box setting, we introduce an indirect objective function that blinds the watermark detector by minimizing the standard deviation of features extracted by a convolutional network, thus rendering the feature maps uninformative. To manage the vast search space of 3DGS models, we employ a group-based optimization strategy to partition the model into multiple, independent sub-optimization problems. Experiments demonstrate that our framework effectively removes both 1D and 2D watermarks from mainstream 3DGS watermarking methods while maintaining high visual fidelity. This work reveals critical vulnerabilities in existing 3DGS copyright protection schemes and calls for the development of more robust watermarking systems.

**Code** — https://github.com/ZengQYuan/GMEA

## Introduction

3D Gaussian Splatting (3DGS) is an emerging 3D scene representation and reconstruction technology. It has the advantages of high fidelity, fast rendering speed, and real-time rendering capabilities (Kerbl et al. 2023; Wu et al. 2024). It has broad application prospects in fields such as film production, game development, virtual reality, and autonomous driving (Zeng and Zhou 2021; Zeng et al. 2021; Liu et al. 2022; Wang et al. 2024d,e; Hong et al. 2024; Tu et al. 2025; Chen et al. 2025b; Wang et al. 2025). Given that creating a 3DGS model represents a significant investment in data acquisition, engineering expertise, and computational resources (Zhang et al. 2024a), and the copyright of 3DGS assets is prone to unauthorized distribution and malicious tampering, it is crucial to effectively protect the copyright of 3DGS assets. To meet this challenge, various invisible watermarking methods for 3DGS have been proposed (Chen et al. 2025a; Jang et al. 2025; Huang et al. 2025). They embed invisible copyright information directly into the Gaussian parameters of 3DGS models, which is subsequently extracted from the rendered images. These approaches encompass different strategies, such as encoding one-dimensional (1D) copyright strings (Chen et al. 2025a) or hiding entire two-dimensional (2D) data like logos or images as watermarks (Zhang et al. 2024b).

These 3DGS invisible watermarking methods (Chen et al. 2025a; Huang et al. 2025) need to be robust enough to truly protect the copyright of 3D assets in reality. That is, they need to maintain the integrity and detectability of watermarks in the face of various potential attacks (Gong et al. 2024; Gong, Huang, and Chen 2022; Zeng et al. 2024a; Zeng, Gong, and Jiang 2024; Zhao et al. 2024). However, so far, no research has explored the robustness of 3DGS invisible watermarks when facing attacks. Therefore, this paper aims to answer the following question: is there an attack method that can destroy these 3DGS invisible watermarks?

There are the following difficulties in destroying 3DGS invisible watermarks. First, visual fidelity must be preserved. The attacker is required to remove the watermark without noticeably degrading the quality of the 3DGS model (Zhang et al. 2020). Second, attacks typically occur in a black-box setting, where attackers often lack knowledge of the watermark's content, embedding process, and detection process (Papernot et al. 2017; Zeng et al. 2024b). This means an attacker cannot accurately locate the watermark's distribution in the Gaussian parameters or rendered images, nor can they use the watermark detector's gradients to guide the optimization process through backpropagation.

*Corresponding author: Min Jiang (e-mail: minjiang@xmu.edu.cn, zengqingyuan2022@163.com).

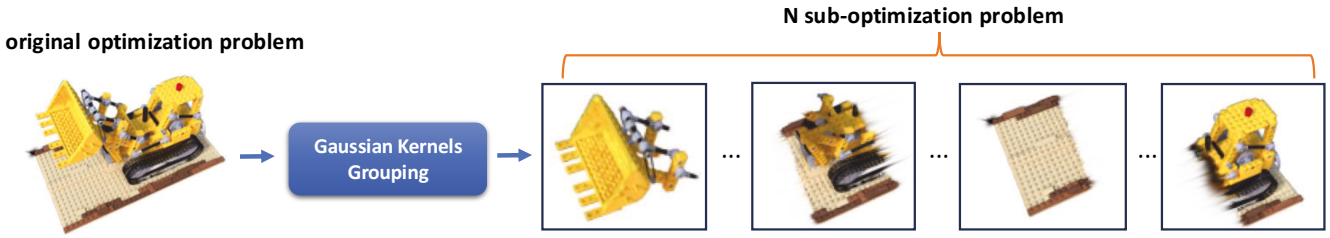


Figure 1: Illustration of the group-based optimization strategy. The original 3DGS model (full Lego object) is partitioned into multiple sub-optimization problems (individual Lego components) to manage the large search space effectively.

To narrow this research gap, we propose GMEA, a universal black-box attack framework against 3DGS invisible watermarks. Our approach formulates the attack as a large-scale multi-objective optimization problem (Wang et al. 2024d,c,a,b), seeking to simultaneously destroy the watermark while preserving the model's visual quality (Wang et al. 2021). The decision variables represent the attacker's two primary actions: selectively pruning some of the model's Gaussian kernels and subtly shifting the color values of others. The optimization objectives are: 1) minimizing visual quality degradation measured by MSE between original and perturbed renders, and 2) maximizing watermark destruction. Critically, in the black-box setting where watermark detectors are inaccessible, we introduce an indirect objective function to evaluate watermark destruction: the standard deviation of convolutional feature maps extracted from rendered images (Lu et al. 2020). Minimizing this deviation reduces discriminative information in feature maps, effectively blinding downstream watermark decoders that rely on convolutional features to extract watermark signals. This breaks the watermark extraction process despite having no detector access.

To solve this large-scale multi-objective problem, we build GMEA upon an evolutionary algorithm, inspired by its unique ability to handle such complex optimizations (Deb et al. 2002; Hong, Jiang, and Yen 2023; Liang et al. 2023; Wang et al. 2024d). However, a direct application of evolutionary algorithms is computationally inefficient. Due to the oversized search space created by the vast number of Gaussian kernels, the process of converging to an effective solution is slow. Therefore, to make the optimization tractable and efficient, we break down the large-scale optimization problem into several smaller sub-optimization problems to be solved independently (Zhang et al. 2018), as shown in Figure 1. Then, we combine the solutions of each sub-optimization problem to obtain the solution to the original optimization problem. Specifically, as shown in Figure 2, we use unsupervised clustering algorithms (such as K-Means (MacQueen 1967; Van Gansbeke et al. 2020)) to cluster the Gaussian kernels of the watermarked 3DGS model into k clusters from the perspective of position, obtaining k sub-3DGS models. Then, we perform multi-objective evolutionary algorithm on each sub-3DGS model to remove the watermark. Finally, we merge all the optimized sub-3DGS models to obtain the complete 3DGS model without watermarks. Our contributions can be summarized as follows:

1. We propose the first universal black-box attack framework GMEA against 3DGS invisible watermarks. It features a model-agnostic objective that disables watermark detection by disrupting convolutional features without requiring any knowledge of the detector.
2. We design a group-based optimization strategy that partitions the immense search space of 3DGS models, significantly improving the search efficiency of our evolutionary algorithm in discovering effective solutions.
3. We conduct extensive experiments to validate our framework's effectiveness and universality, successfully attacking leading methods for both 1D and 2D 3DGS watermarking.

## Related Works

### 3D Gaussian Splatting Watermarking

Modern 3DGS watermarking techniques imperceptibly embed copyright data while preserving visual fidelity. Approaches are typically categorized by message dimensionality: 1D bitstreams for simple copyright strings, and 2D messages for complex data like logos or images.

**1D Bitstream Watermarking.** These methods focus on embedding one-dimensional (1D) binary messages, such as copyright strings, directly into the 3DGS model. GuardSplat (Chen et al. 2025a) embeds messages by modifying existing Gaussian kernels. It utilizes small, learnable offsets to the Spherical Harmonic (SH) features to integrate the watermark, a technique designed to preserve the model's original 3D structure and visual fidelity. 3D-GSW (Jang et al. 2025) takes a refinement approach, preparing the model for watermarking through a process called Frequency-Guided Densification (FGD). This technique first prunes Gaussians that have minimal impact on rendering quality and then splits Gaussians located in high-frequency areas to maintain visual fidelity. The watermark is subsequently embedded into this optimized set of Gaussians via fine-tuning. GaussianMarker (Huang et al. 2025) employs an additive strategy, leaving original Gaussians untouched. It identifies regions of high uncertainty in the model and introduces new, dedicated Gaussian kernels called 'GaussianMarkers' within these areas to carry the watermark information.

**2D Message Watermarking.** Another approach pushes the capacity of steganography further by enabling the embedding of two-dimensional (2D) messages, such as entire

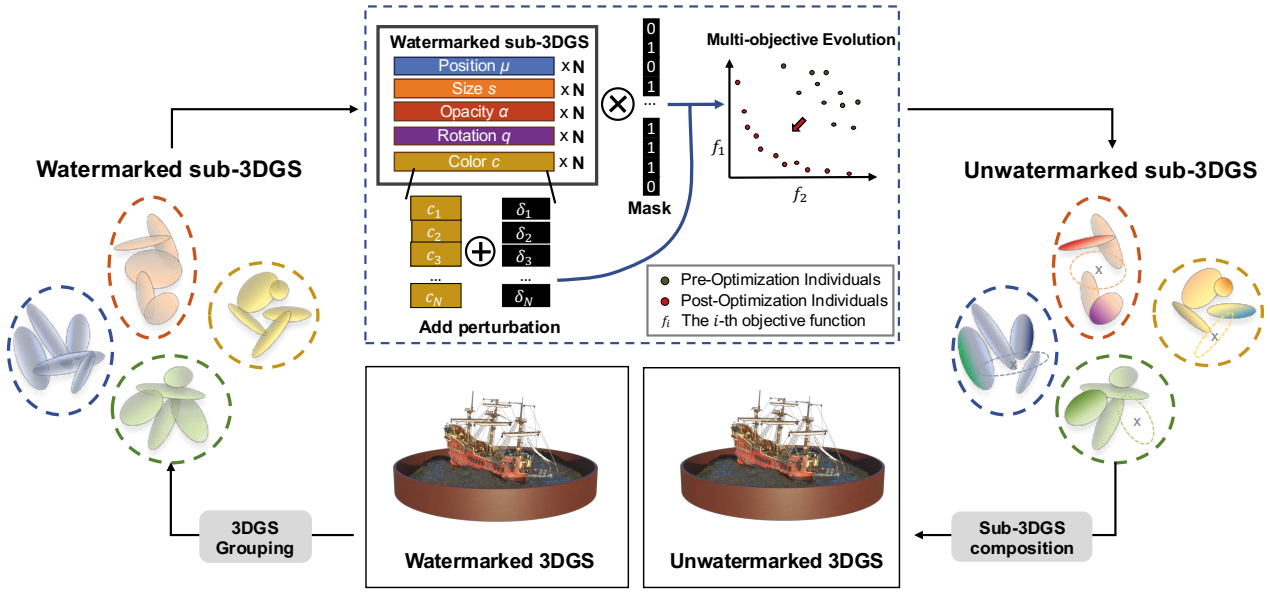


Figure 2: An overview of our proposed Group-based Multi-objective Evolutionary Attack (GMEA) framework. The attack pipeline begins by partitioning the watermarked 3DGS model into several spatially coherent sub-3DGS. Each sub-3DGS then undergoes an independent multi-objective evolutionary optimization. During this phase, potential modifications—represented by a binary mask for pruning Gaussians and a perturbation vector for altering colors—are evolved to simultaneously minimize visual quality loss ($F_1$) and maximize watermark destruction ($F_2$). The resulting optimized and unwatermarked sub-3DGS are then reassembled to form the final unwatermarked 3DGS model.

images or logos. GS-Hider (Zhang et al. 2024b) exemplifies this category. It achieves high-capacity message hiding by fundamentally altering the rendering pipeline. Instead of just modifying SH coefficients, it replaces them entirely with a coupled secured feature attribute. This high-dimensional feature is then rendered into a feature map. Two parallel decoders are employed: a public scene decoder that reconstructs the original visual scene, and a private message decoder that extracts the hidden 2D image from the same feature map. This decoupling allows for the concealment of complex messages without direct interference with the primary scene's rendering.

While the practicality of watermarking methods depends on robustness, the resilience of 3DGS schemes against dedicated attacks remains largely untested. We therefore introduce the first universal attack framework to serve as a security benchmark for the 3DGS ecosystem. Our work aims to not only assess current vulnerabilities but also to catalyze the development of more secure future solutions.

## Methodology

In this section, we detail our proposed black-box attack framework, Group-based Multi-objective Evolutionary Attack (GMEA), designed to remove invisible watermarks from 3DGS models. The overall pipeline of our method is illustrated in Figure 2, while the complete pseudocode and theoretical proofs are provided in the Appendix.

### Problem Formulation

We aim to find an adversarial 3DGS model, $\mathbf{G}_{adv}$, derived from a watermarked 3DGS model $\mathbf{G}_{wm}$. The goal is to simultaneously maintain high visual fidelity and destroy the embedded watermark. This is formulated as a multi-objective optimization problem:

$$\begin{aligned} \min_{\mathbf{G}_{adv}} \quad & F_1(\mathbf{G}_{adv}, \mathbf{G}_{wm}), F_2(\mathbf{G}_{adv}) \\ \text{s.t.} \quad & \mathbf{G}_{adv} \in \mathcal{P}(\mathbf{G}_{wm}) \end{aligned} \tag{1}$$

Here, $F_1$ measures visual quality loss and $F_2$ quantifies watermark destruction. The constraint $\mathbf{G}_{adv} \in \mathcal{P}(\mathbf{G}_{wm})$ specifies that any candidate solution $\mathbf{G}_{adv}$ must be generated by applying the allowed perturbations to $\mathbf{G}_{wm}$.

### Group-Based Optimization Strategy

Our group-based optimization strategy tackles the immense search space of 3DGS models by decomposing the task into smaller, spatially coherent sub-problems. This partitioning makes the search for an effective solution more efficient by simplifying the optimization landscape.

To achieve this partitioning, we employ the K-Means clustering on the spatial coordinates ($\mathbf{xyz}$) of the Gaussian kernels. Let the set of all 3D coordinates for the $N$ Gaussians in the watermarked model $\mathbf{G}_{wm}$ be $\mathbf{P} = \{\mathbf{p}_1, \mathbf{p}_2, \ldots, \mathbf{p}_N\}$, where $\mathbf{p}_j \in \mathbb{R}^3$. The goal of K-Means is to partition this set of kernels $\mathbf{P}$ into $k$ disjoint spatial clusters, denoted by $\mathbf{S} = \{S_1, S_2, \ldots, S_k\}$, by minimizing the within-cluster

sum of squares. The minimization criterion is formalized as:

$$\underset{S}{arg\,min} \sum_{i=1}^{k} \sum_{\mathbf{p}_j \in S_i} \|\mathbf{p}_j - \boldsymbol{\mu}_i\|^2, \tag{2}$$

where $\boldsymbol{\mu}_i$ is the geometric centroid of the coordinates in cluster $S_i$.

Once the optimal coordinate clusters $\{S_1, \ldots, S_k\}$ are determined, we map these spatial groupings back to the Gaussians' original indices. This creates a corresponding partition of the index set $\{1, 2, \ldots, N\}$ into $k$ disjoint sets $\{I_1, \ldots, I_k\}$. Each index set $I_i$ is formalized as

$$I_i = \{j \mid \mathbf{p}_j \in S_i\}. \tag{3}$$

With the indices properly partitioned, we formally construct the 3DGS sub-models. The $i$-th sub-model, $\mathbf{G}_{wm}^{(i)}$, is defined as the collection of Gaussian kernels whose indices fall into the set $I_i$:

$$\mathbf{G}_{wm}^{(i)} = \{g_j \in \mathbf{G}_{wm} \mid j \in I_i\}, \tag{4}$$

where $g_j$ represents the $j$-th Gaussian kernel. This decomposition of the large-scale task into smaller sub-problems significantly reduces search complexity, enabling a more efficient optimization.

## Multi-objective Evolutionary Attack

Each sub-problem defined by a sub-model $\mathbf{G}_{wm}^{(i)}$ is solved with a multi-objective evolutionary algorithm to find an effective adversarial solution , $\mathbf{G}_{adv}^{(i)}$. The algorithm refines a population of solutions to approximate the Pareto-optimal front, balancing the conflicting objectives of visual fidelity and watermark removal.

**Individual Representation.** Each individual in our population represents a potential modification to a sub-model $\mathbf{G}_{wm}^{(i)}$ containing $N^{(i)}$ Gaussians. The individual's genetic representation is a vector

$$\mathbf{x}^{(i)} = [\mathbf{m}^{(i)}, \mathbf{c}^{(i)}], \tag{5}$$

which concatenates a binary mask vector $\mathbf{m}^{(i)} \in \{0, 1\}^{N^{(i)}}$ and a color perturbation vector $\mathbf{c}^{(i)} \in [-\epsilon, \epsilon]^{3N^{(i)}}$. The mask vector determines which Gaussian kernels are pruned ($m_j = 0$), while the color perturbation vector defines additive shifts to the DC color component (RGB) for the remaining Gaussian kernels.

The adversarial sub-model $\mathbf{G}_{adv}^{(i)}$ is constructed by applying these modifications. Let $\mathbf{C}_{dc}$ be the $N^{(i)} \times 3$ matrix of original DC colors. The new color matrix, $\mathbf{C}_{dc,adv}$, is computed as:

$$\mathbf{C}_{dc,adv} = \text{diag}(\mathbf{m}^{(i)}) \left( \mathbf{C}_{dc} + \text{Reshape}(\mathbf{c}^{(i)}) \right). \tag{6}$$

Here, the diagonalized mask $\text{diag}(\mathbf{m}^{(i)})$ filters the Gaussian kernels, while the reshaped perturbation vector modifies the colors of the survivors. All other Gaussian parameters (e.g., position, scale, opacity) are similarly filtered by the mask. The resulting adversarial sub-model $\mathbf{G}_{adv}^{(i)}$ is then used to render images $\mathbf{R}_{adv}^{(i)}$ for fitness evaluation.

**Objective Functions.** We evaluate each individual's fitness using two conflicting objectives designed to preserve visual fidelity while removing the watermark.

The first objective, visual quality loss ($F_1$), measures the perceptual difference between the original watermarked 3DGS and adversarial 3DGS. We calculate this loss as a weighted combination of the L1 distance and the Structural Similarity Index Measure (SSIM) over images rendered from multiple viewpoints $\{v_1, \ldots, v_{N_v}\}$:

$$\begin{aligned} F_1(\mathbf{G}_{adv}^{(i)}) = \frac{1}{N_v} \sum_{v=1}^{N_v} \Big[ \lambda \mathcal{L}_{\text{L1}}(\mathbf{R}_{adv}^{(i,v)}, \mathbf{R}_{wm}^{(i,v)}) \\ + (1-\lambda)(1 - \text{SSIM}(\mathbf{R}_{adv}^{(i,v)}, \mathbf{R}_{wm}^{(i,v)})) \Big], \end{aligned} \tag{7}$$

where $\mathbf{R}_{adv}^{(i,v)}$ and $\mathbf{R}_{wm}^{(i,v)}$ are the rendered images of adversarial 3DGS and original watermarked 3DGS. $\lambda$ is a weighting factor. Minimizing $F_1$ guides solutions to be visually indistinguishable from the original.

The second objective, watermark destruction ($F_2$), provides a model-agnostic attack by neutralizing the convolutional feature extraction common to all decoders. We achieve this by minimizing the feature maps' statistical variance, which flattens their patterns and renders them non-discriminative for watermark detection. Specifically, we pass a rendered adversarial image $\mathbf{R}_{adv}^{(i,v)}$ through a convolutional feature extractor $\Phi$. The dispersion of a single feature channel, $D(\mathbf{F}_c)$, is then quantified by its standard deviation:

$$D(\mathbf{F}_c) = \sqrt{\frac{1}{H'W'} \sum_{h,w} (\mathbf{F}_c(h,w) - \bar{\mathbf{F}}_c)^2}, \tag{8}$$

where $\bar{\mathbf{F}}_c$ is the mean activation of channel $c$. The final objective $F_2$ is the average dispersion over all feature channels and viewpoints:

$$F_2(\mathbf{G}_{adv}^{(i)}) = \frac{1}{N_v} \sum_{v=1}^{N_v} \left[ \frac{1}{C'} \sum_{c=1}^{C'} D(\Phi(\mathbf{R}_{adv}^{(i,v)})_c) \right]. \tag{9}$$

**Evolutionary Process.** The evolutionary process begins with a population $\mathcal{P}_t$ of size $N_{pop}$. At each generation $t$, an offspring population $\mathcal{Q}_t$ is generated from the current population $\mathcal{P}_t$. This involves two main operators. First, a crossover operator produces two new solutions from a pair of parents $(\mathbf{x}_a, \mathbf{x}_b)$, distributing the offspring around the parents' positions in the search space:

$$\mathbf{x}'_{a,b} = 0.5 \left[ (\mathbf{x}_a + \mathbf{x}_b) \mp \beta |\mathbf{x}_b - \mathbf{x}_a| \right], \tag{10}$$

where $\beta$ is a hyperparameter controlling the spread of the offspring. Subsequently, a mutation operator introduces fine-grained perturbations to an individual $\mathbf{x}'$ to enhance local exploration:

$$x''_j = x'_j + \eta_j (ub_j - lb_j), \tag{11}$$

where $\eta_j$ is a small perturbation value, and $[lb_j, ub_j]$ represents the defined lower and upper bounds for the $j$-th decision variable.

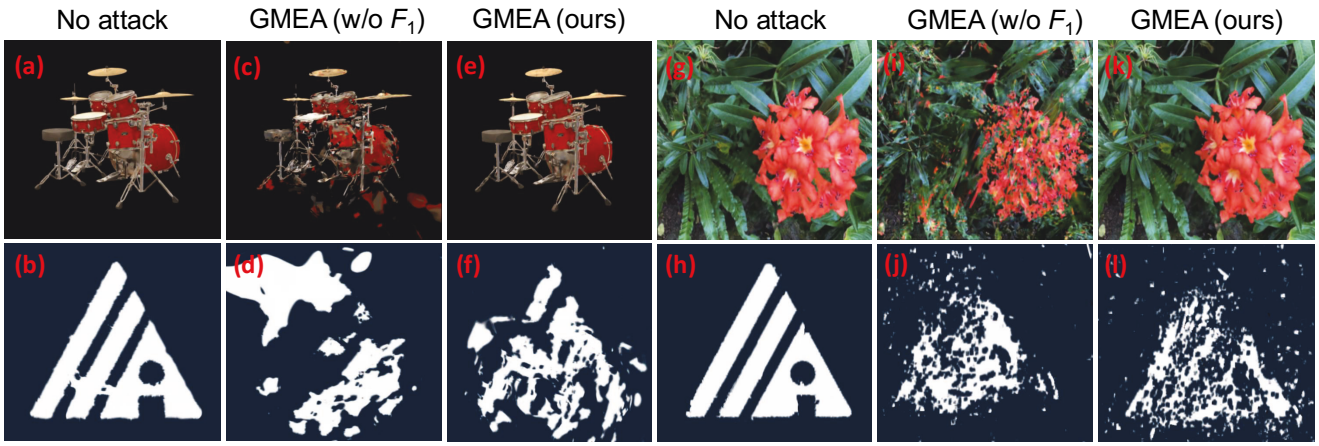


Figure 3: Qualitative results of the GMEA attack on the *Drums* (a-f) and *Flower* (g-l) scenes. While a single-objective attack GMEA (w/o $F_1$) corrupts the watermark at the cost of severe visual degradation (c,d,i,j), our full multi-objective approach GMEA successfully removes the watermark while preserving high visual fidelity (e,f,k,l).

After generating offspring, a rigorous selection process determines which individuals will form the next generation's population. First, the current parent population ($\mathcal{P}_t$) and their offspring ($\mathcal{Q}_t$) are merged into a combined pool, $\mathcal{R}_t = \mathcal{P}_t \cup \mathcal{Q}_t$. This pool is then ranked and partitioned into a hierarchy of non-dominated fronts $\{\mathcal{F}_1, \mathcal{F}_2, \dots\}$ based on Pareto dominance (Hong, Jiang, and Yen 2023).

The next generation is formed by elitism, admitting individuals from the best non-dominated fronts ($\mathcal{F}_1, \mathcal{F}_2, \dots$) until the population capacity ($N_{pop}$) is met. To maintain diversity when the final front ($\mathcal{F}_l$) is truncated, we rank its members by a density metric that prioritizes solutions in less crowded regions. The density score for an individual solution $\mathbf{x}$, denoted $d(\mathbf{x})$, is calculated as:

$$d(\mathbf{x}) = \sum_{o=1}^{M} \frac{F_o(\text{neighbor}^+) - F_o(\text{neighbor}^-)}{F_o^{\max} - F_o^{\min}}, \quad (12)$$

where $M$ is the number of objectives, and $F_o(\text{neighbor}^{\pm})$ are the objective values of the neighbors of solution $\mathbf{x}$ after sorting the front along objective $o$. Individuals with higher density scores are chosen to fill the remaining slots, forming a diverse parent population for the next evolutionary cycle.

**Reconstructing the Adversarial Model.** The final step is to reconstruct the complete adversarial model, $\mathbf{G}_{adv}$. Since our group-based strategy operates on disjoint sets of Gaussian kernels, this reconstruction is a straightforward union of the $k$ optimized sub-models:

$$\mathbf{G}_{adv} = \bigcup_{i=1}^{k} \mathbf{G}_{adv}^{(i)}. \quad (13)$$

The resulting model $\mathbf{G}_{adv}$ aggregates all modifications from the independent optimization runs and represents the final output of our attack framework.

# Evaluation and Results

## Experimental Settings

**Model and dataset.** To demonstrate GMEA's versatility, we target representative systems from two distinct categories of 3DGS watermarking: GaussianMarker (Huang et al. 2025) for 1D bitstream watermarks and GS-Hider (Zhang et al. 2024b) for 2D image watermarks. The evaluation was performed on two datasets: Blender dataset (Mildenhall et al. 2021), comprising objects without backgrounds, and the more challenging LLFF dataset (Mildenhall et al. 2019), which features complex real-world scenes.

**Evaluation metrics.** We assess our framework's performance based on two criteria: visual fidelity and watermark removal efficacy. Visual quality is quantified using standard image metrics: Peak Signal-to-Noise Ratio (PSNR), the Structural Similarity Index Measure (SSIM), and Mean Squared Error (MSE) (Setiadi 2021). For evaluating 1D watermark removal, we use the standard Bit Accuracy Rate (BAR) and our proposed Watermark Uncertainty Score (WUS) and Information Destruction Score (IDS). Detailed definitions for these metrics, along with further experimental settings and results, are deferred to the Appendix.

## Experiment Results

**Attack Performance Evaluation.** We assessed our GMEA framework's effectiveness through extensive experiments on 1D and 2D watermarking systems, as shown in Table 1. Our full GMEA attack significantly disrupts the near-perfect watermark extraction of the *No Attack* baseline in the 1D watermarking system, reducing the Bit Accuracy Rate (BAR) to an average of approximately 65% and substantially increasing the Watermark Uncertainty Score (WUS) and Information Destruction Score (IDS). This renders the extracted bitstream highly unreliable.

In the 2D watermarking system, our full GMEA method's success is evident in the degradation of the extracted watermark image's quality, with a drastic drop in both SSIM and

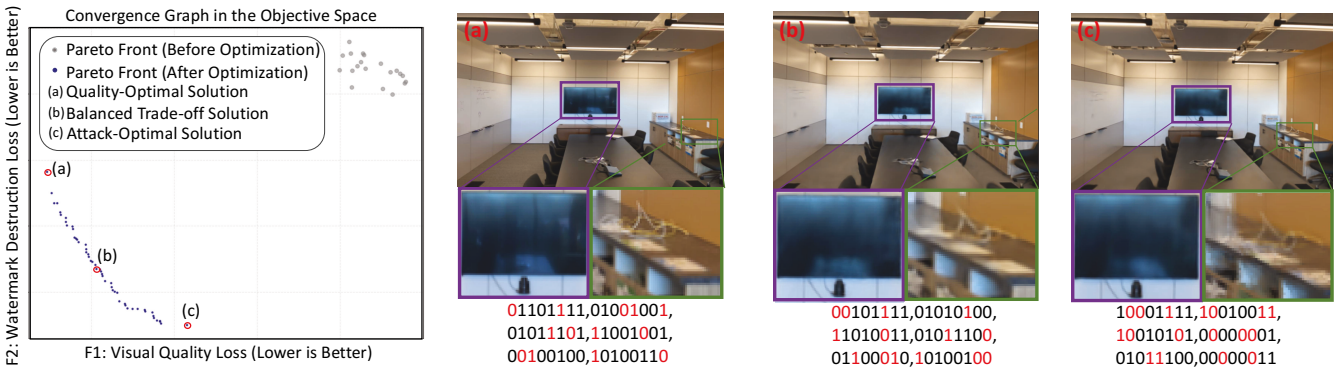


Figure 4: The GMEA attack's Pareto front, showing the trade-off between visual fidelity and attack success. The main plot highlights the optimization's improvement. Three solutions are visualized—(a) quality-optimal, (c) attack-optimal, and (b) balanced—showing their resulting render and the corrupted watermark with errors marked in red.

| | | 1D Watermark (GaussianMarker) | | | | | | 2D Watermark (GS-Hider) | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | Blender | | | LLFF | | | Blender | | | LLFF | | |
| Scene | Method | BAR↓ | WUS↑ | IDS↑ | BAR↓ | WUS↑ | IDS↑ | SSIM↓ | PSNR↓ | MSE↑ | SSIM↓ | PSNR↓ | MSE↑ |
| Chair / Fern | No Attack | 99.95 | 0.1 | 0.1 | 100 | 0.0 | 0.0 | 91.18 | 15.74 | 0.026 | 97.86 | 26.43 | 0.002 |
| | GMEA (w/o $F_1$) | 45.99 | 89.1 | 89.03 | 50.33 | 98.21 | 97.78 | 52.31 | 5.16 | 0.305 | 67.17 | 9.37 | 0.115 |
| | GMEA (ours) | 67.44 | 65.12 | 64.87 | 56.94 | 86.11 | 85.98 | 74.63 | 10.26 | 0.094 | 70.61 | 9.51 | 0.111 |
| Drums / Flower | No Attack | 99.71 | 0.58 | 0.57 | 100 | 0.0 | 0.0 | 93.12 | 18.12 | 0.015 | 96.53 | 23.67 | 0.004 |
| | GMEA (w/o $F_1$) | 45.58 | 89.96 | 90.07 | 57.92 | 84.17 | 83.9 | 52.81 | 5.03 | 0.314 | 62.48 | 8.49 | 0.141 |
| | GMEA (ours) | 65.0 | 70.0 | 69.87 | 60.83 | 78.33 | 78.25 | 75.4 | 9.92 | 0.102 | 65.42 | 11.86 | 0.065 |
| Ficus / Fortress | No Attack | 96.06 | 7.87 | 7.87 | 100 | 0.0 | 0.0 | 95.17 | 22.71 | 0.005 | 94.18 | 19.2 | 0.012 |
| | GMEA (w/o $F_1$) | 54.32 | 89.02 | 88.67 | 57.45 | 80.15 | 80.33 | 64.88 | 7.2 | 0.19 | 30.32 | 5.93 | 0.254 |
| | GMEA (ours) | 71.98 | 56.04 | 55.84 | 64.58 | 70.83 | 71.1 | 74.88 | 8.63 | 0.137 | 56.24 | 8.99 | 0.126 |
| Hotdog / Horns | No Attack | 98.61 | 2.77 | 2.73 | 100 | 0.0 | 0.0 | 95.65 | 22.71 | 0.005 | 97.86 | 26.43 | 0.002 |
| | GMEA (w/o $F_1$) | 58.16 | 82.77 | 82.85 | 60.36 | 77.81 | 77.63 | 60.36 | 6.63 | 0.217 | 60.09 | 8.7 | 0.135 |
| | GMEA (ours) | 58.21 | 83.17 | 83.11 | 61.98 | 76.04 | 75.9 | 73.7 | 9.66 | 0.108 | 75.75 | 12.68 | 0.053 |
| Lego / Leaves | No Attack | 99.99 | 0.02 | 0.02 | 100 | 0.0 | 0.0 | 94.33 | 21.19 | 0.007 | 98.54 | 29.93 | 0.001 |
| | GMEA (w/o $F_1$) | 45.19 | 88.21 | 88.03 | 60.13 | 78.21 | 78.15 | 68.98 | 6.76 | 0.211 | 71.73 | 11.39 | 0.072 |
| | GMEA (ours) | 68.53 | 67.35 | 65.88 | 68.75 | 62.5 | 62.47 | 74.45 | 10.24 | 0.094 | 80.2 | 13.26 | 0.047 |
| Materials / Trex | No Attack | 99.43 | 1.15 | 1.13 | 100 | 0.0 | 0.0 | 91.67 | 17.75 | 0.016 | 97.44 | 28.03 | 0.001 |
| | GMEA (w/o $F_1$) | 55.05 | 83.69 | 83.75 | 59.45 | 77.25 | 77.41 | 60.22 | 6.25 | 0.237 | 64.96 | 9.15 | 0.121 |
| | GMEA (ours) | 63.98 | 71.96 | 71.93 | 65.63 | 68.75 | 68.4 | 79.07 | 10.33 | 0.092 | 69.88 | 10.95 | 0.081 |
| Mic / Room | No Attack | 99.21 | 1.58 | 1.57 | 100 | 0.0 | 0.0 | 92.7 | 18.64 | 0.013 | 98.28 | 29.74 | 0.001 |
| | GMEA (w/o $F_1$) | 56.22 | 86.44 | 86.46 | 66.14 | 65.43 | 65.58 | 32.52 | 2.19 | 0.603 | 70.4 | 9.59 | 0.109 |
| | GMEA (ours) | 67.19 | 65.58 | 65.51 | 71.53 | 56.94 | 56.62 | 77.79 | 8.95 | 0.127 | 75.07 | 11.26 | 0.074 |
| Ship / Orchids | No Attack | 99.4 | 1.21 | 1.19 | 100 | 0.0 | 0.0 | 95.54 | 21.41 | 0.007 | 98.68 | 31.89 | 0.001 |
| | GMEA (w/o $F_1$) | 56.65 | 84.25 | 84.32 | 53.18 | 93.84 | 94.01 | 63.47 | 6.56 | 0.22 | 61.18 | 8.06 | 0.156 |
| | GMEA (ours) | 64.9 | 70.12 | 70.04 | 55.95 | 88.1 | 88.19 | 72.56 | 8.67 | 0.135 | 67.72 | 9.15 | 0.121 |

Table 1: This table quantitatively evaluates the attack's performance on 1D and 2D watermarking systems. Each row pairs a Blender scene with an LLFF scene, presenting their respective results in corresponding columns.

PSNR values compared to the *No Attack* baseline. This severe degradation, detailed in Table 1, confirms our attack effectively renders the 2D visual watermark unrecognizable. These findings validate our GMEA's capability to successfully compromise the detectability of different mainstream watermarking schemes.

**Ablation Study.** To further analyze the components of our framework, we conducted an ablation study by removing the visual quality objective ($F_1$) and creating a single-objective variant, GMEA (w/o $F_1$), which solely optimizes for watermark destruction ($F_2$). As shown in Table 1, this ablation variant exhibits even more potent attack capabilities.

For the 1D watermark, GMEA (w/o $F_1$) achieves a BAR closer to 50% (the theoretical value for random guessing) and higher WUS/IDS scores than our full two-objective method GMEA. For instance, in the Materials / Trex scene, the WUS score for GMEA (w/o $F_1$) attack reaches 83.69, compared to 71.96 for the full GMEA. For the 2D watermark, the GMEA (w/o $F_1$) attack results in a significantly lower SSIM/PSNR for the extracted watermark image, indicating more severe corruption. This demonstrates that by focusing exclusively on maximizing watermark destruction without the constraint of preserving visual fidelity, the attack can achieve a higher degree of watermark removal. However, it completely disregards maintaining the visual quality of the 3DGS model, resulting in the 3DGS being unusable after watermark removal. A detailed analysis of the resulting visual degradation is presented in the next section.

**Justification for the Multi-objective Approach.** While the single-objective attack GMEA (w/o $F_1$) offers superior watermark destruction, it comes at the unacceptable cost of visual degradation to the 3DGS model. This trade-off is

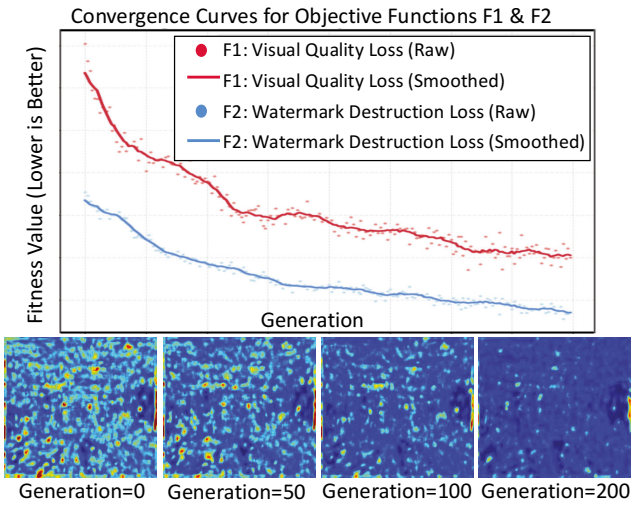


Figure 5: Visualization of GMEA's optimization. The top plot shows the convergence of objectives $F_1$ and $F_2$. The bottom panels show the feature map becoming uniform, which visually confirms the destruction of the watermark.

| Watermarking Target | Attack Method | SSIM↑ | PSNR↑ | MSE↓ |
|---|---|---|---|---|
| 1D (GaussianMarker) | GMEA (w/o $F_1$) | 67.19 | 20.31 | 0.0112 |
| | GMEA (ours) | **95.13** | **30.66** | **0.0011** |
| 2D (GS-Hider) | GMEA (w/o $F_1$) | 60.51 | 17.19 | 0.0263 |
| | GMEA (ours) | **98.22** | **37.76** | **0.0002** |

Table 2: Assessing the visual distortion of 3DGS introduced by our attack via an ablation study.

qualitatively illustrated in Figure 3. As the figure shows, although the watermark extracted by GMEA (w/o $F_1$) is more severely distorted, the rendered image is also riddled with visual artifacts, unlike the pristine render from our full GMEA.

The quantitative data presented in Table 2 corroborates this visual evidence. Our full GMEA method maintains high SSIM (above 95%) and PSNR values, indicating minimal distortion. In contrast, the GMEA (w/o $F_1$) variant causes a drastic drop in these metrics. Since an attack that destroys an asset's visual value defeats the purpose of the theft, this ablation study validates the necessity of our multi-objective formulation, which effectively balances potent watermark removal with the preservation of high visual fidelity.

**Optimization Dynamics and Convergence.** Figure 5 visualizes the operational dynamics of our GMEA. The top graph shows the steady convergence of both visual quality loss ($F_1$) and watermark destruction loss ($F_2$), demonstrating the effectiveness of our multi-objective evolutionary algorithm. The bottom panels qualitatively validate our watermark destruction objective ($F_2$) by displaying visualizations of the intermediate convolutional feature map. At Generation 0, the feature map exhibits distinct spatial patterns that are essential for watermark decoders. As the optimization progresses, these patterns are suppressed, and by Generation 200, the feature map is largely homogeneous, erasing its discriminative features. Since any potential downstream decoder must rely on these upstream convolutional features, this flattening effectively blinds the entire watermark extraction pipeline. This visual collapse of feature information directly corresponds to the convergence of the $F_2$ curve, proving that minimizing feature variance is a successful and universal indirect attack strategy in a black-box setting.

| Groups ($k$) | Time (min) | GPU Memory (GB) |
|---|---|---|
| 1 | 74.2 | 24.0 |
| 5 | 32.5 | 33.6 |
| 10 | 23.7 | 43.2 |
| 20 | 20.1 | 55.4 |
| 50 | 18.5 | 84.9 |

Table 3: Time-memory trade-off analysis for our group-based strategy. The table shows the computational time and peak GPU memory required to reach a fixed target fitness ($F_1 \approx 1.9, F_2 \approx 1.7$) as the number of groups ($k$) varies.

**Pareto Front Analysis and Solution Trade-offs.** A key strength of our multi-objective approach is its ability to find a set of optimal solutions representing the trade-offs between conflicting objectives. Figure 4 analyzes the trade-off between attack effectiveness (low $F_2$) and visual fidelity (low $F_1$). The main plot shows a significant improvement, with solutions evolving from a suboptimal initial state (top-right, gray) to a superior final Pareto front (bottom-left, blue).

For a more granular analysis of the solution space, we visualize three representative solutions from the final front. The Attack-Optimal Solution (c) achieves the most effective watermark corruption at the cost of slight visual artifacts. Conversely, the Quality-Optimal Solution (a) yields the highest visual fidelity but with a less damaged watermark. The Balanced Trade-off Solution (b) presents an ideal compromise, achieving significant watermark disruption with negligible visual distortion. This analysis highlights GMEA's superiority: it discovers a range of effective attack solutions and offers the flexibility to select one based on the desired balance between efficacy and fidelity.

**Analysis of the Group-Based Strategy.** We analyzed the impact of the number of groups ($k$) on optimization efficiency, with results shown in Table 3. The data reveals a clear time-memory trade-off: increasing $k$ accelerates convergence by parallelizing the search, but at the cost of higher peak GPU memory. Considering this trade-off, we identify $k = 10$ as a balanced setting for our main experiments, as it provides a significant speedup while maintaining a manageable memory footprint. More analysis is in the Appendix.

## Conclusion

This paper introduced GMEA, the first universal black-box framework for assessing the robustness of 3DGS watermarking systems. By formulating the attack as a group-based multi-objective optimization problem, GMEA effectively balances watermark removal with visual fidelity preservation, using a novel feature-variance objective to operate without detector knowledge. Our experiments show that GMEA effectively compromises both 1D and 2D watermarking schemes while maintaining high visual fidelity.

## Acknowledgments

This work was supported in part by the National Natural Science Foundation of China under Grant No. 62276222; in part by the Fundamental Research Funds for the Central Universities under Grant No. 20720250164; in part by the Xiamen Natural Science Foundation under Grant No. 3502Z202571027.

## References

Chen, Z.; Wang, G.; Zhu, J.; Lai, J.; and Xie, X. 2025a. GuardSplat: Efficient and Robust Watermarking for 3D Gaussian Splatting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 16325–16335.

Chen, Z.; Yang, J.; Huang, J.; de Lutio, R.; Esturo, J. M.; Ivanovic, B.; Litany, O.; Gojcic, Z.; Fidler, S.; Pavone, M.; Song, L.; and Wang, Y. 2025b. OmniRe: Omni Urban Scene Reconstruction. In *Proceedings of the International Conference on Learning Representations*.

Deb, K.; Pratap, A.; Agarwal, S.; and Meyarivan, T. 2002. A fast and elitist multiobjective genetic algorithm: NSGA-II. *IEEE Transactions on Evolutionary Computation*, 6(2): 182–197.

Gong, Y.; Huang, L.; and Chen, L. 2022. Person re-identification method based on color attack and joint defence. In *Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition*, 4313–4322.

Gong, Y.; Zhong, Z.; Qu, Y.; Luo, Z.; Ji, R.; and Jiang, M. 2024. Cross-modality perturbation synergy attack for person re-identification. In *Proceedings of Neural Information Processing Systems*, volume 37, 23352–23377.

Hong, H.; Jiang, M.; and Yen, G. G. 2023. Improving performance insensitivity of large-scale multiobjective optimization via Monte Carlo tree search. *IEEE Transactions on Cybernetics*, 54(3): 1816–1827.

Hong, S.; Xiao, L.; Zhang, X.; and Chen, J. 2024. ArgMed-Agents: explainable clinical decision reasoning with LLM disscusion via argumentation schemes. In *Proceedings of the IEEE International Conference on Bioinformatics and Biomedicine*, 5486–5493. IEEE.

Huang, X.; Li, R.; Cheung, Y.-m.; Cheung, K. C.; See, S.; and Wan, R. 2025. Gaussianmarker: Uncertainty-aware copyright protection of 3d gaussian splatting. In *Proceedings of Neural Information Processing Systems*.

Jang, Y.; Park, H.; Yang, F.; Ko, H.; Choo, E.; and Kim, S. 2025. 3D-GSW: 3D Gaussian Splatting for Robust Watermarking. arXiv:2409.13222.

Kerbl, B.; Kopanas, G.; Leimkhler, T.; and Drettakis, G. 2023. 3D Gaussian Splatting for Real-Time Radiance Field Rendering. *ACM Transactions on Graphics*, 42(4).

Liang, J.; Ban, X.; Yu, K.; Qu, B.; Qiao, K.; Yue, C.; Chen, K.; and Tan, K. C. 2023. A Survey on Evolutionary Constrained Multiobjective Optimization. *IEEE Transactions on Evolutionary Computation*, 27(2): 201–221.

Liu, B.; Zeng, Q.; Huang, J.; Zhang, J.; Zheng, Z.; Liao, Y.; Deng, K.; Zhou, W.; and Xu, Y. 2022. IVIM using convolutional neural networks predicts microvascular invasion in HCC. *European Radiology*, 32(10): 7185–7195.

Lu, Y.; Jia, Y.; Wang, J.; Li, B.; Chai, W.; Carin, L.; and Velipasalar, S. 2020. Enhancing cross-task black-box transferability of adversarial examples with dispersion reduction. In *Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition*, 940–949.

MacQueen, J. 1967. Some methods for classification and analysis of multivariate observations.

Mildenhall, B.; Srinivasan, P. P.; Ortiz-Cayon, R.; Kalantari, N. K.; Ramamoorthi, R.; Ng, R.; and Kar, A. 2019. Local light field fusion: Practical view synthesis with prescriptive sampling guidelines. *ACM Transactions on Graphics*, 38(4): 1–14.

Mildenhall, B.; Srinivasan, P. P.; Tancik, M.; Barron, J. T.; Ramamoorthi, R.; and Ng, R. 2021. Nerf: Representing scenes as neural radiance fields for view synthesis. *Communications of the ACM*, 65(1): 99–106.

Papernot, N.; McDaniel, P.; Goodfellow, I.; Jha, S.; Celik, Z. B.; and Swami, A. 2017. Practical black-box attacks against machine learning. In *ACM on Asia Conference on Computer and Communications Security*, 506–519.

Setiadi, D. R. I. M. 2021. PSNR vs SSIM: imperceptibility quality assessment for image steganography. *Multimedia Tools and Applications*, 80(6): 8423–8444.

Tu, X.; Radl, L.; Steiner, M.; Steinberger, M.; Kerbl, B.; and de la Torre, F. 2025. VRSplat: Fast and Robust Gaussian Splatting for Virtual Reality. *Proceedings of the ACM on Computer Graphics and Interactive Techniques*, 8(1).

Van Gansbeke, W.; Vandenhende, S.; Georgoulis, S.; Proesmans, M.; and Van Gool, L. 2020. Scan: Learning to classify images without labels. In *Proceedings of European Conference on Computer Vision*, 268–285. Springer.

Wang, J.; Wang, Y.; Wang, H.; and Zhang, J. 2021. A survey on evolutionary computation for adversarial machine learning. *IEEE Transactions on Evolutionary Computation*, 26(5): 994–1009.

Wang, Z.; Cao, L.; Feng, L.; Jiang, M.; and Tan, K. C. 2024a. Evolutionary multitask optimization with lower confidence bound-based solution selection strategy. *IEEE Transactions on Evolutionary Computation*, 28(4): 1053–1067.

Wang, Z.; Lin, Z.; Lin, W.; Yang, M.; Zeng, M.; and Tan, K. C. 2024b. Explainable molecular property prediction: Aligning chemical concepts with predictions via language models. arXiv:2405.16041.

Wang, Z.; Xu, D.; Jiang, M.; and Tan, K. C. 2024c. Spatial-temporal knowledge transfer for dynamic constrained multiobjective optimization. *IEEE Transactions on Evolutionary Computation*, 28(6): 1653–1667.

Wang, Z.; Zeng, Q.; Lin, W.; Jiang, M.; and Tan, K. C. 2024d. Generating Diagnostic and Actionable Explanations for Fair Graph Neural Networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 21690–21698.

Wang, Z.; Zeng, Q.; Lin, W.; Jiang, M.; and Tan, K. C. 2024e. Multiview Subgraph Neural Networks: Self-Supervised Learning With Scarce Labeled Data. *IEEE Transactions on Neural Networks and Learning Systems*.

Wang, Z.; Zhang, X.; Liao, J.; and Jiang, M. 2025. Cross-Field Interface-Aware Neural Operators for Multiphase Flow Simulation. arXiv:2511.08625.

Wu, T.; Yuan, Y.-J.; Zhang, L.-X.; Yang, J.; Cao, Y.-P.; Yan, L.-Q.; and Gao, L. 2024. Recent advances in 3D Gaussian splatting. *Computational Visual Media*, 10(4): 613–642.

Zeng, Q.; Gong, Y.; and Jiang, M. 2024. Cross-task attack: A self-supervision generative framework based on attention shift. In *Proceedings of the International Joint Conference on Neural Networks*, 1–8. IEEE.

Zeng, Q.; Liu, B.; Xu, Y.; and Zhou, W. 2021. An attention-based deep learning model for predicting microvascular invasion of hepatocellular carcinoma using an intra-voxel incoherent motion model of diffusion-weighted magnetic resonance imaging. *Physics in Medicine Biology*, 66(18): 185019.

Zeng, Q.; Wang, Z.; Cheung, Y.; et al. 2024a. Ask, attend, attack: An effective decision-based black-box targeted attack for image-to-text models. In *Proceedings of Neural Information Processing Systems*, volume 37, 105819–105847.

Zeng, Q.; Wang, Z.; Cheung, Y.-m.; and Jiang, M. 2024b. Ask, attend, attack: An effective decision-based black-box targeted attack for image-to-text models. In *Proceedings of Neural Information Processing Systems*, volume 37, 105819–105847.

Zeng, Q.; and Zhou, W. 2021. An attention based deep learning model for direct estimation of pharmacokinetic maps from DCE-MRI images. In *Proceedings of the IEEE International Conference on Bioinformatics and Biomedicine*, 2368–2375. IEEE.

Zhang, R.-Z.; Wang, S.-Z.; Liu, T.-C.; and Zhong, S.-M. 2020. A survey on adversarial attacks on deep-learning based steganography. *IEEE Access*, 8: 189518–189537.

Zhang, T.; Yu, H.-X.; Wu, R.; Feng, B. Y.; Zheng, C.; Snavely, N.; Wu, J.; and Freeman, W. T. 2024a. Physdreamer: Physics-based interaction with 3d objects via video generation. In *Proceedings of European Conference on Computer Vision*, 388–406. Springer.

Zhang, X.; Meng, J.; Li, R.; Xu, Z.; Zhang, Y.; and Zhang, J. 2024b. GS-Hider: Hiding Messages into 3D Gaussian Splatting. In Globerson, A.; Mackey, L.; Belgrave, D.; Fan, A.; Paquet, U.; Tomczak, J.; and Zhang, C., eds., *Proceedings of Neural Information Processing Systems*, volume 37, 49780–49805.

Zhang, X.; Tian, Y.; Cheng, R.; and Jin, Y. 2018. A Decision Variable Clustering-Based Evolutionary Algorithm for Large-Scale Many-Objective Optimization. *IEEE Transactions on Evolutionary Computation*, 22(1): 97–112.

Zhao, X.; Zhang, K.; Su, Z.; Vasan, S.; Grishchenko, I.; Kruegel, C.; Vigna, G.; Wang, Y.-X.; and Li, L. 2024. Invisible image watermarks are provably removable using generative ai. In *Proceedings of Neural Information Processing Systems*, volume 37, 8643–8672.