# STAD: TRANSFERABLE ACTIVE DEFENSE AGAINST GENERATIVE IMAGE STEGANOGRAPHY VIA STRUCTURAL FEATURE PERTURBATIONS

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

Generative image steganography (GIS) directly generates stego images from secret information, making conventional passive defense methods that rely on detecting modification traces difficult to apply. Most existing active defense methods disrupt covert communication through image degradation followed by reconstruction. However, these approaches are ineffective against GIS, where the carrier image is generated directly rather than modified to embed secret information. We propose Structure-guided Transferable Active Defense (STAD), which perturbs stego images using a single frozen encoder. Neither the hidden message nor the target decoder is required. STAD optimizes differences between neighboring features at several encoder layers, avoiding a loss tied to a specific message representation. It combines gradients from nearby images with edge weighting and multiscale smoothing, then constrains distortion in the quantized output. Tests on five SOTA GIS methods give the lowest bit extraction accuracy among the evaluated defenses and an average 8-bit message accuracy of 17.06%. The surrogate and optimization settings remain fixed across targets, and output PSNR stays at or above 30 dB. Experimental results demonstrate that the proposed STAD effectively defends against GIS methods while maintaining high visual quality.

## 1 INTRODUCTION

Image steganography aims to conceal secret information within images while preserving their visual fidelity. Conventional methods typically embed messages by modifying existing cover images, inevitably introducing statistical or residual artifacts that can be exploited by passive steganalysis (Fridrich & Kodovský, 2012; Xu et al., 2016; Ye et al., 2017; Boroumand et al., 2019). In contrast, generative image steganography (GIS) directly synthesizes images carrying secret messages, largely avoiding the embedding traces left by conventional cover modification and thereby exhibiting strong resistance to traditional steganalysis. Existing methods encode messages through disentangled representations or generative modulation (Liu et al., 2022; 2026; Su et al., 2024), as well as diffusion-based synthesis (Rombach et al., 2022; Kim et al., 2025; Hu et al., 2024). Even among diffusion-based methods, secret messages can be represented through Gaussian sampling intervals, latent signs, grilles, conditioning signals, or variance noise (Peng et al., 2024; Yuan et al., 2026; Xu et al., 2025; Yang et al., 2024; Jois et al., 2024). Such images can resemble ordinary synthetic images, creating a potential channel for data exfiltration and malicious payload delivery (Frants & Agaian, 2025).

Passive steganalysis detects suspicious images from residual statistics or learned features (Fridrich & Kodovský, 2012; Xu et al., 2016; Ye et al., 2017; Boroumand et al., 2019), but these cues can become ineffective when secrets are integrated into the generative process rather than embedded through post-hoc image modification. Detection also leaves the message recoverable, motivating active defense that disrupts recovery while preserving visual content. Existing approaches intervene before embedding (Tang & Liu, 2024) or sanitize encoded images, including those transmitted through social networks (Zhu et al., 2023) and those produced by robust steganographic systems (Zhang et al., 2026). Other strategies include erasing and repair (Liu et al., 2023), generative reconstruction (Robinette et al., 2023; 2024), self-supervised sterilization or residual estimation (Liu et al., 2024; 2025), overwriting followed by restoration (Tan et al., 2025), and curvelet- or drift-aware



Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

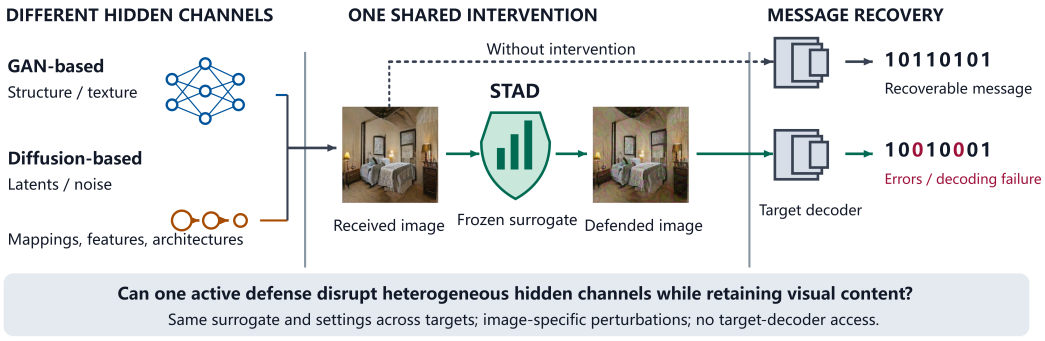



Figure 1: Transferable active defense against generative image steganography.

residual defenses (Hemalatha J et al., 2026; Banerjee et al., 2026). However, most existing defenses are still designed for cover-modification steganography, whereas generative image steganography involves no explicit embedding modification and may encode secrets through heterogeneous generative representations. As a result, their transferability across architectures and message mappings remains unclear: a perturbation effective for one representation may leave another largely intact. This motivates a transferable defense that does not depend on the target message representation or decoder.

Adversarial transfer provides a natural foundation for this goal. Bounded gradient attacks (Goodfellow et al., 2015; Madry et al., 2018) can be strengthened by momentum and acceleration (Dong et al., 2018; Lin et al., 2020), input or gradient transformations (Xie et al., 2019; Dong et al., 2019; Athalye et al., 2018; Long et al., 2022), and intermediate-feature objectives (Wang et al., 2021; Zhang et al., 2022). Intermediate perturbation decay and guided neighborhood sampling further improve transferability by modifying feature propagation and gradient aggregation (Li et al., 2023; Niu et al., 2025). However, conventional adversarial attacks usually rely on a known task objective, while the message representation, decoder, and decision boundaries of the target steganographic model are unavailable. The key challenge is therefore to construct a surrogate objective that transfers across both generative architectures and message mappings.

To this end, we propose *Structure-guided Transferable Active Defense* (STAD), as illustrated in Figure 1. Without access to the secret message, key, or target decoder, STAD uses a frozen VQGAN encoder (Esser et al., 2021) as a fixed surrogate representation and perturbs its horizontal and vertical feature relationships. Neighborhood aggregation, edge weighting, and multiscale smoothing guide the optimization, while a final constraint limits distortion after quantization. The encoder and defense settings remain fixed across all target methods, with only an image-specific perturbation optimized for each input.

The contributions are as follows:

- STAD generates transferable defensive perturbations with one frozen surrogate and no access to target decoders or hidden messages.
- A multilayer spatial-difference objective directs structure-guided optimization, with distortion checked on the quantized output.
- Experimental results demonstrate that the proposed STAD provides effective defense against SOTA GIS methods while maintaining high visual quality.
- To the best of our knowledge, this is the first work to provide effective defense against both GAN-based and diffusion-based GIS.

## 2 Problem Formulation

Let $m \in \{0,1\}^L$ be a secret message of $L$ bits and $\mathcal{X} \subseteq \mathbb{R}^N$ the image space, where $N$ counts scalar pixel values across channels. The sender generates image $x$ with steganographic encoder $G$,

and the receiver extracts $\hat{m} \in \{0,1\}^L$ with decoder $D$:

$$x = G(m, z, \kappa), \qquad \hat{m} = D(x, \kappa), \tag{1}$$

Here $z$ denotes the remaining generation inputs and $\kappa$ the shared key, when one is used. We consider direct generation and extraction, without intervening image processing or transmission losses.

The defender modifies $x$ by adding an image-specific perturbation $\delta \in \mathbb{R}^N$, producing $x' = x + \delta$. For distortion measure $d$ and budget $\varepsilon \geq 0$, define the feasible set

$$\mathcal{C}(x) = \{\delta \in \mathbb{R}^N : x + \delta \in \mathcal{X},\ d(x + \delta, x) \leq \varepsilon\}. \tag{2}$$

Let $\mathcal{E}(\hat{m}, m)$ measure recovery error, so a larger value means less accurate extraction. Ideally, the defender would solve

$$\max_{\delta \in \mathcal{C}(x)} \mathcal{E}\big(D(x + \delta, \kappa), m\big). \tag{3}$$

In practice, the defender sees only $x$. The message, key, target parameters, decoder outputs and gradients, and a corresponding message-free cover are unavailable. The objective in Eq. (3) therefore cannot be evaluated when generating a perturbation.

We instead use a frozen pretrained encoder $F$ to define a differentiable loss $\mathcal{J}_F(u, x)$ between candidate $u \in \mathcal{X}$ and reference $x$. Larger values indicate greater disruption of their feature relationships. Perturbation generation approximately solves

$$\max_{\delta \in \mathcal{C}(x)} \mathcal{J}_F(x + \delta, x) \tag{4}$$

through constrained gradient updates to the image. Let $\theta$ denote the objective settings and optimization hyperparameters. The transfer requirement is to disrupt recovery across different generative models and message mappings while keeping $F$ and $\theta$ fixed, without target-decoder feedback.

# 3 METHOD

## 3.1 METHOD OVERVIEW

STAD perturbs a stego image by changing its spatial feature relationships in a frozen VQGAN encoder $F$ (Esser et al., 2021). Figure 2 shows the three stages: multilayer spatial differences, neighborhood and structure guidance, and distortion control. We normalize $x$ to $[-1, 1]^N$ and compute reference features and a spatial weight map $W(x)$ once. Let $u^t$ be the candidate after $t$ updates, with $t \in \{0, \ldots, T\}$ and $T$ the total update count. Each update aggregates nearby gradients, applies structure guidance, and controls distortion. Quantization and verification give the defended output $x_{\mathrm{d}}^T$, where $\mathrm{d}$ marks the defended image. Iterations use superscripts; layer and neighbor indices use subscripts.

## 3.2 MULTILAYER SPATIAL-DIFFERENCE MODELING

The loss compares spatial relationships in the surrogate features, without referring to a target message mapping. Let $\mathcal{L}$ be the selected encoder layers, $l \in \mathcal{L}$ a layer index, and $h_l(u)$ the feature tensor for image $u$. The reference features $a_l = h_l(x)$ are cached before optimization.

We apply intermediate-level perturbation decay (ILPD) (Li et al., 2023) during the candidate forward pass. At layer $l_\star$, we multiply the candidate–reference activation difference by $\beta \in (0, 1]$ and add it back to the reference. Subsequent layers receive this modified activation. We write the resulting features as $h_l^\beta(u; x)$, where $\beta$ denotes the decay setting rather than an iteration or exponent. This operation changes activations but leaves the encoder weights fixed.

Let $D_{\mathrm{h}}$ and $D_{\mathrm{v}}$ compute horizontal and vertical forward differences between valid adjacent features within each channel. The index $j \in \{\mathrm{h}, \mathrm{v}\}$ selects a direction. Let $d_{\cos}$ be one minus cosine similarity after flattening the difference tensors. The structural loss is

$$L_{\mathrm{str}}(u; x) = \frac{1}{2|\mathcal{L}|} \sum_{l \in \mathcal{L}} \sum_{j \in \{\mathrm{h}, \mathrm{v}\}} d_{\cos}\Big(D_j h_l^\beta(u; x), D_j a_l\Big), \tag{5}$$

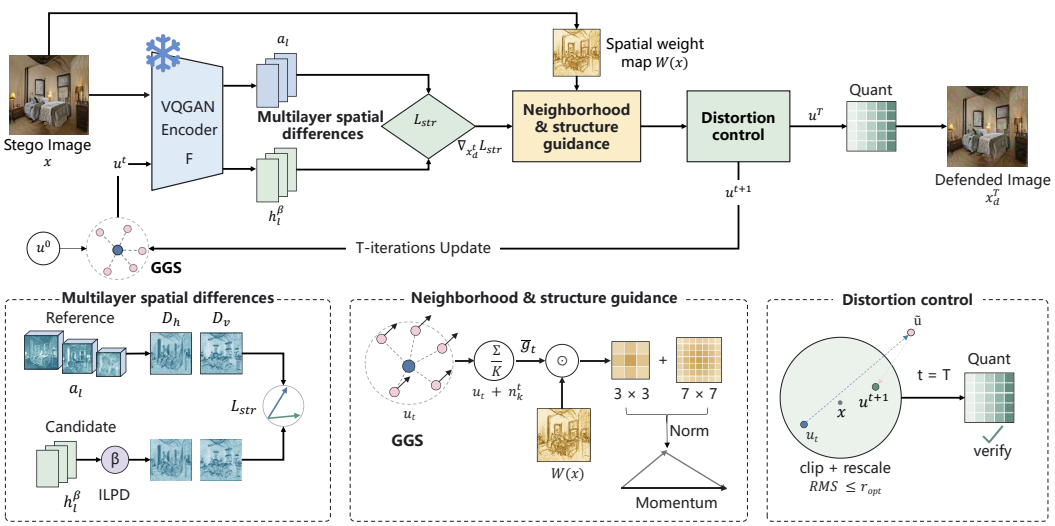



Figure 2: **Overview of STAD.** The optimization loop couples multilayer spatial differences, neighborhood and structure guidance, and distortion control; the lower panels detail each module.

where $|\mathcal{L}|$ is the number of selected layers. We use $L_{\mathrm{str}}(u; x)$ as $\mathcal{J}_F(u, x)$ in Eq. (4). Maximizing it increases angular differences between candidate and reference spatial-difference features at the selected encoder depths. Numerical stabilization is specified in Appendix A.

The objective depends on neighboring feature relationships rather than absolute activations, without assuming a generator-specific correspondence between bits and features.

### 3.3 Neighborhood Aggregation and Structure Guidance

At iteration $t < T$, gradient-guided sampling (GGS) (Niu et al., 2025) draws $K$ neighbors around $u^t$. Neighbor $k \in \{1, \ldots, K\}$ has offset $n_k^t \in \mathbb{R}^N$ and sampling radius $\eta > 0$. Entries of the first offset are sampled independently and uniformly from $[-\eta, \eta]$. Later offsets use independent magnitudes from $[0, \eta]$ and signs taken from the previous neighbor's input gradient. The center stays at $u^t$ throughout this process, and the sampling state resets at each iteration. We average the input gradients as

$$\bar{g}^t = \frac{1}{K} \sum_{k=1}^{K} \nabla_{u^t} L_{\mathrm{str}}\big(\mathrm{clip}(u^t + n_k^t, -1, 1); x\big)\,, \tag{6}$$

where clip acts elementwise and $\nabla_{u^t}$ differentiates with respect to $u^t$. Offsets are held constant during differentiation, while gradients propagate through clipping.

We weight $\bar{g}^t$ with an edge-strength map $E(x)$ whose entries lie in $[0, 1]$. To obtain this map, we convert $x$ to grayscale and sum absolute horizontal and vertical pixel differences. We divide by the map's 90th percentile and clip to $[0, 1]$, giving spatial weights

$$W(x) = w_0 + w_1 E(x), \tag{7}$$

where $w_0 > 0$ allows updates in weak-edge regions and $w_1 \geq 0$ adds weight at stronger edges. The map is shared across color channels and remains fixed during optimization. Appendix A gives boundary handling and numerical floors.

We then smooth the weighted gradient at two scales, following translation-invariant gradient convolution (Dong et al., 2019). For $q \in \{3, 7\}$, let $\mathcal{G}_{q,\sigma_q}$ denote a unit-sum Gaussian kernel of size $q \times q$ and standard deviation $\sigma_q$. Nonnegative mixture weights $\lambda_q$ satisfy $\lambda_3 + \lambda_7 = 1$. The smoothed gradient is

$$s^t = \sum_{q \in \{3,7\}} \lambda_q \mathcal{G}_{q,\sigma_q} * \big(W(x) \odot \bar{g}^t\big)\,, \tag{8}$$

where $\odot$ is elementwise multiplication and $*$ is zero-padded convolution applied independently to each color channel. Momentum (Dong et al., 2018) then accumulates the shaped gradients:

$$v^{t+1} = \mu v^t + \mathrm{Norm}(s^t), \qquad v^0 = 0, \tag{9}$$

where $v^t$ is the momentum tensor and $\mu$ its retention coefficient. Per-image $\mathrm{Norm}$ divides first by the Euclidean norm and then by the mean absolute value, with numerical floors given in Appendix A.

### 3.4 Distortion-constrained updates and output

We bound the root-mean-square (RMS) displacement from $x$ by $r_{\mathrm{opt}} > 0$. RMS is $\|\cdot\|_2/\sqrt{N}$, where $\|\cdot\|_2$ denotes the Euclidean norm. The operator $\mathcal{P}_{r_{\mathrm{opt}}}$ first clips a candidate to $[-1, 1]$. If its displacement RMS exceeds $r_{\mathrm{opt}}$, it rescales that displacement before adding it back to $x$. With step size $\alpha > 0$, the tentative candidate $\widetilde{u}^{t+1}$ and the constrained update are

$$\widetilde{u}^{t+1} = u^t + \alpha\,\mathrm{sign}(v^{t+1}), \qquad u^{t+1} = \mathcal{P}_{r_{\mathrm{opt}}}(\widetilde{u}^{t+1}; x), \tag{10}$$

where sign returns elementwise signs in $\{-1, 0, 1\}$. The updated candidate satisfies $u^{t+1} \in [-1, 1]^N$ and $\|u^{t+1} - x\|_2/\sqrt{N} \le r_{\mathrm{opt}}$ and becomes the next sampling center. We initialize it with the same constraint operator:

$$u^0 = \mathcal{P}_{r_{\mathrm{opt}}}(x + \xi; x), \tag{11}$$

where $\xi \in \mathbb{R}^N$ is bounded, zero-centered uniform noise with independent entries. Appendix A gives its bound. A peak signal-to-noise ratio (PSNR) threshold $\tau_{\mathrm{opt}}$, in decibels, sets $r_{\mathrm{opt}} = 2\,10^{-\tau_{\mathrm{opt}}/20}$.

After $T$ updates, elementwise quantization converts the candidate to 8-bit intensities:

$$\mathrm{Quant}(u) = \mathrm{round}\left[\frac{255}{2}\left(\mathrm{clip}(u, -1, 1) + 1\right)\right], \tag{12}$$

where round returns the nearest integer. Define the quantized reference $b = \mathrm{Quant}(x)$ and an output displacement scale $\gamma \in [0, 1]$. The defended image $x_{\mathrm{d}}^T \in \{0, \ldots, 255\}^N$ is

$$x_{\mathrm{d}}^T = b + \mathrm{round}\big[\gamma\left(\mathrm{Quant}(u^T) - b\right)\big], \tag{13}$$

and must satisfy the output PSNR threshold $\tau_{\mathrm{out}}$, in decibels:

$$\frac{\|x_{\mathrm{d}}^T - b\|_2^2}{N} \le 255^2\, 10^{-\tau_{\mathrm{out}}/10}. \tag{14}$$

If the quantized candidate satisfies the bound, we retain $\gamma = 1$. Otherwise, bisection finds a feasible scale and returns its feasible lower endpoint. We save the verified image losslessly and evaluate extraction from that file. In Section 2, this output corresponds to $x' = 2x_{\mathrm{d}}^T/255 - 1$ and $\delta = x' - x$.

**Algorithm 1** STAD: Structure-guided Transferable Active Defense

**Require:** Stego image $x$, frozen encoder $F$, fixed settings (Appendix A)
**Ensure:** Defended integer image $x_{\mathrm{d}}^T$

```
Cache a_l = h_l(x) for l ∈ L and compute W(x)
Initialize u^0 using Eq. (11); set v^0 = 0
for t = 0, ..., T − 1 do
    Reset GGS state and sample K offsets n_k^t around u^t
    Compute ḡ^t from ILPD-modified features using Eq. (6)
    Shape ḡ^t and update momentum using Eqs. (8)–(9)
    Construct and constrain u^{t+1} using Eq. (10)
end for
Quantize u^T and x using Eq. (12)
Select γ so that Eq. (13) satisfies Eq. (14); save and return x_d^T
```

The surrogate is an ImageNet-pretrained VQGAN with three selected layers; ILPD is applied at the second selected layer. We select settings on a development split of RDI-SWE (Liu et al., 2026) and keep them fixed across targets. Decoder feedback is used only for development evaluation, never to generate a perturbation. Algorithm 1 gives the procedure; Appendix A lists the settings. Each image requires one cached reference pass and $TK$ surrogate backward evaluations.

# 4 EXPERIMENTS AND ANALYSIS

## 4.1 EXPERIMENTAL SETUP

**Targets and defenses.** The targets are IDEAS (Liu et al., 2022), RDI-SWE (Liu et al., 2026), ADPM (Yuan et al., 2026), mas-GRDH (Hu et al., 2024), and LDStega (Peng et al., 2024). The first two generative image steganography methods are GAN-based, while the remaining three are diffusion-based. We compare STAD with EBRA (Liu et al., 2023), Light (Zhang et al., 2026), Overwriting (Tan et al., 2025), and SCNet (Zhu et al., 2023). All defenses use the same images and sample counts per target, with their evaluated configurations. Appendices A and B list fixed STAD settings and target protocols. We evaluate recovery, image quality, and robustness on five state-of-the-art GIS methods. We further select RDI-SWE and ADPM as representative GAN-based and diffusion-based methods, respectively, for the parameter and ablation studies.

**Dataset.** For the datasets, IDEAS uses the FFHQ dataset (Karras et al., 2019), RDI-SWE uses LSUN Bedroom for indoor scenes, and the diffusion targets use the MS-COCO dataset(Lin et al., 2014). We measure quality and extraction accuracy from the saved 8-bit PNGs.

**Metrics.** **Bit-level extraction accuracy (EA)** is the percentage of correctly extracted bits, averaged over images. For **message-level extraction accuracy (MA)**, each consecutive 8-bit secret segment is one message, which counts as correct only if all 8 bits are recovered. We average these percentages over images, exclude incomplete trailing segment, and count undecodable messages as incorrect. The same definitions apply to every defense. Lower EA and MA indicate stronger disruption, which indicates more effective defense performance. The main visual comparison uses **PSNR** (dB, ↑) and **LPIPS** (↓) (Zhang et al., 2018). Supplementary ablations also report SSIM (Wang et al., 2004). Average gives equal weight to the five targets and is calculated before rounding.

## 4.2 EXTRACTION PERFORMANCE AND VISUAL QUALITY

Following (Geng et al., 2020), an EA below 80% is generally considered indicative of a successful defense. As shown in Table 1, STAD gives the lowest EA on all five targets, averaging 72.84%, and the lowest average MA, 17.06%. Its MA ranges from 4.60% to 26.70%: many bits remain correct, but most 8-bit messages cannot be recovered intact. EBRA is the strongest baseline and obtains lower MA on LDStega despite slightly higher EA. Bit errors are therefore distributed differently across messages. Light leaves extraction largely unchanged, with EA between 96.56% and 100%; Overwriting and SCNet also preserve most bits.

Table 1: Extraction performance (%). EA measures correctly extracted bits; MA measures correctly recovered 8-bit messages. Average weights targets equally. Bold marks the best values.

| Defense | ADPM | | mas-GRDH | | LDStega | | IDEAS | | RDI-SWE | | Average | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | EA | MA | EA | MA | EA | MA | EA | MA | EA | MA | EA | MA |
| EBRA | 96.49 | 77.03 | 79.50 | 18.92 | 74.60 | **7.67** | 68.70 | 5.80 | 90.49 | 55.87 | 81.95 | 33.06 |
| Light | 99.52 | 96.55 | 96.56 | 78.45 | 97.26 | 50.06 | 97.79 | 83.66 | 100.00 | 100.00 | 98.23 | 81.74 |
| Overwriting | 99.08 | 93.53 | 93.61 | 63.67 | 95.76 | 43.87 | 97.17 | 79.42 | 99.16 | 96.27 | 96.96 | 75.35 |
| SCNet | 98.98 | 92.83 | 92.93 | 60.38 | 98.71 | 53.38 | 95.45 | 68.92 | 100.00 | 100.00 | 97.21 | 75.10 |
| STAD | **73.50** | **22.54** | **72.39** | **13.18** | **74.20** | 18.26 | **67.37** | **4.60** | **76.73** | **26.70** | **72.84** | **17.06** |

**Visual quality and extraction disruption.** Table 2 and Figure 3 illustrate the visual quality of different defenses. Light achieves the highest fidelity but provides limited extraction disruption, while SCNet shows substantial variation across targets (29.54–48.80 dB PSNR). In contrast, STAD maintains a stable PSNR of about 30 dB across all five GIS methods and achieves better average PSNR and LPIPS than EBRA. As shown in Figure 3, STAD produces spatially distributed perturbations rather than the localized residual patterns observed for EBRA and Overwriting, while largely preserving the semantic content of the image.

Table 2: Visual quality. Each target has separate PSNR (dB) and LPIPS columns. Average weights targets equally. Bold marks the best values.

| Defense | ADPM | | mas-GRDH | | LDStega | | IDEAS | | RDI-SWE | | Average | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | PSNR | LPIPS | PSNR | LPIPS | PSNR | LPIPS | PSNR | LPIPS | PSNR | LPIPS | PSNR | LPIPS |
| EBRA | 23.86 | 0.209 | 24.06 | 0.228 | 23.63 | 0.177 | 29.33 | 0.113 | 25.83 | 0.244 | 25.34 | 0.194 |
| Light | **43.33** | **0.004** | **43.27** | **0.005** | 43.35 | **0.003** | 43.30 | 0.004 | 43.31 | 0.004 | **43.31** | **0.004** |
| Overwriting | 31.14 | 0.048 | 31.64 | 0.050 | 31.12 | 0.040 | 37.42 | 0.031 | 29.51 | 0.072 | 32.17 | 0.048 |
| SCNet | 29.54 | 0.118 | 30.04 | 0.124 | **43.68** | **0.003** | **48.80** | **0.001** | **46.23** | **0.003** | 39.66 | 0.050 |
| STAD | 30.39 | 0.223 | 30.35 | 0.208 | 30.37 | 0.137 | 30.30 | 0.182 | 30.32 | 0.185 | 30.35 | 0.187 |

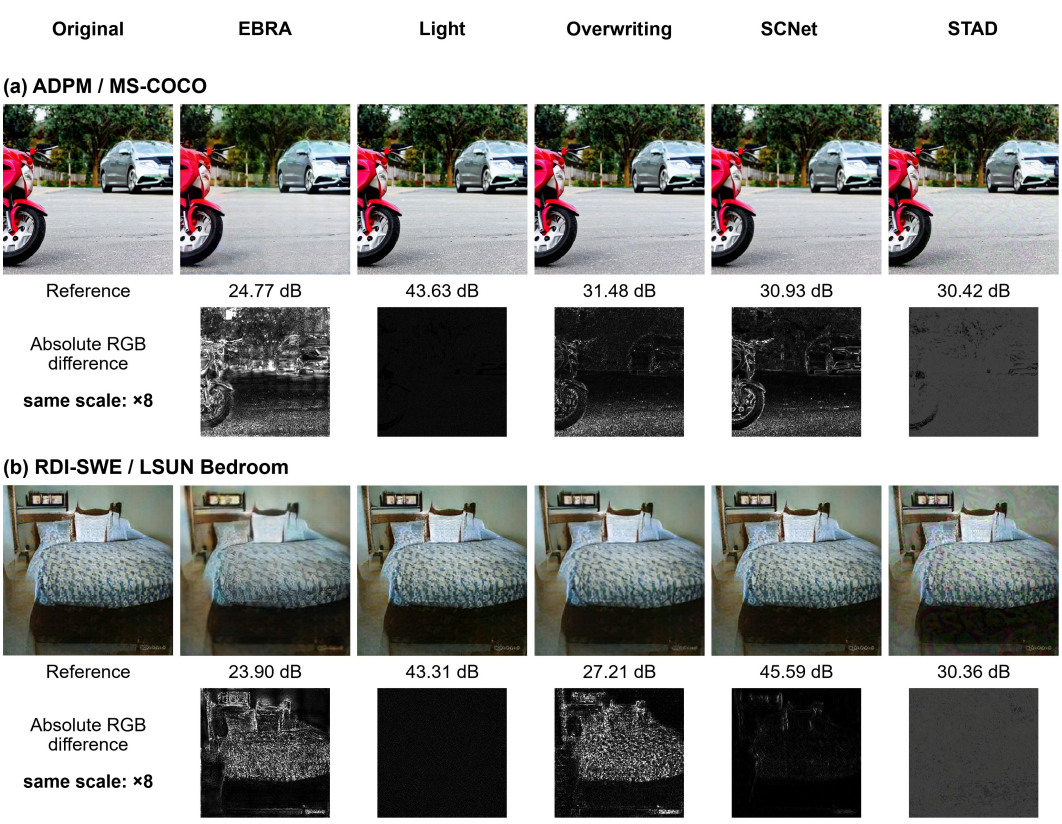



Figure 3: Visual comparison of different defense methods against generative image steganography. Results are shown for ADPM on MS-COCO dataset and IDEAS on FFHQ dataset. EBRA, Light, Overwriting, and SCNet are baseline defenses, while STAD is our proposed method. The bottom row of each example visualizes the absolute RGB difference from the reference image, magnified by $8\times$ using the same scale. PSNR values are reported above the corresponding residual maps.

### 4.3 Robustness to Image Processing

To evaluate the effectiveness of the proposed STAD in practical network transmission scenarios, we measure EA after applying image processing to STAD outputs and their original stego images. As shown in Table 3, STAD remains effective after various image processing operations, including JPEG compression, resizing, blurring, median filtering, and additive noise. For ADPM, mas-GRDH, LDStega, and IDEAS, the defended EA consistently remains below 80% under all evaluated transformations, demonstrating the robustness of STAD to common post-processing. STAD also effectively suppresses secret extraction from RDI-SWE under most conditions, with degradation observed only under stronger additive noise. Overall, these results demonstrate that STAD provides

robust defense across different generative image steganography methods and remains effective under diverse image processing operations.

Table 3: Bit-level extraction accuracy (EA) (%) after image processing. Orig. and STAD denote original and defended stego images receiving the same processing. Bold marks the lower paired mean.

| Processing | ADPM | | mas-GRDH | | LDStega | | IDEAS | | RDI-SWE | |
|---|---|---|---|---|---|---|---|---|---|---|
| | Orig. | STAD | Orig. | STAD | Orig. | STAD | Orig. | STAD | Orig. | STAD |
| None | 98.70 | **73.50** | 97.66 | **72.39** | 99.46 | **74.20** | 98.35 | **67.37** | 100.00 | **76.73** |
| JPEG $Q = 90$ | 96.90 | **73.67** | 95.25 | **72.62** | 98.66 | **74.86** | 95.87 | **67.75** | 99.99 | **70.31** |
| JPEG $Q = 70$ | 92.09 | **73.21** | 90.06 | **72.20** | 95.24 | **74.10** | 90.33 | **67.46** | 99.77 | **70.06** |
| JPEG $Q = 50$ | 88.10 | **72.62** | 86.07 | **71.60** | 90.55 | **72.90** | 85.24 | **67.03** | 98.86 | **71.10** |
| Resize $0.5\times$ | 96.67 | **72.22** | 93.87 | **70.85** | 92.83 | **70.62** | 98.14 | **67.45** | 100.00 | **63.00** |
| Resize $1.5\times$ | 98.61 | **73.49** | 97.50 | **72.43** | 99.29 | **74.76** | 98.32 | **67.41** | 100.00 | **66.03** |
| Blur $\sigma = 0.5$ | 98.62 | **73.47** | 97.55 | **72.42** | 99.14 | **74.58** | 98.34 | **67.45** | 100.00 | **66.40** |
| Blur $\sigma = 1$ | 97.42 | **72.55** | 95.09 | **71.20** | 89.23 | **68.85** | 98.29 | **67.68** | 100.00 | **75.07** |
| Blur $\sigma = 1.5$ | 95.20 | **71.36** | 91.03 | **69.46** | 78.07 | **63.88** | 98.11 | **67.98** | 99.96 | **77.70** |
| Median $3 \times 3$ | 97.26 | **71.51** | 94.61 | **70.35** | 95.74 | **70.52** | 98.29 | **66.98** | 99.99 | **61.87** |
| Median $5 \times 5$ | 93.34 | **69.01** | 88.09 | **67.27** | 84.48 | **64.71** | 98.09 | **66.64** | 97.13 | **56.94** |
| Noise $\sigma = 1$ | 98.21 | **73.54** | 97.01 | **72.42** | 98.78 | **74.17** | 98.06 | **67.36** | 100.00 | **78.73** |
| Noise $\sigma = 2$ | 96.74 | **73.60** | 95.42 | **72.43** | 95.23 | **73.97** | 97.43 | **67.34** | 100.00 | **83.80** |
| Noise $\sigma = 4$ | 93.32 | **73.63** | 91.81 | **72.37** | 87.15 | **72.87** | 95.80 | **67.30** | 100.00 | **88.95** |

Table 4: Component ablations on RDI-SWE and ADPM. EA and MA are percentages; PSNR is in dB. The best results are highlighted in bold, and the second-best results are underlined.

| Variant | RDI-SWE | | | | ADPM | | | |
|---|---|---|---|---|---|---|---|---|
| | EA | MA | PSNR | LPIPS | EA | MA | PSNR | LPIPS |
| **Full STAD** | **76.73** | **26.70** | <u>30.32</u> | 0.185 | **73.50** | <u>22.54</u> | 30.39 | 0.223 |
| Global feature loss | 88.05 | 56.53 | **30.41** | 0.213 | 76.67 | 24.03 | **30.46** | 0.249 |
| Pooled local loss | 99.97 | 99.77 | 30.10 | **0.120** | 94.11 | 68.24 | 30.21 | **0.103** |
| Without smoothing | 95.88 | 80.38 | 30.31 | 0.158 | 75.30 | 23.07 | 30.43 | 0.256 |
| Middle layer only | <u>80.86</u> | <u>29.59</u> | 30.17 | <u>0.126</u> | <u>74.29</u> | 27.29 | 30.27 | <u>0.196</u> |
| Deep layer only | 84.19 | 46.13 | **30.41** | 0.201 | 74.35 | **22.26** | <u>30.45</u> | 0.238 |

### 4.4 Parameter sensitivity

Figure 4 varies one parameter at a time using the same images, fixed seed, and original baseline. Here $K$ counts neighbors, $T$ counts updates, $\zeta = \eta / r_{\text{opt}}$ scales the sampling radius, and $\alpha_{255} = 255\alpha/2$ gives the step size in 8-bit levels. Defaults are $K = 5$, $\zeta = 2$, $T = 40$, and $\alpha_{255} = 1$.

**Neighborhood size and radius.** Increasing $K$ from 1 to 10 raises RDI-SWE EA by 3.48 points but changes ADPM by only $-0.28$ points. A nonzero radius helps ADPM, not RDI-SWE. Unlike $\zeta = 0$, which makes neighbors coincide, $K = 1$ still samples a displaced input.

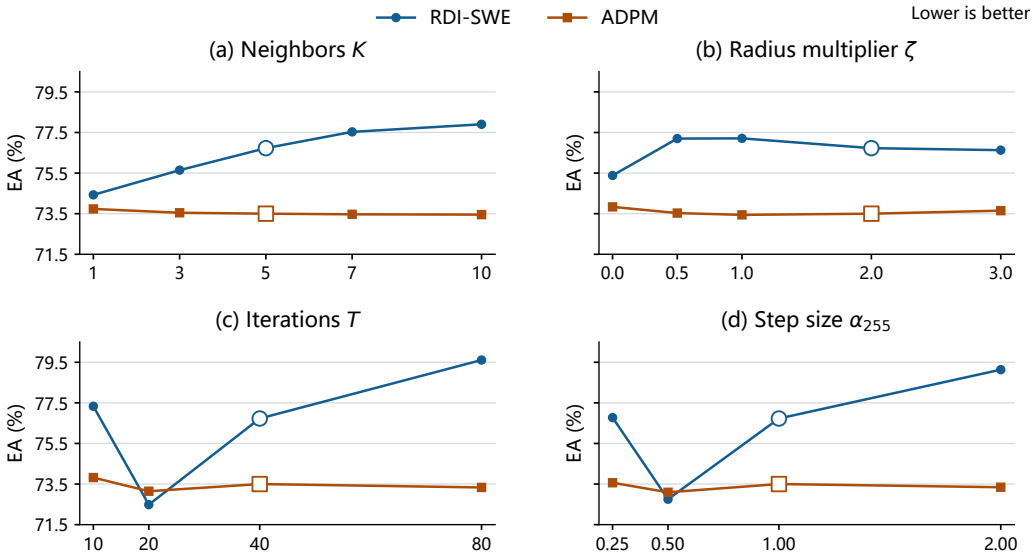



Figure 4: Parameter sensitivity on RDI-SWE and ADPM. Image-averaged EA (%) uses a shared vertical scale. Rings mark the original configuration; lines connect tested settings.

**Iterations and step size.** The lowest EA occurs at $T = 20$ and, separately, $\alpha_{255} = 0.5$ for both targets. More iterations or larger steps do not consistently improve transfer. These fixed-seed, one-factor sweeps do not establish a joint optimum; we retain the original configuration.

### 4.5 Component Ablations

Table 4 evaluates the main design choices of STAD on RDI-SWE and ADPM. Replacing the proposed spatial-difference objective with either a global feature loss or a pooled local loss substantially increases both EA and MA, confirming the importance of modeling local spatial relationships. Removing multiscale smoothing also degrades defense performance, particularly on RDI-SWE, where EA increases from 76.73% to 95.88%. Using only the middle or deep layer generally results in higher extraction accuracy than the full configuration, suggesting that combining features across multiple depths improves transferability. Overall, these results support the effectiveness of the spatial-difference objective and gradient smoothing, while the full STAD provides a favorable balance between extraction disruption and visual quality.

## 5 Conclusion

In this work, we proposed STAD, a transferable active defense against generative image steganography. STAD uses a single frozen surrogate encoder to disrupt spatial feature relationships without requiring access to hidden messages or target decoders. Experiments on five GAN- and diffusion-based GIS methods show that STAD effectively reduces secret extraction accuracy while maintaining high visual quality, and remains robust under common image processing operations. These results demonstrate the effectiveness and transferability of STAD across heterogeneous GIS systems. Future work will explore adaptive receivers, unseen generative models, and more efficient optimization.

## REFERENCES


Anish Athalye, Logan Engstrom, Andrew Ilyas, and Kevin Kwok. Synthesizing robust adversarial examples. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 284–293, 2018.

Abhisek Banerjee, Sreeparna Ganguly, Imon Mukherjee, and Nabanita Ganguly. Towards secure digital media: A drift-aware mlops framework for adaptive stego content sterilization. *Future Generation Computer Systems*, 183:108546, 2026. doi: 10.1016/j.future.2026.108546.

Mehdi Boroumand, Mo Chen, and Jessica Fridrich. Deep residual network for steganalysis of digital images. *IEEE Transactions on Information Forensics and Security*, 14(5):1181–1193, 2019. doi: 10.1109/TIFS.2018.2871749.

Yinpeng Dong, Fangzhou Liao, Tianyu Pang, Hang Su, Jun Zhu, Xiaolin Hu, and Jianguo Li. Boosting adversarial attacks with momentum. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 9185–9193, 2018.

Yinpeng Dong, Tianyu Pang, Hang Su, and Jun Zhu. Evading defenses to transferable adversarial examples by translation-invariant attacks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4312–4321, 2019.

Patrick Esser, Robin Rombach, and Björn Ommer. Taming transformers for high-resolution image synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12873–12883, 2021.

Vladimir Frants and Sos Agaian. Training-free color-aware adversarial diffusion sanitization for diffusion stegomalware defense at security gateways. arXiv preprint arXiv:2512.24499, 2025. URL `https://arxiv.org/abs/2512.24499`.

Jessica Fridrich and Jan Kodovský. Rich models for steganalysis of digital images. *IEEE Transactions on Information Forensics and Security*, 7(3):868–882, 2012. doi: 10.1109/TIFS.2012.2190402.

Linfeng Geng, Weiming Zhang, Haozhe Chen, Han Fang, and Nenghai Yu. Real-time attacks on robust watermarking tools in the wild by cnn. *Journal of Real-Time Image Processing*, 17(3):631–641, 2020.

Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. In *International Conference on Learning Representations*, 2015.

Hemalatha J, Senthil Maharaj Kennedy, Sekar M, and Kaliraj M. Curvelet denoising firewall for universal removal of steganographic image content. *Results in Engineering*, 29:109837, 2026. doi: 10.1016/j.rineng.2026.109837.

Xiaoxiao Hu, Sheng Li, Qichao Ying, Wanli Peng, Xinpeng Zhang, and Zhenxing Qian. Establishing robust generative image steganography via popular stable diffusion. *IEEE Transactions on Information Forensics and Security*, 19:8094–8108, 2024. doi: 10.1109/TIFS.2024.3444311.

Tushar M. Jois, Gabrielle Beck, and Gabriel Kaptchuk. Pulsar: Secure steganography for diffusion models. In *Proceedings of the 2024 ACM SIGSAC Conference on Computer and Communications Security*, pp. 4703–4717, 2024. doi: 10.1145/3658644.3690218.

Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *2019 IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pp. 4396–4405. IEEE, 2019.

Daegyu Kim, Chaehun Shin, Jooyoung Choi, Dahuin Jung, and Sungroh Yoon. Diffusion-Stego: Training-free diffusion generative steganography via message projection. *Information Sciences*, 718:122358, 2025. doi: 10.1016/j.ins.2025.122358.

Qizhang Li, Yiwen Guo, Wangmeng Zuo, and Hao Chen. Improving adversarial transferability via intermediate-level perturbation decay. In *Advances in Neural Information Processing Systems*, volume 36, pp. 32900–32912, 2023. doi: 10.52202/075280-1425.

Jiadong Lin, Chuanbiao Song, Kun He, Liwei Wang, and John E. Hopcroft. Nesterov accelerated gradient and scale invariance for adversarial attacks. In *International Conference on Learning Representations*, 2020.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pp. 740–755. Springer, 2014.

Hangcheng Liu, Tao Xiang, Shangwei Guo, Han Li, Tianwei Zhang, and Xiaofeng Liao. Erase and repair: An efficient box-free removal attack on high-capacity deep hiding. *IEEE Transactions on Information Forensics and Security*, 18:5229–5242, 2023. doi: 10.1109/TIFS.2023.3305839.

Jinjin Liu, Fuyong Xu, Yingao Zhao, Xianwei Xin, Keren Liu, and Yuanyuan Ma. Sterilization of image steganography using self-supervised convolutional neural network. *PeerJ Computer Science*, 10:e2330, 2024. doi: 10.7717/peerj-cs.2330.

Jinjin Liu, Sa Xue, Xinyu Zhang, Fengyuan Xiang, and Yuanyuan Ma. Steganography defense network based on simulation of steganography information distribution. *International Journal of Intelligent Systems*, 2025(1):9958912, 2025. doi: 10.1155/int/9958912.

Xiyao Liu, Ziping Ma, Junxing Ma, Jian Zhang, Gerald Schaefer, and Hui Fang. Image disentanglement autoencoder for steganography without embedding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2303–2312, 2022. CVF open-access version.

Xiyao Liu, Ziping Ma, Jian Zhang, Gerald Schaefer, Kehua Guo, Yuesheng Zhu, and Shichao Zhang. Robust and diversified image steganography without embedding through a disentanglement autoencoder. *IEEE Transactions on Dependable and Secure Computing*, 23(3):6037–6053, 2026. doi: 10.1109/TDSC.2026.3660747.

Yuyang Long, Qilong Zhang, Boheng Zeng, Lianli Gao, Xianglong Liu, Jian Zhang, and Jingkuan Song. Frequency domain model augmentation for adversarial attack. In *Computer Vision – ECCV 2022*, volume 13664 of *Lecture Notes in Computer Science*, pp. 549–566, 2022. doi: 10.1007/978-3-031-19772-7_32.

Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. In *International Conference on Learning Representations*, 2018.

Zenghao Niu, Weicheng Xie, Siyang Song, Zitong Yu, Feng Liu, and Linlin Shen. Enhancing adversarial transferability by balancing exploration and exploitation with gradient-guided sampling. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 3885–3894, 2025.

Yinyin Peng, Yaofei Wang, Donghui Hu, Kejiang Chen, Xianjin Rong, and Weiming Zhang. LDStega: Practical and robust generative image steganography based on latent diffusion models. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pp. 3001–3009, 2024. doi: 10.1145/3664647.3681635.

Preston K. Robinette, Hanchen D. Wang, Nishan Shehadeh, Daniel Moyer, and Taylor T. Johnson. SUDS: Sanitizing universal and dependent steganography. In *Proceedings of the 26th European Conference on Artificial Intelligence*, pp. 1978–1985, 2023. doi: 10.3233/FAIA230489.

Preston K. Robinette, Daniel Moyer, and Taylor T. Johnson. Sanitizing hidden information with diffusion models. In *Proceedings of the 27th European Conference on Artificial Intelligence*, pp. 778–785, 2024. doi: 10.3233/FAIA240562.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10684–10695, 2022.

Wenkang Su, Jiangqun Ni, and Yiyan Sun. StegaStyleGAN: Towards generic and practical generative image steganography. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 240–248, 2024. doi: 10.1609/aaai.v38i1.27776. Issue 1.

Mingyao Tan, Dongsheng Guo, Lei Huang, Zhaorui Gu, Xiaodong Wang, and Haiyong Zheng. Image sterilization via overwriting. *Information Sciences*, 690:121521, 2025. doi: 10.1016/j.ins.2024.121521.

Weixuan Tang and Yadong Liu. Active defense against image steganography. In *Artificial Intelligence Security and Privacy*, volume 14509 of *Lecture Notes in Computer Science*, pp. 131–141, 2024. doi: 10.1007/978-981-99-9785-5_10.

Zhibo Wang, Hengchang Guo, Zhifei Zhang, Wenxin Liu, Zhan Qin, and Kui Ren. Feature importance-aware transferable adversarial attacks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 7639–7648, 2021.

Zhou Wang, Alan C. Bovik, Hamid R. Sheikh, and Eero P. Simoncelli. Image quality assessment: From error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4): 600–612, 2004. doi: 10.1109/TIP.2003.819861.

Cihang Xie, Zhishuai Zhang, Yuyin Zhou, Song Bai, Jianyu Wang, Zhou Ren, and Alan L. Yuille. Improving transferability of adversarial examples with input diversity. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2730–2739, 2019.

Guanshuo Xu, Han-Zhou Wu, and Yun-Qing Shi. Structural design of convolutional neural networks for steganalysis. *IEEE Signal Processing Letters*, 23(5):708–712, 2016. doi: 10.1109/LSP.2016.2548421.

Zihao Xu, Dawei Xu, Zihan Li, and Chuan Zhang. MDDM: Practical message-driven generative image steganography based on diffusion models. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 69832–69848, 2025.

Yiwei Yang, Zheyuan Liu, Jun Jia, Zhongpai Gao, Yunhao Li, Wei Sun, Xiaohong Liu, and Guangtao Zhai. DiffStega: Towards universal training-free coverless image steganography with diffusion models. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence*, pp. 1579–1587, 2024. doi: 10.24963/ijcai.2024/175. URL https://www.ijcai.org/proceedings/2024/175.

Jian Ye, Jiangqun Ni, and Yang Yi. Deep learning hierarchical representations for image steganalysis. *IEEE Transactions on Information Forensics and Security*, 12(11):2545–2557, 2017. doi: 10.1109/TIFS.2017.2710946.

Chengsheng Yuan, Zhaonan Ji, Qi Cui, Zhili Zhou, Xinting Li, and Zhihua Xia. Towards provably secure and highly robust generative image steganography leveraging latent diffusion model. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 1498–1506, 2026. doi: 10.1609/aaai.v40i2.37125. Issue 2.

Jianping Zhang, Weibin Wu, Jen-tse Huang, Yizhan Huang, Wenxuan Wang, Yuxin Su, and Michael R. Lyu. Improving adversarial transferability via neuron attribution-based attacks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14993–15002, 2022.

Qianqian Zhang, Yinhong Huang, Duo Ma, Xiangyang Luo, Yuanyuan Ma, and Ning Ruan. Lightweight steganalysis with low distortion for countering robust image steganography. *IEEE Transactions on Circuits and Systems for Video Technology*, 2026. doi: 10.1109/tcsvt.2026.3701492. URL https://doi.org/10.1109/tcsvt.2026.3701492. Early Access.

Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 586–595, 2018. doi: 10.1109/CVPR.2018.00068.

Zhiying Zhu, Ping Wei, Zhenxing Qian, Sheng Li, and Xinpeng Zhang. Image sanitization in online social networks: A general framework for breaking robust information hiding. *IEEE Transactions on Circuits and Systems for Video Technology*, 33(6):3017–3029, 2023. doi: 10.1109/tcsvt.2022.3224243. URL https://doi.org/10.1109/tcsvt.2022.3224243.