Article | Received 12 November 2025; Revised 8 April 2026; Accepted 13 April 2026; Published 29 June 2026
https://doi.org/10.55092/aic20260009

BS-ADV: a secure black-box image steganography framework enhanced by adversarial sample attack

Shichen Yang[1], Haiyu Xu[1], Xingxing Jia[1,*], Guodong Ye[2], Chengsheng Yuan[3] and Huiyu Zhou[4]

[1] School of Mathematics and Statistics, Lanzhou University, Lanzhou 730000, China

[2] School of Mathematics and Computer Science, Guangdong Ocean University, Zhanjiang 524088, China

[3] School of Cyber Science and Engineering, Nanjing University of Information Science and Technology, Nanjing 210044, China

[4] School of Computing and Mathematical Sciences, University of Leicester, Leicester LE1 7RH, UK

* Correspondence author; E-mail: jiaxx@lzu.edu.cn.

Highlights:

- Proposes BS-ADV, a transferable adversarial framework that enhances image steganography security in black-box settings without requiring access to encoder internals.
- Introduces FGSM-adv and PGD-adv strategies that effectively improve resistance against both feature-based and deep learning-based steganalyzers while preserving image quality.
- Demonstrates strong generalization across conventional, image-in-image, and diffusion-based coverless steganography systems, achieving substantial security gains with reliable payload recovery.

Abstract: Mainstream image steganography approaches struggle to evade increasingly accurate steganalyzers, while many security-enhancement techniques require white-box access to the steganographic encoder, hindering deployment in black-box settings. We address this gap with a framework that achieves strong security and high image quality without access to encoder internals. Our method Black-box Steganography via Transferable Adversarial Attack (BS-ADV) adds small, transferable adversarial perturbations to stego images using gradients from a chosen steganalysis model, yet remains independent of the steganographic encoder. Building on this idea, we instantiate two variants: FGSM-adv, which applies a single-step Fast Gradient Sign Method to inject fixed-sign perturbations, and PGD-adv, which performs multi-step Projected Gradient Descent to enhance the robustness and security of the resulting adversarial stego images. Experiments on public BOSSBase 1.01 (Break Our Steganographic System Base v1.01) and BOWS (Break Our Watermarking System 2) datasets show that BS-ADV substantially outperforms baseline approaches against both feature-based and convolutional neural network (CNN)-based steganalyzers. Beyond conventional algorithms, we further validate BS-ADV with DeepSteganography, HiNet, and CRoSS (diffusion model makes controllable, robust and secure image steganography), a coverless steganography scheme built on Stable Diffusion, demonstrating broad generality and adaptability. Overall,

BS-ADV improves the security and robust-ness of image steganography while preserving image quality and reliable payload recovery, making it well suited for practical black-box deployment.

Keywords: image steganography; image adversarial; generative adversarial networks; diffusion models; steganography analysis

1. Introduction

Steganography conceals secret messages within digital media so as to evade detection by human observers and automated steganalyzers. With the ubiquity of image sharing on social media, images have become the predominant cover medium. In image steganography, algorithms under the distortion-minimization paradigm are mainstream because of their strong covertness: they preferentially embed payloads in textured or noisy regions to reduce detectability. Within this framework, methods fall into two categories: heuristic approaches such as Highly Undetectable stego (HUGO) [1], Spatial Universal Wavelet Relative Distortion (S-UNIWARD) [2], High-pass, Low-pass, and Low-pass (HILL) [3], Wavelet Obtained Weights (WOW) [4], and deep-learning-based approaches such as Hiding Images within Images Using Conditional Generative Adversarial Networks (StegGAN) [5], High Capacity Image Steganography with GANs (SteganoGAN) [6], Unpaired Image-to-Image Translation Using Cycle-Consistent Adversarial Networks (CycleGAN) [7], and Controllable, Robust and Secure Steganography (CRoSS) [8].

Heuristic methods construct a distortion function and employ encoding strategies such as Syndrome-Trellis Codes [9], thereby turning embedding into a distortion minimization problem. Pevný *et al*. [1] proposed HUGO, an early algorithm within this framework that assigns pixel costs via a weighted-norm function to concentrate embedding in regions with minimal impact on statistical properties. Holub *et al*. [4] proposed WOW, which computes directional filter responses via multiresolution wavelet transforms and concentrates distortion in highly textured areas to reduce detection risk. Li *et al*. [3] introduced HILL, which fuses high and low-pass filtering to optimize cost allocation and substantially improve security. Holub *et al*. [2] proposed S-UNIWARD, a universal distortion function applicable in both spatial and transform domains, offering high flexibility and generality. These methods largely rely on expert-designed cost functions and models and achieved notable early successes in steganographic security and resistance to steganalysis, laying the theoretical foundation for modern techniques.

However, as steganalysis advances rapidly, handcrafted distortion functions encounter bottlenecks. Manually designed costs struggle to capture complex higher-order statistics in images, limiting performance under strong detection scenarios. With the rise of deep learning, data-driven end-to-end learning has brought new breakthroughs. This paradigm reframes steganography as a distribution-matching problem: neural networks learn statistical differences between cover and stego images that are least detectable, rather than relying on fixed mathematical assumptions. The introduction of Generative Adversarial Networks (GANs) further optimizes the goal of "minimizing statistical detectability" via adversarial training. GANs, consisting of a generator and a discriminator trained in a game-theoretic manner, are widely used for high-quality image synthesis.

Given their ability to model cover-image distributions, GANs have been widely explored for steganography. StegGAN [5] is among the earliest attempts: its unsupervised adversarial training

optimizes embedding strategies and improves distribution matching, yet its security and extraction accuracy still lag behind heuristic methods. Subsequent work [6] designed deep convolutional architectures to embed arbitrary binary data. Zhu *et al*. [10] proposed Hiding Data with Deep Networks (HiDDeN), which enables flexible trade-offs among payload, security, and robustness to diverse perturbations. Nevertheless, owing to task complexity, current end-to-end deep steganography continues to face challenges in security and generalization, particularly against black-box analyzers.

To further enhance security, several studies learn distortion costs automatically within GAN-based schemes. For example, SPA-RL [11] combines reinforcement learning with adversarial training to resist convolutional neural network (CNN) based steganalysis. Yao *et al*. [12] proposed an Layout Diffusion Generative Model (LDGM)-based adaptive steganographic coding algorithm that fits arbitrary distortion functions and accommodates arbitrary cover lengths, users can select decimation methods to trade off time complexity against security performance. Chen *et al*. [13] proposed a steganographic immunoprocessing (IP) framework based on the artificial immune system (AIS) that improves security by rationally defining distortion costs. However, the security of GAN-based steganography is often constrained by discriminator capacity, The Generative Multi-Adversarial Network (GMAN) [14] addresses this limitation with a multi-discriminator training strategy designed to improve security.

Meanwhile, image steganography has progressed toward two emerging directions—image-in-image steganography and coverless image steganography to increase payload, expressiveness, and flexibility. In image-in-image steganography, a secret image is embedded into another image to achieve higher hiding capacity.Typical models include Deep Steganography [15] and Invisible Steganography via Generative Adversarial Networks (ISGAN) [16]. The latter adopts a luminance-channel embedding strategy to improve imperceptibility and security.

Coverless steganography directly generates secret-bearing images, eliminating reliance on a traditional cover and exhibiting distinctive properties. Latent-diffusion models [17] and CycleGAN [7] have been widely adopted. Bengio *et al*. [18] discussed encoder–decoder architectures; contemporary coverless methods commonly adopt such designs and leverage cycle-consistency to produce stego images. These approaches partially evade steganalysis yet still face notable limitations. In many pipelines, container images are randomly sampled from the generative model, limiting user control over semantic content. Moreover, image-in-image methods typically hide information in a single cover image and lack the capacity to conceal richer or more complex visual content.

Built on latent diffusion [17], the text-to-image model Stable Diffusion [19] iteratively denoises in latent space and then decodes to the image domain, greatly lowering the barrier to high-resolution synthesis. Representative diffusion models Stable Diffusion [19] and Midjourney [20] have achieved widespread adoption, with Midjourney reporting a user base exceeding 14.5 million [21]. Images produced by Stable Diffusion are now ubiquitous, making such models attractive candidate carriers for steganography. Yu *et al*. [8] introduced CRoSS, a diffusion-based coverless framework that offers improved controllability, robustness, and security relative to cover-based methods. GTSD [22] proposes a diffusion-based generative text steganography method that embeds secret information via prompt mapping and batch mapping, thereby alleviating the generation-time bottleneck of prior methods. Nevertheless, existing image-in-image and coverless approaches still struggle to jointly balance controllability, robustness, and security; a unified,

effective solution has yet to emerge.

The rapid advancement of deep learning has recently led to significant breakthroughs in image steganalysis, further intensifying the threat to steganographic security. To improve detection efficiency and mitigate spatiotemporal overhead, recent studies have proposed steganalysis feature selection methods utilizing multidimensional evaluation and dynamic threshold allocation [23]. Expanding evolutionary computation into this domain, researchers have also introduced enhanced image steganalysis frameworks via opposition-based evolutionary algorithms, effectively concentrating the model's attention on critical feature representations [24]. Concurrently, advanced neural architectures are continuously being tailored for precise steganalysis. For instance, recent works have developed dual-path enhancement and fractal downsampling mechanisms to capture hidden statistical artifacts across both spatial and JPEG domains [25]. Similarly, selective pooling strategies coupled with enhanced transformers have been designed to achieve state-of-the-art detection on arbitrary-sized color stego images [26]. Furthermore, self-distillation frameworks with feature pyramids have been introduced to refine feature extraction for steganalysis without relying on computationally heavy architectures [27]. These increasingly sophisticated and accurate steganalyzers make it exceedingly difficult for conventional image steganography approaches to evade detection, underscoring the urgent need for robust, black-box security-enhancement techniques such as transferable adversarial attacks.

To address these issues, recent research introduces adversarial attacks to improve the undetectability of stego images. Adversarial examples add perturbations imperceptible to humans to force machine-learning models into erroneous decisions [28,29]. This technique is widely used in classification, segmentation, and detection. In steganography, adversarial examples can enhance security by keeping visual consistency while making images harder for steganalyzers to identify. Specifically, most attack methods [30,31] focus on adding a noise word or rewriting the entire prompt to craft the adversarial prompt. This significant added noise can be easily detected, thus reducing the attack imperceptibility. However, most existing methods assume white-box access and rely on the target steganographic model's structure, parameters, and gradients. This limits applicability in realistic black-box scenarios, particularly for steganographic encoders based on deep generative or diffusion architectures whose internals are not public.

Against this backdrop, black-box adversarial attacks have gained attention. They include query-based attacks [32] and transfer-based attacks [33,34]. The latter require no knowledge of model structure: adversarial perturbations crafted on a surrogate can transfer across multiple targets and induce misclassification. Such methods are especially suitable when the encoder architecture is unknown or inaccessible, as in diffusion-based coverless steganography.

Accordingly, we propose BS-ADV, a general security-enhancement framework for steganography that leverages transfer-based black-box adversarial techniques to improve the resilience of diverse stego images under black-box analysis. The framework does not depend on the target steganographic model's structure or gradients. Instead, it generates transferable perturbations on a surrogate steganalyzer to boost resistance to detection while preserving high visual quality. Concretely, we provide two implementations:

(1) FGSM-adv: a single-step Fast Gradient Sign Method that injects small, fixed-sign perturbations into stego images to mislead steganalyzers.

(2) PGD-adv: a multi-step Projected Gradient Descent procedure that performs iterative projected

updates to produce more robust adversarial stego images.

Both methods apply to different steganographic generators, including DeepSteganography [15] and the coverless diffusion-based CRoSS [8]. Using gradients from the surrogate YeNet [35], we add perturbations to stego images and obtain high-quality adversarial stego images that resist feature-based steganalysis [36] and CNN-based analyzers such as YeNet [35], ZhuNet [37], XuNet [38], and SiaStegNet [39]. Experiments show that the proposed adversarial steganography methods significantly outperform existing baselines against both (Spatial Rich Model (SRM) and CNN families while maintaining image quality, validating the effectiveness and applicability of the framework.

The contributions of this paper are summarized as follows:

(1) We introduce BS-ADV, the first (to our knowledge) transfer-based, black-box adversarial framework for strengthening steganographic security, which substantially improves resistance to detection without requiring access to encoder internals and is applicable to black-box generative settings.

(2) We instantiate two concrete variants FGSM-adv and PGD-adv that offer complementary efficiency robustness trade-offs, and we apply this approach, for the first time to our knowledge, to diffusion-based coverless steganography systems.

(3) Extensive experiments show that, relative to baseline schemes, BS-ADV improves security against the SRM steganalyzer by 26.40% and against CNN-based detectors (YeNet, ZhuNet, XuNet, SiaStegNet) by up to 34.95%, while preserving visual quality (and reliable payload recovery), thereby enhancing both the security and practicality of stego images.

2. Prerequisite knowledge

This section reviews the background needed for BS-ADV. We first outline the principles of adversarial perturbations for images, with emphasis on transfer-based black-box attacks. We then summarize two canonical algorithms for generating adversarial examples: Fast Gradient Sign Method (FGSM) and Projected Gradient Descent (PGD)—which underlie our subsequent implementations.

2.1. Image adversarial attack

An image adversarial attack adds imperceptible perturbations to an input image so as to induce a deep neural network to make an incorrect prediction. Formally, given an image y and a label c, an attacker seeks a perturbation δ such that, for a model $f(\cdot)$, the following holds:

$$P = \begin{cases} 1, & \text{if } \operatorname{argmax}(f(y_{\text{adv}})) = c' \\ 0, & \text{otherwise} \end{cases} \tag{1}$$

where $y_{\text{adv}} = y + \delta$ is misclassified by the model as the target class c'. The attacker's goal is to maximize the success probability, which is commonly expressed as the constrained minimization:

$$\min_{\delta} \|\delta\|_p \quad \text{s.t.} \quad \operatorname{argmax}(f(y+\delta)) = c' \tag{2}$$

where $\|\cdot\|_p$ denotes the L_p norm, and p is typically 2 (the L_2 norm) or ∞ (the L_∞ norm). Among existing strategies, transfer-based methods are particularly effective: one first crafts adversarial examples on a surrogate model and then applies them to a different model. The key is to exploit transferability so that

adversarial examples remain effective across models.

2.2. Transfer-based black-box adversarial attacks

In practice, black-box adversarial attacks are especially relevant because the internal structure of the target model is often unavailable. Transfer-based black-box attacks use adversarial examples generated on a source model $f_s(\cdot)$ to induce misclassification on another target model $f_t(\cdot)$ with unknown architecture. The success condition can be written as

$$P_{\text{attack}} = \begin{cases} 1, & \text{if } \operatorname{argmax}(f_t(y_{\text{adv}})) = c' \\ 0, & \text{otherwise} \end{cases} \tag{3}$$

and the objective becomes

$$\min_{\delta} \|\delta\|_p \quad \text{s.t.} \quad \operatorname{argmax}(f_s(y+\delta)) = \operatorname{argmax}(f_s(y)) \tag{4}$$

This paradigm does not require querying the target model f_t; it only needs to construct perturbations with sufficient transferability, which aligns well with the constraints of practical black-box steganographic models.

2.3. FGSM

FGSM [40], proposed by Goodfellow, is an efficient technique for generating adversarial examples. FGSM applies a one-step, directionally aligned perturbation by computing the sign of the gradient of the loss with respect to the input image:

$$\delta = \varepsilon \cdot \operatorname{sign}(\nabla_y L(\theta, y, y_{\text{adv}})) \tag{5}$$

where ε controls the perturbation magnitude, and $\nabla_y L(\theta, y, y_{\text{adv}})$ denotes the gradient of the loss L with respect to the input y. Owing to its simplicity and speed, FGSM scales well to large image collections.

2.4. PGD

PGD [41] is a stronger adversarial example generator that uses multi-step iterative optimization. Unlike FGSM's single-step update, PGD recomputes gradients and updates the perturbation at each iteration while projecting it back into the allowed region, thereby yielding more robust adversarial examples. The per-iteration update is

$$\delta_{t+1} = \operatorname{clip}_{\varepsilon}(\delta_t + \alpha \cdot \operatorname{sign}(\nabla_y L(\theta, y, y_{\text{adv}}))) \tag{6}$$

where δ_t is the perturbation at step t, α is the step size, and $\operatorname{clip}_{\varepsilon}$ constrains the perturbation to $[-\varepsilon, \varepsilon]$. By controlling both range and intensity through multiple iterations, PGD typically achieves higher attack success rates and stronger robustness than FGSM, which is desirable for steganographic scenarios with stringent security requirements.

3. Black-box image steganography security enhancement framework based on adversarial examples

This paper proposes BS-ADV, a black-box adversarial example-based steganographic security enhancement framework, which aims to improve the security of stego images against steganalysis detection. BS-ADV

integrates conventional steganographic embedding with black-box adversarial perturbations, enhancing the deception capability of stego images against steganalysis models while preserving visual quality.

3.1. Overview of the BS-ADV framework

In the field of generative image steganography, steganographic algorithms embed secret information into a cover image X to produce a stego image Y, thereby enabling information hiding. However, with advancements in deep learning-based steganalysis, existing steganographic methods still exhibit limited security when confronted with powerful CNN-based steganalyzers. To address this issue, we propose BS-ADV an adversarial steganographic framework that combines steganography with black-box adversarial attack techniques. The core idea is to first embed secret information into the cover image X using steganography, generating an initial stego image Y. Then, black-box adversarial attack techniques are applied to produce a perturbed adversarial stego image Y'. By introducing minimal perturbations, the adversarial stego image remains visually indistinguishable from the original, while effectively deceiving steganalysis models and enhancing security. The BS-ADV framework consists of three stages: the information hiding phase, the adversarial attack phase, and the extraction phase, as illustrated in Figure 1 and Figure 2.

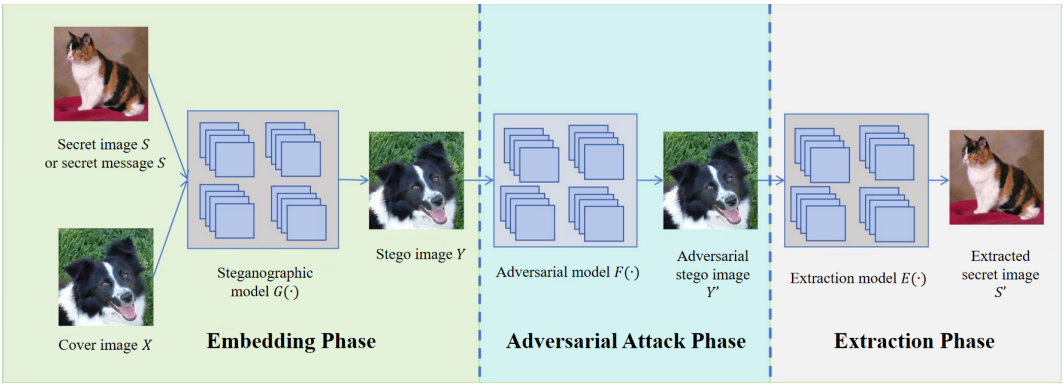


Figure 1. Basic architecture of the adversarial steganography framework.

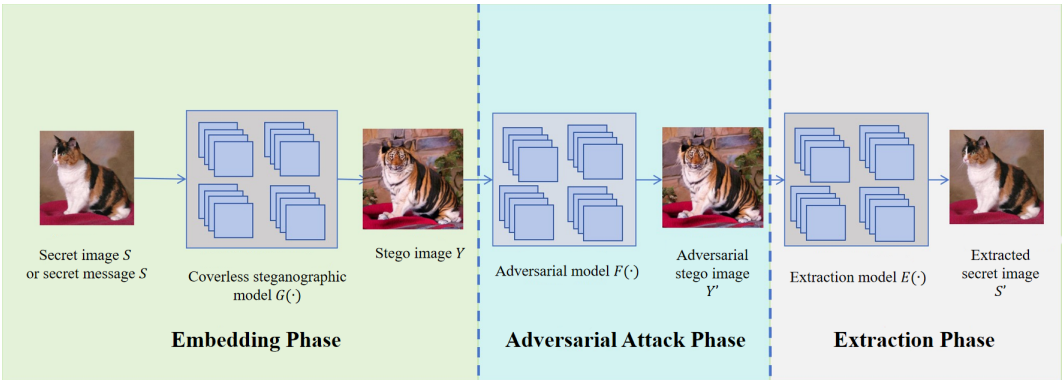


Figure 2. Basic architecture of the adversarial steganography framework for coverless scenarios.

(1) Information Hiding Phase: The steganographic embedding model $G(\cdot)$ is used to embed secret information or a secret image S into the cover image X, producing the stego image Y, *i.e.*,

$$Y = G(X, S) \tag{7}$$

where $G(\cdot)$ can be a steganographic model such as WOW[4], S-UNIWARD[2], HiNet [42], or DeepSteganography [15].

(2) Adversarial Attack Phase: A black-box adversarial attack model $F(\cdot)$ is employed to add imperceptible perturbations to the stego image Y, generating the adversarial stego image Y', *i.e.*,

$$Y' = F(Y) \tag{8}$$

where $F(\cdot)$ can be an adversarial attack model such as FGSM or PGD.

(3) Information Extraction and Verification: An extraction model $E(\cdot)$ is used to recover the secret information S' from the adversarial stego image Y', *i.e.*,

$$S' = E(Y') \tag{9}$$

where $E(\cdot)$ denotes the model responsible for extracting the secret information.

3.2. Loss function design

To enhance the effectiveness and security of adversarial example generation, the proposed BS-ADV framework combines FGSM and PGD methods with the goal of minimizing distortion, forming a unified loss function system. This system integrates classification loss, distortion loss, contrast loss, gradient consistency loss, and pixel smoothness loss to balance the attack performance and visual quality of adversarial examples.

3.2.1. General loss terms

The following loss terms are defined identically in both FGSM-adv and PGD-adv:

(1) Classification Loss

The classification loss measures the discrepancy between the model's prediction on the adversarial example and the true label. Cross-entropy loss is adopted for this purpose. Let $\hat{c}$ be the prediction on adversarial example y_{adv} and c_{true} be the ground truth label. The cross-entropy loss is defined as:

$$\mathscr{L}_{\text{cls}} = -\sum_{m=1}^{M} c_{\text{true}}^{(m)} \log\left(f(y_{\text{adv}})^{(m)}\right) \tag{10}$$

where $f(\cdot)$ denotes the steganalysis model and M is the total number of classes. This term drives the adversarial example to mislead the classifier. By maximizing the classifier's prediction error (*i.e.*, causing the steganalyzer to misclassify the stego image as a cover), it directly enhances the undetectability of the adversarial example at the model level. Its design originates from the fundamental principle of adversarial attacks, which involves perturbing the image along the gradient of the loss function to mislead the target model.

(2) Distortion Loss

The distortion loss constrains the perceptual similarity between the adversarial example and the original image. We use the Structural Similarity Index Measure (SSIM) to quantify this loss:

$$\mathscr{L}_{\text{dis}} = \begin{cases} -\text{SSIM}(x, y_{\text{adv}}), & \text{for cover-based steganography} \\ -\text{SSIM}(y, y_{\text{adv}}), & \text{for coverless steganography} \end{cases} \tag{11}$$

where x is the cover image and y is the stego image. The SSIM loss helps maintain structural consistency through local feature matching. SSIM better aligns with the human visual system's (HVS) perception of structural information than simple pixel-wise errors (e.g., MSE). For cover-based steganography, it constrains the similarity between the adversarial image and the original cover x; for coverless steganography, it constrains the similarity with the initial stego image y. This design is intended to strictly guarantee the imperceptibility of adversarial perturbations to the human eye while pursuing security, thereby avoiding the introduction of noticeable artifacts.

(3) Gradient Consistency Loss

This loss preserves the gradient distribution characteristics of the image. We define the gradient difference between the stego image and the adversarial example as:

$$\mathscr{L}_{\text{grad}} = \|\nabla y - \nabla y_{\text{adv}}\|_2 \tag{12}$$

It suppresses abnormal edge features introduced by adversarial perturbations. The image gradient reflects its edge and texture features. Many steganalyzers, especially rich-model (SRM) ones, utilize gradient or higher-order statistical features for detection. This loss constrains the adversarial perturbations from significantly altering the distribution of the image gradient field. Its purpose is to suppress unnatural edge or texture patterns introduced by the perturbations, making the adversarial image statistically more similar to natural images and further enhancing resistance against feature-based steganalyzers.

3.2.2. Method-specific loss

(1) Contrast Loss in FGSM-adv

To compensate for possible local contrast distortion caused by single-step perturbation, FGSM-adv introduces an L_2 norm constraint:

$$\mathscr{L}_{\text{con}}^{\text{FGSM}} = \|y - y_{\text{adv}}\|_2 \tag{13}$$

This L_2 constraint stabilizes the single-step update by preventing excessive pixel-level deviations, thereby smoothing the perturbation distribution and complementing the structural guidance of the SSIM loss.

(2) Smoothness Loss in PGD-adv

To mitigate high-frequency noise that may accumulate during multi-step iterations, PGD-adv incorporates a neighbor pixel constraint:

$$\mathscr{L}_{\text{smooth}}^{\text{PGD}} = \sum \|y_{\text{adv}}[i,j] - y_{\text{adv}}[i+1,j]\|_2 + \|y_{\text{adv}}[i,j] - y_{\text{adv}}[i,j+1]\|_2 \tag{14}$$

This term enforces spatial coherence between neighboring pixels, effectively suppressing detectable high-frequency noise that can arise from multi-step iterative optimization, thus preserving visual and statistical naturalness.

(3) Loss Function

Both methods share a unified optimization framework but use different loss weight configurations:

$$\mathscr{L}_{\text{total}} = \mathscr{L}_{\text{cls}} + \lambda \mathscr{L}_{\text{dis}} + \gamma \mathscr{L}_{\text{grad}} + \begin{cases} \alpha \mathscr{L}_{\text{con}}^{\text{FGSM}}, & \text{for FGSM-adv} \\ \beta \mathscr{L}_{\text{smooth}}^{\text{PGD}}, & \text{for PGD-adv} \end{cases} \tag{15}$$

The aforementioned loss terms together form a multi-objective optimization problem. $\mathscr{L}_{\text{cls}}$ serves as the dominant term, ensuring attack effectiveness. $\mathscr{L}_{\text{dis}}$ and $\mathscr{L}_{\text{grad}}$ act as core fidelity constraints, preserving image quality from the perspectives of visual perception and statistical characteristics, respectively, preventing the adversarial optimization from deviating excessively. The method-specific $\mathscr{L}_{\text{con}}^{\text{FGSM}}$ or $\mathscr{L}_{\text{smooth}}^{\text{PGD}}$ further fortifies the framework against the inherent weaknesses of each optimization path. The hyperparameters λ, γ, α, and β are used to precisely tune the trade-off among these objectives. Through this design, the BS-ADV framework can significantly enhance its deceptive capability against steganalyzers while ensuring that the generated adversarial stego images maintain high visual imperceptibility and statistical naturalness.

3.3. FGSM-adv and PGD-adv black-box secure image steganography algorithms

The FGSM-adv and PGD-adv black-box secure image steganography models integrate conventional steganography with black-box adversarial attack techniques. The methodology involves embedding secret information into a cover image X using steganography to produce a stego image Y, followed by the application of FGSM and PGD methods to generate an adversarial stego image Y_{adv}. By introducing minimal perturbations, these models ensure that the adversarial stego image remains visually indistinguishable from the original while effectively deceiving steganalysis models, thereby enhancing the security of stego images.

In the proposed framework, adversarial perturbations are generated on a surrogate steganalysis model and transferred to the target detector. This transfer-based design enables black-box deployment, as it does not require access to the architecture, parameters, or gradients of the target steganalyzer, while still preserving effective misclassification capability across heterogeneous detection models.

Both FGSM-adv and PGD-adv operate under an L_∞-norm perturbation constraint ε, which bounds the maximum pixel-wise modification and ensures that the generated adversarial stego image $Y_{\text{adv}} = Y + \delta$ maintains high visual fidelity and reliable payload extraction.

FGSM-adv generates adversarial stego images using single-step gradient updates. It utilizes a composite loss function that includes classification loss, distortion loss, contrast loss, gradient consistency loss, and pixel smoothness loss. Specifically, FGSM-adv computes the gradient of the total loss with respect to the stego image and applies a fixed-sign update $\delta = \varepsilon \cdot \text{sign}(\nabla_Y L)$, which efficiently drives the adversarial example toward maximizing the detection error of the surrogate steganalysis model while incurring minimal computational overhead.

PGD-advemploys multi-step iterative optimization to generate adversarial stego images. In addition to the aforementioned losses, it incorporates an additional pixel smoothness loss to suppress high-frequency noise. Through iterative projection into the L_∞-bounded perturbation space, PGD-adv refines the adversarial perturbation across multiple steps, enabling stronger robustness and improved transferability to unseen steganalysis models at the cost of increased computational complexity.

By introducing minimal perturbations, both models ensure the adversarial stego image remains visually indistinguishable from the original while effectively enhancing security against steganalysis. The detailed step-by-step algorithms for FGSM-adv and PGD-adv are provided in Algorithm 1 and Algorithm 2, respectively.

Algorithm 1 FGSM-adv Black-box Steganographic Attack Algorithm

Require: Cover image X, secret information s, true label c_{true}, perturbation magnitude ε, hyperparameters λ, α, and γ

Ensure: Adversarial steganographic image Y_{adv}

Steganographic Embedding: Use the steganographic embedding model $G(\cdot)$ to embed the secret information s into the cover image X to generate the stego image Y:

$$Y = G(X, s).$$

Adversarial Attack:

1. Initialize the adversarial steganographic image $Y_{\text{adv}} = Y$;
2. Compute the gradient of the loss function with respect to the input: $\nabla_Y \mathscr{L} = \frac{\partial \mathscr{L}}{\partial Y_{\text{adv}}}$;
3. Generate perturbation: $\delta = \varepsilon \cdot \text{sign}(\nabla_Y \mathscr{L})$;
4. Update the adversarial steganographic image: $Y_{\text{adv}} = Y_{\text{adv}} + \delta$;
5. Clip the perturbation to ensure $\|\delta\|_\infty \leq \varepsilon$;
6. Compute classification loss: $\mathscr{L}_{\text{cls}} = -\sum_{m=1}^{M} c_{\text{true}}^m \log(f(Y_{\text{adv}})^m)$;
7. Compute distortion loss: $\mathscr{L}_{\text{dis}} = -\text{SSIM}(X, Y_{\text{adv}})$;
8. Compute contrast loss: $\mathscr{L}_{\text{con}} = \|Y - Y_{\text{adv}}\|_2$;
9. Compute gradient consistency loss: $\mathscr{L}_{\text{grad}} = \|\nabla_Y Y - \nabla_Y Y_{\text{adv}}\|_2$;
10. Compute total loss: $\mathscr{L} = \mathscr{L}_{\text{cls}} + \lambda \cdot \mathscr{L}_{\text{dis}} + \alpha \cdot \mathscr{L}_{\text{con}} + \gamma \cdot \mathscr{L}_{\text{grad}}$;

ReturnAdversarial steganographic image Y_{adv}

Algorithm 2 PGD-adv Black-box Steganographic Attack Algorithm

Require: Cover image X, secret information s, true label c_{true}, perturbation magnitude ε, step size α, number of iterations T, hyperparameters λ, γ, and β

Ensure: Adversarial steganographic image Y_{adv}

Steganographic Embedding: Use the steganographic embedding model $G(\cdot)$ to embed the secret information s into the cover image X to generate the stego image Y:

$$Y = G(X, s).$$

Adversarial Attack:

1. Initialize the adversarial steganographic image $Y_{\text{adv}} = Y$;
2. For $t = 1$ to T do
   (a) Compute the gradient of the loss function with respect to the input: $\nabla_Y \mathscr{L} = \frac{\partial \mathscr{L}}{\partial Y_{\text{adv}}}$;
   (b) Generate perturbation: $\delta_t = \delta_{t-1} + \alpha \cdot \text{sign}(\nabla_Y \mathscr{L})$;
   (c) Project the perturbation into the allowed range: $\delta_t = \text{clip}_\varepsilon(\delta_t)$;
   (d) Update the adversarial steganographic image: $Y_{\text{adv}} = Y + \delta_t$;
   (e) Compute classification loss: $\mathscr{L}_{\text{cls}} = \text{CE}(f(Y_{\text{adv}}), c_{\text{true}})$;
   (f) Compute distortion loss: $\mathscr{L}_{\text{dis}} = -\text{SSIM}(X, Y_{\text{adv}})$;
   (g) Compute gradient consistency loss: $\mathscr{L}_{\text{grad}} = \|\nabla_Y Y - \nabla_Y Y_{\text{adv}}\|_2$;
   (h) Compute pixel smoothness loss: $\mathscr{L}_{\text{smooth}} = \sum_{i,j} \|Y_{\text{adv}}[i,j] - Y_{\text{adv}}[i+1,j]\|_2 + \|Y_{\text{adv}}[i,j] - Y_{\text{adv}}[i,j+1]\|_2$;
   (i) Compute total loss: $\mathscr{L} = \mathscr{L}_{\text{cls}} + \lambda \cdot \mathscr{L}_{\text{dis}} + \gamma \cdot \mathscr{L}_{\text{grad}} + \beta \cdot \mathscr{L}_{\text{smooth}}$;
3. end for

Return: Adversarial steganographic image Y_{adv}

4. Experiments and results analysis

To validate the effectiveness of the proposed BS-ADV framework along with the FGSM-adv and PGD-adv methods, we conduct comprehensive experiments from multiple perspectives. These experiments encompass the datasets, steganographic methods, steganalysis models, evaluation metrics, hyperparameter settings, security analysis, imperceptibility evaluation, and coverless steganography experiments.

4.1. Experimental setup

4.1.1. Implementation details and experimental environment

All experiments were conducted using Python 3.8.13 and PyTorch 1.11.0 within a Conda-managed environment. The training and evaluation pipeline, including adversarial perturbation generation and steganalysis inference, was implemented in a unified codebase to ensure consistent execution across all experiments. We use PyTorch for training on NVIDIA RTX 3090 GPUs. Unless otherwise specified, all steganalysis networks were optimized using the Adam optimizer with a weight decay of 1×10^{-5} and an initial learning rate of 2×10^{-4}. Model checkpoints and experimental logs were saved throughout training and used for subsequent evaluation and result analysis.

4.1.2. Datasets

The experiments employ three benchmark datasets:

(1) BOSSBase 1.01: Consists of 10,000 grayscale images. To improve experimental efficiency, all images are resized to 256×256 pixels. This dataset is primarily used for testing both heuristic and deep learning-based steganographic methods.

(2) BOWS2: Contains 10,000 grayscale images. Similarly, images are resized to 256×256 pixels for generalization experiments.

(3) Dogs *vs.* Cats: For training the diffusion model-based steganography method CRoSS, we utilize 500 cat images from the training set of the 2013 Kaggle Dogs *vs.* Cats dataset. These images are used to embed secret information into tiger images for coverless steganography testing.

4.1.3. Steganographic and steganalysis methods

We evaluate seven steganographic algorithms, including three heuristic schemes: S-UNIWARD [2], WOW [4], and HILL [3]; the GAN-based UT-GAN [43]; image-to-image steganography using DeepSteganography [15] and HiNet [42]; and the diffusion model-based coverless steganography method CRoSS [8] for generating initial target images. Among these, UT-GAN uses a pre-trained 0.4 bpp embedding model from Yang *et al*. [43], trained for 120,000 iterations (equivalent to 72 epochs).

To assess the security of the FGSM-adv algorithm, we employ six advanced steganalysis models, including one handcrafted feature-based model, SRM [36], and five deep learning-based approaches. For SRM, we combine it with an ensemble classifier. The deep learning-based methods include, in addition to the YeNet [35] model used for adversarial example generation, two state-of-the-art CNN-based steganalysis networks, ZhuNet [37] and XuNet [38], enabling a comprehensive security evaluation of the proposed BS-ADV framework.

Based on extensive preliminary experiments, we observed that generating adversarial examples using gradient information from the YeNet model provides more stable transferability and superior attack performance. Consequently, a YeNet model trained for 100 epochs is used as the reference model to generate steganographic adversarial images using its gradient information. For SRM, cover images and their corresponding stego images are randomly split into a 4:1 ratio for training and testing. The security performance of BS-ADV is evaluated using these five steganalysis models on both the BOSSBase 1.01

and BOWS2 datasets. Images are divided into three subsets in an 8:1:1 ratio for training, validation, and testing, respectively.

4.1.4. Evaluation metrics

To quantitatively evaluate the performance of the proposed framework, both steganographic security and image quality metrics are employed:

(1) Detection Error Rate (P_E): Serving as the primary metric for assessing steganalyzers, it is defined as $P_E = \frac{1}{2}(P_{FA} + P_{MD})$, where P_{FA} and P_{MD} denote the probabilities of false alarm and missed detection, respectively. In binary steganalysis, the detection error P_E of 0.5 corresponds to random guessing, representing the ideal scenario where the steganalyzer performs no better than chance—*i.e.*, the theoretical optimum for steganographic security.

(2) Peak Signal-to-Noise Ratio (PSNR): PSNR evaluates pixel-level fidelity on a logarithmic decibel (dB) scale, calculated via the Mean Squared Error (MSE). In this paper, it is employed not only to measure the objective distortion between the cover and adversarial stego images but also to quantify the reconstruction quality between the extracted secret images and the original secret images. A higher PSNR value indicates lower pixel-level deviation.

(3) Structural Similarity Index (SSIM): Compared to PSNR, SSIM better aligns with the human visual system (HVS) by assessing structural integrity across three dimensions: luminance, contrast, and structure. Ranging from 0 to 1, SSIM is utilized in this study both to measure the perceptual fidelity of generated stego images and to verify whether the extracted secret images reliably preserve the edges and textures of the original secret images. An SSIM value closer to 1 implies higher visual consistency.

4.2. Hyperparameter selection

The method proposed in this paper involves five key hyperparameters: the perturbation magnitude ε in adversarial steganographic image generation, and the hyperparameters λ, γ, α, and β in the loss function, which control the weights of distortion loss, gradient consistency loss, contrast loss, and smoothness loss, respectively.

The perturbation magnitude ε is a crucial parameter that determines the degree of perturbation applied to adversarial examples, thus influencing the difference between the generated adversarial steganographic image and the original image. In general, as ε increases, the difference between the adversarial steganographic image and the cover image becomes more pronounced. Figure 3 illustrates the detection accuracy of YeNet [35] on adversarial steganographic images and the SSIM (Structural Similarity Index) between cover images and adversarial steganographic images for various values of ε. We use S-UNIWARD [2] for embedding, with a payload of 0.4 bpp. As ε increases, the accuracy of the steganalysis network in determining whether an image contains hidden information gradually decreases, improving adversarial attack performance, while the image quality degrades. Ultimately, we select $\varepsilon = 0.015$, which provides an attack success rate close to 50% (*i.e.*, making the steganalysis model unable to distinguish between steganographic and cover images) while preserving image quality as much as possible.

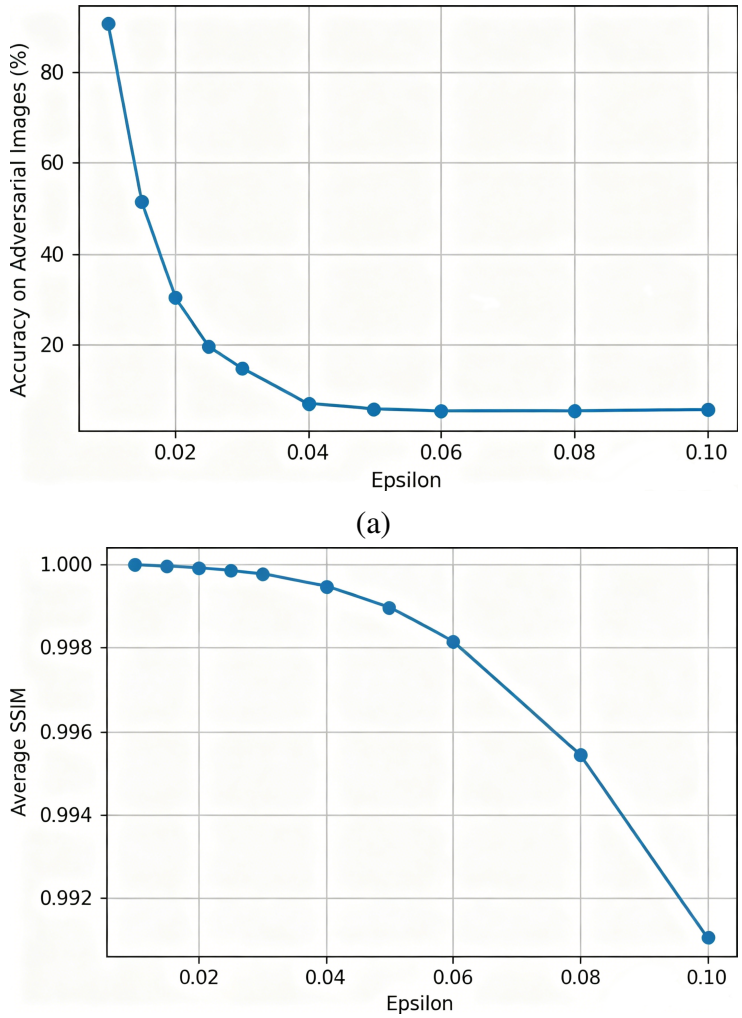


Figure 3. Comparison of accuracy and structural similarity index (SSIM) under different ε values: **(a)** accuracy under different ε values; **(b)** average SSIM under different ε values.

Figure 4 presents the SSIM values between cover images and adversarial steganographic images for different values of λ, α, β, and γ when embedding with S-UNIWARD. These parameters adjust the weights of different components in the loss function. Among them, λ controls the weight of distortion loss, reflecting the model's focus on image quality by quantifying the difference between adversarial steganographic images and original images. As indicated by the plotted curves, variations in these hyperparameters yield continuous and sensitive responses in both visual quality and attack success. This behavior demonstrates that the loss terms—including $\mathscr{L}_{cls}$, $\mathscr{L}_{dis}$, $\mathscr{L}_{grad}$, and the method-specific bounds—do not act in isolation but serve as highly coupled constraints on the multi-objective optimization space. The observed fluctuations confirm that each component actively contributes to the final trade-off, as shifting any single weight directly alters the equilibrium between fidelity and adversarial strength. After considering multiple factors, we select the following parameter configuration: $\lambda = 1$, $\alpha = 0.1$, and $\gamma = 0.1$. This configuration balances the quality of adversarial examples with attack performance effectively. Using the same methodology, we determine the optimal hyperparameters for PGD-adv, ultimately selecting $\varepsilon = 0.03$, $\alpha = 0.003$, $\lambda = 1$, $\gamma = 0.1$, and $\beta = 0.1$ as the default parameters for PGD-adv.

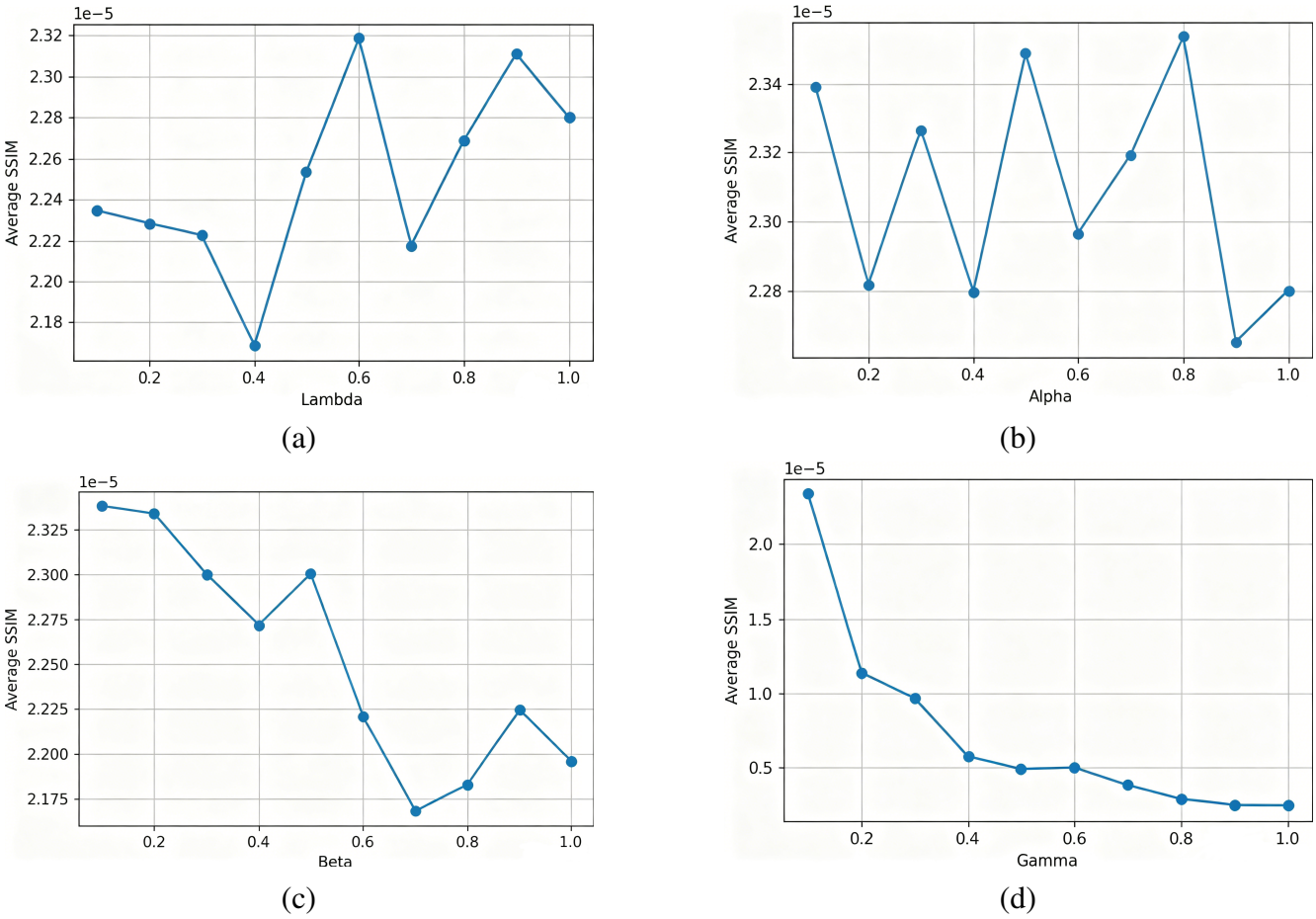


Figure 4. Average SSIM between cover images and adversarial steganographic images under different parameter values: **(a)** different λ values; **(b)** different α values; **(c)** different β values; **(d)** different γ values.

4.3. Performance evaluation of FGSM-adv

4.3.1. Performance on BOSSBase 1.01

This subsection evaluates whether the proposed FGSM-adv outperforms baseline methods across multiple steganalysis models. Three CNN-based steganalyzers are employed: YeNet [35] (used for adversarial example generation), ZhuNet [37], and XuNet [38]. Four steganographic schemes are tested: three heuristic methods S-UNIWARD [2], WOW [4], and HILL [3] and the deep learning-based UT-GAN [43] for generating initial stego images. Due to copyright restrictions, the training data for UT-GAN is not publicly available; thus, we use the author-provided 0.4 bpp model trained for 120,000 iterations (equivalent to 72 epochs).

Security is assessed using detection error rate (PE) and accuracy, while imperceptibility is evaluated using the structural similarity index (SSIM) and peak signal-to-noise ratio (PSNR). Tables 1–5 report the test errors of YeNet, XuNet, and ZhuNet at payloads of 0.1–0.5 bpp on the BOSSBase 1.01 dataset. Methods enhanced by FGSM-adv are denoted as WOW_adv, HILL_adv, S-UNIWARD_adv, and UT-GAN_adv. Under higher payloads, where baseline security is weaker, the proposed framework demonstrates more significant improvements.

Table 1 presents the average detection error rates under SRM analysis for S-UNIWARD, WOW, MIPOD,

HILL, UT-GAN, and the proposed FGSM-adv across five payload levels. The results confirm that FGSM-adv consistently enhances security, especially at higher payloads. For example, at 0.5 bpp, WOW_adv improves the detection error rate by 26.85% compared to the vanilla WOW. However, at 0.1 bpp, limited structural distortion results in modest gains. Notably, security improvement increases steadily with payload capacity.

Table 1. Error rates of steganalysis for SRM on BOSSBase 1.01 with different steganographic methods (%).

2* Steganographic algorithm	Embedded capacity				
	0.1bpp	**0.2bpp**	**0.3bpp**	**0.4bpp**	**0.5bpp**
S-UNIWARD [2]	48.1	39.65	28.15	19.7	14.9
S-UNIWARD_adv	53.3	51.8	47.35	42.15	41.3
WOW [4]	48.3	40.05	30.1	22.95	17.35
WOW_adv	54.7	52.65	44.35	41.7	44.2
HILL [3]	49.2	48.3	45.6	41.1	36.1
HILL_adv	53.6	52.5	51.5	47.55	43.7
UT-GAN [43]	-	-	-	37.65	-
UT-GAN_adv	-	-	-	44.5	-

Table 2 presents the experimental results of S-UNIWARD, WOW, HILL, UT-GAN, and their FGSM-adv enhanced variants under four different payloads, evaluated using YeNet. The average PE is reported for each steganographic algorithm. It can be observed that the FGSM-adv based schemes effectively enhance security, particularly under high payload conditions. For instance, at 0.4 bpp, S-UNIWARD_adv achieves a 27.85% improvement over the original S-UNIWARD. Notably, at a low payload of 0.1 bpp, the performance gain is less pronounced. This is expected, as low payloads introduce minimal modifications to the image, resulting in low distinguishability between cover and stego images. However, as the payload increases, the improvement becomes more substantial.

Table 2. Error rates of steganalysis for YeNet on BOSSBase 1.01 with different steganographic methods (%).

2* Steganographic algorithm	Embedded capacity				
	0.1bpp	**0.2bpp**	**0.3bpp**	**0.4bpp**	**0.5bpp**
S-UNIWARD [2]	48.6	43.3	33.4	26.2	21
S-UNIWARD_adv	54.65	50.4	50.35	54.05	50.5
WOW [4]	49.6	46.55	40.3	33.35	28.15
WOW_adv	57.05	56.15	54.75	53.9	51.75
Hill [3]	49.6	47.7	44.05	42.05	37
Hill_adv	56.95	56.7	55.65	55.45	54.35
UT-GAN [43]	-	-	-	41.45	-
UT-GAN_adv	-	-	-	55.1	-

Table 3 demonstrates that the adversarial-enhanced methods achieve significant improvements when detected by XuNet across various payloads. For example, at 0.5 bpp, the FGSM-adv enhanced versions outperform S-UNIWARD, WOW, and HILL by 34.95%, 27.25%, and 16.55%, respectively. Table 4 further confirms that the proposed method consistently enhances security against the CNN-based steganalyzer ZhuNet at multiple embedding rates. Specifically, at 0.4 bpp, S-UNIWARD_adv exhibits a 4% higher error rate than S-UNIWARD, while HILL_adv surpasses HILL by 0.4%. These results underscore the superiority of the proposed approach in countering CNN-based steganalysis.

Table 3. Error rates of steganalysis for XuNet on BOSSBase 1.01 with different steganographic methods (%).

2* Steganographic algorithm	Embedded capacity				
	0.1bpp	**0.2bpp**	**0.3bpp**	**0.4bpp**	**0.5bpp**
S-UNIWARD [2]	48.6	40.75	30.8	23.7	20.5
S-UNIWARD_adv	54.65	50.4	50.35	52.05	55.45
WOW [4]	48.85	43.35	35.0	27.75	22.5
WOW_adv	54.85	51.65	48.7	48.7	49.75
Hill [3]	49.6	47.7	44.05	38.0	31.65
Hill_adv	55.25	54.0	51.95	49.7	48.2
UT-GAN [43]	-	-	-	34.9	-
UT-GAN_adv	-	-	-	48.0	-

Table 4. Error rates of steganalysis for ZhuNet on BOSSBase 1.01 with different steganographic methods (%).

2* Steganographic algorithm	Embedded capacity				
	0.1bpp	**0.2bpp**	**0.3bpp**	**0.4bpp**	**0.5bpp**
S-UNIWARD [2]	49.1	45.9	36.85	27.75	22.0
S-UNIWARD_adv	49.05	44.9	36.9	31.75	30.3
WOW [4]	49.2	43.35	41.5	33.25	26.4
WOW_adv	49.25	46.9	41.0	35.25	31.25
Hill [3]	49.45	48.9	46.55	42.75	37.7
Hill_adv	49.6	49.05	46.2	42.35	37.8
UT-GAN [43]	-	-	-	39.9	-
UT-GAN_adv	-	-	-	39.9	-

4.3.2. Performance on BOWS2

To evaluate the generalization capability of the proposed method, additional experiments were conducted on the BOWS2 dataset. We compared S-UNIWARD [2], WOW [4], HILL [3], and UT-GAN [43] against their FGSM-adv improved versions using the PE as the security metric. Table 5 reports the PE values of these methods against YeNet at payloads ranging from 0.1 to 0.5 bpp.

Table 5. Error rates of steganalysis for ZhuNet on BOWS with different steganographic methods (%).

2* Steganographic algorithm	Embedded capacity				
	0.1bpp	**0.2bpp**	**0.3bpp**	**0.4bpp**	**0.5bpp**
S-UNIWARD [2]	49.55	47.35	44.05	41.55	38.2
S-UNIWARD_adv	52.4	50.45	46.1	44.2	41.1
WOW [4]	49.85	49.05	47.8	46.45	44.4
WOW_adv	52.6	51.55	49.8	49.0	47.0
HILL [3]	49.8	49.75	49.3	48.6	47.45
HILL_adv	52.7	52.65	51.05	51.6	50.9
UT-GAN [43]	-	-	-	48.4	-
UT-GAN_adv	-	-	-	50.9	-

From Table 5, it is evident that the adversarial-enhanced schemes S-UNIWARD_adv, WOW_adv, HILL_adv, and UT-GAN_adv—consistently achieve higher security across all payloads compared to their original counterparts. For instance, at 0.4 bpp, WOW_adv and HILL_adv improve security over WOW and HILL by 2.11% and 1.96%, respectively. These results demonstrate that the proposed FGSM-adv method significantly enhances steganographic security, even when applied to a different image set, confirming its strong generalization capability.

4.4. PGD-adv generalization performance

To further validate the generalization capability of our proposed framework, we evaluated the PGD-adv based method and compared it with the original steganographic schemes. PGD-adv enhances steganographic security by generating stronger adversarial examples through multi-step iterative perturbation optimization. We refer to the S-UNIWARD, WOW, HILL, and UT-GAN models integrated with PGD-adv as S-UNIWARD_pgd, WOW_pgd, HILL_pgd, and UT-GAN_pgd, respectively. Using these as examples, we demonstrate their performance on the BOSSBASE 1.01 dataset. Table 6 reports the PE at payloads of 0.1, 0.2, 0.3, 0.4, and 0.5 bpp for both the original steganographic methods and their PGD-adv enhanced versions. As shown in the table 6, all PGD-adv enhanced methods significantly outperform their original counterparts across various payloads. For instance, at 0.4 bpp, S-UNIWARD_pgd achieves a PE of 43.10%, which is 23.40% higher than the original S-UNIWARD (19.70%). Similarly, WOW_pgd attains a PE of 42.60%, an improvement of 19.65% over the original WOW (22.95%). Moreover, UT-GAN_pgd reaches a PE of 45.80% at 0.4 bpp, exceeding the original UT-GAN (37.65%) by 8.15%. These results indicate that PGD-adv offers notable advantages in improving steganographic security.

Table 6. Error rates of steganalysis for SRM on BOSSBase 1.01 with different steganographic methods (%).

2* Steganographic algorithm	Embedded capacity				
	0.1bpp	**0.2bpp**	**0.3bpp**	**0.4bpp**	**0.5bpp**
S-UNIWARD [2]	48.1	39.65	28.15	19.7	14.9
S-UNIWARD_pgd	54.5	52.3	48.2	43.1	42.4
WOW [4]	48.3	40.05	30.1	22.95	17.35
WOW_pgd	55.2	53.1	45.8	42.6	45.5
HILL [3]	49.2	48.3	45.6	41.1	36.1
HILL_pgd	54.8	53.2	52.1	48.3	44.9
UT-GAN [43]	-	-	-	37.65	-
UT-GAN_pgd	-	-	-	45.8	-

To further demonstrate the superiority of the proposed PGD-adv method, we compare it with several recent state-of-the-art steganographic approaches, including Generated Image Fluctuation Distortion Learning for Enhancing Steganographic Security (GIFDL) [44], GMAN [14], and GACL [45]. Table 7 reports the detection error rates of these methods against the SRM steganalyzer across payloads from 0.1 to 0.4 bpp. As shown in Table 7, the advantages of WOW_pgd, S-UNIWARD_pgd and HILL_pgd are particularly evident at higher embedding rates, where the detection error rate of the proposed method is much higher than that of GIFDL [44], GMAN [14], and GAN-Based Adaptive Cost Learning for Enhanced Image Steganography Security (GACL) [45].

These results confirm that the adversarial perturbations introduced by PGD-adv not only enhance security beyond that of the original steganographic algorithms but also surpass recent dedicated secure steganography frameworks, highlighting the effectiveness and general applicability of the BS-ADV approach.

Table 7. Error rates of steganalysis for SRM with different state-of-the-art steganographic methods (%).

2* Steganographic algorithm	Embedded capacity			
	0.1bpp	0.2bpp	0.3bpp	0.4bpp
GIFDL [44]	**46.7**	44.1	38.5	34.0
GMAN [14]	45.4	41.6	34.2	30.8
GACL [45]	44.16	38.62	35.57	33.68
S-UNIWARD_pgd	45.5	**47.7**	**48.2**	43.1
WOW_pgd	44.8	46.9	45.8	42.6
HILL_pgd	45.2	46.8	47.9	**48.3**

4.5. Image quality and imperceptibility

Since adversarial stego images are generated via adversarial attacks, it is essential to evaluate the imperceptibility of the proposed FGSM-adv method. We select three images from BOSSBase 1.01 and produce adversarial stego images using S-UNIWARD at a payload of 0.4 bpp within the FGSM-adv framework. The resulting images are shown in Figure 5. No perceptible visual differences are observed between the cover images and the adversarial stego images.

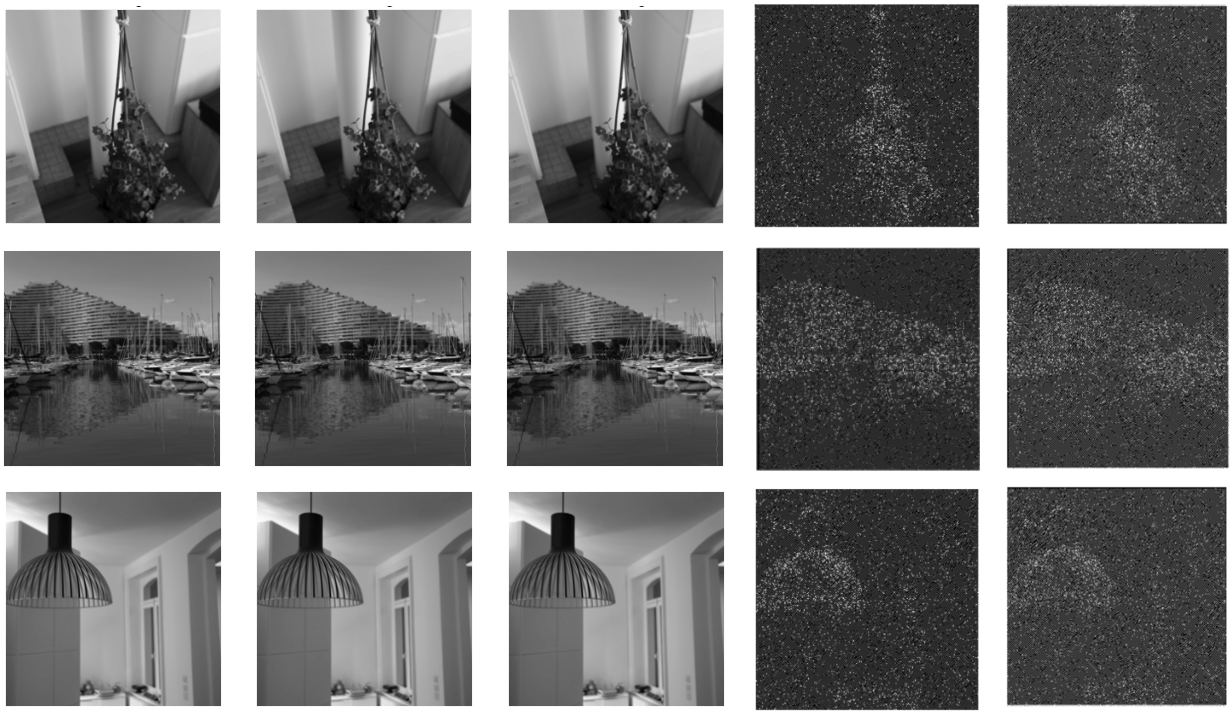


Figure 5. Generation process of adversarial steganography and residual diagram based on S-UNIWARD: **(a)** cover image; **(b)** stego image; **(c)** adversarial image; **(d)** cover-stego; **(e)** cover- adversarial.

To quantitatively assess imperceptibility, we evaluate FGSM-adv using two widely used metrics, PSNR and SSIM. The average PSNR and SSIM are computed over 1000 pairs of cover images and their corresponding adversarial stego images from the BOSSBase 1.01 dataset, and the results are reported in Table 8. As shown in Table 8, the average PSNR and SSIM of the adversarial stego images exceed 36 dB and 0.883, respectively, indicating that the proposed adversarial steganography strategy maintains high imperceptibility.

Table 8. Quantitative metrics between adversarial images and cover images under different bpp.

2* Evaluation indicators	Embedded capacity				
	0.1bpp	**0.2bpp**	**0.3bpp**	**0.4bpp**	**0.5bpp**
PSNR(db)	36.152	36.144	36.151	36.142	36.139
SSIM	0.912	0.898	0.884	0.883	0.885

4.6. Extraction performance analysis: experimental results on deepSteganography and HiNet

We evaluate the steganographic performance of DeepSteganography and HiNet using 1000 images from the BOSSBase 1.01 dataset, evenly split into 500 cover images and 500 secret images. In each case, a secret image is embedded into a cover image, and adversarial stego images are then generated using FGSM-adv to assess robustness against steganalysis models. The PE and image-quality metrics are summarized in Table 8 and Figure 6.

(a) (b) (c) (d) (e) (f)

Figure 6. Generation process of adversarial steganography and residual diagram based on DeepSteganography: **(a)** secret image; **(b)** cover image; **(c)** adversarial image; **(d)** recovered image; **(e)** cover-adversarial; **(f)** secret-recovered.

Considering the PE reported in Table 9, both DeepSteganography [15] and HiNet [42] exhibit substantially improved robustness to steganalysis models (YeNet, ZhuNet, and XuNet) after applying FGSM-adv. For example, under YeNet, the PE of DeepSteganography increases from 35.30% to 49.90%; under ZhuNet, the PE of HiNet rises from 41.80% to 52.10%. These results indicate that the adversarial

perturbations introduced by FGSM-adv substantially increase the detection difficulty for steganalysis models and significantly enhance the security of stego images.

Table 9. Comparison of steganography PE of different steganography models on the BOSSBase 1.01 dataset (%).

2* Steganographic algorithm	Steganalysis algorithm		
	YeNet	ZhuNet	XuNet
DeepSteganography [15]	35.3	37.6	31.0
DeepSteganography_adv	49.9	48.75	40.2
HiNet [42]	40.7	41.8	34.5
HiNet_adv	49.8	52.1	45.7

As shown in Table 10, the PSNR between the secret images and their extracted counterparts remains consistently high with minimal variation across payloads. This indicates only minor pixel-wise deviations from the originals and negligible loss of perceptual image quality. Likewise, SSIM also remains high, confirming that the structural content of the original secret images is well preserved after extraction. Collectively, these results show that, while FGSM-adv improves steganographic security via adversarial perturbations, it has negligible impact on extraction fidelity and enables accurate recovery of the secret information.

Table 10. Comparison of PSNR and SSIM for different model secret images with extracted secret images.

Evaluation indicators	DeepSteganography_adv	HiNet_adv
PSNR (dB)	39.583	38.753
SSIM	0.934	0.953

As illustrated in Figure 6, taking the first image set as an example, DeepSteganography embeds the secret image into the cover image to produce a stego image. The cover and stego images are perceptually indistinguishable, demonstrating the concealment capability of DeepSteganography. Likewise, the adversarial stego image generated by FGSM-adv is visually indistinguishable from both the cover and stego images. The residual map between the cover and stego images highlights the subtle modifications introduced by embedding, whereas the residual map between the cover and the adversarial stego image reflects the effect of FGSM-adv perturbations. Despite these changes, as discussed above, secret recovery remains largely unaffected. The second and third image sets exhibit similar behavior, further validating that FGSM-adv enhances steganographic security without compromising extraction quality.

In summary, FGSM-adv markedly increases the detection error rates of steganalysis models through adversarial perturbations while maintaining the fidelity of secret information extraction. This provides a novel security-enhancement method for image steganography and further demonstrates the potential of adversarial-example techniques in this domain.

4.7. Experimental results on coverless steganography CRoSS

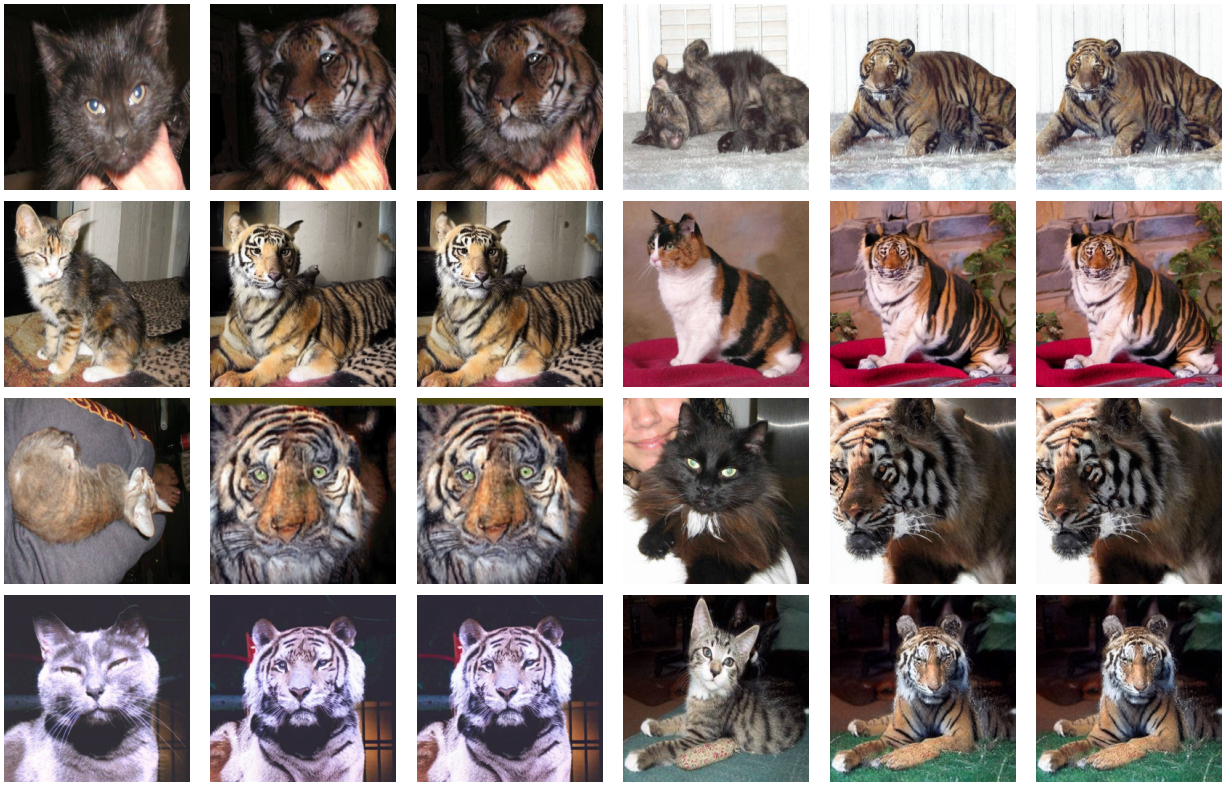
The proposed framework is also applicable to coverless steganography. We conduct experiments on CRoSS [8], a diffusion-based coverless steganography method. Leveraging Denoising Diffusion Implicit

Models (DDIM) deterministic sampling and the intrinsic image-to-image translation capability of diffusion models without additional training Yu *et al*. proposed the coverless framework CRoSS. During hiding, CRoSS transforms a secret image into a container image guided by two text prompts (private key and public key), thereby concealing the secret without a cover image. During revealing, CRoSS reconstructs the secret from the container by reversing the two diffusion processes via DDIM inversion.

We use 1000 cat images from the Dogs *vs.* Cats dataset and conceal them as a similar animal (tiger) with CRoSS. The private key is set to "cat" and the public key to "tiger". After filtering partially failed generations, we apply FGSM-adv to the stego images to produce adversarial stego images, denoted as CRoSS_adv. These are then analyzed using steganalysis models including YeNet [35], ZhuNet [37], and XuNet [38]. The PE is used for quantitative evaluation, with results reported in Table 11. Representative secret images, their stego images, and the corresponding adversarial stego images are shown in Figure 7. The adversarial stego images are perceptually indistinguishable from the stego images, indicating that the proposed adversarial method extends effectively to coverless steganography scenarios without degrading visual quality.

Table 11. PE of different steganalysis algorithms for CRoSS.

2* Steganography	steganalyzer		
	YeNet	ZhuNet	XuNet
CRoSS [8]	55.864	70.268	70.145
CRoSS_adv	49.486	50.103	57.541



(a) (b) (c) (d) (e) (f)

Figure 7. Generation process of adversarial steganography based on CRoSS: **(a, d)** secret image; **(b, e)** stego image; **(c, f)** adversarial image.

5. Conclusion

We present BS-ADV, a secure steganography framework that employs transfer-based black-box adversarial attacks to craft adversarial stego images without access to a steganographic model's internal parameters or architecture. BS-ADV substantially improves robustness to steganalysis models while preserving the visual fidelity of cover images. We instantiate the framework with a Fast Gradient Sign Method variant (FGSM-adv) that couples transfer-based gradient perturbations with a minimal-distortion objective, thereby improving both security and imperceptibility of the resulting adversarial stego images. Extensive experiments demonstrate that the proposed method significantly outperforms existing baselines and comparative approaches against both feature-based and CNN-based steganalysis models. Moreover, the framework extends to coverless image steganography: experiments within the CRoSS diffusion framework show that BS-ADV effectively enhances the security of diffusion-based coverless schemes without degrading visual quality. The current instantiation is primarily tailored to spatial-domain images. Future work will extend the framework to other generative steganography modalities—such as JPEG-domain image steganography and audio steganography—and further investigate cross-domain generalization to strengthen security across heterogeneous data types.

Data availability statement

No supplementary or additional data were generated in this study.

Declaration of generative AI and AI-assisted technologies

During the preparation of this manuscript, the authors used generative AI tools only to improve language and readability. Specifically, the authors used DeepSeek and ChatGPT for language polishing and readability enhancement in limited sections of the manuscript. The authors take full responsibility for the content of the manuscript.

Acknowledgments

This work is supported in part by the National Natural Science Foundation of China No. 62462053, in part by the Key Research and Development Program of Gansu Province No. 24YFGA004.

Authors' contribution

Shichen Yang: Conceptualization, methodology, investigation, writing—original draft preparation, visualization; Haiyu Xu: validation, data curation, writing—review and editing; Xingxing Jia: software, project administration, funding acquisition; Guodong Ye: supervision; Chengsheng Yuan: formal analysis; Huiyu Zhou: resources. All authors have read and agreed to the published version of the manuscript.

Conflicts of interest

The authors declare no conflicts of interest.

References

[1] Pevný T, Filler T, Bas P. Using high-dimensional image models to perform highly undetectable steganography. directional filters. In *Proceedings of the 12th International Conference on Information Hiding*, Alberta, Canada, June 28–30, 2010, pp. 161–177.

[2] Holub V, Fridrich J, Denemark T. Universal distortion function for steganography in an arbitrary domain. *EURASIP J. Inf. Secur.* 2014, 1(1):1.

[3] Li B, Wang M, Huang J, Li X. A new cost function for spatial image steganography. In *Proceedings of the 2014 IEEE International Conference on Image Processing (ICIP)*, Paris, France, October 27–30, 2014, pp. 4206–4210.

[4] Holub V, Fridrich J. Designing steganographic distortion using directional filters. In *Proceedings of the 2012 IEEE International Workshop on Information Forensics and Security (WIFS)*, Tenerife, Spain, December 2–5, 2012, pp. 234–239.

[5] Hayes J, Danezis G. Generating steganographic images via adversarial training. In *Proceedings of the Advances in Neural Information Processing Systems 30 (NIPS 2017)*, California, USA, December 4–9, 2017, pp. 1954–1963.

[6] Zhang K, Cuesta-Infante A, Xu L, Veeramachaneni K. SteganoGAN: high capacity image steganography with GANs. *arXiv* 2019, arXiv:1901.03892.

[7] Zhu J, Park T, Isola P, Efros AA. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the 2017 IEEE International Conference on Computer Vision (ICCV 2017)*, Venice, Italy, October 22–29, 2017, pp. 2223–2232.

[8] Yu J, Zhang X, Xu Y, Zhang J. CRoSS: diffusion model makes controllable, robust and secure image steganography. In *Proceedings of the Advances in Neural Information Processing Systems 36 (NeurIPS 2023)*, Louisiana, USA, December 10–16, 2023, pp. 17–23.

[9] Filler T, Judas J, Fridrich J. Minimizing embedding impact in steganography using trellis-coded quantization, In *Proceedings of the Media Forensics and Security II*, California, USA, January 18–19, 2010, pp. 38–51.

[10] Zhu J, Kaplan R, Johnson J, Li F. Hidden: hiding data with deep networks. In *Proceedings of the European Conference on Computer Vision (ECCV 2018)*, Munich, Germany, September 8–14, 2018, pp. 657–672.

[11] Tang W, Li B, Barni M, Li J, Huang J. An automatic cost learning framework for image steganography using deep reinforcement learning. *IEEE Trans. Inf. Forensics Secur.* 2021, 16:952–967.

[12] Yao Q, Zhang W, Chen K, Yu N. LDGM codes-based near-optimal coding for adaptive steganography. *IEEE Trans. Commun.* 2024, 72(4):2138–2151.

[13] Chen Y, Wang H, Li W, Li W. A steganography immunoprocessing framework against CNN-based and handcrafted steganalysis. *IEEE Trans. Inf. Forensics Secur.* 2024, 19:6055–6069.

[14] Huang D, Luo W, Liu M, Tang W, Huang J. Steganography embedding cost learning with generative multi-adversarial network. *IEEE Trans. Inf. Forensics Secur.* 2024, 19:15–29.

[15] Baluja S. Hiding images in plain sight: deep steganography, In *Proceedings of the Advances in Neural Information Processing Systems 30 (NIPS 2017)*, Long Beach, USA, December 4–9, 2017, pp. 2066–2076.

[16] Zhang R, Dong S, Liu J. Invisible steganography via generative adversarial networks. *Multimedia Tools Appl.* 2019, 78:8559–8575.

[17] Rombach R, Blattmann A, Lorenz D, Esser P, Ommer B. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Louisiana, USA, June 18–24, 2022, pp. 10684–10695.

[18] Bengio Y, Courville A, Vincent P. Representation learning: a review and new perspectives. *IEEE Trans. Pattern Anal. Mach. Intell.* 2013, 35(8):1798–1828.

[19] Rombach R, Blattmann A, Lorenz D, Esser P, Ommer B. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Louisiana, USA, June 18–24, 2022, pp. 10684–10695.

[20] Midjourney, Inc. Midjourney AI Image Generator. 2023. Available: https://www.midjourney.com (accessed on 20 December 2025).

[21] Wilson A. Midjourney statistics: users, polls, and growth. 2023. Available: https://approachableai.com/midjourney-statistics/ (accessed on 20 December 2025).

[22] Wu Z, Wen J, Xue Y, Zhang Z, Zhou Y. GTSD: generative text steganography based on diffusion model. *arXiv* 2025, arXiv:2504.19433.

[23] Ma Y, Xu L, Zhang Y, Zhang T, Luo X. Steganalysis feature selection with multidimensional evaluation and dynamic threshold allocation. *IEEE Trans. Circuits Syst. Video Technol.* 2024, 34(3):1954–1969.

[24] Ma Y, Xu L, Zhang Q, Zhang Y, Xin X, *et al*. EIS-OBEA: enhanced image steganalysis via opposition-based evolutionary algorithm. *IEEE Trans. Inf. Forensics Secur.* 2025, 20:3616–3631.

[25] Fu T, Chen L, Jiang Y, Jia J, Fu Z. Image steganalysis based on dual-path enhancement and fractal downsampling. *IEEE Trans. Inf. Forensics Secur.* 2025, 20:1–16.

[26] Wei K, Luo W, Huang J. Color image steganalysis based on pixel difference convolution and enhanced transformer with selective pooling. *IEEE Trans. Inf. Forensics Secur.* 2024, 19:9970–9983.

[27] Cao Y, Hu Y, Liu M, Wei K, Zhao R, *et al*. A self-distillation framework with feature pyramids and auxiliary classifiers for image steganalysis. *Signal, Image Video Process.* 2025, 19:1453.

[28] Szegedy C, Zaremba W, Sutskever I, Bruna J, Erhan D, *et al*. Intriguing properties of neural networks. *arXiv* 2013, arXiv:1312.6199.

[29] Zhang C, Hu M, Li W, Wang L. Adversarial attacks and defenses on text-to-image diffusion models: a survey. *Inf. Fusion* 2025, 114:102701.

[30] Yang Y, Hui B, Yuan H, Gong N, Cao Y. SneakyPrompt: jailbreaking text-to-image generative models. In *Proceedings of the 2024 IEEE Symposium on Security and Privacy (SP)*, San Francisco, USA, May 19–23, 2024, pp. 897–912.

[31] Zhang Y, Jia J, Chen X, Chen A, Zhang Y, *et al*. To generate or not? Safety-driven unlearned diffusion models are still easy to generate unsafe images... for now. *arXiv* 2023, arXiv:2310.11868.

[32] Brendel W, Rauber J, Bethge M. Decision-based adversarial attacks: reliable attacks against black-box machine learning models. *arXiv* 2017, arXiv:1712.04248.

[33] Papernot N, McDaniel P, Goodfellow I, Jha S, Celik ZB, *et al*. Practical black-box attacks against machine learning. In *Proceedings of the 2017 ACM Asia Conference on Computer and Communications Security*, Abu Dhabi, UAE, 2017, pp. 506–519.

[34] Chen Y, Liu W. A theory of transfer-based black-box attacks: explanation and implications. In *Proceedings of the Advances in Neural Information Processing Systems 36 (NeurIPS 2023)*, Louisiana, USA, December 10–16, 2023, pp. 13887–13907.

[35] Ni J, Ye J, Yi Y. Deep learning hierarchical representations for image steganalysis. *IEEE Trans. Inf. Forensics Secur.* 2017, 12(11):2545–2557.

[36] Fridrich J, Kodovský J. Rich models for steganalysis of digital images. *IEEE Trans. Inf. Forensics Secur.* 2012, 7(3):868–882.

[37] Zhang R, Zhu F, Liu J, Liu G. Depth-wise separable convolutions and multi-level pooling for efficient spatial CNN-based steganalysis. *IEEE Trans. Inf. Forensics Secur.* 2020, 15:1138–1150.

[38] Xu G, Wu H, Shi Y. Structural design of convolutional neural networks for steganalysis. *IEEE Signal Process Lett.* 2016, 23(5):708–712.

[39] Weike Y, Zhang H, Zhao X. A siamese CNN for image steganalysis. *IEEE Trans. Inf. Forensics Secur.* 2021, 16:291–306.

[40] Goodfellow IJ, Shlens J, Szegedy C. Explaining and harnessing adversarial examples. *arXiv* 2015, arXiv:1412.6572.

[41] Madry A, Makelov A, Schmidt L, Tsipras D, Vladu A. Towards deep learning models resistant to adversarial attacks. *arXiv* 2018, arXiv:1706.06083.

[42] Jing J, Deng X, Xu M, Wang J, Guan Z. HiNet: deep image hiding by invertible network. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, Montreal, Canada, October 11–17, 2021, pp. 4733–4742.

[43] Yang J, Ruan D, Huang J, Kang X, Shi Q. An embedding cost learning framework using GAN. *IEEE Trans. Inf. Forensics Secur.* 2020, 15:839–851.

[44] Wang X, Chen K, Qi Y, Liu R, Zhang W, *et al*. GIFDL: generated image fluctuation distortion learning for enhancing steganographic security. *IEEE Trans. Inf. Forensics Secur.* 2025, 20:4581–4594.

[45] Wang D, Yang G, Chen J, Ding X. GAN-based adaptive cost learning for enhanced image steganography security. *Expert Syst. Appl.* 2024, 249:123471.