Removing Hidden Information by Geometrical Perturbation in Frequency Domain

Lin He, Bingwen Feng, Zecheng Peng, Bing Chen, *Member, IEEE*, Zhihua Xia, *Member, IEEE*, and Wei Lu, *Member, IEEE*

***Abstract*— The risk of malicious exploitation of advanced image steganography necessitates the removal of hidden information from images. However, it is crucial to preserve the visual quality of the images undergoing processed. This paper suggests a geometrical attack in frequency domain (GAF) to address this challenge. GAF employs a thin plate spline (TPS) to slightly geometrically perturb the frequency components of the stego image. It incorporates a channel weight estimator and a frequency jammer. The channel weight estimator assigns perturbation strengths to each DCT channel, while the frequency jammer performs the TPS transform on the DCT channels using the assigned perturbation strengths. Experimental results demonstrate that the proposed approach effectively hinders secret image recovery with a little distortion to the stego images. Furthermore, it well preserves the visual quality of clear images that do not contain secret information.**

***Index Terms*— Steganography attack, geometrical perturbation, spatial transformer network (STN), thin plate spline (TPS).**

I. Introduction

IMAGE steganography is an art of covert communication. By using image steganographic tools, Alice can effectively embed secret messages into cover images, as demonstrated in Fig. 1. The resulting stego images are almost indistinguishable from the cover images, and thus can be transmitted via public channels without the risk of being detected. Numerous steganographic methods have been suggested to facilitate covert communication with significant capacity and high security [1], [2], [3], [4], [5], [6]. Nevertheless, this technology could be easily abused for illicit purposes. Terrorists may use image steganography to exchange dangerous messages. As a result, the removal of potential hiding data has gained significant attention.

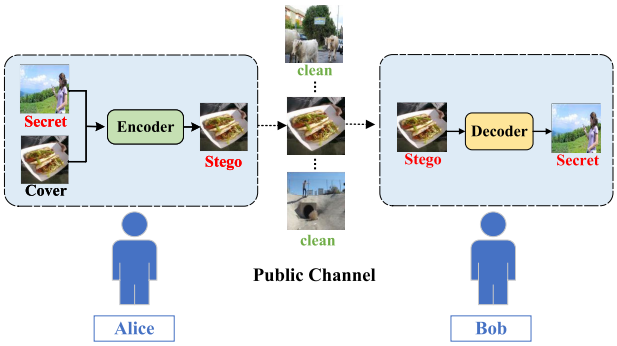


Fig. 1. Covert communication via image steganography.

Traditional image steganographic schemes usually modify the least significant bits (LSBs) to carry message bits for the sake of undetectability [1], [2], [3]. This embedding strategy is known to be fragile to distortions. For example, using JPEG compression can easily erase their embedded message bits [7], [8]. With the development of deep learning, many learning-based steganographic approaches have been suggested. Some of them can even hide color secret images rather than binary secret messages [4], [5], [6]. DS [4] attempted to hide a secret image within a cover image of the same size. ABDH [5] introduced an attention mask to improve stego image quality. DAH [6] proposed a deep adaptive hiding network that progressively extracts and fuses necessary secret and cover information over a range of frequencies and depths (layers). Different from traditional approaches, these schemes do not ask for perfect decoding accuracy. Further, they exploit almost all the bits in a cover image to hide secret information. As a result, while not the original intent, they are able to withstand some distortions [9].

Removing message bits hidden in images are usually carried out by common signal processing such as noising, image filtering, quantization, etc., [10]. Recently, learning-based hiding information removal methods has risen. They try to construct deep network to destruct secret information without degrading

Manuscript received 8 January 2024; revised 8 April 2024 and 17 May 2024; accepted 5 June 2024. Date of publication 10 June 2024; date of current version 27 November 2024. This work was supported in part by the National Key Research and Development Program of China under Grant 2022YFB3103100; in part by the National Natural Science Foundation of China under Grant 61802145, Grant 62261160653, Grant 62102101, Grant 62122032, and Grant U23B2023; in part by the Natural Science Foundation of Guangdong Province, China, under Grant 2023A1515011348; in part by the Fundamental Research Funds for the Central Universities; and in part by the Doctoral Scientific Research Foundation of Guangdong Polytechnic Normal University under Grant 2021SDKYA101. This article was recommended by Associate Editor A. Liu. *(Corresponding author: Bingwen Feng.)*

Lin He, Bingwen Feng, Zecheng Peng, and Zhihua Xia are with the College of Cyber Security, Jinan University, Guangzhou 510632, China (e-mail: hl11685044068@stu2022.jnu.edu.cn; bingwfeng@gmail.com; pzc987048624@163.com; xia_zhihua@163.com).

Bing Chen is with the School of Cyber Security, Guangdong Polytechnic Normal University, Guangzhou 510665, China (e-mail: chenbing@gpnu.edu.cn).

Wei Lu is with the School of Computer Science and Engineering, Guangdong Province Key Laboratory of Information Security Technology, Ministry of Education Key Laboratory of Machine Intelligence and Advanced Computing, Sun Yat-sen University, Guangzhou 510006, China (e-mail: luwei3@mail.sysu.edu.cn).

Color versions of one or more figures in this article are available at https://doi.org/10.1109/TCSVT.2024.3411874.

Digital Object Identifier 10.1109/TCSVT.2024.3411874

stego image quality dramatically. Jung et al. [9] suggested PS that effectively removes secret images by restoring the distribution of original cover images. Li et al. [11] presented a robust watermark steganography attack method, CAGP, based on generative adversarial networks. Xiang et al. [12] proposed a new ProvablE-REmovaL attack (PEEL), which uses image restoration to remove secret images and improves the attacked image quality by exploiting the visual information of containers. These methods outdone common signal processing with more efficient secret message removal and better attacked image quality. However, they begin by evenly treating all image information, while disregarding the fact that secure information is frequently concentrated in the undetectable parts, such as the high frequency component of images. Consequently, these methods may inadvertently damage image components that contain little secret information, causing additional deterioration in the quality of attacked images. Furthermore, as shown in Fig. 1, Alice's harmful stego images are often mixed with a large number of clean images on the public channel. Therefore, it is essential for the removal method to have minimal impact on clean images as all the publicly transmitted images have to be processed to prevent potential covert communication.

On the other hand, adversarial samples are frequently employed to mislead the model to make erroneous decisions by imperceptibly perturbing the original samples [13], [14], [15], [16], [17]. This inspires us to confuse the extraction network of steganographic schemes by performing a black-box attack using the stego image to be attacked. It has been shown that simply spatial transforming the input is sufficient to fool neural networks [14], [15], [16]. ADVFilter [17] models the optical path perturbation by thin plate spline (TPS) to generate adversarial example. It can preserve the pixel distribution while achieving a high attack success compared with those based on adding special noises. As a result, we propose removing the hidden data by using TPS to slightly perturb the stego images. It is expected that a better attacked image quality can be achieved. Moreover, geometrical distortions are known more difficult to be resisted than common signal processing [14], [18]. Therefore, the proposed removal method would be more difficult to defend against.

In this paper, we propose to remove hidden information by geometrically perturbing stego images in the frequency domain. The attack network comprises two modules: channel weight prediction and frequency coefficient perturbation. The former module selects the frequency bands that yield the most effective removal result, while the latter module perturbs the selected frequency bands with the learned TPS. The proposed scheme requires no prior knowledge and ensures the image quality of attacked images. The main contributions of the paper are as follows:

- The geometrical perturbation adaptive to frequency components is employed to interfere with the secret information extraction. This method can maintain the quality of both attacked stego images and attacked clear images.
- The proposed scheme only requires accessing the extraction network of the steganographic scheme to be attacked. Prior knowledge of cover images is not needed, widening its applications in real environments.
- The experimental results on both image hiding schemes and robust data hiding schemes demonstrate that the proposed scheme can effectively remove hidden information without significant degradation in image quality.

The rest of the paper is organized as follows: Section II describes the related work on steganography attacks and TPS. Section III analyzes the property of image hiding networks and the removing effect of geometrical perturbation, and presents the proposed attack network. Section IV reports the experimental results, and Section V gives the conclusion and future work.

II. Related Work

A. Steganography Attack

Steganography attacks can be categorized into passive and active attacks. Passive steganography attacks, also known as steganalysis, try to detect the existence of hidden information [19], [20], [21], [22]. Traditional steganalytic methods manually design elaborate steganalytic features [19], [20]. With the rapid development of deep learning, some methods consider replacing the major parts of traditional approaches with convolutional networks (CNNs) [21], [22]. Recently, an increasing number of methods directly construct steganalytic networks. You et al. [23] use a siamese CNN to capture the non-uniform changes on different image sub-regions caused by adaptive steganography. Zhang et al. [24] enhances the feature representation of the steganalytic network by equipping it with spatial and channel multiple attention modules. Li et al. [25] suggest a SKAttention structure to extract fine-grained features, and construct a multi-scale feature extraction backbone to increase the diversity of features. These approaches excel at steganography detection. However, the process of steganography detection could be time consuming and usually faces the distribution mismatching in training data [26]. Even a single missed stego image would lead to disastrous results.

In contrast, active steganography attacks directly destroy the data hidden in stego images. This can be performed by adding a noise that can disturb the hidden data meanwhile preserving the image quality. Li et al. [11] use a generative adversarial network to learn this noise. On the other hand, some schemes try to restore the distribution of stego images to that of the original cover image. Jung et al. [9] learn the distribution of a dataset having a similar distribution as the original cover images, and then use this distribution to purify the secret image. Xiang et al. [12] explore an inpainting method to restore each block of a stego image. These schemes can effectively remove secret information, thereby preventing potential covert communication.

However, the quality of attacked stego images obtained by many schemes may be unsatisfactory, resulting in a low quality of the public channel. This may be because the distribution of attacked stego images is still different from that of stego images and that of cover images. Additionally, secret information is typically dispersed unevenly across various frequencies,

whereas many schemes, which are primarily focused on the spatial domain, lack the ability to effectively accommodate this variability. For example, in CAGP, a generative adversarial network is employed to introduce noise to the entire stego image. PS aims to purify the stego image by aligning the distribution of the stego image with that of the cover. This process involves iteratively substituting the values of all the pixels. Both of them have the potential to impact the areas that contribute little to information hiding.

Due to these reasons, we launch a desynchronization attack only on the select frequencies of stego images. This strategy can preserve the distribution of attacked images while simultaneously exhibiting compatibility with the varying frequencies employed for secret information transmission. Additionally, the proposed scheme does not rely on prior knowledge of cover images, which expands its applications. Compared with the existing schemes, our scheme obtains the best image quality with similar effect of secret information removal. We will detail the theoretical analysis and experimental comparison in Section IV.

B. Thin Plate Spline and Spatial Transformer Network

Thin Plate Spline (TPS) [27] is a simulation method based on the physical phenomenon of bending of thin metal plates. Its basic idea is to map a set of control points on the plane, known as control points, to other points on the plane using TPS function, thereby realizing plane deformation and interpolation. Due to the ability of accurately aligning the input image with the target or reference image, TPS can enhance the performance and robustness of network models, and thus plays an important role in image synthesis [28], image segmentation [29], target detection [30], and many other applications.

TPS can be incorporated into Spatial Transformer Network (STN) to automatically correct perspective distortion [31]. It comprises a localisation network, a grid generator, and a sampler. The localisation network predicts a set of fiducial points, the grid generator estimates the TPS parameters and generates a sampling grid, and the sampler resamples the pixels in the output image. This module is differentiable and can be seamlessly integrated into existing architectures without extra supervised training. In our approach, we use this STN module to desynchronize the input instead of aligning it, making it impossible to extract the hidden data.

III. Proposed Algorithm

A. Analyze of Steganography Influence on Frequency Domain

1) Hiding Binary Messages: Embedding message bits in the least significant bits (LSBs) of image pixels has been extensively investigated in traditional steganographic techniques due to their undetectable nature [1], [2], [3]. Additionally, given the relative higher complexity of middle and high-frequency components compared to low-frequency components, secret information is preferentially embedded in these components for the purpose of statistical undetectability. In fact, some embedding strategies have been specifically defined in the frequency domain [32], [33], leading to a concentration of secret information primarily in these components.

LSB embedding, however, is susceptible to distortions. Recently, there has been a shift towards robust steganography to resist potential channel distortions, particularly against JPEG compression, as this image processing is commonly encountered in public platforms. JPEG compression splits an image into individual 8×8 pixel blocks and performs discrete cosine transform (DCT), quantization, and entropy coding operations on each block to compress the image. Among them, DCT converts the pixels into their frequency domain representation. The resulting DCT coefficients comprise one DC component and 63 AC components, as depicted in Figure 2(b). In this figure, the coefficient labeled with 1 signifies the DC component, while the remaining coefficients represent the AC components.

Since JPEG compression is performed in DCT domain, several schemes try to embed message bits in DCT coefficients to achieve this robustness [7], [8], [34], [35]. Inevitably, these approaches also concentrate the secret information in the middle and high-frequency components. Herein we evaluate the impact of two such schemes, MRE [34] and DLD [8], on the frequency domain. MRE constructs a robust domain utilizing robust element extraction in DCT domain. Then Reed-Solomon code (RS) and Syndrome Trellis Code (STC) [1] are employed to embed message bits. DLD aligns the embedding distortion with the modification probability of the lossy transmission channel, and proposes a framework of coding scheme extended from Dual-STCs.

The COCO dataset [36], with images sized 256×256, is employed for this evaluation, due to its extensive usage in data hiding applications [35], [37], [38]. We randomly select 100 test images from the COCO dataset and utilize the two steganographic schemes to generate stego images. Figure 2 depicts the distribution of embedding changes across various DCT coefficients. It can be observed that over 99% of these embedding changes occur in the middle and high-frequency components.

Based on the above observations, the embedded secret information can be compromised if certain fractions of DCT coefficients are removed. We remove 1/3 (corresponding to 44 to 64 AC coefficients in Fig. 2(b)), 2/3 (corresponding to 23 to 64 AC coefficients in Fig. 2(b)), and all AC components and attempt to re-extract the embedded message bits. The accuracy of message extraction, as a function of the removed DCT coefficients, is presented in Table I, where the extraction accuracy is measured by bit error rate (BER), defined as

$$BER = \frac{1}{l} \sum_{i=1}^{l} |R(i) - \hat{R}(i)| \tag{1}$$

where l is the length of the secret message, and R and $\hat{R}$ represent the secret messages extracted from the stego image before and after the attack. Furthermore, Peak Signal-to-Noise Ratio (PSNR) [39] and Structural SIMilarity index (SSIM) [40] are used to measure the image quality of attacked stego images. It can be observed that the secret messages have been almost completely erased.

TABLE I
QUALITY OF ATTACKED STEGO IMAGES AND EXTRACTION ACCURACIES WITH DIFFERENT PROPORTIONS OF AC COEFFICIENTS REMOVED

Steganography	PSNR			SSIM			BER		
	delete 1/3	delete 2/3	delete 3/3	delete 1/3	delete 2/3	delete 3/3	delete 1/3	delete 2/3	delete 3/3
MRE [34]	**40.82**	34.58	21.24	**0.987**	0.949	0.416	0.000	0.001	**0.493**
DLD [8]	**44.57**	35.22	21.34	**0.995**	0.957	0.455	0.002	0.011	**0.483**

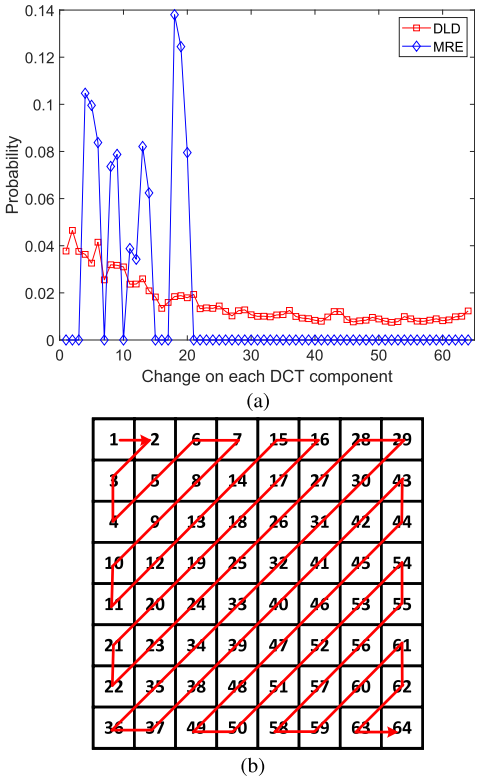


Fig. 2. (a) Distribution of embedding changes across different coefficients in 8 × 8 DCT blocks, where the stego images are obtained by MRE [34] and DLD [8]. The probabilities in (a) are calculated across all 8 × 8 DCT blocks, encompassing all color channels in all test images. Additionally, the scanning order of DCT coefficients adheres to the zigzag order utilized in JPEG compression, as illustrated in (b).

2) Hiding Secret Images: Numerous learning-based steganographic schemes leverage encoder-decoder architectures to embed images within other images [4], [5], [6], [41], [42]. Additionally, various modules have been developed to enhance the quality of both the stego and the recovered images, as well as improve their undetectability. For instance, ABDH [5] uses an attention mask to enhance the quality of stego images. UDH [41] restricts the embedding energy on the residual images. CAIS [42] proposes a self-generated supervision mechanism to ensure both visual quality and undetectability. On the other hand, invertible networks have recently gained popularity in image hiding, where they treat image concealing and revealing as a pair of reversed mappings [37], [38], [43]. All the above schemes have to minimize the perceptual distance between the cover and stego images, with some also considering the statistical distance. As a result, similar to traditional steganographic methods, they often conceal secret information in the middle and high frequency. HiNet [37] is a notable example, where a low-frequency wavelet loss is specifically imposed to ensure that the secret information is embedded in high-frequency wavelet sub-bands.

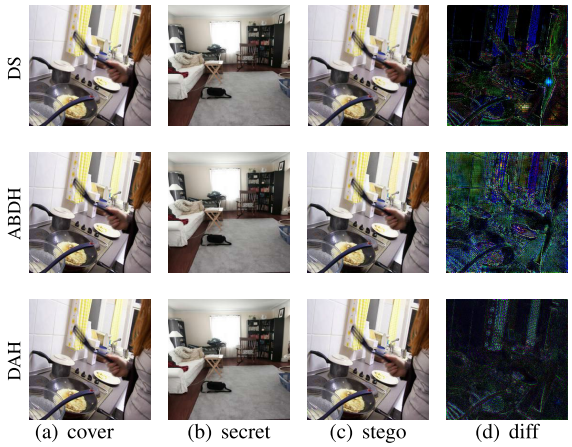


Fig. 3. Demonstration of cover images and corresponding stego images obtained by different schemes.

Herein we experimentally evaluate the steganography effect of these approaches on the frequency domain by using three methods: DS [4], ABDH [5], and DAH [6]. Once again, the COCO dataset is utilized, which is worth noting as it is also employed in ABDH. The training samples are the same as in Section IV-A, we randomly select 100 images from the validation set for testing. Figure 3 showcases the cover, secret, and stego images obtained by these methods. Additionally. we illustrate the difference between cover and stego images in this figure. It can be observed that the difference mainly lies in the edges and textured regions, suggesting that secret data is primarily embedded into the high-frequency components.

To further explore the embedding location, we remove a specific subset of high-frequency components and assess how this affects the recovered images. We employ the Destruction Rate (DT) metric, as defined in [9], to quantitatively evaluate the impact of this removal. It is defined as

$$DT = \frac{\sum_{i=1}^{h}\sum_{j=1}^{w}|R(i,j) - \hat{R}(i,j)|}{h \times w} \tag{2}$$

where h and w represent the length and width of the image, respectively, R and $\hat{R}$ represent the secret images recovered from the stego image before and after the attack. The higher the DT, the better the attack effectiveness. PSNR and SSIM are still used to measure the image quality of attacked stego images.

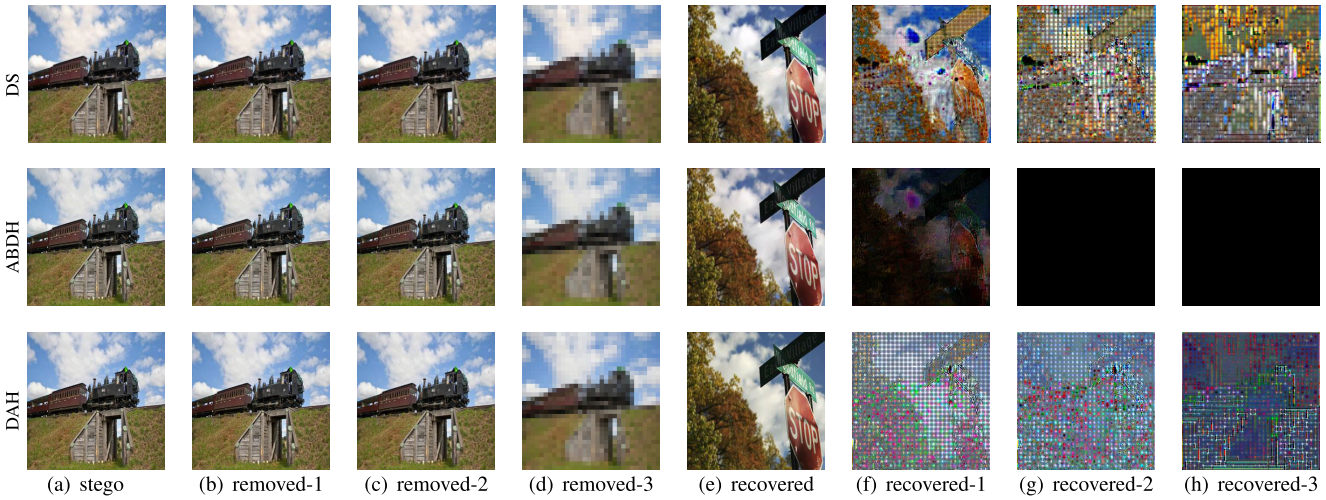


Fig. 4. Demonstration of recovered images from manipulated stego images, where parts of the high-frequency components have been removed. In (b), (c), and (d), 1/3, 2/3, and 3/3 of AC components have been removed, respectively. The corresponding recovered images are shown in (e), (f), and (g).

TABLE II
QUALITY OF RECOVERED AND STEGO IMAGES WITH DIFFERENT PROPORTIONS OF AC COEFFICIENTS REMOVED

Steganography	PSNR			SSIM			DT		
	delete 1/3	delete 2/3	delete 3/3	delete 1/3	delete 2/3	delete 3/3	delete 1/3	delete 2/3	delete 3/3
DS [4]	**40.47**	33.90	21.15	**0.984**	0.938	0.439	0.166	0.230	**0.26**
ABDH [5]	**37.51**	32.44	20.13	**0.979**	0.925	0.413	0.341	0.437	**0.437**
DAH [6]	**38.18**	32.71	19.90	**0.978**	0.924	0.487	0.225	0.261	**0.288**

We apply DCT on the stego images and remove specific fractions of AC components: 1/3, 2/3, and 3/3. The recovered images from the manipulated stego images are shown in Fig. 4. Table II lists the corresponding DT, PSNR, and SSIM scores. It can be observed that the quality of the recovered images experiences a significant decline, which occurs at a more rapid pace than attacking MRE and DLD. Specifically, when 1/3 or 2/3 of AC components are removed, the stego image quality remains relatively high, while the recovered images suffer significant distortion. Therefore, it suffices to alter only the high-frequency components to disturb secret information within stego images.

On the other hand, it suggests that different steganographic methods utilize distinct frequency components to conceal secret information. By eliminating 2/3 of the high-frequency components, almost all of the hidden information in the stego images created by ABDH can be removed. However, this process has less effects on the stego images created by DAH and DS.

B. Analyze of Desynchronization

Rather than removing the secret information as in previous work [9], [11], [12], we attempt to desynchronize the secret recovery by launching geometrical attacks. Previous geometrical attacks are mainly launched in the spatial domain [13], [14], [15], [16], [17], [44]. However, we have shown that steganographic schemes often conceal secret information within the middle and high-frequency components. Therefore, the effectiveness of geometrical attacks in the frequency domain is herein analyzed.

We consider a simple convolution $W \star X$, where W and X represent the convolution kernel and image, respectively, and $\star$ signifies the convolution operator. The influence of geometrical attack can be loosely assessed as

$$\begin{aligned} e_s &= E[W \star X - W \star \tilde{X}] \\ &= W \star E[X - \tilde{X}] \end{aligned} \tag{3}$$

where $\tilde{X}$ is the warped version of X. Suppose the pixel $\tilde{X}[\mathbf{u}]$ in $\tilde{X}$ at location $\mathbf{u}$ is transformed from $X[\mathbf{u} + \delta(\mathbf{u})]$ in X. Equation (3) can then be rewritten as

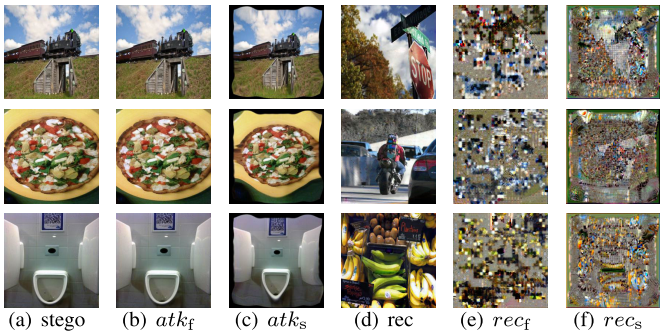
$$e_s = W \star E[P_{s,\delta}(X)] \tag{4}$$

where $P_{s,\delta}(X)$ is a probability matrix whose $\mathbf{u}$-th coefficient $P_{s,\delta}(X[\mathbf{u}]) = Pr\{X[\mathbf{u}] - X[\mathbf{u} + \delta(\mathbf{u})] | X[\mathbf{u}]\}$. Consequently, the attack influence e_{spat} remains small when the image X exhibits smoothness.

On the other hand, if we warp the AC coefficients of X, it can be derived that

$$\begin{aligned} e_f &= E[(W \star AFA^T - W \star A\tilde{F}A^T)] \\ &= W \star A(E[P_{f,\delta}(X)])A^T \end{aligned} \tag{5}$$

where A represents the DCT transform matrix. F represents the DCT coefficient matrix of X, and its coefficient located at $\mathbf{u}$ is denoted by $F[\mathbf{u}]$. The $\mathbf{u}$-th coefficient of $P_{f,\delta}$ can be calculated as $P_{f,\delta}(X[\mathbf{u}]) = Pr\{F[\mathbf{u}] - F[\mathbf{u} + \delta(\mathbf{u})] | F[\mathbf{u}]\}$. Since AC coefficients of an image are less smooth compared



(a) stego (b) atk_{f} (c) atk_{s} (d) rec (e) rec_{f} (f) rec_{s}

Fig. 5. Demonstration of geometrical attacks in the spatial domain and in the frequency domain. Stego images are (a), (b), and (c) are the original versions, images having been attacked in the frequency domain, and images having been attacked in the spatial domain, respectively. In (d), (e), and (f) are the recovered images from (a), (b), and (c), respectively.

TABLE III

QUALITY COMPARISON OF ATTACKED STEGO IMAGE AND RECOVERED IMAGES IN THE CASES OF DESYNCHRONIZATION IN THE SPATIAL DOMAIN AND IN THE FREQUENCY DOMAIN

Steganography	Spatial domain			Frequency domain		
	PSNR	SSIM	DT	PSNR	SSIM	DT
DS [4]	11.08	0.276	0.254	**36.65**	**0.951**	**0.299**
ABDH [5]	8.99	0.148	0.243	**39.43**	**0.981**	**0.247**
DAH [6]	12.32	0.416	**0.296**	**37.04**	**0.908**	0.270

to the pixels of that image, it satisfies that $P_{f,\delta}(X[\mathbf{u}]) > P_{s,\delta}(X[\mathbf{u}])$. Therefore, geometrical attacks in the frequency domain would incur a larger influence.

We experimentally evaluated this by employing Thin Plate Spline (TPS) [27] to implement the geometrical attack. In the first case, pixels of stego images are warped using TPS. In the second case, the AC components of stego images are warped. The influence on both the stego and recovered images in these two cases are demonstrated in Fig. 5, and Table III lists the corresponding scores. It can be observed that both attacks achieve similar DT scores, but the quality of stego images being attacked in the frequency domain is much higher than those being attacked in the spatial domain.

From Tables II and III, it is worth noting that warping AC components can better balance the attacked image quality and attack effectiveness compared to directly removing them. This type of attack can achieve competitive DT scores and, simultaneously, provide a higher quality of attacked stego images than removing 2/3 or 3/3 of AC components, only slightly worse than removing 1/3 of AC components.

C. Attack Realization

Based on the above analysis, we suggest a geometrical attack in the frequency domain (GAF), which only warps the middle and high-frequency components of stego images. We assume that a set of stego images as well as the secret images hidden in them are available to the attacker. It should be noted that neither the original host images nor their distribution are in need. The attack network learns to hinder the recovery of the secret image meanwhile keeping the imperceptibility of the attacked images.

Since images are typically stored as DCT coefficients in the YCbCr color model within the JPEG format, the proposed scheme is implemented on this color model and utilizes DCT to extract the frequency components of images. This strategy can align with the image format and thus alleviate the attacking impact on image quality. Additionally, as different steganographic methods utilize frequency components differently, the attack learns to adaptively warp frequency components for specific steganographic methods. Figure 6 describes the structure of the proposed attack network. It comprises two parts: the channel weight estimator and the frequency jammer. The channel weight estimator module assigns perturbation strengths to each DCT channel according to its contribution to carrying secret information. The frequency jammer module then performs a TPS transform to each frequency channel using the weight determined by the channel weight estimator.

1) Channel Weight Estimator: This module estimates the perturbation strength of each DCT channel of the stego image. Since the proposed scheme is implemented on the YCbCr color model, the Channel weight estimator initially converts the stego image to this color model. Then each component passes through a shared CNN layer to acquire three weight vectors, each with a length of 64. By combining these three vectors, we obtain the final weight vector of length 192. It can be formed as:

$$\mathbf{w} = EST(I_Y) \oplus EST(I_{Cb}) \oplus EST(I_{Cr}) \tag{6}$$

where EST represents the CNN layer, and I_Y, I_{Cb}, and I_{Cr} represent the Y, Cb, and Cr components of image I, respectively. $\oplus$ stands for the concatenation operation.

2) Frequency Jammer: This module performs a weighted TPS on the DCT coefficients of the stego image. It generates the initial and target control points and interpolates the warped field throughout the image using the TPS transform. We employ the Spatial Transformer Network (STN) with $c \times c$ control points, as suggested in [31], to generate the target control points θ and warped field $\mathbf{m}$.

The module is illustrated in Fig. 6. It starts with a 8×8 block-DCT layer to transform the Y, Cb, and Cr components of the stego image into the frequency domain, resulting in a DCT coefficient tensor with 192 channels. Then it is warped using STN. These warped DCT coefficients are subsequently added to the original DCT coefficients, weighted by the channel weight estimator. At last, the modified DCT coefficients are transformed back to the spatial domain, yielding the attacked image. This can be expressed as:

$$\begin{aligned} I_{\mathrm{atk}} = IDCT\,(\mathbf{w} \otimes STN(DCT(I_Y \oplus I_{Cb} \oplus I_{Cr})) \\ +(1-\mathbf{w}) \otimes DCT(I_Y \oplus I_{Cb} \oplus I_{Cr})) \end{aligned} \tag{7}$$

where $\otimes$ denotes the matrix cross product.

3) Loss Function: In order to improve the imperceptibility of attacked images and disturb secret information effectively, our loss function consists of two parts: perceptual loss and disturbing loss. It is formed as

$$L = \lambda_1 L_{\mathrm{perc}} + \lambda_2 * L_{\mathrm{dist}} \tag{8}$$

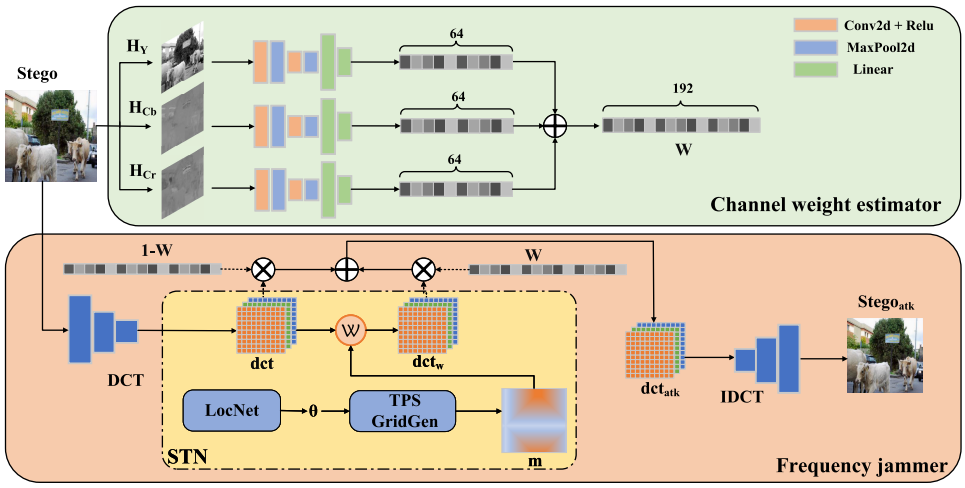


Fig. 6. Overview of the proposed geometrical attack in frequency domain (GAF). In the figure, **w** represents the channel weight, whereas θ and **m** represent the target control points and warped field, respectively, used in the TPS.

where λ_1 and λ_2 are the weights for balancing the individual objective.

We utilize VGG loss as the perceptual loss. Specifically, a pre-trained VGG19 model is applied to extract feature maps from the stego image I and the attacked image I_{atk}, and calculate:

$$L_{\text{perc}} = \sum \frac{1}{N} \|VGG_k(I) - VGG_k(I_{atk})\|_2^2 \tag{9}$$

where $VGG(k)$ represents the features extracted from layer k in VGG19. N represents the total feature neuron numbers.

Regarding the disturbing loss, it maximizes the disparity between the recovered image from the original stego image and that from the attacked image. We employ L1 norm to measure this disparity, as suggested in [9]. Consequently, this loss can be written as:

$$L_{\text{dist}} = \frac{h * w}{\sum |R - \hat{R}|} \tag{10}$$

where $\sum |R - \hat{R}|$ represents the BER defined in Eq. (1) when attacking steganographic schemes that hide binary messages. Conversely, when aiming at steganographic schemes that hide secret images, it signifies the DT defined in Eq. (2).

IV. Experiments

A. Experimental Setup

To assess the effectiveness of our GAF, we conducted a comparative analysis with two learning-based active steganography attacks: CAGP [12] and PS [9]. CAGP utilizes a GAN to eliminate hidden information from the stego images, while PS aims to restore the original distribution of cover images. Besides the COCO dataset described in Section III, we also conduct experiments on ImageNet, MLRSNet, DIV2K and CIFAR-10.

ImageNet [45] is a large visual database for object recognition. This dataset was collected through web crawling, manual annotation, and leveraging Amazon crowdsourcing platforms. For our purposes, we utilized a subset known as ISLVRC2012, which comprises over 1.2 million diverse natural images. Given the variance in image dimensions, we standardized the size for our experiments by uniformly scaling them to a resolution of 256×256.

MLRSNet [46] comprises high-resolution optical satellite and aerial images, offering an overhead perspective of Earth's objects captured through satellite or aerial sensors. The dataset comprises a collection of 109, 161 samples, categorized into 46 distinct scene types. The size of each RGB image is 256×256.

DIV2K [47] consists from 1000 DIVerse 2K resolution RGB images. These images exhibit exceptional aesthetic quality and little noise and other corruptions. The dataset encapsulates a broad range of content, encompassing people, handmade objects, environments, flora, fauna, and natural sceneries, including underwater scenes. While the images vary in size, we have uniformly resized them to 256×256 for our experiments.

CIFAR-10 [48] is a small dataset for object recognition. It contains a total of 10 categories of RGB color images gathered from the Internet. There are in total 60, 000 images in the dataset, which are of size 32×32.

These datasets are compiled from a large number of natural images spanning various scenarios, which can serve to enhance the validation of the experimental results' generalizability in real-world contexts. Due to the small number of images in the DIV2k dataset, we randomly select 800 images for training and 200 images for testing. For the other datasets we randomly select 50, 000 images for training and 5, 000 images for testing.

The experiments are conducted on a GPU GeForce RTX 3090 24G in the environment of Python 3.8.8 and Pytorch 1.10.1+cu111. For the proposed network, we employ the Adam optimizer with a learning rate of 0.001. The default value of k is 2 for images of size 32×32 and 10 for images of size 256×256. c used for TPS is set with 10. The values of λ_1 and λ_2 in Eqn. (8) are set to 6 and 4, respectively. The value of batch size is set to 8.

B. Theoretical Analysis

We first evaluates the effectiveness of the proposed scheme by delving into the impact of desynchronization attacks. It is

worth mentioning that there exist theoretical analyses of desynchronization attacks [18], [49]. However, our focus in this context is on the superiority of desynchronization attacks over modification-based approachess.

Consider a generalized data hiding scheme that creates a stego image I by

$$I = Gen(X, S) = X + M \tag{11}$$

where X represents the cover image, and M denotes the necessary image modifications for retrieving the concealed information S. Alongside the generator, this scheme incorporates a decoder capable of extracting the secret through $R = rec(I)$.

On the opposing side, an attacker $atk(I)$ endeavors to maximize the recovery error $L_R(rec(I), rec(atk(I)))$, while concurrently minimizing the distortion inflicted upon the image content $L_X(X, atk(I))$. Here, L_R and L_X represent specific loss functions tailored for their respective tasks. It should be noted that certain schemes assess distortion specifically on the stego image I. Nevertheless, the signal M contributes marginally to the image content and is typically significantly lower than X. Given that X and S are independent, it's reasonable to presume that the entirety of the secret information is encoded within M. Consequently, the attacker's objective is effectively equivalent to maximizing $L_M(M, atk(M))$ for some specified loss function L_M.

We investigate a straightforward scenario in which both X and M adhere to Gaussian distributions characterized by zero means and their respective variances, σ_X^2 and σ_M^2. Subsequently, the aforementioned loss functions can be expressed in terms of capacity, which exhibits a loose inverse relationship with the signal-to-noise ratio (SNR)

$$1/L_M(M, atk(M)) \propto \sigma_M^2/\sigma_{N_M}^2 \tag{12}$$

$$1/L_X(X, atk(I)) \propto \sigma_X^2/\sigma_{N_X}^2 \tag{13}$$

Here, $\sigma_{N_M}^2$ represents the variance of noise associated with the channel for concealing information, while $\sigma_{N_X}^2$ denotes the variance of noise pertaining to the image channel. For simplicity, we refer to the SNR in Eqs. (12) and (13) as SNR_M and SNR_X, respectively.

Many attack strategies, including CAGP and PS, rely on directly altering the pixel values of the stego image. This manipulation can be formed as:

$$\tilde{I}[\mathbf{u}] = I[\mathbf{u}] + N[\mathbf{u}] = X[\mathbf{u}] + M[\mathbf{u}] + N[\mathbf{u}] \tag{14}$$

where $N[\mathbf{u}]$ represents the introduced noise, which follows $N[\mathbf{u}] \sim \mathcal{N}(0, \sigma_N^2)$. In this context, the two SNRs can be derived as

$$SNR_M = \sigma_M^2/(\sigma_X^2 + \sigma_N^2) \tag{15}$$

$$SNR_X = \sigma_X^2/(\sigma_M^2 + \sigma_N^2) \tag{16}$$

Contrasting with the aforementioned approaches, the proposed scheme performs a desynchronization attack. Following the channel model described in [18], we introduce a constant offset $\delta > 0$ to I and utilize a simple linear interpolation between $I[\mathbf{u}]$ and $I[\mathbf{u} + \mathbf{t}]$ at a distance $\mathbf{t}$ to compute the desynchronized signal

$$\tilde{I}'[\mathbf{u}] = \frac{t-\delta}{t} I[\mathbf{u}] + \frac{\delta}{t} I[\mathbf{u} + \mathbf{t}] \tag{17}$$

Taking into account the inherent relationship within image content, the value of $X[\mathbf{u} + \mathbf{t}]$ can be expressed as $X[\mathbf{u} + \mathbf{t}] = X[\mathbf{u}] - P_\delta(X[\mathbf{u}])$, where $P_\delta(X[\mathbf{u}]) = Pr\{X[\mathbf{u}] - X[\mathbf{u} + \delta]|X[\mathbf{u}]\}$. Loosely, we assume that $P_\delta(X[\mathbf{u}]) \sim \mathcal{N}(0, \sigma_\delta^2)$, with $\sigma_{P_\delta}^2$ typically increasing as δ increases. Subsequently, Eq. (17) can be reformed as

$$\tilde{I}'[\mathbf{u}] = X[\mathbf{u}] + \frac{t-\delta}{t} M[\mathbf{u}] - \frac{\delta}{t} P_\delta(X[\mathbf{u}]) + \frac{\delta}{t} M[\mathbf{u} + \mathbf{t}] \tag{18}$$

It can be observed that only the first term contributes to the original image sample, while solely the second term accounts for the original hidden information sample. Consequently, the two SNRs are now computed as

$$SNR'_M = \frac{(\mathbf{t} - \delta)^2 \cdot \sigma_M^2}{\mathbf{t}^2 \cdot \sigma_X^2 + \delta^2 \cdot (\sigma_{P_\delta}^2 + \sigma_M^2)} \tag{19}$$

$$SNR'_X = \frac{\mathbf{t}^2 \cdot \sigma_X^2}{(\mathbf{t} - \delta)^2 \cdot \sigma_M^2 + \delta^2 \cdot (\sigma_{P_\delta}^2 + \sigma_M^2)} \tag{20}$$

Assuming that both types of attacks impact message recovery equivalently, that is, $SNR_M = SNR'_M$, we can derive that

$$SNR'_X/SNR_X = a/b \tag{21}$$

$$a = (\sigma_M^2 + \sigma_N^2) * (\sigma_N^2 + \sigma_X^2 - \sigma_1^2)^2 \tag{22}$$

$$b = \left(\sigma_M^2 * \left(\sigma^2 - \sqrt{(\sigma_1^2 + \sigma_X^2) * \sigma_N^2 + \sigma_X^4}\right)^2 + \sigma_1^2 * \left(\sigma_N^2 + \sigma_X^2 - \sqrt{(\sigma_1^2 + \sigma_X^2) * \sigma_N^2 + \sigma_X^4}\right)^2\right) \tag{23}$$

$$\sigma_1^2 = \sigma_{P_\delta}^2 + \sigma_M^2 \tag{24}$$

Observing the expressions a and b, it is evident that both are convex functions with respect to $\sigma_{P_\delta}^2$. Notably, they intersect solely at the point where $\sigma_{P_\delta}^2 = \sigma_N^2 + \sigma_X^2 - \sigma_M^2$, and $a > b$ in all other cases. Consequently, this implies that $SNR'_X \geq SNR_X$, signifying that desynchronization attacks preserve the quality of the attacked image better than modification-based attacks. On the contrary, when $\sigma_{P_\delta}^2 \to \infty$, SNR'_X becomes equal to SNR_X. This indicates that as the correlation of the cover signal weakens, the advantage of desynchronization attacks diminishes.

It is worth emphasizing that modeling state-of-the-art modification-based attacks with additive white Gaussian noise (AWGN) is not restrictive. Similarly, modeling the proposed scheme with a desynchronization with a constant offset is also non-restrictive. Both modification and desynchronization in these schemes are optimized to strike a balance between attack effectiveness and image quality preservation. Nevertheless, the analysis still offers valuable insights into the effectiveness of desynchronization attacks.

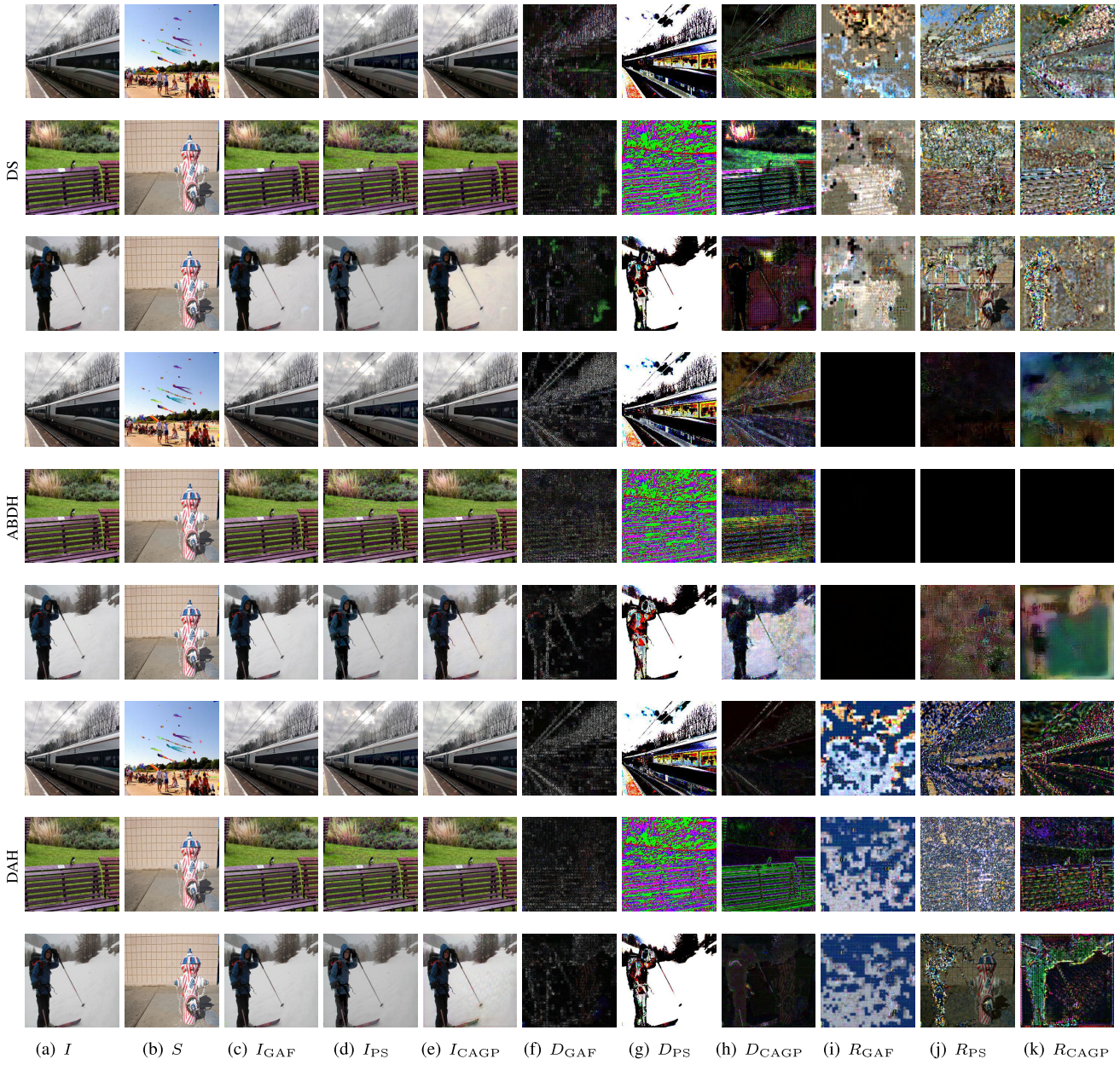


Fig. 7. Demonstration of attack impact of three attack schemes on different steganographic methods. I is the stego image, S is the secret image; I_* represent the results of the three attack models; D_* represents the difference between the stego image and the attacked image, and R_* represents the secret image re-extracted from the attacked image.

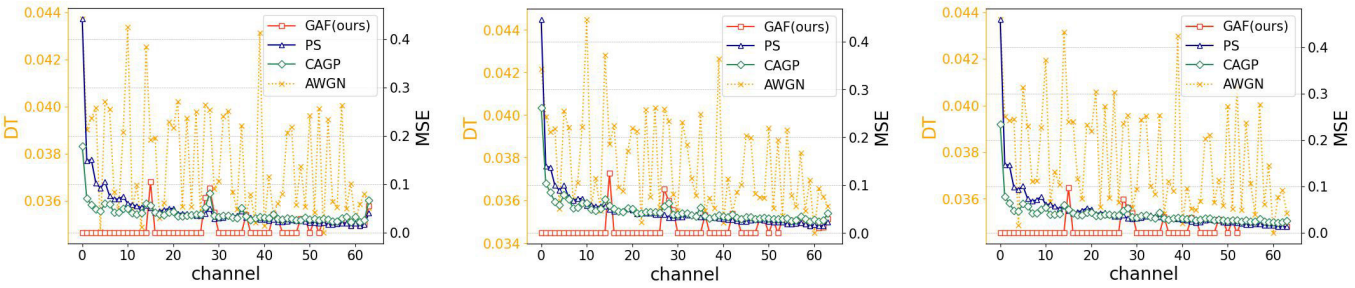
C. Evaluation of Removing Hidden Images

This section evaluates the performance of the proposed approach on attacking image hiding schemes. Three schemes, DS [4], DAH [6] and ABDH [5], are used as the attack target. All of them can hide a secret image within another image.

1) Visual Comparison: The impact of different attack schemes is demonstrated in Fig. 7. It can be seen that all the schemes can effectively hinder the secret image recovery. However, utilizing CAGP against ABDH leaves some of the secret information within certain stego images. For example, in the 4th row of Fig. 7, the content of the secret image can still be faintly discerned within the recovered image. Similarly, PS would fail to eliminate all hidden secret information in the flat regions of stego images. We can still detect the fire hydrant in the recovered image in the second to last column of Fig. 7, which was presumably hidden in the snowfield of the original stego image. In contrast, our scheme consistently performs well across these test stego images.

Furthermore, we evaluated the quality of the attacked images. As shown in Fig. 7, images attacked using CAGP and PS frequently display noticeable color deviations compared to the original stego images. Conversely, our scheme effectively mitigates this issue, ensuring that the attacked images remain highly similar to the original stego images. This is primarily due to the TPS transformation's ability to preserve pixel distribution, thus reducing distortion within the attacked images.

2) Frequency Awareness Comparison: In Section III-A, we have demonstrated that numerous data hiding techniques conceal secret information within the middle and high-frequency components. Herein, we assess whether the schemes



(a) DS (b) DAH (c) ABDH

Fig. 8. Demonstration of the sensitivity of each DCT component towards AWGN in terms of DT. Additionally, the modifications to these DCT components resulting from three attacking schemes are depicted in terms of MSE. The targeted image-hiding schemes are a) DS, b) DAH, and c) ABDH.

TABLE IV
COMPARISON OF THREE ATTACK SCHEMES ON ATTACKING STEGANOGRAPHIC METHODS THAT HIDE SECRET IMAGES

Dataset	Attack	DS [4]			ABDH [5]			DAH [6]		
		PSNR	SSIM	DT	PSNR	SSIM	DT	PSNR	SSIM	DT
COCO	GAF(ours)	**36.16**	**0.946**	**0.308**	**37.05**	**0.960**	**0.418**	**40.21**	**0.977**	**0.362**
	PS [9]	25.51	0.842	0.245	25.46	0.828	0.372	25.51	0.842	0.245
	CAGP [11]	31.80	0.922	0.260	30.83	0.949	0.272	33.39	0.964	0.332
ImageNet	GAF(ours)	**38.47**	**0.959**	**0.319**	**39.19**	**0.973**	**0.377**	**42.01**	**0.980**	0.288
	PS [9]	25.52	0.897	0.213	25.36	0.838	0.369	25.20	0.852	0.260
	CAGP [11]	32.69	0.920	0.304	36.11	0.950	0.220	37.01	0.968	**0.297**
MLRSNet	GAF(ours)	**41.98**	**0.972**	**0.286**	**39.04**	**0.969**	**0.295**	**40.18**	**0.985**	**0.273**
	PS [9]	25.89	0.805	0.254	25.95	0.787	0.287	25.84	0.902	0.259
	CAGP [11]	29.99	0.919	0.263	36.43	0.963	0.294	35.81	0.969	0.229
DIV2K	GAF(ours)	**39.68**	**0.964**	**0.305**	**38.68**	**0.960**	**0.314**	**38.04**	**0.981**	**0.322**
	PS [9]	26.26	0.735	0.201	26.04	0.893	0.316	25.71	0.831	0.264
	CAGP [11]	29.54	0.956	0.281	31.42	0.955	0.283	31.15	0.914	0.311
Cifar-10	GAF(ours)	**34.27**	0.978	**0.279**	**41.31**	**0.993**	**0.714**	**39.57**	**0.991**	**0.381**
	PS [9]	25.46	0.898	0.251	26.04	0.871	0.654	25.87	0.898	0.299
	CAGP [11]	33.35	**0.981**	0.271	35.09	0.889	0.516	35.56	0.990	0.310

being compared are cognizant of this distinction across frequency components.

The experiments on the three image hiding schemes are conducted using the COCO dataset. To investigate the influence of different frequencies, we apply identical AWGN noise, $N \sim \mathcal{N}(0, 1)$, specifically targeting certain DCT components while preserving the remaining ones unchanged. By altering the targeted DCT components, we compared the effectiveness of the attacks. The comparison results in terms of DT are presented in Fig. 8. It can be observed that certain components exhibit greater sensitivity to the attacks compared to others. Consequently, it is justifiable to adjust the attack strength across DCT components.

Next, we compare the modifications induced by the three attack schemes, CAGP, PS, and our GAF. The comparison results in terms of MSE are illustrated Fig. 8. It can be observed that both CAGP and PS primarily concentrate their modifications on the DC component, while the AC components are almost equally influenced. In contrast, our proposed scheme affects only a select few AC components, aligning closely with their sensitivity. It implies that our approach can hinder the secret recovery more effectively.

3) Quantitative Comparison: To facilitate a more intuitive comparison, we then use PSNR and SSIM as the indicators to assess the quality of the attacked images. Further, DT is employed to measure the attack effectiveness. Table IV lists the experimental results averaged over five datasets. It can be observed that our approach exhibits slightly higher DT values compared to the other two attack schemes for all the datasets except ImageNet. Furthermore, our scheme achieves the best imperceptibility among all the attacked schemes on COCO, ImageNet, MLRSNet, DIV2K. However, when attacking DS on the CIFAR-10 dataset, our method's SSIM scores are slightly worse than those of PS, but it still obtains the highest PSNR scores, indicating fewer modifications. On the whole, our scheme outperforms the compared ones in terms of both attacked image quality and attack impact. This suggests that our attack scheme can effectively disrupt secret image recovery while maintaining high attacked image quality.

TABLE V
COMPARISON OF COMMON SIGNAL PROCESSING ON ATTACKING STEGANOGRAPHIC METHODS THAT HIDE SECRET IMAGES

Dataset	Attack	DS [4]			ABDH [5]			DAH [6]		
		PSNR	SSIM	DT	PSNR	SSIM	DT	PSNR	SSIM	DT
COCO	JEPG	34.25	0.924	0.244	33.23	0.946	0.421	34.83	0.967	0.302
	Gaussian noise	22.14	0.457	0.288	36.47	0.890	0.418	36.50	0.956	0.262
	GAF(our)	**36.23**	**0.947**	**0.303**	**37.21**	**0.964**	**0.432**	**40.03**	**0.973**	**0.368**

TABLE VI
EFFECTIVENESS OF STEGANOGRAPHY ATTACK SCHEMES ATTACKING ROBUST STEGANOGRAPHIC SCHEMES THAT HIDE BINARY MESSAGES

Dataset	Attack	MRE [34]			DLD [8]			QEM [11]		
		PSNR	SSIM	BER	PSNR	SSIM	BER	PSNR	SSIM	BER
COCO	GAF(ours)	**35.69**	**0.956**	**0.435**	**35.14**	**0.959**	0.312	**37.69**	**0.966**	**0.495**
	PS [9]	24.13	0.839	0.296	24.10	0.841	**0.358**	24.35	0.873	0.491
	CAGP [11]	30.89	0.936	0.224	30.91	0.945	0.294	28.16	0.886	0.493
ImageNet	GAF(ours)	**41.60**	**0.988**	**0.473**	**36.71**	**0.957**	0.307	**43.65**	**0.989**	0.481
	PS [9]	24.58	0.904	0.368	24.36	0.878	**0.329**	24.65	0.931	0.423
	CAGP [11]	33.93	0.946	0.310	32.21	0.971	0.297	35.69	0.948	**0.487**
MLRSNet	GAF(ours)	**42.17**	**0.929**	**0.423**	**36.12**	**0.971**	**0.322**	**39.91**	**0.978**	**0.519**
	PS [9]	24.36	0.893	0.280	24.36	0.890	0.263	25.23	0.894	0.487
	CAGP [11]	32.15	0.970	0.252	34.46	0.886	0.317	33.42	0.940	0.499
DIV2K	GAF(ours)	**36.59**	**0.976**	**0.457**	**35.44**	**0.967**	**0.323**	**38.02**	**0.955**	**0.496**
	PS [9]	24.71	0.907	0.366	25.64	0.912	0.319	25.52	0.849	0.486
	CAGP [11]	32.13	0.974	0.236	31.40	0.869	0.280	32.55	0.943	0.491

To further validate the effectiveness of our scheme, We compare it with two widely employed image processing methods: Gaussian noise and JEPG compression. Exclusively utilizing the COCO dataset, we present the comparison results in Table V. It shows that our method exhibits higher PSNR and SSIM scores compared to the two image processing methods at similar DT values. These conventional methods often compromise the overall image quality, disregarding the uneven distribution of secret information within the stego image. In contrast, our approach optimizes the perturbations, resulting in a significant enhancement in the quality of attacked images.

Fig. 9. Demonstration of attack impact of three attack schemes on robust data hiding method.

D. Evaluation of Removing Robust Hidden Data

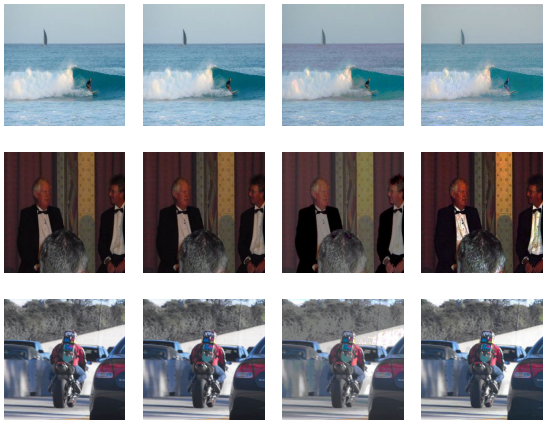
Since robust steganography enhances traditional steganographic techniques with anti-distortion capabilities, it is crucial to assess the efficacy of attacking these enhanced methods. Noting that various robust data hiding techniques can be explored for this purpose, in addition to the robust steganographic schemes MRE [34] and DLD [8], a data hiding technique described in [11], denoted as QEM, is also utilized for a more comprehensive evaluation. This method utilizes quantization index modulation on the quaternion exponent moment of the cover image to ensure robustness against JPEG compression, filtering, noising, and other common distortions.

We apply the three learning-based attack schemes to these steganographic methods and compare the visual quality of attacked images as well as the effectiveness of attacks. PSNR and SSIM are continue employed to assess the attacked image quality, whereas the effectiveness of the attack is now measured by BER defined in Eq. (1).

Figure 9 demonstrates the impact of attacking QEM. It can be observed that all three schemes successfully remove the hidden data. However, our scheme achieves the best attacked image quality. We further compare them on the COCO, ImageNet, MLRSNet, and DIV2K datasets. For DLD and QEM, pseudorandom binary sequences of length 1024 are utilized as message bits, whereas for MRE, pseudorandom binary sequences of length 570 are employed. Table VI presents the average comparison results. It confirms that the PSNR and SSIM scores obtained by our scheme are higher than the

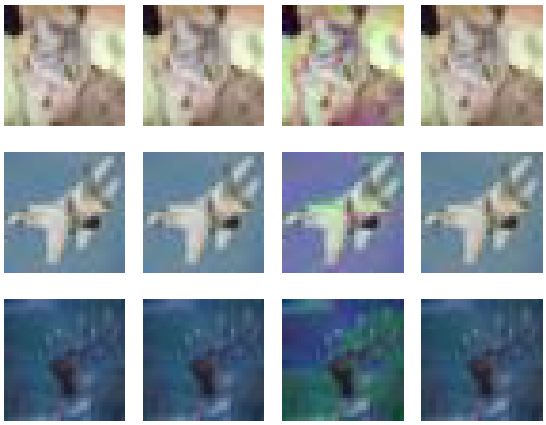
TABLE VII
COMPARISON OF COMMON SIGNAL PROCESSING ON ATTACKING STEGANOGRAPHIC METHODS THAT HIDE BINARY MESSAGES

Dataset	Attack	MRE [34]			DLD [8]			QEM [11]		
		PSNR	SSIM	BER	PSNR	SSIM	BER	PSNR	SSIM	BER
COCO	JEPG	35.08	0.919	0.112	34.98	0.840	0.259	33.16	0.910	**0.501**
	Gaussian noise	35.64	0.956	0.146	34.90	0.929	0.250	36.44	0.912	0.487
	GAF(our)	**35.69**	**0.956**	**0.435**	**35.14**	**0.959**	**0.312**	**37.69**	**0.966**	0.495



(a) *clear* (b) I_{GAF} (c) I_{PS} (d) I_{CAGP}

Fig. 10. Demonstration of three attack schemes directly attacking the images in the COCO dataset.



(a) *clear* (b) I_{GAF} (c) I_{PS} (d) I_{CAGP}

Fig. 11. Demonstration of three attack schemes directly attacking the images in the Cifar-10 dataset.

compared ones. Although the BER scores achieved by our proposed scheme may sometimes be slightly inferior to those of the comparison methods when employing the ImageNet dataset, the quality of the attacked images significantly surpasses that of the competitors. Additionally, Table VII reports attacking results using Gaussian noise and JPEG compression. The attack energy of these two signal processing is adjusted to achieve a similar visual quality of the attacked images as our proposed scheme. The results demonstrate that our method significantly outperforms in terms of attack effectiveness.

E. Evaluation of Harmless on Clear Images

In many instances, it is challenging to determine if a publicly transmitted image contains secret information, making the harmlessness of clean images a concern for attack schemes. We compare three attack methods, GAF, CAGP, and PS, on clear images from datasets with different image sizes. Figure 10 illustrates the impact of directly attacking the images from the COCO dataset, while Fig. 11 illustrates the impact of directly attacking the images from the Cifar-10 dataset. The experimental results on all the datasets are reported in Table VIII. It can be observed that our GAF offers higher PSNR and SSIM scores than the other two schemes on all the datasets used for the experiments. As a result, it can effectively reduce the impact on innocent communication.

TABLE VIII
HARMLESSNESS COMPARISON OF THREE ATTACK SCHEMES ON THE COCO DATASET

DATASET	GAF(ours)		PS [9]		CAGP [11]	
	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
COCO	**40.04**	**0.982**	24.60	0.907	30.57	0.936
MLRSNet	**43.67**	**0.987**	25.34	0.823	35.20	0.970
DIV2K	**37.77**	**0.972**	24.35	0.876	32.07	0.939
ImageNet	**39.14**	**0.977**	24.56	0.852	32.23	0.916
Cifar-10	**39.69**	**0.988**	25.08	0.840	35.29	0.979

TABLE IX
ABLATION EXPERIMENTS FOR CHANNEL WEIGHT ESTIMATOR

	Steganography	PSNR	SSIM	DT
GAF	DS [4]	**37.08**	**0.957**	**0.255**
	ABDH [5]	**35.97**	**0.956**	0.456
	DAH [6]	**39.79**	**0.975**	0.305
GAF^{-}	DS [4]	34.69	0.932	0.235
	ABDH [5]	33.16	0.931	**0.458**
	DAH [6]	33.56	0.921	**0.308**

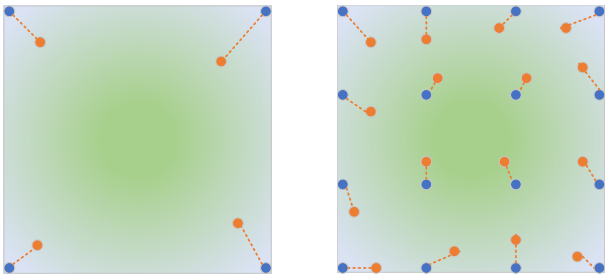
* GAF indicates with channel weight estimator, and GAF^{-} indicates without channel weight estimator.

F. Ablation Study

1) Channel Weight Estimator: We first investigate the influence of the channel weight estimator. Let GAF represent the proposed framework that incorporates this module, and GAF^{-} represent the framework without it. We randomly select 30 images from the test dataset for the experiment and then compare the PSNR, SSIM, and DT metrics for both

TABLE X
PERFORMANCE OF GAF WITH DIFFERENT NUMBERS OF CONTROL POINTS

Steganography	PSNR							SSIM							DT						
	n=2	n=4	n=6	n=8	n=10	n=12	n=14	n=2	n=4	n=6	n=8	n=10	n=12	n=14	n=2	n=4	n=6	n=8	n=10	n=12	n=14
DS [4]	34.29	36.84	36.74	37.10	**37.19**	37.09	37.12	0.934	**0.961**	0.957	0.959	0.961	0.956	0.960	0.160	0.228	0.258	0.241	0.237	**0.259**	0.242
ABDH [5]	31.49	34.26	36.31	35.26	35.59	35.97	**36.34**	0.898	0.944	**0.966**	0.958	0.959	0.956	0.965	0.457	0.454	0.457	0.457	**0.457**	0.453	0.457
DAH [6]	37.25	39.63	38.17	34.32	**40.95**	39.80	40.87	0.962	0.975	0.967	0.936	**0.980**	0.974	0.980	0.196	0.292	0.300	0.304	**0.339**	0.305	0.316



(a) (b)

Fig. 12. Demonstration of distribution of control points. The total number of control points is $2 \times 2 (c = 2)$ in (a), while the number is $4 \times 4 (c = 4)$ in (b).

frameworks. The experimental results are listed in Table IX. It can be found that using a channel weight estimator generally enhances performance. However, it slightly compromises the attack effectiveness against ABDH and DAH. Nevertheless, it significantly enhances imperceptibility. This may be attributed to the channel weight estimator's ability to sidestep frequency components that contain little secret information but play a crucial role in stego image quality.

2) Number of Control Points: To generate the warped field, the STN demands the selection of control points across a grid of $c \times c$. A larger c signifies a more dense distribution of control points, leading to more precise control. Herein we assess how the number of control points impacts the performance of our approach. We vary c in $\{2, 4, 6, 8, 10, 12, 14\}$ and retrain our scheme on attacking three steganographic methods. Table X presents the experimental results in terms of PSNR, SSIM, and DT. It can be found that increasing c generally enhances both attacked image quality and attack effectiveness. This is likely due to the fact that a higher density of control points ensures a greater number of points are situated within regions containing secret information. For instance, Fig. 12 displays a green-colored region where secret information lies. In Fig. 12(a), only four points are present, leaving the region unoccupied. Conversely, Fig. 12(b) exhibits a more dense distribution of control points, ensuring points are present within the region and thereby enhancing TPS performance. However, it is also observable that an overabundance of control points can undermine performance. Empirically, $c = 10$ can achieve good performance and thus is chosen in our approach.

3) Loss Weight Setting: There are two super parameters in Eq. (8), λ_1 and λ_2, that are used to balance attack effectiveness and attacked image quality. This section tests the selection of these two parameters. Table XI gives the experimental results by using different λ_1 and λ_2. As can be seen, a larger λ_1 can lead to better attacked image quality, but it also dramatically compromises attack effectiveness. It can be observed that a balanced approach can be achieved with $\lambda_1 = 6$ and $\lambda_2 = 4$. As a result, this setting has been chosen for our scheme.

TABLE XI
PERFORMANCE OF THE PROPOSED SCHEME WITH DIFFERENT λ_1 AND λ_2

	$\lambda_1 = 3$, $\lambda_2 = 7$	$\lambda_1 = 4$, $\lambda_2 = 6$	$\lambda_1 = 5$, $\lambda_2 = 5$	$\lambda_1 = 6$, $\lambda_2 = 4$	$\lambda_1 = 7$, $\lambda_2 = 3$
PSNR	36.19	36.96	37.08	**38.07**	37.68
SSIM	0.943	0.955	0.956	0.961	**0.964**
DT	**0.283**	0.254	0.256	0.262	0.239

V. CONCLUSION AND DISCUSSION

In this paper, we propose a geometrical attack in the frequency domain (GAF) to thwart the secret image recovery of advanced steganographic methods, particularly those learning-based approaches. Our scheme filters the DCT coefficients by a channel weight estimator and then uses a frequency jammer to subtly warp the DCT coefficients with the TPS transform. Observing that simply geometrically perturbing the high frequency of stego images can effectively disrupt the recovery capabilities of numerous steganographic methods, the proposed scheme only warps the DCT coefficients instead of manipulating the pixel values directly as in prior works. Furthermore, the suggested channel weight estimator focuses the perturbation on DCT channels that carry significant secret information but contribute little to the overall image quality. These strike a favorable balance between attacked image quality and attack effectiveness. Additionally, it does not require any prior knowledge of cover images and can preserve the image quality well when attacking clear images, making it a good strategy to hinder potential covert communication.

However, our method has certain limitations. It becomes less effective when a significant portion of the secret information is concealed within the low-frequency components of the cover image. For example, DLD uses some of the low-frequency information, which leads to large perturbations and ineffective attacks when we attack the scheme. As shown in Table VI, the image quality after attacking the DLD is lower than that after attacking the MRE on multiple datasets, and the effectiveness of the attack on the DLD is also reduced relative to the MRE. Additionally, robust data hiding schemes that are resistant to geometrical attacks exist [50], [51], [52]. Our scheme does

not perform well against these schemes. When attacking [51] and [52], we encountered issues related to vanishing gradients, resulting in the inability of the training process to converge. Nevertheless, these schemes were not originally designed for steganography purposes. The capacity and undetectability of these schemes cannot meet the steganography requirement.

In future research, we aim to develop a reversible perturbation approach that would enable the recovery of secret images embedded within attacked images when necessary. Furthermore, the potential to replace secret information within stego images could be valuable in specific scenarios.

REFERENCES

[1] T. Filler, J. Judas, and J. Fridrich, "Minimizing embedding impact in steganography using trellis-coded quantization," *Proc. SPIE*, vol. 7541, pp. 38–51, Jan. 2010.

[2] B. Feng, W. Lu, and W. Sun, "Secure binary image steganography based on minimizing the distortion on the texture," *IEEE Trans. Inf. Forensics Security*, vol. 10, no. 2, pp. 243–255, Feb. 2015.

[3] W. Tang, B. Li, M. Barni, J. Li, and J. Huang, "An automatic cost learning framework for image steganography using deep reinforcement learning," *IEEE Trans. Inf. Forensics Security*, vol. 16, pp. 952–967, 2021.

[4] S. Baluja, "Hiding images in plain sight: Deep steganography," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, 2017, pp. 1–11.

[5] C. Yu, "Attention based data hiding with generative adversarial networks," in *Proc. AAAI Conf. Artif. Intell.*, vol. 34, 2020, pp. 1120–1128.

[6] L. Zhang, Y. Lu, J. Li, F. Chen, G. Lu, and D. Zhang, "Deep adaptive hiding network for image hiding using attentive frequency extraction and gradual depth extraction," *Neural Comput. Appl.*, vol. 35, no. 15, pp. 10909–10927, May 2023.

[7] J. Tao, S. Li, X. Zhang, and Z. Wang, "Towards robust image steganography," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 29, no. 2, pp. 594–600, Feb. 2019.

[8] Q. Guan, P. Liu, W. Zhang, W. Lu, and X. Zhang, "Double-layered dual-syndrome trellis codes utilizing channel knowledge for robust steganography," *IEEE Trans. Inf. Forensics Security*, vol. 18, pp. 501–516, 2023.

[9] D. Jung, H. Bae, H.-S. Choi, and S. Yoon, "PixelSteganalysis: Pixel-wise hidden information removal with low visual degradation," *IEEE Trans. Dependable Secure Comput.*, vol. 20, no. 1, pp. 331–342, Jan. 2023.

[10] F. A. Petitcolas, M. Steinebach, F. Raynal, J. Dittmann, C. Fontaine, and N. Fates, "Public automated web-based evaluation service for watermarking schemes: Stirmark benchmark," *Proc. SPIE*, vol. 4314, pp. 575–584, Aug. 2001.

[11] Q. Li et al., "Concealed attack for robust watermarking based on generative model and perceptual loss," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 32, no. 8, pp. 5695–5706, Aug. 2022.

[12] T. Xiang, H. Liu, S. Guo, and T. Zhang, "PEEL: A provable removal attack on deep hiding," 2021, *arXiv:2106.02779*.

[13] X. Jia, X. Wei, X. Cao, and X. Han, "Adv-watermark: A novel watermark perturbation for adversarial examples," in *Proc. 28th ACM Int. Conf. Multimedia*, Oct. 2020, pp. 1579–1587.

[14] C. Xiao, J.-Y. Zhu, B. Li, W. He, M. Liu, and D. Song, "Spatially transformed adversarial examples," 2018, *arXiv:1801.02612*.

[15] L. Engstrom, B. Tran, D. Tsipras, L. Schmidt, and A. Madry, "A rotation and a translation suffice: Fooling CNNs with simple transformations," in *Proc. ICLR*, 2018, pp. 1–21.

[16] C. Li, Y. Yang, J. Lin, and R. Zhan, "An improved adversarial example generating method with optimized spatial transfrom," in *Proc. IEEE Int. Conf. Prog. Informat. Comput. (PIC)*, Dec. 2020, pp. 173–178.

[17] L. Zhang and X. Wang, "ADVFilter: Adversarial example generated by perturbing optical path," in *Proc. Asian Conf. Comput. Vis.*, 2022, pp. 29–40.

[18] R. Bauml, J. J. Eggers, R. Tzschoppe, and J. Huber, "Channel model for watermarks subject to desynchronization attacks," *Proc. SPIE*, vol. 4675, pp. 281–292, Apr. 2002.

[19] J. Fridrich and J. Kodovsky, "Rich models for steganalysis of digital images," *IEEE Trans. Inf. Forensics Security*, vol. 7, no. 3, pp. 868–882, Jun. 2012.

[20] J. Kodovsky, J. Fridrich, and V. Holub, "Ensemble classifiers for steganalysis of digital media," *IEEE Trans. Inf. Forensics Security*, vol. 7, no. 2, pp. 432–444, Apr. 2012.

[21] Y. Qian, J. Dong, W. Wang, and T. Tan, "Deep learning for steganalysis via convolutional neural networks," *Proc. SPIE*, vol. 9409, pp. 171–180, Mar. 2015.

[22] J. Ye, J. Ni, and Y. Yi, "Deep learning hierarchical representations for image steganalysis," *IEEE Trans. Inf. Forensics Security*, vol. 12, no. 11, pp. 2545–2557, Nov. 2017.

[23] W. You, H. Zhang, and X. Zhao, "A Siamese CNN for image steganalysis," *IEEE Trans. Inf. Forensics Security*, vol. 16, pp. 291–306, 2021.

[24] X. Zhang, X. Zhang, and G. Feng, "Image steganalysis network based on dual-attention mechanism," *IEEE Signal Process. Lett.*, vol. 30, pp. 1287–1291, 2023.

[25] H. Li, X. Luo, and Y. Zhang, "Improving CoatNet for spatial and JPEG domain steganalysis," in *Proc. IEEE Int. Conf. Multimedia Expo (ICME)*, Jul. 2023, pp. 1241–1246.

[26] J. Jia, M. Luo, S. Ma, L. Wang, and Y. Liu, "Consensus-clustering-based automatic distribution matching for cross-domain image steganalysis," *IEEE Trans. Knowl. Data Eng.*, vol. 35, no. 6, pp. 5665–5679, Jun. 2023.

[27] F. L. Bookstein, "Principal warps: Thin-plate splines and the decomposition of deformations," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 11, no. 6, pp. 567–585, Jun. 1989.

[28] X. Han, Z. Wu, Z. Wu, R. Yu, and L. S. Davis, "VITON: An image-based virtual try-on network," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 7543–7552.

[29] C. Chen, D. Freedman, and C. H. Lampert, "Enforcing topological constraints in random field image segmentation," in *Proc. CVPR*, 2011, pp. 2089–2096.

[30] K. Xu et al., "Adversarial T-shirt! Evading person detectors in a physical world," 2019, *arXiv:1910.11099*.

[31] B. Shi, X. Wang, P. Lyu, C. Yao, and X. Bai, "Robust scene text recognition with automatic rectification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 4168–4176.

[32] V. Holub and J. Fridrich, "Digital image steganography using universal distortion," in *Proc. 1st ACM Workshop Inf. Hiding Multimedia Secur.*, Jun. 2013, pp. 59–68.

[33] B. Li, M. Wang, J. Huang, and X. Li, "A new cost function for spatial image steganography," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Oct. 2014, pp. 4206–4210.

[34] Y. Zhang, X. Luo, Y. Guo, C. Qin, and F. Liu, "Multiple robustness enhancements for image adaptive steganography in lossy channels," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 30, no. 8, pp. 2750–2764, Aug. 2020.

[35] Y. Lan, F. Shang, J. Yang, X. Kang, and E. Li, "Robust image steganography: hiding messages in frequency coefficients," in *Proc. AAAI Conf. Artif. Intell.*, vol. 37, 2023, pp. 14955–14963.

[36] T.-Y. Lin et al., "Microsoft COCO: Common objects in context," in *Computer Vision—ECCV*. Zürich, Switzerland: Springer, Sep. 2014, pp. 740–755.

[37] J. Jing, X. Deng, M. Xu, J. Wang, and Z. Guan, "HiNet: Deep image hiding by invertible network," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 4733–4742.

[38] Z. Guan et al., "DeepMIH: Deep invertible network for multiple image hiding," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 1, pp. 372–390, Jan. 2023.

[39] Q. Huynh-Thu and M. Ghanbari, "Scope of validity of PSNR in image/video quality assessment," *Electron. Lett.*, vol. 44, no. 13, pp. 800–801, 2008.

[40] W. Zhou, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: From error visibility to structural similarity," *IEEE Trans. Image Process.*, vol. 13, no. 4, pp. 600–612, Apr. 2024.

[41] C. Zhang, P. Benz, A. Karjauv, G. Sun, and I. S. Kweon, "UDH: Universal deep hiding for steganography, watermarking, and light field messaging," in *Proc. NIPS*, 2020, pp. 10223–10234.

[42] Z. Zheng, Y. Hu, Y. Bin, X. Xu, Y. Yang, and H. T. Shen, "Composition-aware image steganography through adversarial self-generated supervision," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 34, no. 11, pp. 9451–9465, Nov. 2023.

[43] S.-P. Lu, R. Wang, T. Zhong, and P. L. Rosin, "Large-capacity image steganography based on invertible neural networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 10816–10825.

[44] S. Voloshynovskiy, S. Pereira, V. Iquise, and T. Pun, "Attack modelling: Towards a second generation watermarking benchmark," *Signal Process.*, vol. 81, no. 6, pp. 1177–1214, Jun. 2001.

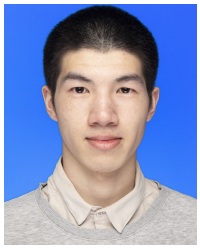
[45] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2009, pp. 248–255.
[46] X. Qi et al., "MLRSNet: A multi-label high spatial resolution remote sensing dataset for semantic scene understanding," *ISPRS J. Photogramm. Remote Sens.*, vol. 169, pp. 337–350, Nov. 2020.
[47] R. Timofte et al., "NTIRE 2017 challenge on single image super-resolution: Methods and results," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops*, Jul. 2017, pp. 114–125.
[48] A. Torralba, R. Fergus, and W. T. Freeman, "80 million tiny images: A large data set for nonparametric object and scene recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 30, no. 11, pp. 1958–1970, Nov. 2008.
[49] P. Moulin, A. Briassouli, and H. Malvar, "Detection-theoretic analysis of desynchronization attacks in watermarking," in *Proc. 14th Int. Conf. Digit. Signal Process.*, 2002, pp. 77–84.
[50] T. Zong, Y. Xiang, I. Natgunanathan, S. Guo, W. Zhou, and G. Beliakov, "Robust histogram shape-based method for image watermarking," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 25, no. 5, pp. 717–729, May 2015.
[51] M. Tancik, B. Mildenhall, and R. Ng, "StegaStamp: Invisible hyperlinks in physical photographs," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 2117–2126.
[52] J. Jia et al., "RIHOOP: Robust invisible hyperlinks in offline and online photographs," *IEEE Trans. Cybern.*, vol. 52, no. 7, pp. 7094–7106, Jul. 2022.

Lin He received the B.S. degree from Guangxi University of Science and Technology, Liuzhou, China, in 2022. She is currently pursuing the M.S. degree with Jinan University, Guangzhou, China.

Her research interests include multimedia security and steganography attack.

Bingwen Feng received the B.E. and Ph.D. degrees from Sun Yat-sen University, Guangzhou, China, in 2008 and 2014, respectively.

He is currently an Associate Professor with the College of Cyber Security, Jinan University, Guangzhou. His research interests include multimedia security, AI security, and privacy protection.



Zecheng Peng received the B.S. degree from the Jinling Institute of Technology, Nanjing, China, in 2021. He is currently pursuing the M.S. degree with Jinan University, Guangzhou, China.

His research interests include multimedia security and multimedia forensics.

Bing Chen (Member, IEEE) received the Ph.D. degree in computer science and technology from Sun Yat-sen University, Guangzhou, China, in 2020.

He is currently a Lecturer with the School of Cyber Security, Guangdong Polytechnic Normal University, Guangzhou. His research interests include multimedia security, data hiding, and secret sharing.

Zhihua Xia (Member, IEEE) received the Ph.D. degree in computer science and technology from Hunan University, China, in 2011.

He was a Visiting Scholar with New Jersey Institute of Technology, USA, in 2015, and a Visiting Professor with Sungkyunkwan University, South Korea, in 2016. He was a Lecturer, an Associate Professor, and a Professor with the College of Computer and Software, Nanjing University of Information Science and Technology. He is currently a Professor with the College of Cyber Security, Jinan University, China. His research interests include AI security, cloud computing security, and digital forensic. He serves as the Managing Editor for IJAACS.

Wei Lu (Member, IEEE) received the B.S. degree in automation from Northeast University, China, in 2002, and the M.S. and Ph.D. degrees in computer science from Shanghai Jiao Tong University, China, in 2005 and 2007, respectively.

From 2006 to 2007, he was a Research Assistant with The Hong Kong Polytechnic University. He is currently a Professor with the School of Computer Science and Engineering, Sun Yat-sen University, Guangzhou, China. His research interests include multimedia forensics and security, data hiding and watermarking, and privacy protection. He is an Associate Editor of *Signal Processing* and the *Journal of Visual Communication and Image Representation*.