Composition-Aware Image Steganography Through Adversarial Self-Generated Supervision

Ziqiang Zheng, Yuanmeng Hu, Yi Bin, Xing Xu, *Member, IEEE*, Yang Yang, *Senior Member, IEEE*, and Heng Tao Shen, *Fellow, IEEE*

***Abstract*— Steganography is an important and prevailing information hiding tool to perform secret message transmission in an open environment. Existing steganography methods can mainly fall into two categories: predefined rule-based and data-driven methods. The former is susceptible to the statistical attack, while the latter adopts the deep convolution neural networks to promote security. However, deep learning-based methods suffer from perceptible artificial artifacts or deep steganalysis. In this article, we introduce a novel composition-aware image steganography (CAIS) to guarantee both visual security and resistance to deep steganalysis through the self-generated supervision. The key innovation is an adversarial composition estimation module, which has integrated the rule-based composition method and generative adversarial network to help synthesize steganographic images with more naturalness. We first perform a rule-based image blending method to obtain infinite synthetically data–label pairs. Then, we utilize an adversarial composition estimation branch to recognize the message feature pattern from the composite image based on these self-generated data–label pairs. Through the adversarial training, we force the steganography function to synthesize steganographic images, which can fool the composition estimation network. Thus, the proposed CAIS can achieve better information hiding and higher security to resist deep steganalysis. Furthermore, an effective global-and-part checking is designed to alleviate visual artifacts caused by hiding secret information. We conduct a comprehensive analysis of CAIS from various aspects (e.g., security and robustness) to verify the superior performance of the proposed method. Comprehensive experimental results on three large-scale widely used datasets have demonstrated the superior performance of our CAIS compared with several state-of-the-art approaches.**

***Index Terms*— Generative adversarial network (GAN), image steganography, self-generated supervision.**

Manuscript received 18 February 2021; revised 28 August 2021 and 26 January 2022; accepted 24 February 2022. Date of publication 9 June 2022; date of current version 30 October 2023. This work was supported in part by the National Natural Science Foundation of China under Grant U20B2063 and in part by the Dongguan Songshan Lake Introduction Program of Leading Innovative and Entrepreneurial Talents. *(Corresponding author: Yang Yang.)*

Ziqiang Zheng, Yi Bin, Xing Xu, and Yang Yang are with the Center for Future Multimedia and the School of Computer Science and Engineering, University of Electronic Science and Technology of China, Chengdu 610041, China (e-mail: dlyyang@gmail.com).

Yuanmeng Hu is with the FFM Center, Pusan National University, Busan 46241, South Korea.

Heng Tao Shen is with the Center for Future Multimedia and the School of Computer Science and Engineering, University of Electronic Science and Technology of China, Chengdu 610041, China, and also with the Peng Cheng Laboratory, Shenzhen 518066, China.

Color versions of one or more figures in this article are available at https://doi.org/10.1109/TNNLS.2022.3175627.

Digital Object Identifier 10.1109/TNNLS.2022.3175627

I. Introduction

In THE past decades, considerable attention has been paid to information security since the fast development of the online community and cloud computing has led to information leakage problems [1], [2]. To perform a secure transmission of private information in an open community, information hiding [3] and cryptography [4] algorithms have been proposed. Cryptography targets encrypting the private message to meaningless or irrelevant output to prevent private information leakage. Sometimes, the yielded encrypted outputs usually result in the attention of the attackers. The security of cryptography mainly relies on high computation and time costs to decode private secret information. Different from cryptography algorithms, information hiding methods aim to hide the secret information to a cover, and the composite outputs are required not to be recognized as much as possible [3], [5]. In this way, the approach named steganography [3], [6], [7] performs private message transmission in the open community while not being noticed.

Steganography is regarded as the art and science of invisible communication, which is accomplished by hiding messages into a cover to secure the existence of messages. Previous steganography algorithms mainly focus on hiding text information and binary data to one cover image [8]–[12], which has a strong capacity to hide the secret. The steganography techniques can mainly be divided into two kinds: those in the image domain and those in the transform domain. The most common, simple approach to hide the message in the image domain, is the least significant bit (LSB) [13], [14] insertion by placing the secret message to the LSBs. Hussain *et al.* [15] conducted steganography in the spatial domain and incorporated image processing methods to determine which part of a cover to place the private information. For the latter, researchers [16], [17] proposed to hide the message in the discrete cosine transform domain. Despite those impressive decoding results achieved by these methods, they were relatively vulnerable to statistical attack and brute-force attack once their steganographic methods had been known by the attacker.

Image steganography that aims to hide one secret image to another irrelevant cover image is much more challenging due to the relatively high capacity compared with hiding the text messages. A simple illustration of image steganography is shown in Fig. 1. Alice wants to send her portrait to Bob by uploading it to an open environment. She places her portrait

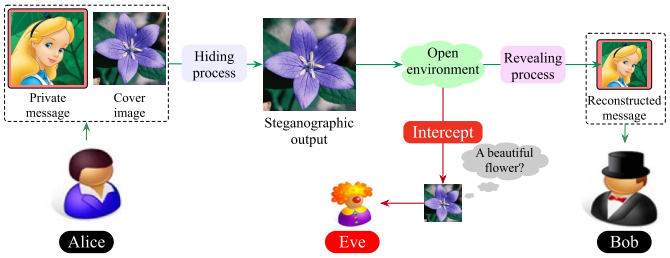

Fig. 1. Alice intends to send her portrait to Bob by transmitting it in an open community, while she does not want anyone other than Bob access to her photograph. Alice randomly selects a cover image and hides her portrait image into the cover to generate a steganographic image, which is visually identical to the cover image. Even though this steganographic image is intercepted by an attacker like Eve, no personal information of Alice will be recognized.

image ("message image") to the same size beautiful flower image ("cover image") and yields a steganographic output that can be decoded by Bob. One thing worth mentioning is that the steganographic image not only hides the information by changing the original cover but also covers up the fact that there is a secret message in the steganographic image due to its friendly visual effect compared with image encryption methods. Recently, convolutional neural networks (CNNs) have revolutionized the computer vision community because of their extraordinary performance on various tasks [18]–[20]. A lot of attempts [9], [21]–[24] adopted two reverse networks to hide and reveal the message image. Baluja [21], [25] and Rahim *et al.* [23] introduced the autoencoder architecture to achieve message image hiding and revealing through an end-to-end manner, which has shown remarkable steganography performance. Considering the homogeneous property between image synthesis and image steganography, the generative adversarial network (GAN) [26] is also applied in [6], [7], [9], [22], [24], and [27]–[31] to achieve information hiding through conditional image generation.

As above discussed, unlike text messages, message images contain a larger amount of information. It is essential to find an appropriate way to place the message image on the plain cover image. Inspired by image processing methods such as watermarking [32]–[35] and image blending [36], which provide a feasible manner to combine two images, we propose to incorporate the image fusion techniques [32]–[34], [36] into the deep generative adversarial model [26]. The proposed adversarial composition estimation module can promote the ability to generate a more natural steganographic image with higher robustness and security. In detail, we first yield infinite composite images (self-generated data) through the rule-based image blending method by introducing a random variable α (self-generated label) from a uniform distribution to control the message composition in the composite images. By formulating the self-generated data–label pairs shown in Fig. 2, we design one auxiliary estimation task to recognize the message composition from the composite images to conduct self-supervision. On the one hand, by approximately exploring all the possible combinations to fuse the cover and message images through saturation sampling, the auxiliary estimation branch can learn how to distinguish the feature patterns from the message image under self-generated supervision. On the other hand, by minimizing the estimated value of the synthesized steganographic image generated by the steganography function, the adversarial training can drive the generative model to synthesize steganographic images without recognizable message feature patterns.

The overall flowchart of the proposed composition-aware image steganography (CAIS) is shown in Fig. 2. Two reverse functions are responsible to perform steganography and reconstruction, separately. We adopt alpha blending for the adversarial composition estimation module for generating composite images. An effective global-and-part checking is also combined to alleviate artificial artifacts caused by hiding the message image. To enhance the reconstruction performance, both pixelwise and perceptual losses are included. To prove the effectiveness of our CAIS, we have constructed extensive experiments on different large-scale image datasets. A comprehensive analysis of the security, robustness, and capacity of the proposed method is provided. Besides, we have simulated the noise, compression (JPEG compression), and image distortions (cropping, flipping, and blurring) attacks to demonstrate the robustness of CAIS. To sum up, the major contributions of this work are given as follows.

1) We propose a novel CAIS method based on adversarial training, which integrates the rule-based image fusion method and the deep generative model through an auxiliary self-generated task. The self-generated supervision can result in a stronger ability of the discriminator to recognize the message pattern from the synthesized steganographic image. Through adversarial training, the steganography function could generate more natural steganographic images with high robustness and security.
2) An effective global-and-part checking is developed to alleviate the artificial artifacts caused by hiding secret message images. Both the pixelwise and perceptual losses are constructed to boost the steganography and reconstruction performance.
3) Comprehensive analysis regarding the issues of security and robustness has been conducted. The experimental results have demonstrated the superior security and robustness of our method.

The rest part is organized as follows. Section II briefly introduces the related work and Section III elaborates the proposed approach. Section IV presents the extensive experimental results on various datasets, followed by the conclusion in Section V. The codes and pretrained models of our CAIS are available at https://github.com/zhengziqiang/CAIS.

II. Related Work

A. Information Hiding

Traditional information hiding methods can mainly fall into two categories: watermarking [34] and steganography [5], [37]. Watermarking aims to create the translucent message image on the cover image to provide authenticity. Watermarking can perform robust fingerprint generation through both imperceptible and visible ways. Steganography aims to conceal private messages to make secret information

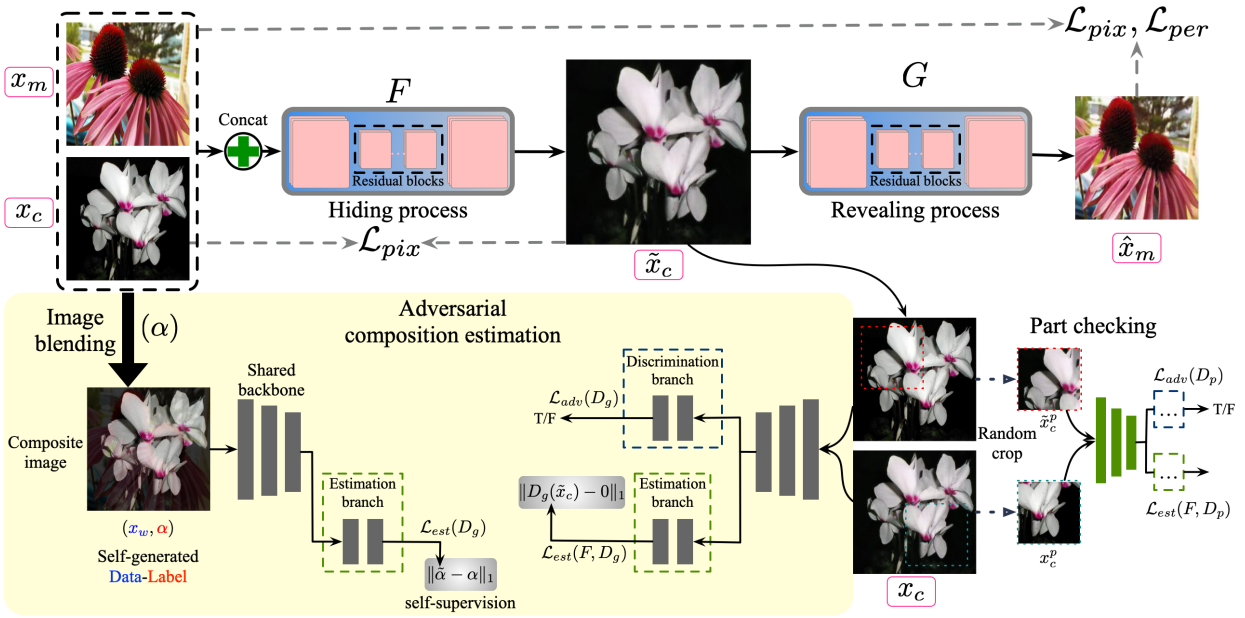

Fig. 2. Four components of the full system. Top-left corner: hiding message image x_m in cover image x_c through steganography function F and synthesizing steganographic output $\tilde{x}_c$. Bottom-left corner: generating composite image x_w by Alpha blending and using self-generated data–label: (x_w, α) to optimize the estimation of D_g. Bottom-right corner: randomly cropping part regions and performing part checking through part discriminator D_p. Top-right corner: uncovering the steganographic image with the revealing network G to a reconstructed image $\hat{x}_m$.

unrecognizable. The ultimate objectives of steganography are undetectability (security), robustness, and capacity. Previous steganography methods attempted to hide the text information to cover image based on different rules [10], [11], [16], [17], [38], [39]. By converting the text message to specific encoding formats, works [13], [14], [16] combined the secret message and cover in spatial [13], [14] and frequency [16] domain. At the revealing stage, the decoded message recovered by the rule-based steganography methods is usually lossless. However, the rule-based methods suffer from high risks to be detected and recovered by the attacker through the statistical attack or brute-force attack. Recently, plenty of work [6], [7], [21], [25], [28], [29], [40], [41] adopted deep learning to perform image steganography and got remarkable performance. Qian *et al.* [40] first introduced a customized CNN to learn the feature representation for image steganography automatically. The representative work of Baluja [21], [25] adopted two reverse neural networks to hide the message image into a plain image. Rahim *et al.* [23] extended this idea [21] and adopted autoencoder architecture to perform end-to-end image steganography. To boost the steganography performance, the multiscale feature fusion strategy is conducted in [23]. UDH [42] proposed a general framework for image steganography and watermarking. A comprehensive analysis of robustness was also conducted in UDH paper [42]. In this article, we target to integrate the rule-based image composition method and DCNN to synthesize steganographic images with more naturalness and security.

B. GAN-Based Image Steganography

Since developed, the GANs [26] had achieved the state-of-the-art performance on various vision tasks, such as image synthesis [43], domain translation [18], image encryption [27], and image steganography [37], [44]. The earliest application of GAN-based image steganography was introduced in [45], which first adopted the deep convolutional generative adversarial network (DCGAN) [46] to generate image-like containers. Chu *et al.* [24] further analyzed the potential of unpaired image-to-image translation methods to achieve symmetric image steganography. The cycle-consistency constraint [18] was introduced to ensure the reconstruction of the message image among two reverse functions. Furthermore, Yang *et al.* [22] explored using an embedding simulator to place a smaller message image to the cover image. Zhang *et al.* [47] targeted to boost the steganography performance using the U and V channels and concealed a grayscale message image. Chang [7] proposed to combine GAN to predict bit planes that had been applied to carry the message information to boost image steganography performance. Differently, our method is the first one that incorporates the previous image blending method and the deep generative model through an adversarial composition estimation module, which leads to a better steganography performance.

C. Self-Generated Supervision

Self-supervised learning is designed to obtain effective feature representations based on unlabeled data [48]–[50]. The key innovation of self-supervised learning is to design self-generated supervision and obtain learned feature representations through various pretext tasks [49], [51]. The self-supervision has been widely adopted in various vision tasks, such as domain adaptation [52], image classification [51], and semantic segmentation [49], [53], to boost vision performance and reduce the labor to collect the data annotations.

Inspired by this annotation-free training scheme, we introduce an adversarial composition estimation module to boost the image steganography performance. The proposed module first randomly samples the self-generated labels from a uniform distribution and performs an image blending method to synthesize infinite desired composite images. The composite images and the self-generated labels formulate the self-generated data–label pairs. Through self-generated supervision, the ability to distinguish the private message pattern from the steganographic images can be promoted heavily so that the proposed method could generate steganographic images with higher security.

III. Proposed Method

A. Preliminary

Image steganography is to generate a steganographic image that slightly alters the appearance of the cover image and obtains a reconstructed message image that is as similar as possible to the original message image. To estimate the composition of the message in the steganographic image $\tilde{x}_c$, we propose adversarial composition estimation, which integrates the rule-based image blending method and the adversarial training. This module conceptually consists of two operations.

1) The composition estimation task is optimized based on self-generated data–label pairs.
2) Minimize the estimated message composition from $\tilde{x}_c$ to formulate adversarial training.

Through the supervised manner, we can teach the adversarial composition estimation module to distinguish the message composition from the composite images. Besides, adversarial training can also promote the ability to synthesize steganographic images with more naturalness. The whole framework of CAIS is shown in Fig. 2, which contains two stages: hiding process and reverse revealing process. In the hiding stage, the cover image and the message image are fused by the steganography function F to generate the steganographic output. The adversarial composition estimation module is conducted to boost the steganography performance. To alleviate the artificial artifacts caused by hiding the message image, the dual-path discriminators are designed for global-and-part checking. At the revealing stage, we adopt a reverse reconstruction function G to obtain the reconstructed message. Both the pixelwise and feature-level losses are adopted between the message image and the reconstructed message.

B. Adversarial Composition Estimation

In the hiding process, the steganography function F targets to hide a private message image x_m into a cover image x_c to generate a steganographic image $\tilde{x}_c$, which looks similar to the cover image. This procedure can be described as follows:

$$\tilde{x}_c = F(\mathcal{C}(x_c, x_m)) \tag{1}$$

where $\mathcal{C}$ indicates the concatenate operation. Similarly, we can also obtain one composite image x_w using the image blending method [36] and adopt alpha blending as

$$x_w = \alpha x_m + (1-\alpha)x_c \tag{2}$$

where α is a hyperparameter from a uniform distribution that controls the composition ratio of the message image in x_w. In theory, we can approximately explore all the possible combinations of x_c and x_m and obtain infinite self-generated data–label pairs (x_w, α) for D_g (global discriminator) and D_p (part discriminator) shown in Fig. 2. D_g and D_p share the same network architecture. For simplicity, we only illustrate the adversarial composition estimation procedure based on D_g, and the same procedure is also conducted for D_p. D_g has two branches: one primary adversarial branch for steganography discrimination and an auxiliary estimation branch for supervised regression. Given the data–label pair: (x_w, α), the estimation branch targets to reproject the composite image x_w to the label space and yield $\tilde{\alpha}$. Then, we compute the distance between α and $\tilde{\alpha}$

$$\mathcal{L}_{\text{est}}(D_g) = \|\tilde{\alpha} - \alpha\|_1, \quad \text{with } \tilde{\alpha} = D_g(x_w) \tag{3}$$

we compute the one-norm distance between the output and the corresponding sampled label. Through this self-generated supervision, we can teach D_g how to distinguish the message patterns from the generated composite images. Besides, to force the generator to be composition-aware, we formulate an additional adversarial composition estimation to the synthesized steganographic image $\tilde{x}_c$

$$\mathcal{L}_{\text{est}}(F, D_g) = \|D_g(\tilde{x}_c) - 0\|_1 \tag{4}$$

by minimizing $\mathcal{L}_{\text{est}}(F, D_g)$ to zero, and we constitute the adversarial training between D_g and F. As $\alpha \to 0$, the synthesized steganographic image $\tilde{x}_c$ can be very close to the cover image visually, which indicates a better steganography performance. Through the adversarial training, F learns to synthesize the steganographic output without discriminative message patterns to fool D_g. Besides, the adversarial composition estimation could be regarded as a steganography detection function with more stronger constraint than the binary classification. With the parallel training, F could synthesize steganographic images with more naturalness and higher security.

C. Global-and-Part Checking

To alleviate the artificial artifacts caused by embedding the message image, we adopt a global-and-part checking process, as shown in Fig. 2. Iizuka *et al.* [54] first proposed a global–local adversarial training to effectively promote the global and local consistency for image completion. We introduce the part checking procedure to image steganography to synthesize steganographic images with more naturalness. To be noted, different from [54], in which the cropped part is from the ground truth mask for completion, our part discriminator is fed with a random part cropped from the corresponding images. The global discriminator D_g is responsible for the global consistency, while the part checking implemented by the part discriminator D_p can enhance local harmony. For the global checking, the adversarial loss $\mathcal{L}_{\text{adv}}(F, D_g)$ is computed as follows:

$$\begin{aligned}\mathcal{L}_{\text{adv}}(F, D_g) = {} & \mathbb{E}_{x_c \sim P_{\text{data}}(x_c)}\big[\log D_g(x_c)\big] \\ & + \mathbb{E}_{\tilde{x}_c \sim P_{\text{data}}(\tilde{x}_c)}\big[\log(1 - D_g(\tilde{x}_c))\big]\end{aligned} \tag{5}$$

and through the adversarial training, we can reduce the distance between the "real" image distribution $P_{\text{data}}(x_c)$ and the "fake" steganographic sample distribution $P_{\text{data}}(\tilde{x}_c)$.The adversarial composition estimation loss $\mathcal{L}_{\text{est}}(F, D_g)$ is defined as (4).

As for the part checking, we randomly crop regions: $\tilde{x}_c^p$, x_c^p, and x_w^p from the steganographic image $\tilde{x}_c$, the cover image x_c, and the composite image x_w, respectively. This adversarial loss $\mathcal{L}_{\text{adv}}(F, D_p)$ for part checking can be described as follows:

$$\mathcal{L}_{\text{adv}}(F, D_p) = \mathbb{E}_{x_c^p \sim P_{\text{data}}(x_c^p)}\left[\log D_p(x_c^p)\right] + \mathbb{E}_{\tilde{x}_c^p \sim P_{\text{data}}(\tilde{x}_c^p)}\left[\log(1 - D_g(\tilde{x}_c^p))\right] \quad (6)$$

and the adversarial composition estimation loss $\mathcal{L}_{\text{est}}(F, D_p)$ for part checking is computed as follows:

$$\mathcal{L}_{\text{est}}(F, D_p) = \|D_p(\tilde{x}_c^p) - 0\|_1 \quad (7)$$

and we also compute

$$\mathcal{L}_{\text{est}}(D_p) = \|D_p(x_w^p) - \alpha\|_1 \quad (8)$$

to optimize D_p. In this article, the cropped region is half the size of the original image. Further investigation about the part size can be found in Section IV-G1. The additional part checking procedure could boost local content consistency and reduce artificial artifacts.

D. Loss Functions

We also conduct a pixelwise loss between the steganographic image $\tilde{x}_c$ and the cover image x_c, which is widely adopted for different image steganography methods [6], [7], [31]. For the revealing stage, the reconstruction function G is to synthesize $\hat{x}_m$, which is the reconstruction of the message image. To boost the reconstruction performance and recover the detailed information of the message image, we perform both pixelwise and perceptual losses between the recovered message and the original message. These two loss functions are described as follows. Pixelwise loss is one necessary constraint for both the hiding and revealing processes. For the hiding process, we compute the pixelwise information residuals between x_c and $\tilde{x}_c$ to improve the undetectability of the steganographic images. For the revealing procedure, we compute the pixelwise distance between $\hat{x}_m$ and x_m to guarantee that the secret information could be recovered. The pixelwise loss $\mathcal{L}_{\text{pix}}$ is described as follows:

$$\mathcal{L}_{\text{pix}} = \|\tilde{x}_c - x_c\|_1 + \beta\|\hat{x}_m - x_m\|_1 \quad (9)$$

where β is the hyperparameter to control the two components. The first term contributes to the steganography performance, while the second term is to guarantee the reconstruction of the message information. Perceptual loss is also applied to promote the reconstruction performance of the message image. The perceptual loss [55] can provide G multiple hierarchical constraints. Different from the pixelwise loss which compares two images pixel by pixel and each pixel contributes to the loss equally, the perceptual loss measures the similarity between the reconstructed message and the original message at the feature level. Concretely, we follow Chen and Koltun's work [55] and incorporate an extra pretrained VGG-19 network on ImageNet. The trained network could effectively extract some semantic representations and introduce some prior knowledge for the message reconstruction. This loss constitutes the distance at multiple scales of feature representations, which represents both low- and high-level information of images

$$\mathcal{L}_{\text{per}}(G) = \sum_n^N \lambda_n \mathbb{E}_x[||\Phi_n(x_m) - \Phi_n(G(F(x_m)))||_1] \quad (10)$$

where Φ_n is the nth feature extraction layer of the pretrained VGG-19 network. We compute the perceptual loss at the defined $N = 5$ selective layers. The hyperparameter λ_n controls the influence of different scales. We regard different scales as equally important and set all λ_n to 1. Note that the parameters of the pretrained model are frozen during the training process.

1) Final Objective Function: The total loss of our method is a weighted sum of all the losses mentioned above

$$\begin{aligned}\mathcal{L}(F, G) &= \mathcal{L}_{\text{adv}} + \mathcal{L}_{\text{est}} + \lambda\mathcal{L}_{\text{pix}} + \gamma\,\mathcal{L}_{\text{per}} \\ \mathcal{L}_{\text{adv}} &= \mathcal{L}_{\text{adv}}(F, D_g) + \mathcal{L}_{\text{adv}}(F, D_p) \\ \mathcal{L}_{\text{est}} &= \mathcal{L}_{\text{est}}(F, D_g) + \mathcal{L}_{\text{est}}(F, D_p)\end{aligned} \quad (11)$$

where λ and γ are the hyperparameters to balance the different loss terms.

IV. Experiment

A. Experimental Setup

1) Datasets: Following the previous work [21], [47], we adopt three widely used datasets in our experiment to evaluate both steganography and reconstruction performance. The brief introduction of these datasets is depicted as follows.

1) 102Flowers [56] is a large-scale flower image dataset captured in diverse environments. This dataset contains 102 different categories of flowers and each category consists of around 40–258 images. The images of this dataset have a large diversity of scale, pose, and illumination.
2) Caricature[1] contains thousands of cartoon images of six public figures. This dataset consists of 4942 images training set, 2060 images cross validation set, and 856 images test set. We inherit the train/test split of this dataset in all our experiments.
3) ImageNet [57] is a large-scale image database, which has 1000 different categories. Each category has an average of over 500 images and all images are captured at different scenes and have a large diversity.

2) Compared Methods: We mainly perform four mostly related image steganography methods.

1) Steganography [21] is the first work that applies the deep learning model to perform image steganography task.
2) Deep-stegano [23] extended the Baluja *et al.*'s work [21] to perform the end-to-end image steganography based on autoencoder architectures.

[1]https://www.kaggle.com/ranjeetapegu/caricature-image

3) ISGAN [47] combined the adversarial training and split the U and V channels to hide the grayscale message image. To make a fair comparison, we modified the official codes to achieve the RGB image steganography.
4) UDH [42] proposed a novel general cover-agnostic framework to embed the message image, which is robust to light field messaging (LFM) and other distortions.

3) Evaluation Metrics: To quantitatively evaluate the effectiveness of image steganography methods, we compare all the methods from two aspects: steganography performance and reconstruction performance.

Steganography performance is an essential evaluation for image steganography algorithms, which guarantees the security of transporting private messages. We mainly evaluate image steganography performance by three aspects: image quality, TMQI, and human-marked accuracy considering both the automatic and human evaluations.

1) Image Quality: It is measured through three aspects. First, mean square error (MSE) and root mean square error (RMSE) are employed to measure the information residuals since we need to compute the pixelwise distance between the steganographic image and the cover image. The lower the information residuals, the harder it is to detect and recover the original message information. Second, we compute the peak signal-to-noise ratio (PSNR) to measure the quality of reconstruction of lossy compression. The structural similarity index (SSIM) is also computed to evaluate the structural similarity. The higher SSIM and PSNR scores, the better performance.
2) *TMQI:* We adopt an objective quality named TMQI as our assessment metric to evaluate the naturalness of the steganographic images. TMQI is an algorithm for tone-mapped images that combines a multiscale signal fidelity measure based on a modified SSIM and a naturalness measure based on intensity statistics of natural images [58]. Three evaluation items in TMQI marked as TMQI-S, TMQI-N, and TMQI-Q represent for fidelity score, naturalness score, and final score, respectively. All of these scores range from 0 to 1. The larger the score be, the higher quality an image possesses.
3) *Human-Marked Accuracy:* Considering that the steganographic images could be intercepted, it requires the synthesized images to fool humans. To simulate this attack, we randomly generate 2000 steganographic images for every method. Twenty volunteers were asked to recognize the artificial image from the original cover image and steganographic image. The accuracy is counted statistically for various methods. The accuracy close to 50% indicates the high naturalness of synthesized steganographic images, which cannot be recognized by humans.

Second, reconstruction performance is significant to evaluate the stability to decode the private message image from the steganographic image. Following the same setting to evaluate the steganography performance, we adopt four metrics: MSE, RMSE, PSNR, and SSIM defined in image quality as the main evaluation criteria.

TABLE I
NETWORK ARCHITECTURE OF F AND G. KS, S, AND OC INDICATE KERNEL SIZE, STRIDE SIZE, AND OUTPUT CHANNEL NUMBER, RESPECTIVELY

F (steganography)				G (reconstruction)			
Type	KS	S	OC	Type	KS	S	OC
Padding	3	1	6	Padding	3	1	3
CIR	7	1	64	CIR	7	1	64
CIR	3	2	128	CIR	7	1	64
CIR	3	2	256	CIR	7	1	64
RB	3	1	256	RB	3	1	256
RB	3	1	256	RB	3	1	256
RB	3	1	256	RB	3	1	256
RB	3	1	256	RB	3	1	256
RB	3	1	256	RB	3	1	256
RB	3	1	256	RB	3	1	256
RB	3	1	256	RB	3	1	256
RB	3	1	256	RB	3	1	256
RB	3	1	256	RB	3	1	256
DIR	3	2	128	DIR	3	2	128
DIR	3	2	64	DIR	3	2	64
Deconv	7	1	3	Deconv	7	1	6
Tanh	-	-	3	Tanh	-	-	6

4) Implementation Details: The detailed network architecture of F and G is shown in Table I. We first adopt the reflection padding for the concatenation of x_c and x_m. Then, three Conv-InstanceNorm-ReLU (CIR) blocks with kernel size 4 and stride size 2 to achieve downsampling are conducted. To enlarge the information capacity, F combines nine residual blocks (RBs) to stuck residual information. For the RBs, the kernel size is 3 and the stride size is 1. As for the reverse upsampling stage, another three Deconv-InstanceNorm-ReLU (DIR) blocks are adopted. Finally, a Tanh activation is adopted to obtain the normalized steganographic output. G has a symmetric-like architecture with F. D_g and D_p have the same network architecture shown in Table II. The shared backbone of the discrimination branch and the estimation branch contains one Conv-LeakyReLU layer and three Conv-InstanceNorm-LeakyReLU (CILR) layers. The slope for the Leaky RelU is 0.2. Then, for the discrimination branch, we obtain the logit output to perform the real/fake discrimination after one Conv layer. For the estimation branch, we apply another three downsampling layers to obtain the estimation output. We optimize D_g with the sum of $\mathcal{L}_{\text{est}}(D_g)$ and $\mathcal{L}_{\text{adv}}(D_g)$, while we optimize D_p with the sum of $\mathcal{L}_{\text{est}}(D_p)$ and $\mathcal{L}_{\text{adv}}(D_p)$. We choose the Adam optimizer [59] in our all experiments and set the initial learning rate to 0.0002. To balance the steganography and reconstruction performance, we set β to 1. For (11), we set $\lambda = 10$ and $\gamma = 0.1$ in our experiments. More detailed experimental results about the hyperparameter selection are provided in Section IV-G4.

B. Overall Comparison

In this section, we perform different image steganography methods on various image datasets. We first evaluate the effectiveness of the proposed method on the 102Flowers dataset [56]; 7000 images from the 102Flowers dataset are randomly sampled for training and the left 1189 images for evaluation. There are 7000 $\times$ 7000 different combinations

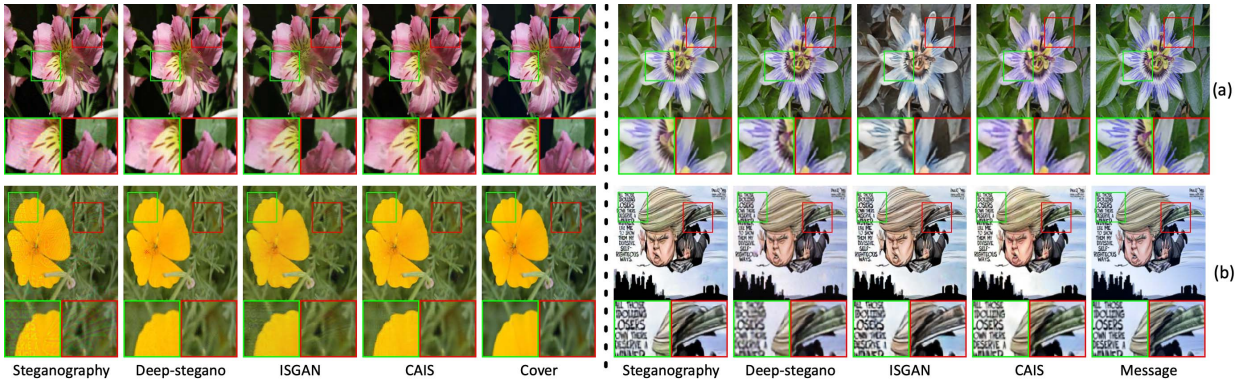

Fig. 3. Visual image steganography comparison of different methods: (a) both the cover and message images are flower images and (b) cover image is a flower image, while the message image is a caricature image. The images at the left side of black dotted show the steganography comparison, while the images at the right of black dotted show the reconstruction comparison.

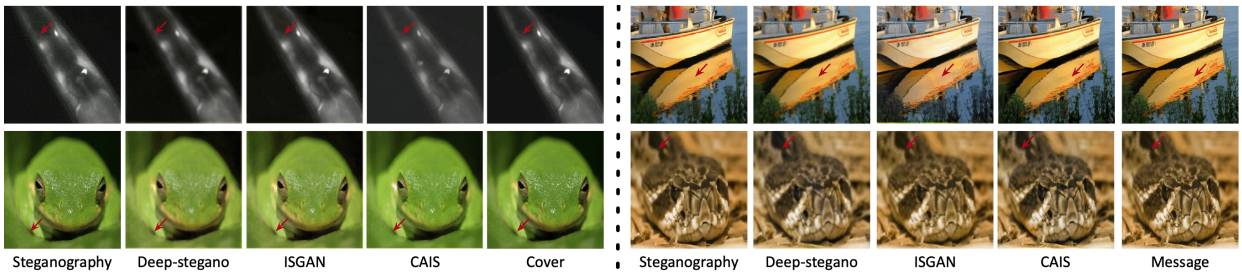

Fig. 4. Visual large-scale image steganography results of different methods. The images at the left of black dotted show the steganography comparison, while the images at the right of black dotted show the reconstruction comparison.

TABLE II

NETWORK ARCHITECTURE OF D_g AND D_p. THE GREEN PART INDICATES THE SHARED PARAMETERS OF THE ADVERSARIAL BRANCH AND THE ESTIMATION BRANCH

Estimation branch				Adversarial branch			
Type	KS	S	OC	Type	KS	S	OC
Conv+LR	4	2	64	Conv+LR	4	2	64
CILR	4	2	128	CILR	4	2	128
CILR	4	2	256	CILR	4	2	256
CILR	4	2	512	CILR	4	2	512
CILR	4	2	512	Conv	1	1	1
CILR	4	2	256				
CILR	4	1	128				
Conv	1	1	1				

to perform at the training stage. We randomly sample 2000 cover–message pairs from the testing images to measure the steganography performance. To be noted, all the methods are evaluated by using the same image pairs. The visual comparison of both steganographic and reconstructed outputs is shown in Fig. 3. As illustrated, steganography [21] and Deep-stegano [23] cannot hide the message perfectly (with obvious artificial artifacts covered by the red and green boxes). In contrast, our CAIS can synthesize natural-looking steganographic outputs with negligible textile changes compared with the raw cover image. Besides, we also consider that the cover images and the message images share different content representations (e.g., the message images are caricature images, while the cover images are flower images). Following the same setup, we perform a comparison based on the images from the Caricature dataset and 102Flowers dataset [56].

The quantitative comparison under the two settings is reported in Tables III and IV. The steganographic images generated by CAIS have a lower distance from the corresponding original cover images, which indicates less risk to be detected. To further investigate the steganography performance, we have also computed the TMQI scores. The proposed method outperforms all the other methods at all the metrics. Besides, the human-marked classification accuracy is 64.3%, while the counterpart scores of other methods are over 70%. It is extremely difficult for humans to judge whether the image carries private visual information. For the reconstruction performance comparison, our method also gains a better reconstruction of the message images (the highest PSNR and SSIM scores).

1) Large-Scale Image Steganography: To prove that CAIS can be adopted for the real-world image steganography application, we conduct the image steganography experiment on the large-scale ImageNet dataset [57]. The cover and message images are both from the same dataset. To make a fair comparison, we follow the official train/test split, and all methods are trained until convergence with the same number of iterations. Similarly, 2000 cover–message pairs from the test images are randomly chosen to perform the comparison. Both the visual and quantitative comparison are shown in Fig. 4 and Table V, respectively. The visual and quantitative comparison has demonstrated that our method could preserve more image statistics of the cover image even for the complex

TABLE III

QUANTITATIVE COMPARISON OF BETWEEN DIFFERENT METHODS ON 102FLOWERS DATASET [56]. ↑ (↓) INDICATES THAT THE LARGER (SMALLER) THE VALUE IS, THE BETTER THE PERFORMANCE

Methods	Steganography								Reconstruction			
	MSE ↓	RMSE ↓	PSNR ↑	SSIM ↑	TMQI-S↑	TMQI-N↑	TMQI-Q↑	Human↓	MSE ↓	RMSE ↓	PSNR ↑	SSIM ↑
Steganography [21]	0.0022	0.0462	26.78	0.9822	0.1997	0.5860	0.6210	94.7	0.0022	0.0467	26.66	0.9742
Deep-stegano [23]	0.0015	0.0370	29.01	0.9821	0.2016	0.5917	0.6237	78.9	0.0026	0.0493	26.37	0.9788
ISGAN [47]	0.0014	0.0395	29.16	0.9815	0.2031	0.5978	0.6243	71.3	0.0037	0.0596	25.17	0.9742
CAIS	**0.0011**	**0.0297**	**31.07**	**0.9836**	**0.2178**	**0.6287**	**0.6417**	**64.3**	**0.0014**	**0.0347**	**29.58**	**0.9861**

TABLE IV

QUANTITATIVE COMPARISON OF IMAGE STEGANOGRAPHY BETWEEN DIFFERENT METHODS ON THE CARICATURE (MESSAGE IMAGES) AND 102FLOWERS (COVER IMAGES) DATASETS

Methods	Steganography								Reconstruction			
	MSE ↓	RMSE ↓	PSNR ↑	SSIM ↑	TMQI-S↑	TMQI-N↑	TMQI-Q↑	Human↓	MSE ↓	RMSE ↓	PSNR ↑	SSIM ↑
Steganography [21]	0.0027	0.0512	25.90	0.9876	0.2034	0.5790	0.6227	98.4	0.0032	0.0565	25.00	0.9755
Deep-stegano [23]	0.0025	0.0497	27.76	0.9856	0.2087	**0.5842**	0.6276	83.1	0.0029	0.0521	26.65	0.9765
ISGAN [47]	0.0014	0.0384	29.18	0.9814	0.2093	0.5813	0.6218	76.3	0.0029	0.0542	26.14	0.9713
CAIS	**0.0010**	**0.0151**	**30.73**	**0.9845**	**0.2172**	0.5829	**0.6331**	**61.6**	**0.0023**	**0.0439**	**27.70**	**0.9853**

TABLE V

QUANTITATIVE COMPARISON OF LARGE-SCALE IMAGE STEGANOGRAPHY BETWEEN DIFFERENT METHODS

Methods	Steganography							Reconstruction			
	MSE ↓	RMSE ↓	PSNR ↑	SSIM ↑	TMQI-S↑	TMQI-N↑	TMQI-Q↑	MSE ↓	RMSE ↓	PSNR ↑	SSIM ↑
Steganography [21]	0.0036	0.0791	24.67	0.9546	0.1817	0.5782	0.6014	0.0041	0.0904	24.16	0.9572
Deep-stegano [23]	0.0032	0.0687	25.76	0.9644	0.1954	0.5835	0.6076	0.0037	0.0783	25.23	0. 9623
ISGAN [47]	0.0030	0.0576	26.05	0.9723	0.2045	0.5912	0.6094	**0.0034**	**0.0689**	25.67	0.9656
CAIS	**0.0029**	**0.0553**	**26.12**	**0.9785**	**0.2076**	**0.5932**	**0.6105**	0.0035	0.0695	**25.74**	**0.9678**

TABLE VI

QUANTITATIVE COMPARISON OF BOTH RECONSTRUCTION PERFORMANCE OF DIFFERENT SETTINGS

Method	APD ↓	SSIM ↑	PSNR ↑	LPIPS ↓
UDH (Hiding)	**2.73**	**34.21**	**0.9791**	0.0048
CAIS (Hiding)	2.81	33.82	0.9732	**0.0043**
UDH (Revealing)	**3.74**	**33.72**	**0.9781**	**0.0065**
CAIS (Revealing)	3.78	33.56	0.9765	0.0068

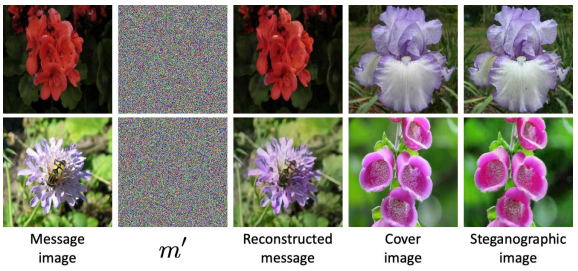

Fig. 5. Qualitative results of the proposed CAIS under the cover-agnostic setting, where m' indicates the transformed message image after H.

message image within complicated content information and diverse backgrounds. Besides, CAIS can generate natural steganographic images, which shows great similarity to the cover images. Both the qualitative and quantitative results have demonstrated the effectiveness of the proposed CAIS.

C. General Framework

1) Comparison With UDH: The recent UDH [42] is a representative general framework to perform image steganography, which has done a comprehensive analysis under different settings. In this section, we first make a fair comparison with UDH under the same setting. Then, we conduct the corresponding experiments under the same setup of UDH [42] to prove that our CAIS could be also extended to a general framework with some small modifications. We adopt the official pretrained models on the ImageNet dataset[2] from UDH [42] (the models were trained under 128 × 128 image resolution). To make a fair comparison, the image resolution is set to 128 × 128 for our CAIS to perform image steganography and message reconstruction. Following the setup of UDH [42], the overall comparison of both steganography and revealing (message reconstruction) is reported in Table VI. As illustrated, the proposed CAIS and UDH are neck-and-neck.

2) Cover Agnostic: We have performed experiments under the cover-agnostic setting proposed in UDH [42]. An additional network H is introduced to perform message transformation. We perform a Tanh activation to obtain the normalized outputs. The qualitative and quantitative results on the 102Flowers dataset [56] are reported in Fig. 5 and Table VII. For a better illustration, we map the normalized outputs (denoted as m' in Fig. 5) to $[0, 255]$ to make m' consistent with the natural images. From Table VII, we can observe that the message reconstruction performance could be

[2]The average pixel distance (APD) and LPIPS [60] are computed in UDH [42] to evaluate the steganography and reconstruction performance.

TABLE VII

QUANTITATIVE RESULTS OF OUR CAIS UNDER DIFFERENT SETTINGS. ALL THE EXPERIMENTS ARE PERFORMED ON THE 102FLOWERS DATASET [56] FOLLOWING THE SAME EXPERIMENTAL SETTING IN SECTION IV-B

Setting	Steganography							Reconstruction			
	MSE ↓	RMSE ↓	PSNR ↑	SSIM ↑	TMQI-S↑	TMQI-N↑	TMQI-Q↑	MSE ↓	RMSE ↓	PSNR ↑	SSIM ↑
Cover-agnostic	0.0014	0.0386	29.45	0.9802	0.2103	0.5983	0.6305	**0.0013**	**0.0341**	**29.79**	**0.9884**
Noisy cover	0.0012	0.0306	30.79	0.9827	-	-	-	0.0015	0.0431	29.41	0.9857
Multiple messages (3)	0.0013	0.0356	30.45	0.9811	0.2145	0.5812	0.6231	0.0015	0.0478	28.91	0.9812
CAIS	**0.0011**	**0.0297**	**31.07**	**0.9836**	**0.2178**	**0.6287**	**0.6417**	0.0014	0.0347	29.58	0.9861

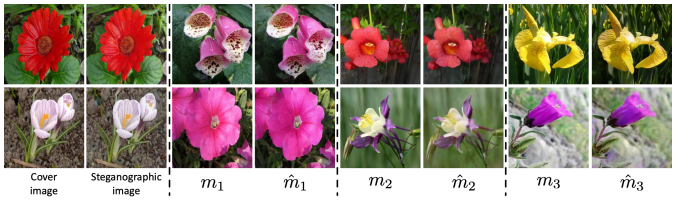

Fig. 6. Qualitative results of the proposed CAIS. We target hide three different message images into one same cover image. m_1, m_2, and m_3 indicate different message images and $\hat{m}_1$, $\hat{m}_2$, and $\hat{m}_3$ are the corresponding message image reconstructions.

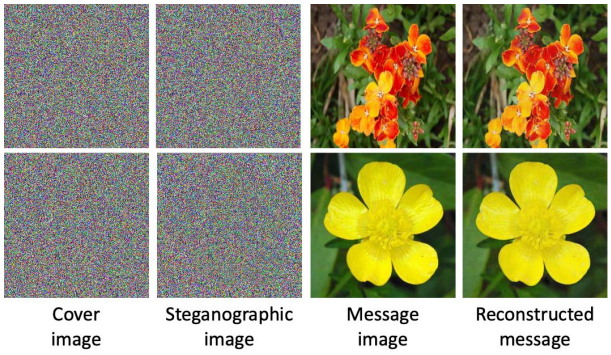

Fig. 7. Qualitative results of the proposed CAIS, where the cover image is random noise, while the message image is the natural flower image.

enhanced under the cover-agnostic setting, while there is slight image steganography performance degradation.

3) Hiding m Images in n Images: It is first proposed in UDH [42], which targets to hide multiple message images into one or several cover images by introducing different neural networks for decoding the message. Following the same setup, we perform experiments on hiding three different messages into one same cover image. The visual results of our CAIS are shown in Fig. 6. The experimental results have demonstrated that the proposed CAIS could hide more message information into one cover image with some performance degradation reported in Table VII.

4) Noisy Cover: Considering that the cover images and message images are from different distributions (e.g., the cover image is random noise and the message image is the natural image, it is meaningless to set the message image as random noise), we conduct experiments to hide the flower image into the meaningless random noise (cover image). The qualitative and quantitative results are reported in Fig. 7 and Table VII, respectively. Since the cover image is meaningless, we do not compute the TMQI scores to evaluate the steganography performance.

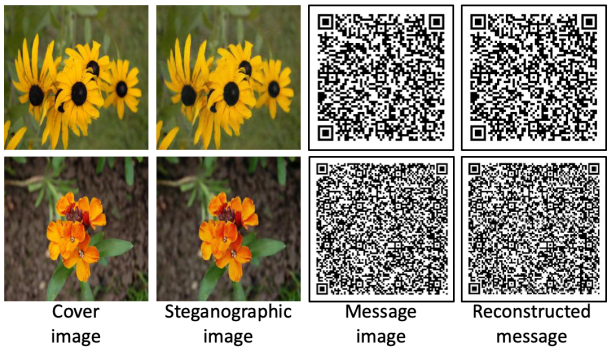

Fig. 8. Qualitative results of the proposed CAIS, where the message image is the OR code image covering meaningful information. The information can be decoded accurately from the reconstructed QR code image.

Furthermore, we have also explored choosing the OR code image as the message image to convey the message information. The qualitative image steganography and message reconstruction results are shown in Fig. 8. We can observe that the proposed CAIS can reconstruct QR code images accurately while preserving the meaningful message information after decoding.

Finally, to provide a fair and comprehensive comparison with UDH under various settings, we perform experiments on the ImageNet dataset [57]. We adopt the official codes and the pretrained models from UDH to conduct the experiments. Following the experimental setting of UDH, we report the experimental results of UDH and our CAIS in Table VIII. As reported, our CAIS and UDH are comparable under the cover-agnostic and noisy cover settings. Since our model has a large network capacity than UDH, our method is better than UDH under the multiple messages setting.

D. Security Analysis

1) Steganalysis: The steganography algorithms should have a strong ability to evade detection by steganalysis tools. We adopt a popular open-source steganalysis tool StegExpose [61] to detect the steganographic images. Following the setup in SteganoGAN [9], we randomly use 2000 cover–steganographic image pairs to examine StegExpose. Considering that StegExpose is specially designed for detecting LSB steganography, StegExpose failed to detect the steganographic images. Thus, we further perform deep steganalysis based on recent steganalysis detection approaches.

TABLE VIII

QUANTITATIVE RESULTS OF UDH OUR CAIS UNDER DIFFERENT SETTINGS. ALL THE EXPERIMENTS ARE PERFORMED ON THE IMAGENET DATASET [57]

Setting	Cover-agnostic				Noisy cover				Multiple messages (3)			
	APD ↓	SSIM ↑	PSNR ↑	LPIPS ↓	APD ↓	SSIM ↑	PSNR ↑	LPIPS ↓	APD ↓	SSIM ↑	PSNR ↑	LPIPS ↓
UDH (Hiding)	**2.87**	**33.74**	0.9678	0.0045	**3.15**	**33.12**	**0.9605**	**0.0056**	3.75	32.98	0.9579	0.0063
CAIS (Hiding)	2.93	33.69	**0.9685**	**0.0042**	3.17	33.09	0.9601	0.0061	**3.63**	**33.03**	**0.9589**	**0.0057**
UDH (Revealing)	**3.87**	**33.58**	**0.9715**	**0.0075**	4.56	**33.38**	**0.9587**	0.0078	4.78	33.06	0.9524	0.0083
CAIS (Revealing)	3.91	33.53	0.9706	0.0071	**4.55**	33.31	0.9578	**0.0075**	**4.68**	**33.25**	**0.9557**	**0.0077**

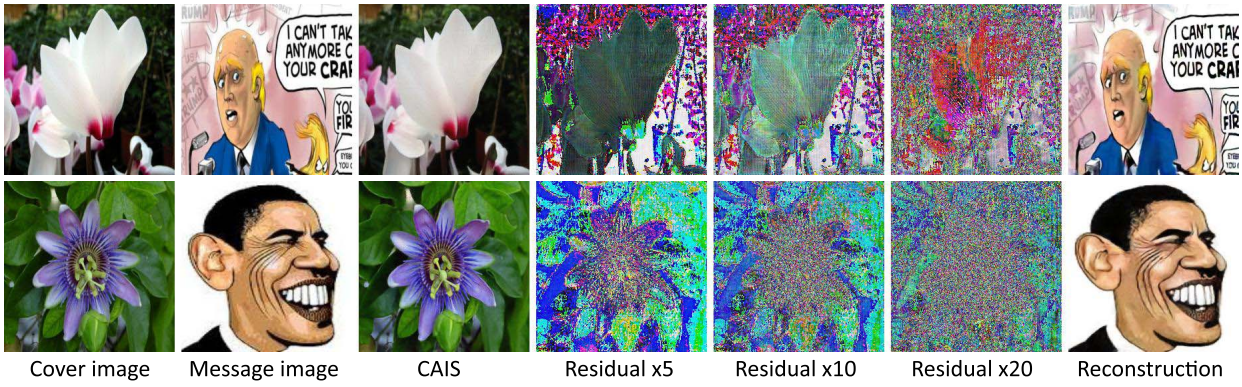

Fig. 9. First three columns: the cover image, message image, and steganographic image. Middle three columns: residual can be computed if the cover image is leaked and is subtracted from the steganographic image. Even with enough enhancement (20×), in most cases, the message image cannot be revealed. Last column: reconstructed image from a steganographic image through our proposed CAIS method.

TABLE IX

OVERALL STEGANOGRAPHY DETECTION PERFORMANCE COMPARISON OF DIFFERENT IMAGE STEGANOGRAPHY METHODS BASED ON VARIOUS BACKBONES

Method	Inception-v3	SRNet	CovNet
Steganography	98.7	99.8	100
Deep-stegano	86.5	98.4	98.3
ISGAN	74.6	92.1	95.2
UDH	61.6	84.2	91.4
CAIS	**56.1**	**75.3**	**83.5**

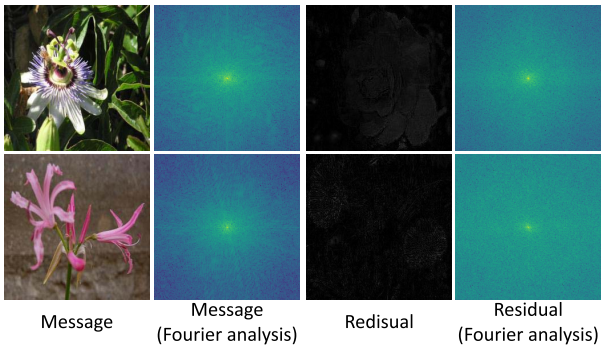

Fig. 10. Fourier analysis of the message image and the information residual between the cover image and the corresponding steganographic image.

We first adopt a general classification network (Inception-v3 [62]) to perform binary classification. To perform a fair comparison, 7000 cover–steganographic image pairs synthesized from different image steganography methods are chosen for training to obtain the corresponding steganalysis models and 2000 unseen pairs for evaluation. The classification accuracies of different image steganography methods are reported in Table IX. Our CAIS could achieve the lowest score 56.1% among all the methods, which indicates that the steganographic images generated by our CAIS have the best steganalysis performance. Besides, two specially designed deep steganalysis networks, CovNet [63] and SRNet [64], are performed to verify the security of different image steganography methods. Following the same setup of the above Inception-v3 [62] steganography detection experiment, we conduct the steganalysis experiments based on SRNet and CovNet. For SRNet, the classification accuracy comparison of 2000 random cover–steganographic image pairs is also included in Table IX. Our CAIS can also achieve the lowest 75.3% accuracy among all the steganography methods. As for the more efficient CovNet, the steganalysis results of all the other image steganography methods are above 90%, while our score is 83.5%. From the overall comparison, our CAIS has a stronger ability to evade steganography detection. The proposed adversarial composition estimation module targets regressing the message composition from the synthetic steganographic images, which provides a stronger constraint for steganography detection than the binary classification loss. With parallel adversarial training, our CAIS can resist various types of deep steganalysis.

2) Information Residuals: In most cases, we can assume that our original cover image would never be exposed to the public to guarantee security or it is difficult to obtain the original image based on the steganographic image. However, what if it is leaked for some reason? In that case, what could then be ascertained about the message image, even without

Fig. 11. Pix2pix model using brute-force attack cannot reconstruct meaningful message images correctly. For comparison, the reconstructed message images by CAIS are also illustrated.

the revealing network? In Fig. 9, we provided the difference between the cover image and the steganographic image as information residual image. We examine the residual image under the enhancement $5\times$, $10\times$, and even $20\times$, and almost no private message information is visible, which indicates that our method can resist high security when the original cover images are available. Besides, we have also performed the Fourier analysis on the message image and the information residual in Fig. 10 following the same setup in UDH [42]. As shown, there is a clear frequency discrepancy between the message image and the information residual between the cover image and the steganographic image, which has demonstrated that it is difficult for the attacker to reconstruct the message information from the information residual.

3) Utility Analysis on Brute-Force Reverse Engineer: To reveal a secret message from the steganographic images, the attacker may perform a brute-force reverse engineer. In detail, the attacker could obtain a lot of paired message images and steganographic images (e.g., 5000 pairs) by uploading different message images to synthesize corresponding steganographic images. We have optimized a paired message image reconstruction network based on Pix2pix [65] and evaluated this model using another 1000 unseen steganographic images. The visual reconstruction results are shown in Fig. 11. The average MSE loss between the reconstructions and raw message images is 0.1278 and the PSNR is 9.922, which indicates that the synthesized steganographic images from CAIS can resist this kind of brute-force attack.

E. Robustness Analysis

1) Robustness to Compression and Noise: Considering that there may be some information loss and compression during the transmission procedure, we also explored the robustness of the steganographic outputs to the compression and the noise. First, we regard the JPEG compression as the main compression during the transmission procedure and perform experiments with different levels of JPEG compression. In our settings, we choose JPEG quality with 85, 90, and 95 for JPEG compression, while 95 is the default value of JPEG compression. As shown in Fig. 12, the revealing network G can still decode the visual private message with some content loss from the compressed JPEG images. The average quantitative results of 1000 samples are provided in Table X. At the default JPEG compression setting (quality $=$ 95), the PSNR score changed from 27.70 to 20.29. Besides, we add different scales of Gaussian noise to the steganographic outputs shown in Fig. 12. Our CAIS cannot recover a clear message image from the noise steganography when the noise level is larger than $N(0, 0.015)$. The corresponding average quantitative results of 1000 samples when adding different scales of Gaussian noise to the steganographic images are also shown in Table X. Since the distance between the cover and the steganographic output is small, the noise attack does indeed harm the reconstruction performance.

TABLE X
QUANTITATIVE COMPARISON OF RECONSTRUCTION PERFORMANCE OF DIFFERENT SETTINGS

Method	MSE ↓	RMSE ↓	PSNR ↑	SSIM ↑
$N(0, 0.005)$	0.0063	0.0765	22.59	0.9522
$N(0, 0.01)$	0.0211	0.1360	17.81	0.8936
$N(0, 0.015)$	0.0378	0.1856	14.97	0.8442
$N(0, 0.02)$	0.0512	0.2192	13.39	0.8127
JPEG (95)	0.0109	0.1001	20.29	0.9415
JPEG (90)	0.0238	0.1479	16.88	0.8798
JPEG (85)	0.0360	0.1834	14.95	0.8417
Original	**0.0023**	**0.0439**	**27.70**	**0.9853**

2) Robustness to LFM and Image Distortions: To make CAIS robust to LFM, we follow the setup mentioned in [12] and [42] to apply the geometric transform to the steganographic images. The visual results are shown in Fig. 13. Following UDH [42], we compute the average bit error rate (BER). Our CAIS has achieved 4.28%, which is comparable with 4.41% achieved by UDH. We considered other types of image distortions such as blurring, flipping, and random cropping. We perform these operations on the steganographic images and try to reconstruct the message information from the modified images. All the visual results are shown in Fig. 13. We can observe that CAIS can resist to the random flipping and slight Gaussian blur. The proposed method achieves a blur reconstructed message output under the random cropping setting. We attribute this failure to the reason that the random crop operation has broken the relationship between image parts, which leads to information loss.

3) Robustness to Photoshop Transformations: We have also considered some special transformations from Photoshop: Country [66], Starlight [67], Solarize [66], and Crayon [67]. The Country operation aims to change the color of the entire image similarly. The Starlight operation is to add the starlike lights into the images. The Solarize transformation lets dark areas appear light and light areas appear dark. The Crayon operation is to add evenly distributed crayon lines to the images. The qualitative results are shown in Fig. 14. As shown, our method is robust to the Country transformation and has adequate resistance to the Starlight and Solarize transformations. In contrast, our method failed to reconstruct meaningful information under the Crayon transformation.

4) Failure Cases: Finally, we have provided the failure reconstructions of our CAIS under an extreme compression when JPEG quality is 50 in Fig. 15. The receivers cannot

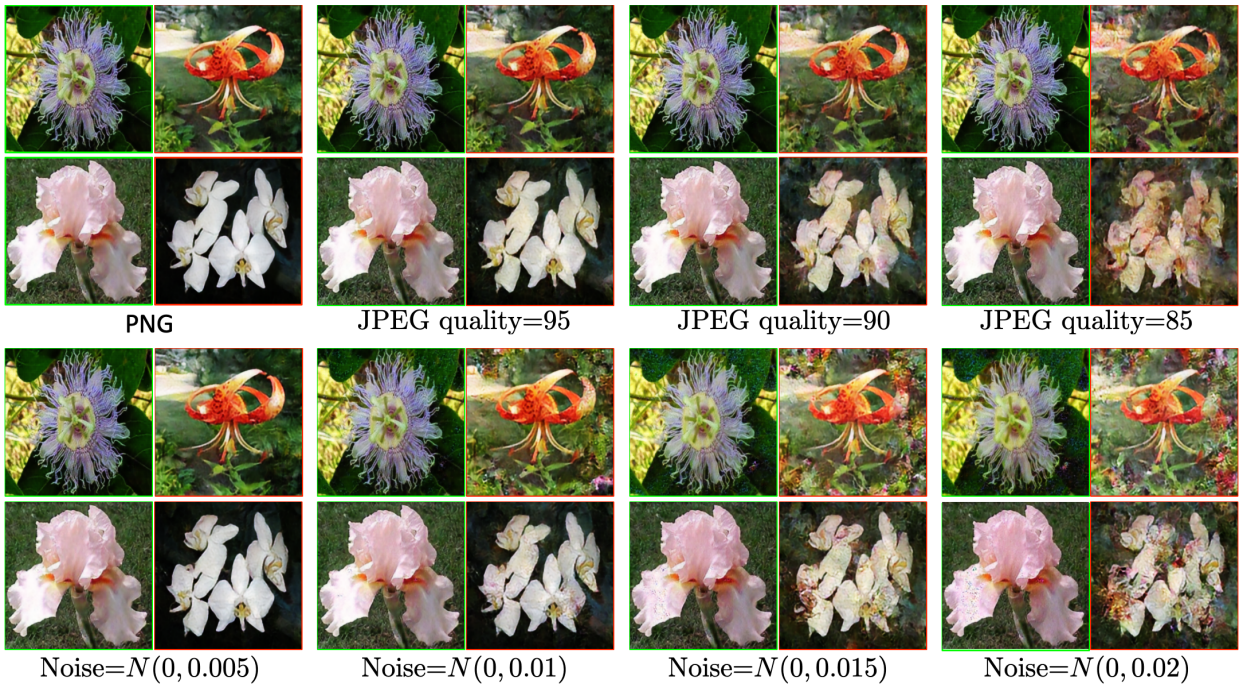

Fig. 12. We exhibit the visual reconstruction results at different JPEG compression qualities and adding different scales of Gaussian noise to the steganographic outputs. The images in green boxes are the modified steganographic images and the images in red boxes are the corresponding reconstructed message images.

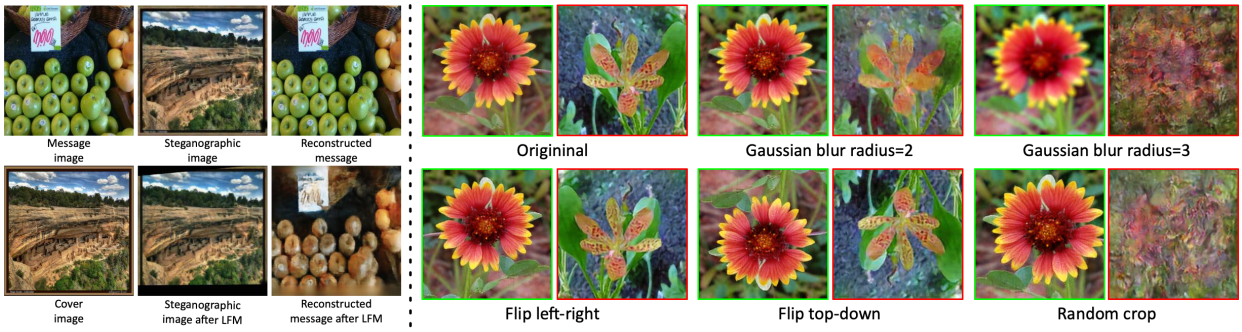

Fig. 13. Qualitative results of CAIS under LFM and image distortions. The images in green boxes are the modified steganographic images and the images in red boxes are the reconstructed message images.

Fig. 14. Qualitative results under various Photoshop transformations. The images in green boxes are the modified steganographic images and the images in red boxes are the corresponding reconstructed message images.

obtain any meaningful information from the degraded reconstructed message images.

F. Capacity Analysis

We have also provided a discussion about the model capacity of different methods. Reed–Solomon bits-per-pixel (BPP) proposed in [9] was designed to measure the average number of bits that can be reliably transmitted in an image. A higher value indicates a greater capacity of the embedded information that the algorithm can carry. SteganoGAN [9] only targeted to hide the binary code into a large cover image with much more information amount. Each pixel only hides D (the depth of the

TABLE XI
QUANTITATIVE COMPARISON AMONG DIFFERENT SETTINGS

Method	Steganography							Reconstruction			
	MSE ↓	RMSE ↓	PSNR ↑	SSIM ↑	S↑	N↑	Q↑	MSE ↓	RMSE ↓	PSNR ↑	SSIM ↑
w/o D_p	0.0014	0.0194	29.85	0.9823	0.2152	0.5815	0.6316	**0.0021**	**0.0412**	**27.91**	0.9848
x_p(96×96)	0.0011	0.0167	30.23	0.9814	0.2168	**0.5832**	**0.6341**	0.0022	0.0431	27.81	**0.9887**
x_p(128×128)	**0.0010**	**0.0151**	**30.73**	**0.9845**	**0.2172**	0.5829	0.6331	0.0023	0.0439	27.70	0.9853
x_p(160×160)	0.0016	0.0201	29.94	0.9823	0.2145	0.5821	0.6301	0.0022	0.0425	27.74	0.9876
x_p(192×192)	0.0015	0.0197	29.67	0.9819	0.2123	0.5803	0.6267	0.0024	0.0423	27.87	0.9834
w/o $\mathcal{L}_{per}$	0.0013	0.176	29.71	0.9831	0.2132	0.5784	0.6287	0.0026	0.0487	26.97	0.9804
w/o $\mathcal{L}_{est}$	0.0017	0.216	28.82	0.9812	0.2091	0.5765	0.6217	**0.0022**	**0.0421**	**27.82**	**0.9856**
CAIS	**0.0010**	**0.0151**	**30.73**	**0.9845**	**0.2172**	**0.5829**	**0.6331**	0.0023	0.0439	27.70	0.9853

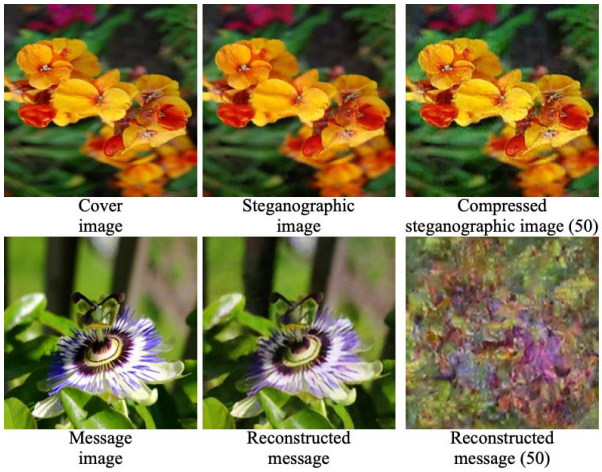

Fig. 15. Failure cases of our CAIS under the JPEG compression with JPEG quality 50.

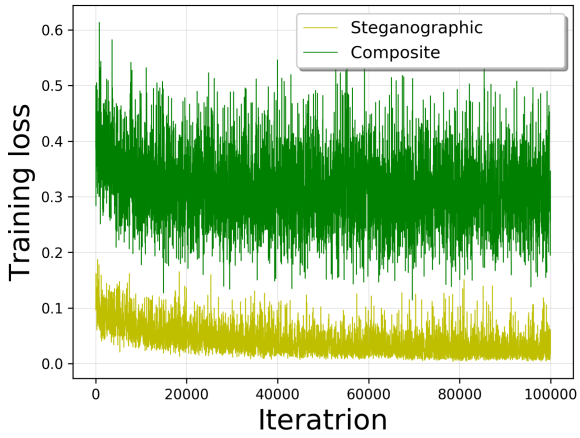

Fig. 16. Training loss curve of $\mathcal{L}_{\text{est}}(F, D_g)$ and $\mathcal{L}_{\text{est}}(D_g)$.

private message) binary information (either 0 or 1, typically lower than 0.5 BPP). Similarly, HiDDeN [12] was less than 0.002 BPP. However, for hiding the same size message image into another plain image task proposed in [21] and [25], each pixel has to hide one Uint8 value for an 8-bit digital image. Referring to UDH [42], the proposed CAIS and UDH have a message capacity of 24 BPP. According to this comparison, hiding the private message image into another plain image was extremely challenging.

G. Further Analysis

1) Effectiveness of Part Checking: To show the importance of part checking, we removed D_p and evaluated the improvement of the part checking. We reported the quantitative comparison in Table XI. As shown, the part checking based on the random cropped regions can promote the steganography performance. We have also explored the influence of the input image size of the part discriminator on steganography performance. We designed four different part sizes: 96 × 96, 128 × 128, 160 × 160, and 192 × 192 for the part checking, and the corresponding quantitative results are reported in Table XI.

2) Effectiveness of Adversarial Composition Estimation: The adversarial composition estimation procedure was introduced to promote image steganography performance. By forcing the generator to be composition-aware, our method could promote the ability to hide message information and generate steganographic images with more naturalness (higher TMQI scores) shown in Table XI. Besides, we have also checked intermediate results ($\mathcal{L}_{\text{est}}(F, D_g)$ and $\mathcal{L}_{\text{est}}(D_g)$) of the adversarial composition estimation in Fig. 16. As illustrated, $\mathcal{L}_{\text{est}}(D_g)$ (not a constant value) could teach D_g how to recognize the message information from the composite images and guarantee the adversarial training could be continuously conducted. With the adversarial training, $\mathcal{L}_{\text{est}}(F)$ was approaching 0, which indicates that the proposed CAIS could synthesize steganographic images that can fool the estimation branch of D_g.

3) Effectiveness of the Perceptual Loss: We evaluated the improvement of the perceptual loss $\mathcal{L}_{\text{per}}$. $\mathcal{L}_{\text{per}}$ provided a better reconstruction performance of the reconstructed message images in Table XI. $\mathcal{L}_{\text{per}}$ could improve the PSNR score of the reconstruction performance from 26.97 to 27.70.

4) Hyperparameter Selection: To show the influence of choosing different hyperparameters: λ and γ in our method, we conducted different experiments using different combinations of these two hyperparameters. For λ, we choose four different options: 0.1, 1.0, 10.0, and 100.0. Considering that the perceptual loss computes the distance from five different layers, the perceptual loss is larger than the pixelwise loss. To better balance the two losses, we set two different values: 0.1 and 1.0 for γ. To provide an intuitive comparison, we visualized the PSNR scores of both steganography and reconstruction performance under different settings in Fig. 17. From this figure, we observed that if λ was too large (e.g., 100), $\mathcal{L}_{\text{pix}}$ played the most important role, and this change of $\mathcal{L}_{\text{pix}}$ had not brought about obvious improvement to the PSNR score of the reconstruction performance.

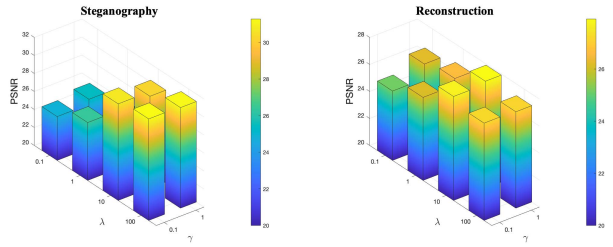

Fig. 17. Visual quantitative PSNR scores of different combinations of the hyperparameters: λ and γ. λ is the weight of the pixelwise loss and γ controls the influence of the perceptual loss.

An approximate γ (e.g., 0.1) could boost the reconstruction performance.

V. Conclusion

In this article, we proposed a novel image steganography method called CAIS to achieve impressive image steganography and message reconstruction performance. The adversarial composition estimation has boosted the synthesis of steganographic images and promoted steganalysis performance through self-generated supervision. To further reduce the visual artifacts, we combined the global-and-part checking to yield steganographic images with more naturalness. The perceptual and the pixel-level losses were introduced to achieve both better steganography and reconstruction performance. Comprehensive experiments considering the security, robustness, and capacity analysis have been performed on various datasets. The proposed CAIS can also be extended to a general framework with some small modifications. The experimental results demonstrated the superior performance of our CAIS than current state-of-the-art image steganography methods.

References

[1] Y.-Q. Zhang and X.-Y. Wang, "A symmetric image encryption algorithm based on mixed linear–nonlinear coupled map lattice," *Inf. Sci.*, vol. 273, pp. 329–351, Jul. 2014.

[2] M. Zanin and A. N. Pisarchik, "Gray code permutation algorithm for high-dimensional data encryption," *Inf. Sci.*, vol. 270, no. 20, pp. 288–297, 2014.

[3] B. Li, J. He, J. Huang, and Y. Q. Shi, "A survey on image steganography and steganalysis," *J. Inf. Hiding Multimedia Signal Process.*, vol. 2, no. 2, pp. 142–172, 2011.

[4] S. Wen, Z. Zeng, T. Huang, Q. Meng, and W. Yao, "Lag synchronization of switched neural networks via neural activation function and applications in image encryption," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 26, no. 7, pp. 1493–1502, Jul. 2015.

[5] A. Cheddad, J. Condell, K. Curran, and P. M. Kevitt, "Digital image steganography: Survey and analysis of current methods," *Signal Process.*, vol. 90, no. 3, pp. 727–752, Mar. 2010.

[6] J. Liu *et al.*, "Recent advances of image steganography with generative adversarial networks," *IEEE Access*, vol. 8, pp. 60575–60597, 2020.

[7] C.-C. Chang, "Adversarial learning for invertible steganography," *IEEE Access*, vol. 8, pp. 198425–198435, 2020.

[8] N. Provos and P. Honeyman, "Detecting steganographic content on the internet," in *Proc. NDSS*, San Diego, CA, USA, Feb. 2002.

[9] K. Alex Zhang, A. Cuesta-Infante, L. Xu, and K. Veeramachaneni, "SteganoGAN: High capacity image steganography with GANs," 2019, *arXiv:1901.03892*.

[10] R. A. A. Campbell, R. W. Eifert, and G. C. Turner, "Openstage: A low-cost motorized microscope stage with sub-micron positioning accuracy," *PLoS ONE*, vol. 9, no. 2, Feb. 2014, Art. no. e88977.

[11] X. Luo, F. Liu, C. Yang, S. Lian, and D. Wang, "On F5 steganography in images," *Comput. J.*, vol. 55, no. 4, pp. 447–456, Apr. 2012.

[12] J. Zhu, R. Kaplan, J. Johnson, and L. Fei-Fei, "Hidden: Hiding data with deep networks," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 657–672.

[13] K. Joshi and R. Yadav, "A new LSB-S image steganography method blend with cryptography for secret communication," in *Proc. 3rd Int. Conf. Image Inf. Process. (ICIIP)*, Dec. 2015, pp. 86–90.

[14] P. Yadav, N. Mishra, and S. Sharma, "A secure video steganography with encryption based on LSB technique," in *Proc. IEEE Int. Conf. Comput. Intell. Comput. Res. (ICCIC)*, Dec. 2013, pp. 1–5.

[15] M. Hussain, A. W. A. Wahab, Y. I. B. Idris, A. T. S. Ho, and K.-H. Jung, "Image steganography in spatial domain: A survey," *Signal Process. Image Commun.*, vol. 65, pp. 46–66, Jul. 2018.

[16] K. B. Raja, C. R. Chowdary, K. R. Venugopal, and L. M. Patnaik, "A secure image steganography using LSB, DCT and compression techniques on raw images," in *Proc. 3rd Int. Conf. Intell. Sens. Inf. Process. (ICISIP)*, 2005, pp. 170–176.

[17] E. Walia, P. Jain, and N. Navdeep, "An analysis of LSB & DCT based steganography," *Global J. Comput. Sci. Technol.*, vol. 10, no. 1, Apr. 2010.

[18] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2223–2232.

[19] Y. Ren-Er, Z. Zhiwei, T. Shun, and D. Shilei, "Image steganography combined with DES encryption pre-processing," in *Proc. 6th Int. Conf. Meas. Technol. Mechatronics Autom.*, Jan. 2014, pp. 323–326.

[20] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter, "GANs trained by a two time-scale update rule converge to a local Nash equilibrium," in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, 2017, pp. 6626–6637.

[21] S. Baluja, "Hiding images in plain sight: Deep steganography," in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, 2017, pp. 2069–2079.

[22] J. Yang, K. Liu, X. Kang, E. K. Wong, and Y.-Q. Shi, "Spatial image steganography based on generative adversarial network," 2018, *arXiv:1804.07939*.

[23] R. Rahim and S. Nadeem, "End-to-end trained CNN encoder-decoder networks for image steganography," in *Proc. Eur. Conf. Comput. Vis. Workshops (ECCVW)*, 2018, pp. 1–6.

[24] C. Chu, A. Zhmoginov, and M. Sandler, "CycleGAN, a master of steganography," 2017, *arXiv:1712.02950*.

[25] S. Baluja, "Hiding images within images," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 7, pp. 1685–1697, Jul. 2020.

[26] I. Goodfellow *et al.*, "Generative adversarial nets," in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, 2014, pp. 2672–2680.

[27] Z. Zheng *et al.*, "EncryptGAN: Image steganography with domain transform," 2019, *arXiv:1905.11582*.

[28] W. Tang, B. Li, S. Tan, M. Barni, and J. Huang, "CNN-based adversarial embedding for image steganography," *IEEE Trans. Inf. Forensics Security*, vol. 14, no. 8, pp. 2074–2087, Aug. 2019.

[29] W. Tang, S. Tan, B. Li, and J. Huang, "Automatic steganographic distortion learning using a generative adversarial network," *IEEE Signal Process. Lett.*, vol. 24, no. 10, pp. 1547–1551, Oct. 2017.

[30] C.-C. Chang, "Cryptospace invertible steganography with conditional generative adversarial networks," *Secur. Commun. Netw.*, vol. 2021, pp. 1–14, Mar. 2021.

[31] C. Zhang, P. Benz, A. Karjauv, and I. S. Kweon, "Universal adversarial perturbations through the lens of deep steganography: Towards a Fourier perspective," 2021, *arXiv:2102.06479*.

[32] N. Cauvery, "Water marking on digital image using genetic algorithm," *Int. J. Comput. Sci. Issues*, vol. 8, no. 6, p. 323, 2011.

[33] A. Babu and S. Ayyappan, "A reversible crypto-watermarking system for secure medical image transmission," in *Proc. Annu. IEEE India Conf. (INDICON)*, Dec. 2015, pp. 1–6.

[34] V. M. Potdar, S. Han, and E. Chang, "A survey of digital image watermarking techniques," in *Proc. IEEE Int. Conf. Ind. Informat. (INDIN)*, Aug. 2005, pp. 709–716.

[35] Y. Quan, H. Teng, Y. Chen, and H. Ji, "Watermarking deep neural networks in image processing," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 32, no. 5, pp. 1852–1865, May 2021.

[36] P. Ganesan and R. Bhavani, "A high secure and robust image steganography using dual wavelet and blending model," *J. Comput. Sci.*, vol. 9, no. 3, pp. 277–284, Mar. 2013.

[37] C. Zhang, C. Lin, P. Benz, K. Chen, W. Zhang, and I. S. Kweon, "A brief survey on deep learning based data hiding," 2021, *arXiv:2103.01607*.

[38] A. Mansurov, "A CTF-based approach in information security education: An extracurricular activity in teaching students at Altai State University, Russia," *Modern Appl. Sci.*, vol. 10, no. 11, p. 159, Aug. 2016.

[39] A. M. Fard, M.-R. Akbarzadeh, and F. Varasteh, "A new genetic algorithm approach for secure JPEG steganography," in *Proc. IEEE Int. Conf. Eng. Intell. Syst.*, Apr. 2006, pp. 1–6.
[40] Y. Qian, J. Dong, W. Wang, and T. Tan, "Deep learning for steganalysis via convolutional neural networks," *Proc. SPIE*, vol. 9409, Mar. 2015, Art. no. 94090J.
[41] L. Pibre, J. Pasquet, D. Ienco, and M. Chaumont, "Deep learning is a good steganalysis tool when embedding key is reused for different images, even if there is a cover sourcemismatch," *Electron. Imag.*, vol. 2016, no. 8, pp. 1–11, 2016.
[42] 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. Adv. Neural Inf. Process. Syst. (NeurIPS)*, vol. 33, 2020, pp. 10223–10234.
[43] R. Abdal, Y. Qin, and P. Wonka, "Image2StyleGAN: How to embed images into the StyleGAN latent space?" in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 4432–4441.
[44] J. Hayes and G. Danezis, "Generating steganographic images via adversarial training," 2017, *arXiv:1703.00371*.
[45] D. Volkhonskiy, B. Borisenko, and E. Burnaev, "Generative adversarial networks for image steganography," Tech. Rep., 2016. [Online]. Available: https://openreview.net/pdf?id=H1hoFU9xe
[46] A. Radford, L. Metz, and S. Chintala, "Unsupervised representation learning with deep convolutional generative adversarial networks," 2015, *arXiv:1511.06434*.
[47] R. Zhang, S. Dong, and J. Liu, "Invisible steganography via generative adversarial networks," *Multimedia Tools Appl.*, vol. 78, no. 7, pp. 8559–8575, Apr. 2019.
[48] M. Caron, P. Bojanowski, A. Joulin, and M. Douze, "Deep clustering for unsupervised learning of visual features," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 132–149.
[49] Y. Zou, Z. Yu, X. Liu, B. V. K. V. Kumar, and J. Wang, "Confidence regularized self-training," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 5982–5991.
[50] K. Saito, Y. Ushiku, T. Harada, and K. Saenko, "Strong-weak distribution alignment for adaptive object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 6956–6965.
[51] K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick, "Momentum contrast for unsupervised visual representation learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 9729–9738.
[52] S. Zhao *et al.*, "A review of single-source deep unsupervised visual domain adaptation," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 33, no. 2, pp. 473–493, Feb. 2022.
[53] Y.-H. Tsai, W.-C. Hung, S. Schulter, K. Sohn, M.-H. Yang, and M. Chandraker, "Learning to adapt structured output space for semantic segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2018, pp. 7472–7481.
[54] S. Iizuka, E. Simo-Serra, and H. Ishikawa, "Globally and locally consistent image completion," *ACM Trans. Graph.*, vol. 36, no. 4, pp. 1–14, 2017.
[55] Q. Chen and V. Koltun, "Photographic image synthesis with cascaded refinement networks," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (CVPR)*, Oct. 2017, pp. 1511–1520.
[56] M.-E. Nilsback and A. Zisserman, "Automated flower classification over a large number of classes," in *Proc. 6th Indian Conf. Comput. Vis., Graph. Image Process. (ICVGIP)*, Dec. 2008, pp. 722–729.
[57] 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. (CVPR)*, Jun. 2009, pp. 248–255.
[58] H. Yeganeh and Z. Wang, "Objective quality assessment of tone-mapped images," *IEEE Trans. Image Process.*, vol. 22, no. 2, pp. 657–667, Feb. 2013.
[59] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014, *arXiv:1412.6980*.
[60] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, "The unreasonable effectiveness of deep features as a perceptual metric," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2018, pp. 586–595.
[61] B. Boehm, "StegExpose—A tool for detecting LSB steganography," 2014, *arXiv:1410.6656*.
[62] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 2818–2826.
[63] X. Deng, B. Chen, W. Luo, and D. Luo, "Fast and effective global covariance pooling network for image steganalysis," in *Proc. ACM Workshop Inf. Hiding Multimedia Secur.*, Jul. 2019, pp. 230–234.
[64] M. Boroumand, M. Chen, and J. Fridrich, "Deep residual network for steganalysis of digital images," *IEEE Trans. Inf. Forensics Security*, vol. 14, no. 5, pp. 1181–1193, May 2018.
[65] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 1125–1134.
[66] C. Zhang, A. Karjauv, P. Benz, and I. S. Kweon, "Towards robust deep hiding under non-differentiable distortions for practical blind watermarking," in *Proc. 29th ACM Int. Conf. Multimedia (ACM MM)*, Oct. 2021, pp. 5158–5166.
[67] Y. Liu, M. Guo, J. Zhang, Y. Zhu, and X. Xie, "A novel two-stage separable deep learning framework for practical blind watermarking," in *Proc. 27th ACM Int. Conf. Multimedia (ACM MM)*, Oct. 2019, pp. 1509–1517.

Ziqiang Zheng received the B.Eng. degree in communication engineering from the Ocean University of China, Qingdao, China, in 2019.

He is currently with the Center for Future Media, School of Computer Science and Engineering, University of Electronic Science and Technology of China, Chengdu, China. His research interests include multimedia content analysis and computer vision.

Yuanmeng Hu received the B.S. degree in information and computational science from the Ocean University of China, Qingdao, China, in 2019. She is currently pursuing the M.S. degree with the FFM Center (Applied mathematics in Big data/Industries/Finance), Pusan National University, Busan, South Korea.

Her research interests include matrix completion, matrix decomposition, and machine learning.

Yi Bin received the B.Eng. degree in electronic engineering from the Civil Aviation University Of China (CAUC) in 2013. He is currently pursuing the Ph.D. degree with the School of Computer Science and Engineering, University of Electronic Science and Technology of China, Chengdu, China.

His current research interests include multimedia analysis, vision understanding, and deep learning.

Xing Xu (Member, IEEE) received the B.E. and M.E. degrees from the Huazhong University of Science and Technology, Wuhan, China, in 2009 and 2012, respectively, and the Ph.D. degree from Kyushu University, Fukuoka, Japan, in 2015.

He is currently with the Center for Future Media and the School of Computer Science and Engineering, University of Electronic Science and Technology of China, Chengdu, China. His current research interests mainly focus on multimedia information retrieval, pattern recognition, and computer vision.

Yang Yang (Senior Member, IEEE) received the bachelor's degree from Jilin University, Changchun, China, in 2006, the master's degree from Peking University, Beijing, China, in 2009, and the Ph.D. degree from The University of Queensland, Brisbane, QLD, Australia, in 2012, all in computer science.

He is currently with the University of Electronic Science and Technology of China, Chengdu, China. His current research interests include multimedia content analysis, computer vision, and social media analytics.

Heng Tao Shen (Fellow, IEEE) received the B.Sc. (Hons.) and Ph.D. degrees from the Department of Computer Science, National University of Singapore, Singapore, in 2000 and 2004, respectively.

He is currently the Dean of the School of Computer Science and Engineering and the Executive Dean of the AI Research Institute, University of Electronic Science and Technology of China (UESTC), Chengdu, China. His research interests mainly include multimedia search, computer vision, artificial intelligence, and big data management.

Dr. Shen is a fellow of Association for Computing Machinery (ACM) and Optica (formerly known as The Optical Society) (OSA). He is/was an Associate Editor of *ACM Transactions of Data Science*, IEEE TRANSACTIONS ON IMAGE PROCESSING, IEEE TRANSACTIONS ON MULTIMEDIA, IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, and *Pattern Recognition*.