Anti-Rounding Image Steganography With Separable Fine-Tuned Network

Xiaolin Yin, Shaowu Wu, Ke Wang, Wei Lu, *Member, IEEE*, Yicong Zhou, *Senior Member, IEEE*, and Jiwu Huang, *Fellow, IEEE*

***Abstract*— Image steganographic methods based on encoder-decoder model with end-to-end network architecture recently have been proposed. However, in steganographic applications, the feature map (called stego matrix) generated by the encoder needs to be rounded as a real stego image for the receiver. The loss of precision by rounding stego matrix leads to the decline in the accuracy of extracted secret messages. The challenge of using end-to-end network to preserve robustness against rounding operation is that it is non-differentiable. In this paper, we propose an anti-rounding image steganography method with separable fine-tuning network architecture which includes the joint training stage (JT-stage) and the separable fine-tuning stage (SF-stage). Firstly, in JT-stage, an embedded generator and a stego matrix extractor are jointly learned without rounding operation. Utilizing concatenation in embedded generator can realistically fuse cover image and secret messages. And the multi-scale fusion block and residual dense block in stego matrix extractor can make secret messages more correctly decoded. Moreover, the discriminator is constructed by generative adversarial nets (GAN) in JT-stage to effectively improve the authenticity and steganalysis security. Then, in SF-stage, the embedded generator is frozen, and the stego matrix is obtained and rounded as a stego image. A stego image extractor is constructed by fine-tuning the layers of the stego matrix extractor to improve the accuracy of message extraction. As the loss will not backpropagate in the embedded generator, the non-differentiability of rounding operation can be offset. Experiments show that the proposed separation fine-tuning network is robust to rounding operation, and effectively reduces the degradation of the image quality and steganalysis performance.**

Manuscript received 5 November 2022; revised 11 February 2023; accepted 19 April 2023. Date of publication 24 April 2023; date of current version 30 October 2023. This work was supported in part by the National Natural Science Foundation of China under Grant 62261160653, Grant U19B2022, Grant 62102101, and Grant U1636202; and in part by the Alibaba Group through Alibaba Innovative Research (AIR) Program. This article was recommended by Associate Editor P. Bestagini. *(Corresponding author: Wei Lu.)*

Xiaolin Yin, Shaowu Wu, Ke Wang, and Wei Lu are with the Guangdong Province Key Laboratory of Information Security Technology and the Ministry of Education Key Laboratory of Machine Intelligence and Advanced Computing, School of Computer Science and Engineering, Sun Yat-sen University, Guangzhou 510006, China (e-mail: yinxl6@mail2.sysu.edu.cn; wushw25@mail2.sysu.edu.cn; wangk237@mail2.sysu.edu.cn; luwei3@mail.sysu.edu.cn).

Yicong Zhou is with the Department of Computer and Information Science, University of Macau, Macau, China (e-mail: yicongzhou@um.edu.mo).

Jiwu Huang is with the Guangdong Key Laboratory of Intelligent Information Processing, the Key Laboratory of Media Security, and the Guangdong Laboratory of Artificial Intelligence and Digital Economy (SZ), Shenzhen University, Shenzhen 518060, China, and also with the Shenzhen Institute of Artificial Intelligence and Robotics for Society, Shenzhen 518055, China (e-mail: jwhuang@szu.edu.cn).

Color versions of one or more figures in this article are available at https://doi.org/10.1109/TCSVT.2023.3269468.

Digital Object Identifier 10.1109/TCSVT.2023.3269468

***Index Terms*— Image steganography, anti-rounding, separable fine-tuned network, precision loss.**

I. Introduction

IMAGE steganography is an art of covert communication to reveal the suspicious trace for the existence of secret messages in stego images. It plays an important role in information security and data communication, which is of great significance to ensure data security. Steganography algorithms improve the statistical security against steganalysis methods. In order to resist the high dimensional statistical characterization, the traditional content-adaptive steganography methods [1], [2], [3], [4], [5], [6], [7] are proposed to embed secret messages by slightly modifying the high complex textures, so as to improve the concealment of steganography. The steganography distortion cost function is designed to measure the embedding cost of each pixel, so as to find the appropriate embedding position and minimize the steganography distortion.

With the rapid development of deep learning (DL) technology, many DL-based steganalysis methods have been proposed [8], [9], [10], and the traditional content-adaptive steganography methods are gradually difficult to resist detections. Because of the strong ability of feature learning in DL, inspiredly, DL-based steganography methods [11], [12], [13], [14], [15], [16] are proposed rather than manually designed embedding. The embedding location and the embedding mode can be learned independently through the network.

There are three typical categories of DL-based image steganography [17]. The first one is the cover image acquisition technology based on generative adversarial network (GAN) [18], [19] and adversarial samples [20]. The second one is based on learning the steganographic distortion design. In [21] and [22], steganographic distortion costs are designed based on GAN. In adversarial training, the embedding probability of each pixel is learned, and then it is transformed into a modification map. In this way, the stego image is generated by adding a cover image with the modification map. However, it takes time for the network to learn the embedding probability. In [23], Tang et al. proposed the steganographic distortion cost based on adversarial examples to determine the modification direction. To mislead the steganalyzer, stego images with the features of adversarial examples are generated.

The third category is based on encoder-decoder model with end-to-end network, which is the technology we focus

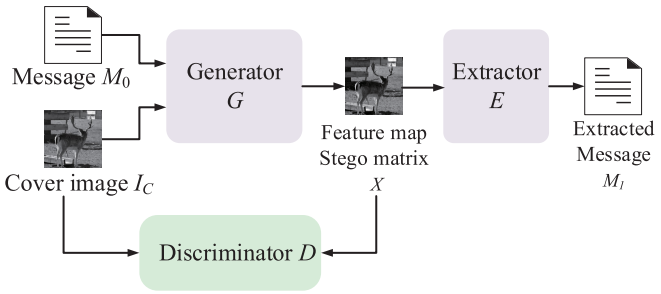


Fig. 1. The general training procedure for image steganography based on one-stage encoder-decoder model.

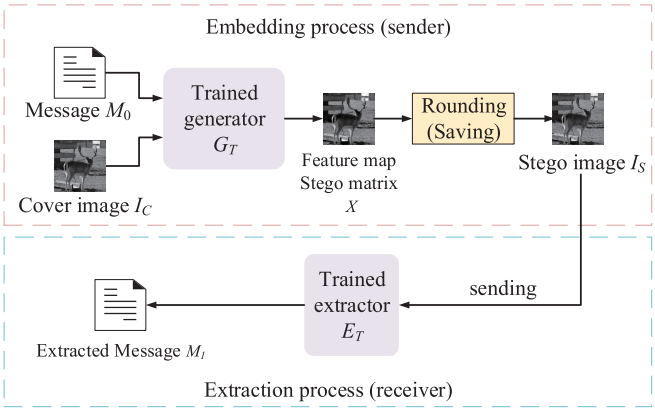


Fig. 2. The general testing procedure for image steganography based on one-stage encoder-decoder model.

on in this paper. The encoder-decoder models are widely used to hide information in images. As a branch of data hiding, digital watermarking is the technology to robustly encode secret messages. In [24], in order to improve the robustness, HiDDeN (Hiding Data with Deep Networks) is proposed to add the noise layer in training to simulate the noise attack. In [25], ABDH (Attention Based Data Hiding) proposes two generative models and two discriminative models for both images and secret payload. In [26], a two-stage blind watermarking network is proposed to train the decoder with arbitrary noise to improve the robustness of the watermark. Different from digital watermarking, steganography is the technology for secret communication. The encoder-decoder-based methods of stego images generation are included hiding texts within images (bit-level information) and hiding images within images (image-level information). In texts hiding methods, the secret messages are encoded to meet with the input of networks. In [27], the adversarial training technique named HayesGAN is proposed for the discriminative task of learning a steganographic algorithm. In [28], Zhang et al. proposed the SteganoGAN (Steganography with Generative Adversarial Networks) to optimize the perceptual quality, and it achieves great embedding payloads. In [29], CHAT-GAN (Channel Attention Mechanisms based on GAN) proposes a channel attention module to learn the channel interdependencies, which improves the accuracy of message extraction. In images hiding methods, the secret image is preprocessed to be embedded into cover image by the encoder, and reconstructed from the stego image by the decoder. In [30], the secret image can be realistically restore by utilized the upsampling and downsampling with connection operation based on U-Net. ISGAN (Invisible Steganography via Generative Adversarial Networks) [31] embeds the secret image in the Y channel and proposes a SSIM-based mixed loss function to improve the quality of the restored secret image. In a word, steganography based on encoder-decoder model can embed and extract the secret messages without the prior knowledge of images when owning the encoder network and decoder network.

The existing steganographic one-stage encoder-decoder networks [25], [27], [28], [29] do not describe in detail the impact of precision loss caused by the rounding operation on the accuracy performance of steganography. In these networks, the generator, the extractor, and the discriminator are jointly trained, as shown in Fig. 1. When the samples are tested in the trained model, the general testing procedure for image steganography based on one-stage encoder-decoder model is shown in Fig. 2. The applicable steps are listed as followed:

1) The sender will input the cover image and message into the trained generator, and the feature map stego matrix will be outputted.
2) The feature map stego matrix will be rounded and saved as a stego image.
3) The stego image will be sent to the receiver.
4) The receiver will input the stego image into the trained extractor, and the extracted message can be obtained.

In the encoder-decoder model, the output feature map generated by the CNN-based encoder is usually stored as the floating-point number, which we call it stego matrix. Floating-point arithmetic provides a larger range of values and higher precision, which are favored in CNN weights to handle real numbers [32]. However, in steganographic applications, a real stego image is needed for the receiver. It can be seen that the significant difference is the input of the extractor in different procedures. The input of the extractor in the training procedure is the stego matrix, and the input of the trained extractor in the testing procedure is the stego image. It means that the generated stego matrix must be saved as a stego image by rounding operation. When the stego image is input into the trained decoder of the receiver, the loss of precision will lead to a sharp decline in the accuracy of message extraction. It requires the joint training of the encoder and the decoder can resist the loss of precision. Nevertheless, the rounding operation is non-differentiable for backpropagation. It cannot simulate the saving process of stego image by directly rounding the stego matrix when training the encoder and decoder. In this way, the gradient will not be obtained, so the parameters of the network model cannot be updated. We believe that a careful and delicate design of the network architecture to offset the non-differentiability of rounding operation is a non-trivial work.

Some existing robust watermarking methods deal with the precision loss caused by the non-differentiability problem, mainly through derivative approximation methods. In [24], the non-differentiability problem of quantization in JPEG compression is solved by two differentiable approximations: JPEG-Mask and JPEG-Drop, which zeros the

high-frequency coefficients. In the field of image compression, the non-differentiability of rounding (the rounding is discontinuous and its derivative is zero or infinite anywhere) has attracted attention. The most common methods are randomized approximation [33], [34] and smooth approximation [35]. However, these methods add global differentiable functions to compensate for the precision loss caused by rounding. In real rounding applications, the accuracy of the extractor in steganography can still be improved. Therefore, we propose a two-stage separable training method to solve the non-differentiable problem and precision loss caused by the real rounding operation.

In this paper, a separable fine-tuned network (SFTN) architecture is proposed which is robust to rounding operation in steganographic applications. The proposed SFTN architecture includes the joint training stage (JT-stage) and the separable fine-tuning stage (SF-stage). In JT-stage, a stego matrix extractor is trained to correctly extract secret messages while an embedded generator is jointly trained to ensure the image quality. A discriminator is constructed by GAN to effectively improve the authenticity and steganalysis security of stego image. In SF-stage, all layers of the embedded generator are copied and kept frozen during training. The stego matrix is rounded as a stego image. By retraining the stego matrix extractor and fine tuning the layers, a stego image extractor is constructed to increase the accuracy of message extraction. In two extractors, the multi-scale fusion block and residual dense block are utilized to fuse cover image and secret messages. The main contributions of this paper are as follows:

1) The proposed SFTN architecture can offset the non-differentiability of rounding operation.
2) The concatenations can realistically fuse cover image and secret messages. In addition, the multi-scale fusion block and residual dense block are utilized to improve the accuracy of message extraction.
3) Based on GAN, the discriminator is constructed to evaluate the authenticity of stego in the joint training stage, so as to effectively improve the quality of stego image and the security of steganalysis.

The remainder of this paper is organized as follows. In Section II, we discuss the challenges of precision loss and rounding operation in the existing steganographic one-stage encoder-decoder networks. In Section III, we introduce the proposed anti-rounding image steganography with separable fine-tuned network. It includes the discussion about the overall model architecture, and the details about the JT-stage and the SF-stage, as well as the design of the loss functions. Section IV analyzes the impact of the proposed separable fine-tuned network. Andthe experimental verifications are discussed. Finally, the conclusion is presented in Section V.

II. Non-Derivable Rounding in One-Stage Encoder-Decoder Model

In Fig. 1, the general procedure for image steganography based on one-stage encoder-decoder model is presented. There are general three components including the generator G with parameters θ_G, extractor E with parameters θ_E, and the discriminator D with parameters θ_D. The CNN-based embedding generator G is utilized to fuse the cover image I_C and secret messages M_0, and the floating-point feature map called stego matrix X is output. Floating-point arithmetic provides a larger range of values and higher precision, which are favored in CNN weights to handle real numbers [32]. The embedding process and extraction process of existing one-stage encoder-decoder steganographic network architecture [24], [28], [29] are presented as:

$$\begin{aligned} X &= G(M_0, I_C; \theta_G), \\ M_1 &= E(X; \theta_E) \end{aligned} \tag{1}$$

where M_1 is the extracted messages. The discriminator D is utilized to determine the authenticity of the stego matrix X. Through joint training, a high accuracy of message extraction can be obtained by extracting from the floating-point stego matrix X.

In steganographic applications, secret messages are imperceptibly embedded into cover image, and then the stego image with high visual quality is transmitted through the public channel. Using the encoder-decoder model, it is inevitable to transfer the floating-point stego matrix X into the stego image I_S by using the rounding operation. The embedding process and extraction process of steganographic practical applications are presented as:

$$\begin{aligned} X &= G(M_0, I_C; \theta_G), \\ I_S &= \lfloor X \rceil, \\ M_1 &= E(I_S; \theta_E) \end{aligned} \tag{2}$$

where $\lfloor \cdot \rceil$ is the rounding operation. After receiving the stego image I_S, the extractor E is utilized to extract the secret messages M_1. The precision loss between the stego matrix X and the stego image I_S brings the challenge of the robustness in the trained extractor. Thus, the precision loss caused by the rounding operation lead to a decline of the accuracy of message extraction.

In order to fully consider the rounding operation, the joint training of the generator and extractor need to resist the precision loss. However, the rounding operation is non-derivable [36]. The stego image I_S has derivative 0 nearly everywhere, which is not compatible with the gradient-based methods. Assume a joint training of encoder-decoder model by directly rounding the stego matrix X to stego image I_S, shown in Fig. 3. Denote the differences between message M_0 and the extracted message M_1 as the message loss L_m. The differences between cover image I_C and the stego matrix X are denoted as the image quality loss L_q and adversarial loss L_d. The training objective is to minimize:

$$L = \lambda_m L_m + \lambda_q L_q + \lambda_d L_d \tag{3}$$

where L is the overall loss, and λ_m, λ_q, λ_d control the relative weight of each item. By jointly training three components, a set of parameters is searched to minimize the overall loss L, so as to construct the generator which ensures the visual quality of stego and security performance. The gradients of G are utilized to optimize the parameters of the generator G.

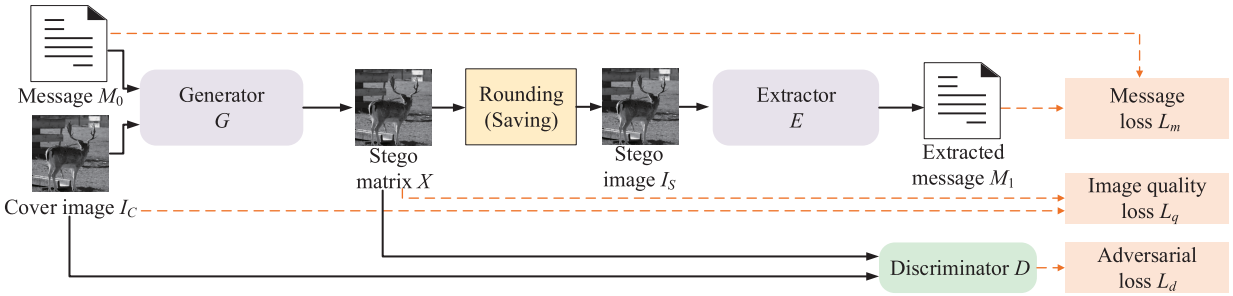


Fig. 3. The joint training of encoder-decoder model by directly rounding the stego matrix X to stego image I_S. The failure of the generator G parameter optimization is caused by the rounding operation.

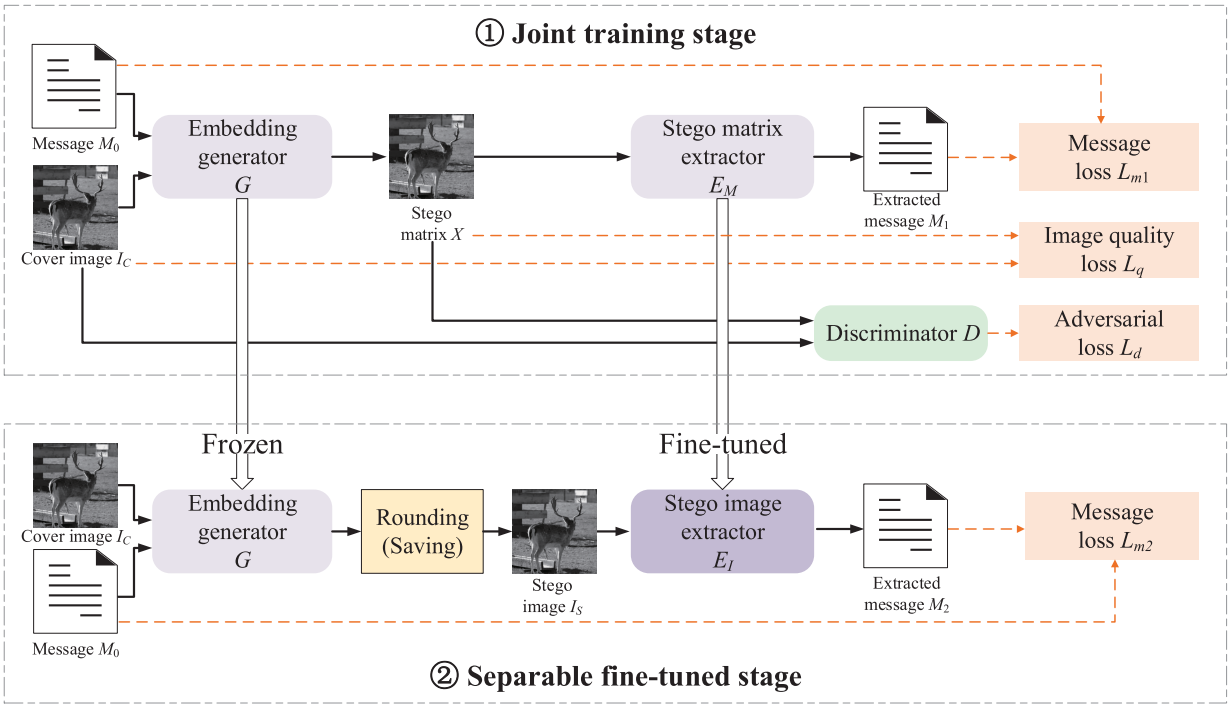


Fig. 4. The model architecture of the proposed anti-rounding image steganography with separable fine-tuned network.

The gradient of generator G is denoted as:

$$\begin{aligned}\triangle\theta_G &= \frac{\partial L_m}{\partial G} + \frac{\partial L_q}{\partial G} + \frac{\partial L_d}{\partial G} \\ &= \frac{\partial L_m}{\partial I_S}\cdot\frac{\partial I_S}{\partial X}\cdot\frac{\partial X}{\partial G} + \frac{\partial L_q}{\partial X}\cdot\frac{\partial X}{\partial G} + \frac{\partial L_d}{\partial X}\cdot\frac{\partial X}{\partial G}\end{aligned} \tag{4}$$

Since the rounding operation is non-differentiable, $\frac{\partial I_S}{\partial X} = \frac{\partial \lfloor X \rceil}{\partial X}$ does not exist. The gradient of G will not be backpropagated to update the parameters of the generator G. It leads to the failure of the training of the generator G.

When focusing on extractor E in the one-stage encoder-decoder network, the gradient of extractor E is denoted as:

$$\triangle\theta_E = \frac{\partial L_m}{\partial E} \tag{5}$$

Since $\triangle\theta_E$ does not contain $\frac{\partial I_S}{\partial X}$, the rounding operation does not affect the optimization of the extractor E. However, in the one-stage joint training, due to the failure of the parameter optimization of the generator G, the better embedding method will no longer be learned. The optimization of the extractor E is worthless. In a word, the joint training of one-stage encoder-decoder model by directly rounding the stego matrix X to stego image I_S shown in Fig. 3 is unreasonable. It is a non-trivial work to design the network architecture to offset the non-differentiability of rounding operation and the precision loss.

III. Proposed Method

To resist the rounding operation and the precision loss, the anti-rounding image steganography with separable fine-tuned network (SFTN) is proposed. The SFTN architecture is presented in Fig. 4, including the joint training stage (JT-stage) and the separable fine-tuning stage (SF-stage). Without rounding the stego matrix X, the main purpose of JT-stage is to ensure the end-to-end training is differentiable. In SF-stage, all layers of the embedded generator G are copied and kept frozen during training. The stego matrix X is generated from the frozen embedded generator G and rounded as a stego image I_S. By fine tuning the layers of the trained stego matrix

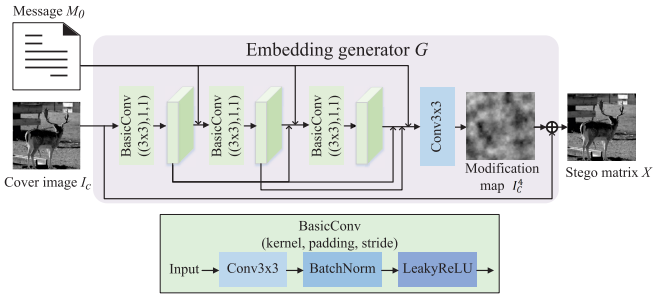


Fig. 5. The model of the embedding generator G using concatenations between multi-layer CNN cover image features and message.

extractor E_M in JS-stage, the stego image extractor E_I needs to be retrained to increase the accuracy of message extraction. The main purpose of SF-stage is to prevent the losses from propagating back to the embedding generator. By separating the training of the stego image extractor E_I and embedded generator G, the non-differentiability of the rounding operation can be offset. The separable training can further improve the robustness of the stego image extractor E_I against the precision loss so that enhance the message extraction ability from stego image I_S.

A. Model Architecture

Three components constitute the whole model in JT-stage, including the embedding generator G with parameters θ_G, the stego matrix extractor extractor E_M with parameters θ_{E_M}, and the Discriminator D with parameters θ_D.

1) Embedding Generator: The model of the embedding generator G is presented in Fig. 5. As input with C channels and $H \times W$ size, the cover image $I_C \in P^{C\times H\times W}$ with is denoted, where $P \in \{0, 1, \ldots, 255\}$ is the pixel values in spatial domain. The message $M_0 \in \{0, 1\}^{C\times H\times W}$ is also input in G, which is a randomized binary stream with the same size as the cover image I_C. The embedding process combining cover image I_C and message M_0 in the embedding generator G can be denoted as:

$$X = G(M_0, I_C; \theta_G) \tag{6}$$

where the output X is the stego matrix with $C \times H \times W$ size. A basic CNN component called BasicConV is constructed by a convolution layer (Conv3 $\times$ 3), a Batch Normalization layer (BatchNorm), and the Leaky Rectified Linear Unit (LeakyReLU). The Conv3 $\times$ 3 is with a kernel size of 3 $\times$ 3, a padding of 1, and a stride of 1. The BatchNorm provides the predictive and stable behavior of gradients for faster training [37]. The LeakyReLU with a negative slope of 0.1 is the activation function that assigns a non-zero slope.

Utilizing multiple CNNs can capture diverse information in image [38]. The channel numbers of the cover image can be enlarged by using a CNN with multiple convolution filters, so that richer embeddable features can be captured. Compared with a single CNN, concatenation of multi-layer CNN feature maps integrates the embeddable position from different-layer CNNs together to obtain the fusion between I_C and M_0 with a high visual quality. In embedding generator G, we design the embedding process to obtain the stego matrix X by using concatenation between multi-layer CNN cover image features and message. Referring to the concatenation in [29], the channel attention module is not adopted in the proposed method, but deepens the model of extractors to improve the accuracy of message extraction. The embedding generator G consists of three BasicConV with 32 filters and a Conv3 $\times$ 3 with C filters. Firstly, the cover image I_C is input in the first BasicConV to capture the first residual features I_C^1 in 32 channels. After that, the input is constructed by the concatenation with the current residual features, all preceding residual features, and the message M_0. For example, the concatenation (I_C^1, M_0) with $(32+C)$ channels is input in the second BasicConV to capture the second residual features I_C^2. The concatenation (I_C^1, M_0) with $(32+C)$ channels is input in the second BasicConV to capture the second residual features I_C^2. Similarly, the concatenation (I_C^1, I_C^2, M_0) with $(64 + C)$ channels is input in the third BasicConV to capture the third residual features I_C^3. Then, the concatenation $(I_C^1, I_C^2, I_C^3, M_0)$ with $(96 + C)$ channels is input in the Conv3 $\times$ 3 capture the final residual features I_C^4. The final residual features I_C^4 can be regarded as the modification map. The modification map is the result learned by the multi-layer BasicConV. The modification map is the embedding noise that determines the modification direction and the magnitude of pixel values. In this way, the stego matrix $X \in P^{C\times H\times W}$ can be obtained by adding I_C^4 to the cover image I_C. Through the concatenation and residual learning, the embedding generator G effectively fuses the message M_0 and cover image I_C, and the modification of the cover image can be reduced.

2) Stego Matrix Extractor: The model of the stego matrix extractor E_M is presented in Fig. 6. The output stego matrix X from the embedding generator G is usually stored as the floating-point number. It is directly input in the stego matrix extractor E_M. The extraction process can be denoted as:

$$M_1 = E_M(X; \theta_{E_M}) \tag{7}$$

where M_1 is the extracted message. Two blocks are utilized to fully capture the features of stego matrix X to generate the extracted message M_1:

1) Multi-scale fusion block (MFB): Using different padding standards on different spatial scales, the integrated features are obtained by fusing different convolution features [39]. The preceding feature map is input in three convolution layers with padding of 1, 2, and 3, respectively. The multi-scale features are concatenated, and then the integrated features are processed in a BasicConV (shown in Fig. 5) to obtain more detailed spatial context information.
2) Residual dense block (RDB): The RDB consists dense connected layers and local feature fusion [40]. It integrates the residual block and the dense block. The contiguous memory (CM) mechanism is constructed by utilizing all layers with local dense connections. Each convolution layer in RDB passes on the retained extraction information to all subsequent layers. The local feature fusion concatenated the current residual dense

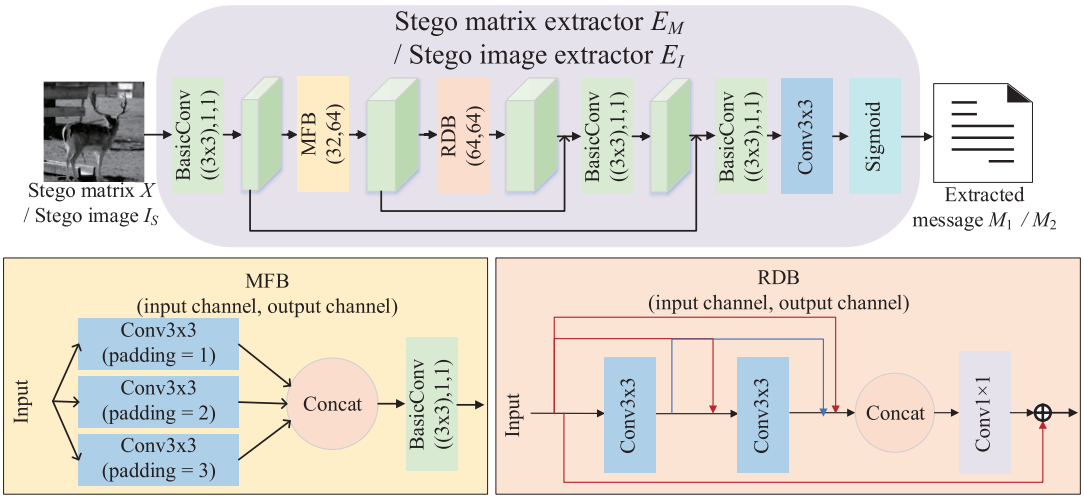


Fig. 6. The model of the stego matrix extractor E_M and stego image extractor E_I using MFB and RDB with layers concatenation.

layer with all preceding layers, and then input in a convolution layer with a kernel size 1×1 to extract local dense features.

Firstly, the BasicConV is used to enlarge the number of channels of the stego matrix to capture richer modification features. Then, an MFB is used to make full use of hierarchical complementary features to obtain the embedding position information. The hierarchical complementary features is input in an RDB to further learn the local modification features. The embedding position and the direction of pixel modification can be accurately obtained. The deep modification features are captured through two BasicConV and a convolution layer. The precision of learning is increased by concatenating the feature maps using the MFB and the RDB with the deep modification features learned by BasicConV. Finally, the deep modification features are normalized to the range from 0 to 1 by using Sigmoid as activation, which are regarded as the extracted message M_1.

3) Discriminator: The discriminator is constructed by generative adversarial nets (GAN) in JT-stage. The discriminator is used to classify the authenticity of the stego matrix to effectively improve the authenticity and steganalysis security [41]. The key to image steganalysis is to judge the existence of weak image steganographic signals. It is critical to extract effective steganographic signal features in digital image steganalysis. In the proposed method, the discriminator is similar to steganalysis to explore whether secret messages are hidden in the generated stego matrix. Therefore, the discriminator needs to capture the steganographic noise residual effectively. Generally, by utilizing simple designed multi-layer CNN and fully-connected layers, the ability to extract the steganographic features is limited. Therefore, we choose XuNet [42] which is a specific steganalysis network as the discriminator. XuNet employs the absolute value layer and hyperbolic tangent at the early stages of the networks, batch normalization, and 1×1 convolutions to reduce the strength of the model, which provides better detection.

4) Stego Image Extractor: Through the training of the JT-stage, the embedded generator is obtained which ensures the minimal differences of cover image I_C and stego matrix X. In SF-stage, the central issue is further training of extraction ability. All layers and parameters of the embedded generator G are copied and kept frozen during training. Using the fixed embedded generator G, the stego matrix X is generated, and then it is rounded as a stego image I_S. The stego image extractor E_I keeps the same model as the stego matrix extractor E_M. The extraction process of stego image extractor E_I is denoted as:

$$M_2 = E_I(I_S; \theta_{E_I}) \tag{8}$$

By fixing the embedded generator G, the message loss L_{m2} only propagates back only through the stego image extractor E_I. So the parameter θ_{E_I} of the stego image extractor E_I can be optimized. In summary, the non-differentiability of rounding operation can be offset by using the proposed separable fine-tuned network. And the decline of the message extraction accuracy caused by precision loss can be compensated.

B. Loss Function

In JT-stage, the discriminator D is optimized alternately with the embedding generator G and stego matrix extractor E_M. The discriminator D aims to classify the cover image I_C and the stego matrix X. Thus, the difference between the prediction scores of I_C and the prediction scores of X can be used to update the discriminator D. We use XuNet as the steganalyzer to predict I_C and X. The details of discriminator D refer to [29]. The adversary loss L_d is denoted as:

$$L_d = |D(I_C; \theta_D) - D(X; \theta_D)|. \tag{9}$$

By jointly training, the embedding generator G ensures slight differences between I_C and X. So that the stego image I_S generated by rounding X can own a better objective visual quality. Thus, the image quality loss L_q is calculated by mean square error (MSE), defined as:

$$\begin{aligned} L_q &= MSE(I_C, X) \\ &= \frac{1}{CHW} \sum_{j=1}^{CHW} (i_{c_j} - x_j)^2 \end{aligned} \tag{10}$$

where i_{c_j} and x_j are the pixel value elements of the cover image I_C and the stego matrix X, respectively. After the stego matrix X is input in the stego matrix extractor E_M, the floating-point extracted messages M_1 in range of [0, 1] is obtained. We hope the stego matrix extractor E_M is optimized to minimize the distance between the extracted messages M_1 and the original message M. The message loss L_{m1} is calculated by the binary cross entropy (BCE) loss, denoted as:

$$\begin{aligned} L_{m1} &= BCE(M, M_1) \\ &= \frac{1}{CHW}\sum_{j=1}^{CHW} -m_j \log(m_{1_j}) - (1-m_j)\log(1-m_{1_j}) \end{aligned} \tag{11}$$

where m_j and m_{1_j} are the 0 and 1 elements of the original messages M and the extracted messages M_1, respectively. Thus, the training objective of the JT-stage is to minimize the overall loss L_{JT}, denoted as:

$$\begin{aligned} (\theta_D, \theta_G, \theta_{E_S}) &= \arg\max_{\theta_D,\theta_G,\theta_{E_S}} L_{JT} \\ &= \arg\max_{\theta_D,\theta_G,\theta_{E_S}} (\lambda_d L_d + \lambda_q L_q + \lambda_{m1} L_{m1}) \end{aligned} \tag{12}$$

where the λ_d, λ_q, and λ_{m1} are the hyperparameters to control the relative weight of each item.

In SF-stage, the layers of the discriminator D and the embedding generator G are frozen. The stego image extractor E_I is optimized to minimize the distance between the extracted messages M_2 and the original message M. The overall loss L_{ST} is constructed only by the message loss L_{m2}, which is calculated by BCE loss as well, denoted as:

$$\begin{aligned} L_{ST} = L_{m2} &= BCE(M, M_2) \\ &= \frac{1}{CHW}\sum_{j=1}^{CHW} -m_j \log(m_{2_j}) - (1-m_j)\log(1-m_{2_j}) \end{aligned} \tag{13}$$

where m_j and m_{2_j} are the 0 and 1 elements of the original messages M and the extracted messages M_2, respectively.

IV. Experiments and Results

A. Experimental Setting

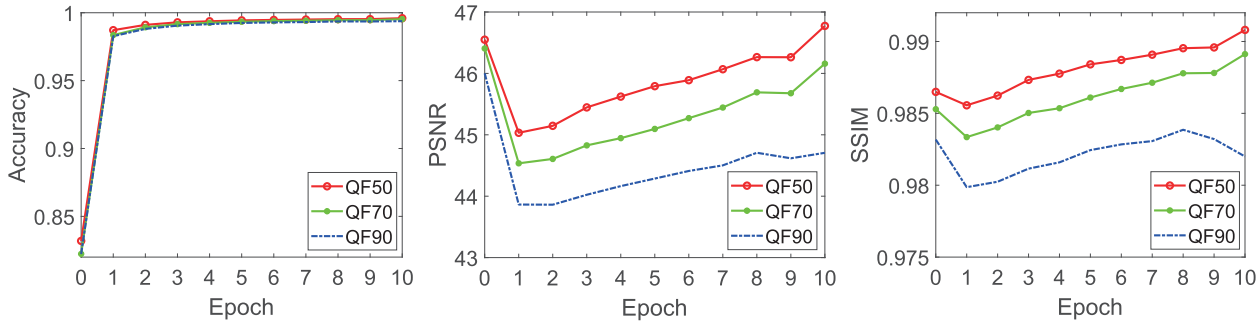
The experiments are conducted on the BossBase-1.01 image database [43]. To enrich the database, all images in BossBase-1.01 are compressed with quality factors (QF) 50, 70, 90. Then, the spatial pixel values are saved as the lossless PNG images to construct 3 cover databases (called 'QF50', 'QF70', and 'QF90' in the following) containing 10000 images, respectively. In each cover database, 9000 images are used for training while 1000 images are utilized for testing. The proposed network models are implemented by PyTorch and executed on NVIDIA GeForce RTX 2080 Ti. All images are gray-scale of size $C \times H \times W = 1 \times 512 \times 512$. In order to remain good image quality, the embedding rate in the proposed method is kept at 1 bpp. The length of message M_0 is also of size $C \times H \times W = 1 \times 512 \times 512$, which is randomly generated following the Bernoulli distribution with a 0.5 probability. A better trade-off between image quality and the accuracy of message extraction in JT-stage can be obtain by setting the weight factors $\lambda_d = 1$, $\lambda_q = 1$, and $\lambda_{m1} = 100$. The 3 components in JT-stage and the stego image extractor E_I are optimized by the Adam [44] optimizer with default hyperparameters. The learning rate decay process is utilized in the training of embedding generator G and 2 extractors. The initial learning rate is 10^{-3} for G, E_M, and E_I, which is decayed by 0.1 every 10 epochs. The learning rate for the discriminator is 10^{-4}. The 3 components in JT-stage are trained for 10 epochs with a batch of size 8. And the stego image extractor E_I in SF-stage is trained for 40 epoches with a batch of size 8.

The accuracy is denoted as $\frac{N_{acc}}{C \times H \times W}$, where N_{acc} denotes the number of the corrected extracted messages. The secure performances are measured by the steganalyzers using the ensemble classifier for training and testing to obtain the average classification error rate. The steganalysis methods used to evaluate statistical security are the SPAM (Subtractive Pixel Adjacency Matrix) [45] and SRNet [46]. To further evaluate the visual quality of stego images, two objective visual quality measures are employed, including the peak signal-to-noise ratio (PSNR) and structural similarity (SSIM). PSNR measures the modification between the cover image and the stego image. SSIM measures the luminance, contrast, and structure between 2 images. A high PSNR and a high SSIM mean better imperceptivity, and the stego image is perceived more similar to the cover image.

B. Ablation Study

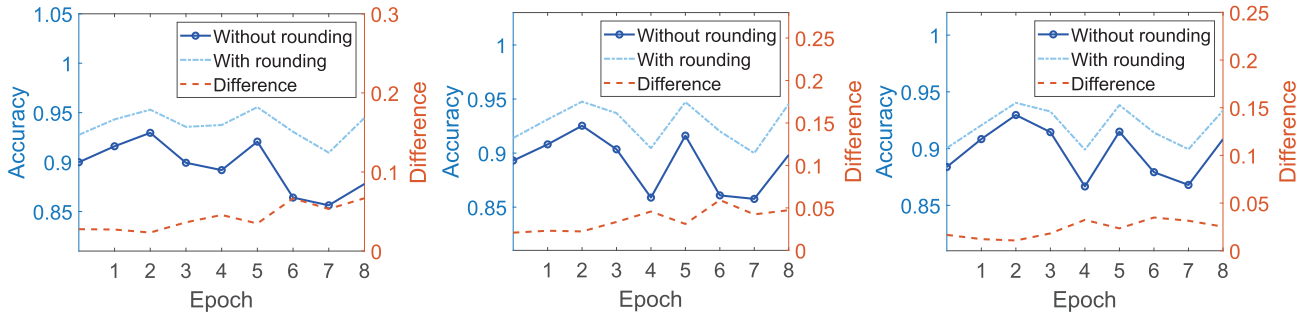
1) Impact of Rounding Operation: As mentioned in Section II, the precision loss brought by rounding the stego matrix X to the stego image I_S leads to the decline of the accuracy of message extraction. Firstly, we focus on the JT-stage to verify whether the accuracy performances of message extraction without rounding (inputting the stego matrix X) and that with rounding (inputting the stego image I_S) are different, so as to illustrate the robust challenge against the rounding operation of the trained extractor. The training performances of JT-stage in 3 different databases are shown in Fig. 7. In Fig. 7(a), starting from epoch 1, the accuracy of message extraction can reach more than 98% in 3 databases. And the accuracy plots are converged from epoch 5, which are greater than 99.3%. In Fig. 7(b) and Fig. 7(c), on the premise of maintaining high accuracy, PSNR and SSIM are gradually improved from epoch 1. It means that taking long training time, the accuracy performances and the visual qualities can be effectively increased in JT-stage. It also verifies the feasibility of the encoder-decoder steganographic network architecture [24], [28], [29] in the training.

When testing the trained models in JT-stage, it can be seen that the accuracy performances are decreased to 90% to 95% even without rounding operation in Fig. 8. Moreover, when the stego matrix X generated by the embedding generator G is rounding as the stego image I_S, and then the I_S are input into the trained stego matrix extractor E_M, the accuracy decreases



(a) (b) (c)

Fig. 7. The training performances of JT-stage in 3 different databases. (a) Accuracy of message extraction. (b) PSNR. (c) SSIM.



(a) (b) (c)

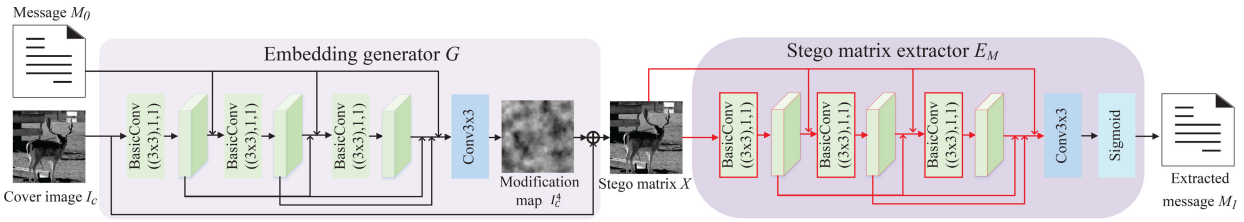
Fig. 8. The accuracy performances of message extraction in JT-stage without rounding and with rounding. Test in (a) QF50, (b) QF70, (c) QF90.

significantly. It shows that the robustness of the stego matrix extractor E_M to offset the precision loss needs to be improved. In addition, without rounding, epoch 2 and epoch 8 can achieve high 95% accuracy. However, the decline of accuracy in epoch 8 is significant after rounding. The red line in Fig. 8 represents the differences between the accuracy with rounding and the accuracy without rounding. With the increment of training, the differences are increasing, which shows that the longer-time training in JT-stage often makes the stego matrix extractor E_M more dependent on the precision. Thus, the accuracy performances with rounding are decreased greatly in longer-time trained models. Therefore, although the higher accuracy can be obtained using the stego matrix X in the stego matrix extractor E_M in JT-stage, the overfitting unfortunately reduces the robustness of the stego matrix extractor E_M against rounding operation. Experimentally, in the later SF-stage training, we fixed the embedding generator G and training stego image extractor E_I with better anti-rounding ability in shorter-time trained models.

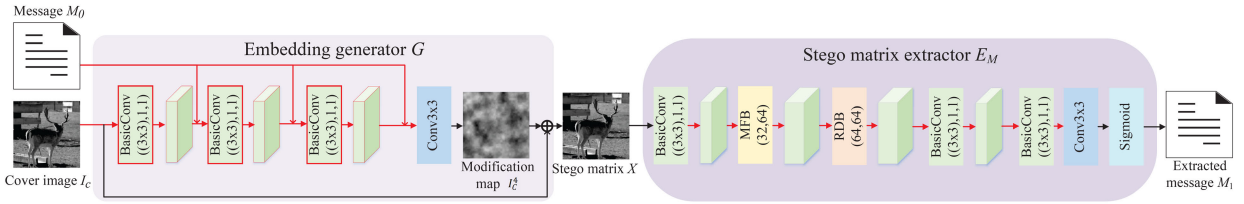
2) Impact of Different Structure: As shown in Fig. 5 and Fig. 6, different structures are designed between the generator (encoder) and the extractor (decoder). For some tasks such as image generation and image compression, the decoder is a reverse component for the encoder to assist the training of encoder. Generally, the parameters of the decoder are the inverse function for the parameters of the encoder. In this way, the structures of encoder and decoder are generally similar, while the parameters are opposite. In steganography, the senders use encoder, so-called generator, to embed messages into the cover image and generate the stego image. The receivers use decoder, so-called extractor, to extract messages from the stego image. Therefore, the tasks of the generator and the extractor are completely different. The task of the generator is to realistically fuse cover image and secret messages. The task of the extractor is to accurately extract the messages. It is noted that relying on image visual redundancy, a shallow CNN-based network can achieve the fusion task with good visual quality. However, the input of the extractor is only a stego image. In order to improve the accuracy of the extracted message, the design of the extractor with a deeper network structure is necessary.

In this section, the accuracy impact of the different structure in the proposed method on the accuracy of message extraction is discussed. To conduct the confirmatory experiments, the same structures of the embedding generator and the stego matrix/image extractor are presented Fig. 9(a). All other experimental settings are kept the same as the proposed model. The accuracy performances of Structure A in the QF50 test database are shown in Table II. Compared with the proposed method, Structure A achieves lower accuracy performances below 90% in testing in JT-stage with rounding. After training the SF-stage, the accuracy performances are improved, but they are obviously not as good as that of the proposed method. It is verified that the extractor, which keeps the same structure as the generator, cannot precisely capture the steganographic noise from the stego image.

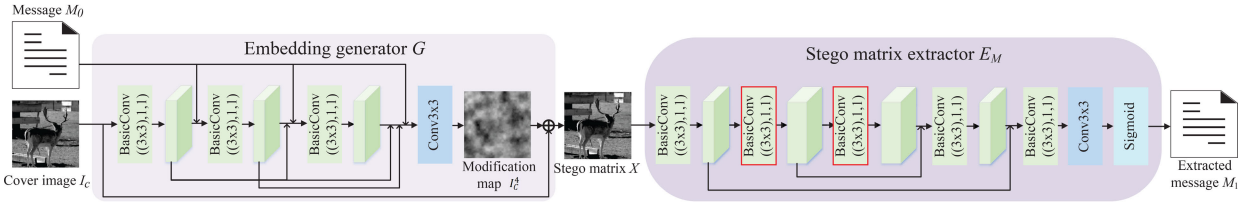
3) Impact of Concatenation: As shown in Fig. 5, the concatenations are used to fuse the cover image I_C and



(a) Structure A



(b) Structure B



(c) Structure C

Fig. 9. Three structures are conducted for ablation study. (a) Structure A: keeping the same structure between generator and extractor. (b) Structure B: without the concatenation in the embedding generator G. (c) Structure C: without the MFB and RDB in stego matrix extractor E_M and stego image extractor E_I.

messages M_0 in the embedding generator G. Similarly, the concatenations are utilized to obtain the extracted messages in the 2 extractors. In [24] and [29], the experimental results are verified that the concatenations in the designed models can realistically fuse cover image I_C and messages M_0, and the distances between the generated stego matrix X and the cover image I_C are effectively reduced. In this section, the accuracy impact of the concatenations in the proposed separable fine-tuned network on the accuracy of message extraction is discussed.

The detailed structures of the embedding generator and the stego matrix/image extractor with concatenations and without concatenations are presented in Table I and Fig. 9(b). Except that the concatenations input of layers are different, all other experimental settings are kept the same as the proposed model. The accuracy performances of Structure B in the QF50 test database are shown in Table II. When training the models, the accuracy in JT-stage is trained to over 99%. When testing in JT-stage without rounding, the accuracy performances of the model without concatenations are declined to about 92%. However, in JT-stage with rounding, the accuracy performances of the model without concatenations are decreased significantly. Moreover, as shown in Table II, the test results using the model without concatenations are reduced compared with the test results using the proposed model. It means that the robustness of the model to offset the rounding operation is weakened when the concatenations are removed. As mentioned in Section IV-B.1, the shorter-time trained models without concatenations are utilized in SF-stage. The test results in SF-stage are improved compared with that in JT-stage, manifesting that the accuracy performances are effectively

TABLE I

THE DETAILED STRUCTURES OF THE EMBEDDING GENERATOR AND THE STEGO MATRIX/IMAGE EXTRACTOR WITH CONCATENATIONS AND WITHOUT CONCATENATIONS

	Method	Input (channel)	Output (channel)
Embedding generator	Proposed	I_C (C)	I_C^1 (32)
		I_C^1,M_0 (32+C)	I_C^2 (32)
		I_C^1,I_C^2,M_0 (64+C)	I_C^3 (32)
		I_C^1,I_C^2,I_C^3,M_0 (96+C)	I_C^4 (C)
		I_C^4+I_C (1)	X (1)
	Without concatenation	I_C (C)	I_C^1 (32)
		I_C^1,M_0 (32+C)	I_C^2 (32)
		I_C^2,M_0 (32+C)	I_C^3 (32)
		I_C^3,M_0 (32+C)	I_C^4 (C)
		I_C^4+I_C (1)	X (1)
Stego matrix / image extractor	Proposed	X / I_S (C)	X^1 (32)
		X^1 (32)	X^2 (64)
		X^2 (64)	X^3 (64)
		X^3,X^2 (128)	X^4 (32)
		X^4,X^1 (64)	X^5 (1)
		X^5 (1)	X^6 (1)
		X^6 (1)	M_1 / M_2 (1)
	Without concatenation	X / I_S (C)	X^1 (32)
		X^1 (32)	X^2 (64)
		X^2 (64)	X^3 (64)
		X^3 (64)	X^4 (32)
		X^4 (32)	X^5 (1)
		X^5 (1)	X^6 (1)
		X^6 (1)	M_1 / M_2 (1)

TABLE II

ABLATION STUDY OF THE MESSAGE EXTRACTION ACCURACY COMPARED WITH STRUCTURE A (KEEPING THE SAME STRUCTURE) SHOWN IN FIG. 9(A), STRUCTURE B (WITHOUT CONCATENATIONS) SHOWN IN FIG. 9(B), STRUCTURE C (WITHOUT MFB AND RDB) SHOWN IN FIG. 9(C) IN QF50 TEST DATABASE

Test in JT-stage with rounding					Test in SF-stage				
Epoch	Structure A	Structure B	Structure C	Proposed	Epoch	Structure A	Structure B	Structure C	Proposed
0	0.8509	0.8901	0.8903	**0.8999**	4	0.8959	0.9072	0.9225	**0.9283**
1	0.8964	0.9075	**0.9192**	0.9160	9	0.9195	0.9191	0.9233	**0.9430**
2	0.8754	0.8934	0.9150	**0.9296**	14	0.9055	0.9170	0.9278	**0.9425**
3	0.8627	0.8905	**0.9049**	0.8993	19	0.9011	0.9149	0.9287	**0.9358**
4	0.8509	0.8405	**0.9122**	0.8919	24	0.9087	0.9194	0.9298	**0.9428**
5	0.8709	0.8315	**0.9234**	0.9207	29	0.8953	0.9146	0.9276	**0.9395**
6	0.8300	0.8668	**0.9245**	0.8641	34	0.8919	0.9163	0.9262	**0.9356**
7	0.8344	0.8561	**0.9219**	0.8563	39	0.8924	0.9210	0.9299	**0.9402**

improved to 92.1%. Although the embedding generator is frozen in SF-stage, the removed concatenations will not lead to the failure of the separably training of stego image extractor. It is verified that the separable fine-tuned network can be applied to different designs of encoder-decoder models and markedly compensate for the precision loss caused by rounding operation. However, removing the concatenations, the accuracy performances are not as good as that of the proposed models. It can be seen that the concatenations can improve the ability of the separately-trained stego image extractor to obtain the most dominant features which can resist the precision loss. Therefore, the proposed model with the concatenations can significantly improve the accuracy performances.

4) Impact of MFB and RDB: In Fig. 6, an MFB and an RDB are utilized to fully capture the features of stego matrix X and stego image I_S to generate the extracted message M_1 and M_2. The accuracy impact of these 2 blocks needs to be discussed. Shown in Fig. 9(c), we replace the MFB and RDB with 2 convolution layers with a kernel size of 3×3, a padding of 1, and a stride of 1. All other experimental settings are kept the same as the proposed model, including the concatenations in G, E_M, and E_I. The accuracy performances of Structure C in the QF50 test database are also shown in Table II. The accuracy in JT-stage is also trained to over 99% when training the models. The results in JT-stage with rounding show that based on the joint training, the precision loss also affects the stego matrix extractor E_M. The accuracy performances are greater than that of the proposed models with MFB and RDB, which are up to 90%. The MFB is utilized to capture the multi-scale features, while the RDB is used to capture local density features by the residual dense layer. The use of these two blocks undoubtedly deepens the model of stego matrix extractor. It is verified that even if the simple short and shallow network is designed in the JT-stage (one-stage), the extractor is less dependent on the precision with the effects of concatenations. However, in the SF-stage, the accuracy performances are just slightly improved which are 92.9%. The accuracy performances of the proposed models are significantly improved to 94%, which are also greater than that of the models without MFB and RDB in the JT-stage with rounding. With the network deepening, when the stego image extractor E_I is fine-tuned and trained separately, the MFB and RDB can effectively capture the anti-rounding features. In this way, the stego image extractor E_I can enhance the robustness to resist the precision loss.

C. Comparative Experiments

In this section, to discuss the robustness of the proposed anti-rounding separable fine-tuned network, we conduct the comparative experiments with the one-stage encoder-decoder model. CHAT-GAN [29] proposes an encoder for data embedding, a discriminator for steganalysis, and a decoder for data extractor. A novel channel attention module has been designed to obtain a stronger representation ability for stego generation or message recovery. ABDH [25] is a watermarking and steganographic scheme, including a target image generative model and a secret image generative model. The attention mechanism is designed to aware of the spotlights and the inconspicuous areas of cover images. The proposed SFTN network is compared with CHAT-GAN [29] and ABDH [25] on the accuracy of message extraction, visual quality, and the steganalysis security. The train database and test database are kept the same, as well as the size of secret messages.

The comparisons between the proposed SFTN and these two different one-stage encoder-decoder models are shown in Table III. First, we focus on whether the precision loss will lead to the failure of these encoder-decoder networks. Since the JT-stage of the proposed SFTN does not conduct the rounding operation when training the models, it can be regarded as a one-stage encoder-decoder network. Therefore, in Table III, the 'Accuracy of JT-stage (Without rounding)' represents the test results which inputs the generated stego matrix to the decoder, while the 'Accuracy of JT-stage (With rounding)' represents the test results which rounds the stego matrix as the stego image, and inputs the stego image to the decoder. The accuracy performances of three methods are trained to over 99%. By this way, the test results in JT-stage without rounding of the proposed SFTN outperform the other two networks in 3 databases. Both the accuracy performances of CHAT-GAN [29] and those of ABDH [25] are over 91%. Then we simulate the real steganography scenario to obtain the accuracy of JT-stage with rounding. Both the sender and the receiver share the trained encoder and the

TABLE III
COMPARISONS BETWEEN THE PROPOSED SFTN AND DIFFERENT STEGANOGRAPHIC METHODS USING SEPARABLE FINE-TUNED ARCHITECTURE

Database	Method	Accuracy of JT-stage (Without rounding)	Accuracy of JT-stage (With ounding)	Accuracy of SF-stage	Detection by SPAM [45]	Detection by SRNet [46]	PSNR	SSIM
QF50	ABDH [25]	0.9254	0.5134	0.5192	0.0366	0.0308	17.3794	0.4009
	CHAT-GAN [29]	0.9507	**0.9317**	0.9407	0.0582	0.0441	40.1284	0.9716
	Proposed	**0.9530**	0.9296	**0.9430**	**0.0584**	**0.0475**	**41.9413**	**0.9806**
QF70	ABDH [25]	0.9390	0.5117	0.5264	0.0224	0.0214	17.3726	0.3973
	CHAT-GAN [29]	0.9288	0.9088	0.9219	0.0504	0.0289	40.5393	0.9741
	Proposed	**0.9476**	**0.9253**	**0.9396**	**0.0530**	**0.0302**	**41.2754**	**0.9776**
QF90	ABDH [25]	0.9238	0.5125	0.5269	0.0206	0.0167	17.3274	0.3918
	CHAT-GAN [29]	0.9177	0.9041	0.9115	0.0483	**0.0283**	40.1749	**0.9705**
	Proposed	**0.9403**	**0.9296**	**0.9373**	**0.0487**	0.0282	**40.2685**	0.9697

trained decoder. The stego matrix generated by the encoder is saved as a PNG-format stego image, and then we send it to the receiver. The receiver extracts the secret messages using the shared trained decoder. The accuracy results of JT-stage with rounding are actually the test results of the one-stage encoder-decoder methods. However, the test results of ABDH [25] are significantly decreased. ABDH proposes an implicit attention mask to build a localizable representation to imitate the human attention mechanism. We find that the attention mask plays a crucial role to understand the sensitivity of cover image and generate a stego matrix (called target image in [25]). The sensitive details provided by the attention mask are destroyed after the stego matrix is rounding, which makes it difficult for the trained decoder to effectively extract the secret messages. Except for QF50 test database, the accuracy performances of the proposed SFTN are better than that of CHAT-GAN [29].

To further verify that the separable fine-tuned network can be applied to different designs of encoder-decoder models, we tried to utilize the SF-stage training for CHAT-GAN [29] and ABDH [25]. The stego image extractors in SF-stage have kept the same designs as their own decoder models. All layers of the trained embedding generator are copied and kept frozen during SF-stage. The attention model in ABDH [25] is also frozen while only the stego image extractor (called secret image generative model in [25]) is retrained in SF-stage. Then, the sender and the receiver share the frozen embedding generator and the stego image extractor. Similarly, the generated stego matrices are saved as the PNG-format stego images. The test results are presented in 'Accuracy of SF-stage' in Table III. Compared with the results in JT-stage with rounding, the accuracy performances of three methods are evidently improved. Moreover, the accuracy performances of CHAT-GAN [29] and the proposed SFTN are close to that of JT-stage without rounding. It is verified that the proposed SFTN can markedly compensate for the precision loss caused by rounding operation. The test results of the proposed SFTN are better than the other two methods using SF-stage, which shows that the model design of the proposed SFTN can achieve significant anti-rounding robustness.

The embedding generator of CHAT-GAN [29] and the proposed SFTN can obtain the residual features regarded as the modification map to determine the modification direction and the magnitude of pixel values. In traditional steganography methods, the messages are encoded by minimizing the distortion function. The modification map is calculated to guide the pixels to be added by 1 or subtracted by 1. In the proposed method, the modification map is truncated. The stego image I_S is defined as:

$$I_S = \lfloor Trunc_0^{255}[Trunc_{-10}^{10}(I_C^4) + I_C] \rceil \tag{14}$$

where $Trunc^b a$ is the truncation function to range value from a to b. If modification maps are truncated to -1, 0, and 1 just like the modification of the traditional steganography methods, it is disadvantageous for the deep learning methods which are sensitive to precision. The rounding and truncation $\lfloor Trunc_0^{255}[Trunc_{-1}^{1}(I_C^4) + I_C] \rceil \approx I_C^4$. Once the stego matrix is rounded, the precision of residual features will be almost destroyed. To achieve a trade-off between visual quality and accuracy, the modification maps are truncated by the factor 10, so the modification range of the pixels is from -10 to 10. The truncation with a factor 10 can convert some floating-point residual features into integers, so as to resist the precision loss caused by rounding.

However, facing the great changes of pixels, the security is still a challenge to resist the steganalysis methods which are based on feature statistics. The steganalyzers are trained on the three Bossbase-1.01 databases and then the test images are tested on the trained steganalyzers. The results show that the detection error rate of the three methods is less than 0.05. Since the labels of cover images and stego images cannot be obtained in the real scenario, the security of steganalysis still needs to be improved.

When concerning the visual qualities, the stego images of CHAT-GAN [29] and the proposed SFTN own PSNR results over 40 dB, while the SSIM results are greater than 0.96. Different from the design of the encoder in ABDH [25], the concatenations of multi-layer CNN feature maps are used in the embedding generator in CHAT-GAN [29] and the proposed SFTN. The deeper network design of the stego matrix / image extractor also ensures the reasonable selection of the

TABLE IV
THE CROSS-DATABASE COMPARISONS BETWEEN THE PROPOSED SFTN AND DIFFERENT STEGANOGRAPHIC METHODS USING SEPARABLE FINE-TUNED ARCHITECTURE

Database	Method	Accuracy of JT-stage (Without rounding)	Accuracy of JT-stage (With ounding)	Accuracy of SF-stage	PSNR	SSIM
QF50	ABDH [25]	**0.9804**	0.5164	0.5238	16.7710	0.5459
	CHAT-GAN [29]	0.9485	**0.9318**	0.9407	39.5193	0.9726
	Proposed	0.9473	0.9314	**0.9410**	**40.9000**	**0.9777**
QF70	ABDH [25]	**0.9881**	0.5145	0.5205	16.7705	0.5465
	CHAT-GAN [29]	0.9209	0.9013	0.9158	40.1308	0.9762
	Proposed	0.9405	**0.9235**	**0.9349**	**40.5035**	**0.9773**
QF90	ABDH [25]	**0.9692**	0.5152	0.5219	16.7612	0.5421
	CHAT-GAN [29]	0.9120	0.8984	0.9062	**39.7218**	**0.9727**
	Proposed	0.9363	**0.9294**	**0.9352**	39.5546	0.9691

embedding positions, which improves the visual quality of the stego image. Therefore, the proposed SFTN can achieve better visual quality results compared with the other 2 methods.

The cross-database experiments are conducted. The models are trained on 9000 images in Bossbase-1.01 database, while 1000 images in BOWS2 [47] are tested on the trained models. The cross-database comparisons between the proposed SFTN and these two different one-stage encoder-decoder models are shown in Table IV. The accuracy performances of three methods are trained to over 99% in Bossbase-1.01 database. The test results in JT-stage without rounding show that ABDH [25] performs better. The implicit attention mask effectively improves the accuracy when the rounding operation is ignored. Focusing on the accuracy of JT-stage with rounding and SF-stage, the proposed method can markedly resist and compensate for the precision loss in the cross-database tests. Similarly, the stego images in the proposed method maintain PSNR results about 40 dB and SSIM results about 0.97, which are considered as good objective image quality.

V. CONCLUSION

In this paper, we discussed the problem that the precision loss caused by rounding operation to generate the stego image in DL-based steganographic applications. And an anti-rounding image steganography method with separable fine-tuning network architecture is proposed. In JT-stage, the embedded generator and stego matrix extractor are trained without rounding operation to ensure the accuracy of message extraction. In SF-stage, the pretrained embedded generator is frozen, and the loss will not backpropagate in the embedded generator. In this way, the non-differentiability of rounding operation can be offset. Inside the design of models, the concatenations are utilized to enhance the authenticity of the stego image, and MFB and RDB are used to effectively extract the embedding features. In addition, based on GAN, the discriminator is constructed to ensure the steganalysis security. The separable fine-tuned network can be applied to different designs of one-stage encoder-decoder models. Experiments show that the proposed SFTN can markedly compensate for the precision loss caused by rounding operation, and achieve better image visual quality compared with the existing one-stage steganography method. In the near future, we will push forward the steganography framework based on deep learning, and further research the robustness in steganography applications. The trade-off between the message extraction accuracy, visual quality of stego images, and the security is still challenging.

REFERENCES

[1] V. Holub and J. Fridrich, "Designing steganographic distortion using directional filters," in *Proc. IEEE Int. Workshop Inf. Forensics Secur. (WIFS)*, Dec. 2012, pp. 234–239.
[2] W. Su, J. Ni, X. Hu, and J. Fridrich, "Image steganography with symmetric embedding using Gaussian Markov random field model," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 31, no. 3, pp. 1001–1015, Mar. 2021.
[3] V. Holub, J. Fridrich, and T. Denemark, "Universal distortion function for steganography in an arbitrary domain," *EURASIP J. Inf. Secur.*, vol. 2014, no. 1, pp. 1–13, 2014.
[4] W. Lu, L. He, Y. Yeung, Y. Xue, H. Liu, and B. Feng, "Secure binary image steganography based on fused distortion measurement," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 29, no. 6, pp. 1608–1618, Jun. 2019.
[5] W. Zhang, Z. Zhang, L. Zhang, H. Li, and N. Yu, "Decomposing joint distortion for adaptive steganography," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 27, no. 10, pp. 2274–2280, Oct. 2017.
[6] W. Su, J. Ni, X. Li, and Y.-Q. Shi, "A new distortion function design for JPEG steganography using the generalized uniform embedding strategy," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 28, no. 12, pp. 3545–3549, Dec. 2018.
[7] X. Qin, B. Li, S. Tan, W. Tang, and J. Huang, "Gradually enhanced adversarial perturbations on color pixel vectors for image steganography," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 32, no. 8, pp. 5110–5123, Aug. 2022, doi: 10.1109/TCSVT.2022.3148406.
[8] B. Li, W. Wei, A. Ferreira, and S. Tan, "ReST-Net: Diverse activation modules and parallel subnets-based CNN for spatial image steganalysis," *IEEE Signal Process. Lett.*, vol. 25, no. 5, pp. 650–654, May 2018.
[9] R. Zhang, F. Zhu, J. Liu, and G. Liu, "Depth-wise separable convolutions and multi-level pooling for an efficient spatial CNN-based steganalysis," *IEEE Trans. Inf. Forensics Security*, vol. 15, pp. 1138–1150, 2020.
[10] M. Chen, V. Sedighi, M. Boroumand, and J. Fridrich, "JPEG-phase-aware convolutional neural network for steganalysis of JPEG images," in *Proc. 5th ACM Workshop Inf. Hiding Multimedia Secur.*, Jun. 2017, pp. 75–84.
[11] N. Zhong, Z. Qian, Z. Wang, X. Zhang, and X. Li, "Batch steganography via generative network," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 31, no. 1, pp. 88–97, Jan. 2021.
[12] J. Yang, H. Zheng, X. Kang, and Y.-Q. Shi, "Approaching optimal embedding in audio steganography with GAN," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2020, pp. 2827–2831.

[13] R. Meng, Q. Cui, Z. Zhou, Z. Li, Q. M. J. Wu, and X. Sun, "High-capacity steganography using object addition-based cover enhancement for secure communication in networks," *IEEE Trans. Netw. Sci. Eng.*, vol. 9, no. 2, pp. 848–862, Mar. 2022.

[14] L. Zhou, G. Feng, L. Shen, and X. Zhang, "On security enhancement of steganography via generative adversarial image," *IEEE Signal Process. Lett.*, vol. 27, pp. 166–170, 2020.

[15] B. Sultan and M. A. Wani, "Multi-data image steganography using generative adversarial networks," in *Proc. 9th Int. Conf. Comput. Sustain. Global Develop. (INDIACom)*, Mar. 2022, pp. 454–459.

[16] A. A. Lopez-Hernandez, R. F. Martinez-Gonzalez, J. A. Hernandez-Reyes, L. Palacios-Luengas, and R. Vazquez-Medina, "A steganography method using neural networks," *IEEE Latin Amer. Trans.*, vol. 18, no. 3, pp. 495–506, Mar. 2020.

[17] Z. Fu, E. Li, X. Cheng, Y. Huang, and Y. Hu, "Recent advances in image steganography based on deep learning," *J. Comput. Res. Develop.*, vol. 58, no. 3, pp. 548–568, 2021.

[18] Y. Wang, K. Niu, and X. Yang, "Information hiding scheme based on generative adversarial network," *J. Comput. Appl.*, vol. 38, no. 10, p. 2923, 2018.

[19] F. Peng, G. Chen, and M. Long, "A robust coverless steganography based on generative adversarial networks and gradient descent approximation," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 32, no. 9, pp. 5817–5829, Sep. 2022, doi: 10.1109/TCSVT.2022.3161419.

[20] S. Li, D. Ye, S. Jiang, C. Liu, X. Niu, and X. Luo, "Anti-steganalysis for image on convolutional neural networks," *Multimedia Tools Appl.*, vol. 79, nos. 7–8, pp. 4315–4331, Feb. 2020.

[21] 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.

[22] J. Yang, D. Ruan, J. Huang, X. Kang, and Y.-Q. Shi, "An embedding cost learning framework using GAN," *IEEE Trans. Inf. Forensics Security*, vol. 15, pp. 839–851, 2020.

[23] 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.

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

[25] C. Yu, "Attention based data hiding with generative adversarial networks," in *Proc. AAAI Conf. Artif. Intell.*, vol. 34, no. 1, 2020, pp. 1120–1128.

[26] 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*, Oct. 2019, pp. 1509–1517.

[27] J. Hayes and G. Danezis, "Generating steganographic images via adversarial training," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2017, pp. 1951–1960.

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

[29] J. Tan, X. Liao, J. Liu, Y. Cao, and H. Jiang, "Channel attention image steganography with generative adversarial networks," *IEEE Trans. Netw. Sci. Eng.*, vol. 9, no. 2, pp. 888–903, Mar. 2021, doi: 10.1109/TNSE.2021.3139671.

[30] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Medical Image Computing and Computer-Assisted Intervention—MICCAI*. Cham, Switzerland: Springer, 2015, pp. 234–241.

[31] 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.

[32] V. Leon, T. Paparouni, E. Petrongonas, D. Soudris, and K. Pekmestzi, "Improving power of DSP and CNN hardware accelerators using approximate floating-point multipliers," *ACM Trans. Embedded Comput. Syst.*, vol. 20, no. 5, pp. 1–21, Sep. 2021.

[33] G. Toderici et al., "Variable rate image compression with recurrent neural networks," 2015, *arXiv:1511.06085*.

[34] J. Ballé, V. Laparra, and E. P. Simoncelli, "End-to-end optimization of nonlinear transform codes for perceptual quality," in *Proc. Picture Coding Symp. (PCS)*, 2016, pp. 1–5.

[35] L. Theis, W. Shi, A. Cunningham, and F. Huszár, "Lossy image compression with compressive autoencoders," 2017, *arXiv:1703.00395*.

[36] R. Shin and D. Song, "JPEG-resistant adversarial images," in *Proc. NeurIPS Workshop Mach. Learn. Comput. Secur.*, vol. 1, 2017, p. 8.

[37] S. Santurkar, D. Tsipras, A. Ilyas, and A. Madry, "How does batch normalization help optimization?" in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 31, 2018, pp. 1–11.

[38] L. D. Nguyen, D. Lin, Z. Lin, and J. Cao, "Deep CNNs for microscopic image classification by exploiting transfer learning and feature concatenation," in *Proc. IEEE Int. Symp. Circuits Syst. (ISCAS)*, May 2018, pp. 1–5.

[39] M. Huang, Z. Liu, L. Ye, X. Zhou, and Y. Wang, "Saliency detection via multi-level integration and multi-scale fusion neural networks," *Neurocomputing*, vol. 364, pp. 310–321, Oct. 2019.

[40] Y. Zhang, Y. Tian, Y. Kong, B. Zhong, and Y. Fu, "Residual dense network for image super-resolution," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 2472–2481.

[41] T. Wu, W. Ren, D. Li, L. Wang, and J. Jia, "JPEG steganalysis based on denoising network and attention module," *Int. J. Intell. Syst.*, vol. 37, no. 8, pp. 5011–5030, Aug. 2022, doi: 10.1002/int.22749.

[42] G. Xu, H.-Z. Wu, and Y.-Q. Shi, "Structural design of convolutional neural networks for steganalysis," *IEEE Signal Process. Lett.*, vol. 23, no. 5, pp. 708–712, May 2016.

[43] P. Bas, T. Filler, and T. Pevnỳ, "Break our steganographic system: The ins and outs of organizing BOSS," in *Proc. Int. Workshop Inf. Hiding*. Cham, Switzerland: Springer, 2011, pp. 59–70.

[44] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014, *arXiv:1412.6980*.

[45] T. Pevný, P. Bas, and J. Fridrich, "Steganalysis by subtractive pixel adjacency matrix," *IEEE Trans. Inf. Forensics Security*, vol. 5, no. 2, pp. 215–224, Jun. 2010.

[46] 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 2019.

[47] P. Bas and T. Furon. (Aug. 2019). *BOWS-2*. [Online]. Available: http://bows2.ec-lille.fr

Xiaolin Yin received the B.S. degree in software engineering from Sun Yat-sen University, China, in 2018, where she is currently pursuing the Ph.D. degree with the School of Computer Science and Engineering. Her research interests include multimedia security and data hiding.

Shaowu Wu received the B.S. degree in mathematics and applied mathematics from the Beijing University of Chemical Technology, China, in 2017, and the M.S. degree in mathematics from the Beijing University of Technology, China, in 2020. He is currently pursuing the Ph.D. degree with the School of Computer Science and Engineering, Sun Yat-sen University, China. His research interests include multimedia security and data hiding.

Ke Wang received the B.S. degree in information security from Sun Yat-sen University, China, in 2021, where she is currently pursuing the M.S. degree with the School of Computer Science and Engineering. Her research interests include multimedia security and data hiding.

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. He was a Research Assistant with The Hong Kong Polytechnic University from 2006 to 2007. 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 AI security.

Yicong Zhou (Senior Member, IEEE) received the B.S. degree in electrical engineering from Hunan University, Changsha, China, and the M.S. and Ph.D. degrees in electrical engineering from Tufts University, Medford, MA, USA.

He is currently a Professor and the Director of the Vision and Image Processing Laboratory, Department of Computer and Information Science, University of Macau, Macau, China. His research interests include image processing, computer vision, machine learning, and multimedia security. He is a fellow of SPIE and was recognized as a "Highly Cited Researcher" in Web of Science in 2020. He received the Third Price of Macao Natural Science Award as a Sole Winner in 2020 and a co-recipient in 2014. He was a recipient of the Best Editor Award for his contributions to *Journal of Visual Communication and Image Representation* in 2020. He has been the leading Co-Chair of Technical Committee on Cognitive Computing in the IEEE Systems, Man, and Cybernetics Society since 2015. He serves as an Associate Editor for IEEE TRANSACTIONS ON NEUTRAL NETWORKS AND LEARNING SYSTEMS, IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY, IEEE TRANSACTIONS ON GEOSCIENCE AND REMOTE SENSING, and four other journals.

Jiwu Huang (Fellow, IEEE) received the B.S. degree from Xidian University, Xian, China, in 1982, the M.S. degree from Tsinghua University, Beijing, China, in 1987, and the Ph.D. degree from the Institute of Automation, Chinese Academy of Sciences, Beijing, in 1998. He is currently a Professor with the College of Information Engineering, Shenzhen University, Shenzhen, China. His current research interests include multimedia forensics and security.