Robust Image Steganography via Color Conversion

Qi Li, Bin Ma, Xianping Fu, Xiaoyu Wang, Chunpeng Wang, *Member, IEEE*,
and Xiaolong Li, *Member, IEEE*

***Abstract*— In this paper, we propose a robust image steganography method utilizing color conversion, leveraging de-colorization and colorization models to achieve covert transmission of secret information. The motivation is to use color conversion of the stego image to conceal steganographic behavior. For the sender, secret information is embedded into the color cover image using a robust embedding algorithm based on quaternion exponent moments. The stego images are then de-colorized to obtain grayscale images, which can be transmitted over public channels. For the receiver, a corresponding colorization network is designed to reconstruct the stego image and extract the secret information. Additionally, an attack module using Gaussian noise is implemented to enhance the robustness of the proposed steganography. Given a color image, its grayscale version can be chosen from various options, making it difficult for attackers to detect steganographic activity as long as the generated grayscale image appears normal and meaningful. Extensive simulation results demonstrate the feasibility and scalability of the proposed steganography method.**

***Index Terms*— Image steganography, de-colorization and colorization, color conversion, robust embedding algorithm.**

I. Introduction

CURRENTLY, the most common methods for ensuring data security are to utilize encryption algorithms to transform sensitive data into meaningless ciphertext information [1], [2]. However, the encryption behavior can expose the existence of sensitive data, which makes it susceptible to attack. Consequently, steganography technology [3], [4], [5], which can both protect the content of secret information and conceal its transmission, is gaining increasing attention.

Steganography achieves covert communication by modifying the details of the cover image to embed secret information. Based on the modification priority of the cover image, steganography can be divided into non-adaptive and adaptive methods. Non-adaptive steganography, most notably represented by the LSB (Least Significant Bits)-based information embedding algorithm [6], [7], modifies image pixels with the same priority across different regions. In contrast, adaptive steganography assigns different embedding distortion costs to pixels in various regions of the cover image, embedding secret information into areas with relatively complex textures [8], [9]. Adaptive steganography based on STC (Syndrome-Trellis Code) [10] further minimizes cover image distortion while embedding secret information, with encoding performance that can even approach the rate-distortion bound. Subsequently, defining a reasonable distortion function became mainstream in the field of steganography, with representative methods such as HUGO [11], WOW [12], UNIWORD [13], and so on. In addition, steganography methods using JPEG images as cover continue to emerge, among which the most representative is the Jsteg proposed by Upham and others, which replaces LSB of DCT coefficients in JPEG images with secret information bits [14], [15]. The steganography based on secret information being carried out in the frequency domain. JPEG images has higher security due to the embedding task of

For the past few years, GAN (Generative Adversarial Network) has gained a great deal of attention due to its powerful image generation capabilities [16]. Therefore, many researchers have noted an opportunity between GAN and image steganography. In 2024, Mao et al. [4] proposed the robust generative video steganography, which can accomplish the embedding task of secret information using semantic feature. Alejandro et al. explored the characteristics of GAN in detail and demonstrated that it can improve the capabilities of spatial steganalysis methods. According to this discovery, they proposed having GAN learn how to adapt to an image, so that it can minimally alter the cover image to accomplish the embedding task of secret information. Sun et al. [17] proposed a synthesis-based generative steganography model, which can leverage features to generate images with specific styles and attributes, thereby concealing changes in the cover image. And the experimental results have proved that their scheme is very robust. Subsequently, RoSteALS [18] is proposed for applying steganography technology to real-world scenarios. In their scheme, frozen pre-trained autoencoders

Received 9 July 2024; revised 1 September 2024; accepted 18 September 2024. Date of publication 24 September 2024; date of current version 13 February 2025. This work was supported in part by the National Natural Science Foundation of China under Grant 62272255, Grant 62302249, Grant 62302248, and Grant 62261160653; in part by the National Key Research and Development Program of China under Grant 2021YFC3340602; in part by the Ability Improvement Project of Science and Technology small and medium-size enterprise (SMES) in Shandong Province under Grant 2022TSGC2485 and Grant 2023TSGC0217; in part by Shandong Post-Doctoral Science Foundation under Grant SDCX-ZG-202400316; in part by the Project of Jinan Introduction of Innovation Team under Grant 202228016; in part by the Project of Jinan City-School Integration Development under Grant JNSX2021030; in part by Taishan Scholar Program of Shandong under Grant tsqn202306251; in part by the Youth Innovation Team of Colleges and Universities under Grant 2022KJ124 and Grant 2023KJ131; and in part by the Qilu Integration Pilot Project of Science Education Industry under Grant 2023PY060. This article was recommended by Associate Editor Z. Qian. *(Corresponding authors: Bin Ma; Xiaoyu Wang; Chunpeng Wang.)*

Qi Li, Bin Ma, Xiaoyu Wang, and Chunpeng Wang are with the Key Laboratory of Computing Power Network and Information Security, Ministry of Education, Shandong Provincial Key Laboratory of Industrial Network and Information System Security, Shandong Computer Science Center (National Supercomputer Center in Jinan), and Shandong Fundamental Research Center for Computer Science, Qilu University of Technology (Shandong Academy of Sciences), Jinan 250000, China (e-mail: sddxmb@126.com; qluwxy@163.com; mpeng1122@163.com).

Xianping Fu is with the School of Information Science & Technology, Dalian Maritime University, Dalian 116026, China.

Xiaolong Li is with the School of Computer and Information Technology, Beijing Jiaotong University, Beijing 100044, China.

Digital Object Identifier 10.1109/TCSVT.2024.3466961

can be designed to free payload embeddings from learning the distribution of cover images. Compared to other state-of-the-art image steganography methods, RoSteALS has strong security and robustness. In addition, some encoder-decoder based image steganography methods have also been pioneering. Baluja et al. [19] proposed the deep steganography, which can hide a full-size color image within another color image of the same size. However, this model has some security problems such as leakage of secret information and distortion of stego image. Zhang et al. [20] proposed an invisible steganography based on generative adversarial network, which hides a grayscale image into the Y channel of the cover image to improve the visual quality of stego image. After reviewing the current steganography methods, we find that almost all of them focus on embedding secret information into the complex texture regions of the cover image for security. However, a significant security risk exists in these methods: if attackers can obtain the original cover images, they can distinguish the stego image from the cover image by analyzing the discrepancies in their histogram distributions. Consequently, embedding secret information into any region of the cover image will inevitably alter the cover image itself. If the original cover image is compromised, steganographic behavior will be easily detected.

In this paper, we propose a robust image steganography scheme, which introduces colorization and de-colorization techniques [21] into the transmission process of secret information. The design concept of the proposed image steganography can be seen in Fig. 1. The brief key contributions can be summarized as follows:

1) To the best of our knowledge, this is the first time that colorization and de-colorization models have been introduced to the field of image steganography. The motivation behind this approach is to exploit color conversion of the stego image to conceal the presence of steganographic behavior.

2) An attack module is incorporated to enhance the robustness of the proposed steganography scheme during the model's training process. Experimental results demonstrate that the model can successfully extract secret information even after being subjected to attacks such as Gaussian noise and salt-and-pepper noise.

3) Extensive experiments prove that color conversion can eliminate the influence of embedding operation for secret information on the appearance of cover images.

The rest of the paper is organized as follows: Section II introduces the related technologies utilized in this paper, including image steganography, image de-colorization and colorization and robust embedding algorithm. Section III mainly introduces the image steganography proposed in this paper. Section IV gives the introduction of experimental results and corresponding experimental analysis. Section V provides a summary of the proposed image steganography scheme and a prospect of future work.

II. Related Work

A. Image Steganography

The purpose of image steganography is to embed the secret information into the cover image in an invisible way. The original concept of image steganography is to design a suitable distortion function or manual extract features to accomplish the embedding task of secret information [22], [23]. Due to the selectable limitation of embedding regions, traditional image steganography has not made a breakthrough in embedding capacity for many years. In recent years, researchers have introduced CNN (convolutional neural networks) into the field of image steganography. Thanks to the powerful feature extraction ability of CNN, the embedding capacity of image steganography methods is greatly improved. In 2021, Lu et al. [24] proposed invertible steganography network (ISN), which regards the embedding and extracting process of image steganography as a pair of inverse problems. The parameter sharing properties can ensure the steg images and reconstructed secret images with high quality. Meanwhile, HiNet [25] is proposed to further improve the visual quality of generated stego image. In addition, a low-frequency loss is designed to hide the secret image in the wavelet domain of cover image. In 2022, Guan et al. [26] proposed deep invertible network for hiding multiple secret images, which is called DeepMIH. Different from the previous image steganography methods, DeepMIH regards the embedding and extraction of secret information as the forward and backward process of the same network, so it can easily accomplish the hiding task of multiple secret images at the same time. From the research work in recent years, we can see that the current image steganography mainly focuses on how to improve the embedding capacity but ignores security. Take the simplest example: if the original cover image is acquired by the attacker, the difference between the stego image and cover image can be distinguished from the histogram distribution. Therefore, the current image steganography based on deep learning still has great security risks.

B. Image De-Colorization and Colorization

Image de-colorization and colorization are very classical techniques in the field of computer vision [27], [28]. Intuitively, image de-colorization converts the color image to a grayscale image, which is a kind of image dimension reduction operation. At present, there is no unified evaluation standard for grayscale images. The most original grayscale processing method is to select a certain channel (RGB or YIQ color space) of a color image as the target grayscale image. It is obvious that the original color-to-grayscale transformation will lose most of the structural information of the color image. Bala et al. [29] proposed a color-to-grayscale method that can preserve the luminance information during the process of color conversion. Rasche et al. [30] designed a linear algorithm to reduce the size of the dataset with color quantization. Compared with the original color-to-grayscale transformations, the proposed algorithm has better real-time performance. Image colorization is the opposite of image de-colorization, which aims to render pleasing and meaningful color information into the target grayscale image. At first, colorization technology is utilized to render old pictures or movies. In recent years, with the rapid development of deep learning, deep neural networks have been applied in the

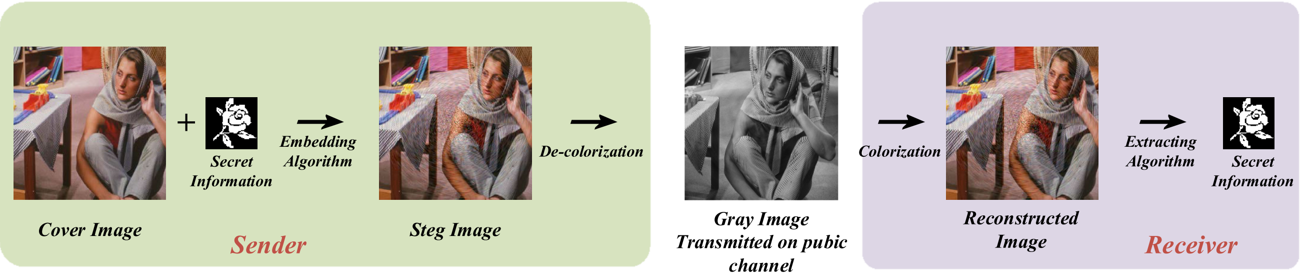


Fig. 1. The schematic diagram for proposed image steganography. Sender: for the sender, the cover image is modified twice to enhance the transmission security for secret information. First, with the help of embedding algorithm, secret information is embedded into the cover image to obtain stego image. Then the de-colorization algorithm is utilized to covert the stego image into the gray image, which can be transmitted on public channels. Receiver: for the receiver, the corresponding colorization algorithm is performed to reconstruct stego image. And the secret information can be extracted from the reconstructed stego image.

field of image colorization, which is called deep colorization. In general, deep colorization utilizes the pre-trained models to complete the coloring task for grayscale images. In [31], the pre-trained network is utilized to extract the color information. Xia et al. [21] proposed invertible grayscale (IG), in which de-colorization and colorization can be accomplished using encoder-decoder architecture at the same time. On the basis of invertible grayscale, JPEG robust invertible grayscale [32] is proposed. In their methods, adversarial loss is introduced to improve the visual quality of colorized images. A simulator of JPEG compression is designed to enhance the robustness of invertible grayscale. In the field of image steganography, the embedding operation of secret information can lead to the color distortion of cover images. Therefore, the security problem (the difference between cover image and stego image) of image steganography can be solved fundamentally by eliminating the color distortion of cover images using the de-colorization technology.

C. Robust Embedding Algorithms

Robust embedding algorithms are commonly used in the field of digital watermarking. Digital watermarking is a key technology of copyright protection for digital images. The identification information (digital watermark) is embedded in the image to be protected, so as to achieve the task of image copyright protection. And robust watermarking focuses on the fact that the identification information can still be extracted from the attacked watermarked images. In recent years, to resist geometric attacks, many robust watermarking algorithms based on geometric invariants [33], synchronous correction [34], local feature regions, or other strategies have been proposed. In addition, a variety of robust watermarking methods are designed to resist gradient attacks, sensitivity attacks, and disturbing attacks. With the development of deep learning, many researchers extend deep neural networks into the field of digital watermarking. Haribabu et al. [35] proposed a digital watermarking algorithm based on auto-encoder network in 2015. The basic idea of their algorithm is to use standard gradient descent backpropagation to learn the weights of auto-encoder network for imperceptible embedding operation of the watermark information. In 2020, Hao et al. [36] proposed a watermarking algorithm based on generative adduction network. The watermarked images obtained by their method have better visual effects and more advantages in anti-noise ability. In our opinion, the robust embedding algorithm is very suitable for the image steganography scheme proposed in this paper. Color conversion can be regarded as a kind of attack. Therefore, the robust embedding algorithm can contribute to the successful extraction of secret information.

III. The Proposed Method

The image steganography process via color conversion is illustrated in Fig. 2 and can be described in three stages: In the first stage, a robust embedding algorithm based on QEM (quaternion exponent moments) [37] is used to embed the secret image into the cover image, resulting in the stego image. In the second stage, to eliminate the embedding impact of the secret information on the cover image, a de-colorization network is employed to convert the stego image into a grayscale image. During training, we simulate Gaussian noise attacks on the grayscale images transmitted over public channels to enhance the robustness of the proposed steganography model. In the final stage, a corresponding colorization network is designed to restore the color information of the attacked stego images. Subsequently, the secret information can be extracted from the reconstructed stego image.

For information sender, the robust embedding method based on QEM is introduced to accomplish the steganography task of secret information. The embedding process for secret information can be described as follows:

$$I_t = Embed_{QEM}(I_c, I_s) \tag{1}$$

where I_s, I_c, and I_t represent the secret image, the cover image (original color image), and the stego image, respectively. In the second stage, the de-colorized network is utilized to conduct secondary operation on the stego image I_t, and the grayscale image I_g is obtained. This process can be expressed as follows:

$$I_g = D(I_t) \tag{2}$$

where D is the de-colorized network. The network architecture of the de-colorized model in Invertible Grayscale is utilized to accomplish the color conversion task of stego image. The purpose of de-colorized network is to implicitly encode the color information into the generated grayscale image to reconstruct stego image more effectively for the information receiver in the extraction stage.

In the transmitted process of generated grayscale image, a series of attacks are simulated to improve the robustness of

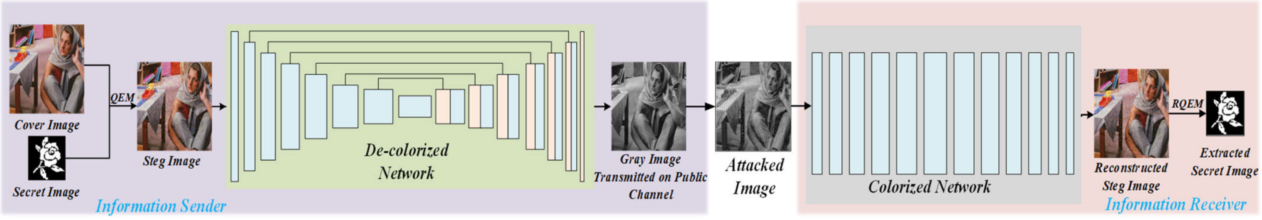


Fig. 2. The architecture for the proposed robust image steganography via color conversion. For information sender, the cover image is processed twice (robust embedding and de-colorization) to improve the security of image steganography. In the transmitted process, Gaussian noise is simulated to attack the grayscale image to enhance the robustness of image steganography. For information receiver, the colorized network is designed to reconstruct the stego image, and the secret image can be extracted from the reconstructed stego image.

the proposed image steganography model. Therefore, in the training process of de-colorized network and colorized network, an attack module A is added to interfere the de-colorized image, which can be expressed as follows:

$$I_a = A(I_g) \tag{3}$$

where I_a represents the attacked de-colorized image. For the information receiver, the main task is to reconstruct the color information of the stego image as much as possible. Therefore, the colorized model in Invertible Grayscale is also exploited to obtain the reconstructed stego image I_r, and this process can be described as follows:

$$I_r = C(I_a) \tag{4}$$

Then the secret image can be extracted from the reconstructed stego image I_r and the extraction algorithm based on QEM is utilized to complete the extracting task of secret image. The extracting process can be shown in Eq. (5):

$$I_{rs} = Extract_{RQEM}(I_r) \tag{5}$$

Note that the embedding and extraction operations of secret image do not participate in the training process of the de-colorized and colorized models. The de-colorized and colorized models are trained in an end-to-end manner. As can be seen from Fig. 2, the pixel-based MSE (Mean Squared Error) loss function is defined to minimize the difference between the original stego image I_t and the reconstructed stego image I_r in distribution and appearance, which can be expressed as follows:

$$\mathcal{L}_{MSE} = ||I_t - I_r||_2 \tag{6}$$

In the proposed steganography scheme, the grayscale image generated by the de-colorized network is transmitted in the common channel. Therefore, how to ensure that the generated grayscale is not recognized as the stego grayscale image is the key to guarantee the security of steganography scheme. For a color image, the corresponding grayscale version can be selected from a variety of options, such as a grayscale image with maximum value, grayscale image with average value, and single-channel grayscale image, et al. Therefore, as long as the generated grayscale image is normal and meaningful, it is difficult for the attacker to determine the existence of steganography behavior, because there are no fixed grayscale images to compare with the generated grayscale image. For obtaining normal and meaningful grayscale image, the luminance loss is firstly designed to reduce the difference between the generated grayscale image and the grayscale version of the stego image in luminance information.

$$\mathcal{L}_{lum} = ||\max|I_g - L(I_t)| - a, b||_1 \tag{7}$$

where L can extract the luminance channel of the stego image I_t and the values of a and b are set to 70 and 0, respectively. Additionally, a contrast loss function is proposed to further reduce the disparity between the generated grayscale image and the grayscale version of the stego image:

$$\mathcal{L}_{con} = ||E_{vgg19}(I_g) - E_{vgg19}(I_c)||_1 \tag{8}$$

where E_{vgg19} represents the pre-trained $VGG19$ network, and the corresponding feature maps of the generated grayscale image and the grayscale version of the stego image can be extracted using "conv4_4" layer in E_{vgg19}, which are selected as the representation of image contrast. The local loss function is defined to represent the local features:

$$\mathcal{L}_{loc} = ||Var(I_g) - Var(I_t)||_1 \tag{9}$$

In Eq. (9), the mean values of local variation for an image can calculated using the function $Var(\cdot)$. In addition, to ensure that each pixel of the generated grayscale image is stored in 8-bit integer bits, the quantization loss is also set to forcibly control the pixel values of the generated grayscale image. The quantization loss function is defined as follows:

$$\mathcal{L}_{qua} = ||\min_{d=0}^{255}\left\{\left|I_g - M_d\right|\right\}| \tag{10}$$

where min $(\cdot)$ represents the element minimum operator. M_d represents the matrix of the same size as the generated grayscale image, where all values in M_d are d. In summary, all the loss functions defined above are exploited to optimize the de-colorized model and colorized model to obtain the optimum parameters:

$$\mathcal{L}_{total} = \lambda_1\mathcal{L}_{mse} + \lambda_2\mathcal{L}_{lum} + \lambda_3\mathcal{L}_{con} + \lambda_4\mathcal{L}_{loc} + \lambda_5\mathcal{L}_{qua} \tag{11}$$

In Eq. (11), different λ is utilized to balance the different loss functions in the training process of de-colorized model and colorized model. And λ_1, λ_2, λ_3, λ_4, and λ_5 are set to 1.0, 1.0, 3.0, 10, and 1.0, respectively.

IV. Experimental Results and Analysis

In this section, a series of corresponding experiments are conducted to verify the performance and feasibility of our

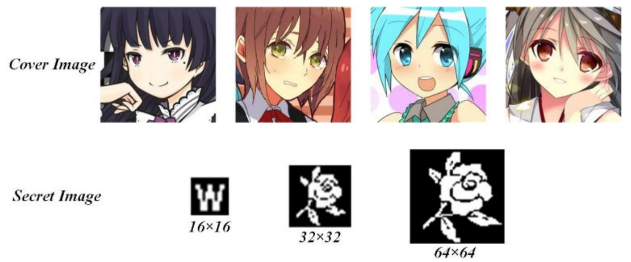


Fig. 3. Some representative cover images and secret images with different sizes.

TABLE I
THE EXPERIMENTAL RESULTS FOR THE AVERAGED PSNR AND SSIM BETWEEN THE ORIGINAL IMAGE AND THE STEGO IMAGE

16×16		32×32		64×64	
PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
39.5440	0.9881	34.0136	0.9364	28.4732	0.7865

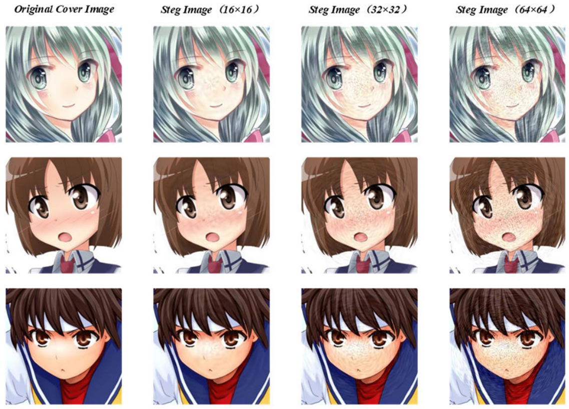


Fig. 4. Some stego images after embedding secret images with different sizes (16 × 16, 32 × 32 and 64 × 64).

proposed method. Our experimental codes are implemented with Pytorch 1.6 and Python 3.6 using the GPU NVIDIA GeForce Tesla V100 32G. The de-colorized and colorized models are trained using the images with the size of 256×256. We select 6000 training images randomly as the training dataset from the anime face dataset [38]. The sizes of secret images are set to 16 × 16, 32 × 32, and 64 × 64, respectively. Some representative cover images and secret images with different sizes are shown in Fig. 3.

A. Imperceptibility

Before decolorizing the original cover image, the robust embedding algorithm based on quaternion exponent moments is utilized to embed the secret image into the original cover image for obtaining the stego image. Some stego images after embedding secret images with different sizes are shown in Fig. 4. In addition, the experimental results for the averaged PSNR (Peak signal-to-noise ratio) and SSIM (Structure Similarity Index Measure) between the original covers image and the stego images are shown in Table I. From the experimental results in Fig. 4 and Table I, the larger the embedded secret images, the more serious the distortion of stego images, which also shows that the traditional embedding algorithms have great security risks.

For robust watermarking algorithms, the watermark information can be extracted from the attacked watermarked images. In fact, the reconstructed stego image also can be regarded as the attacked watermarked image. It is proved that the stego images based on quaternion exponent moments can resist various attacks (median filter, edge sharpening et al), that is, the secret information can be successfully extracted from the attacked stego image. So, we come to a conclusion: as long as the quality metrics of the reconstructed stego images are superior to the images attacked by the traditional interference methods, the secret information can also be extracted with a lower BER (bit error rate) from the reconstructed stego images. The experimental results on imperceptibility are shown in Fig. 5.

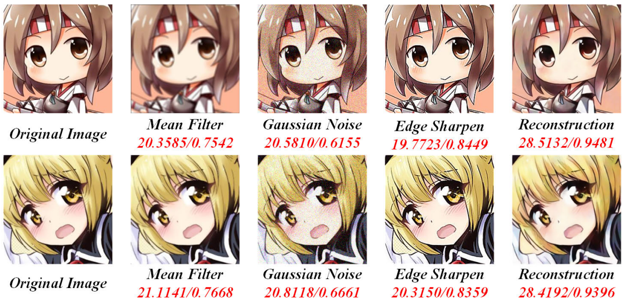


Fig. 5. The experimental results of the imperceptibility for the reconstructed stego images.

To more fully and objectively demonstrate the feasibility and superior imperceptibility of the proposed steganography, we calculated the average values of SSIM and PSNR for 200 attacked images under the condition of embedding secret images with different sizes (16 × 16, 32 × 32 and 64 × 64) and different attack methods, and the specific experimental results can be seen in Table II.

From the experimental results in Table II, compared with the stego images obtained by traditional attack methods, the reconstructed stego images after conducting the de-colorized and colorized operation have obvious advantages in SSIM and PSNR (image quality), which also indirectly indicates that the proposed image steganography scheme can accomplish the covert transmission task for the secret information. In addition, we have also observed a phenomenon: the larger the embedded information, the lower the visual quality of the attacked image.

B. Robustness

Robustness is also crucial to the image steganography methods. Although the steganography scheme proposed in this paper can resist the detection of steganalysis methods, the stego images can not completely avoid some attacks in the transmitted process, such as filtering, noise, and so on. Once the transmitted image is attacked, the information receiver is likely to be unable to extract the secret information from the attacked image, which means that the steganography model cannot accomplish the covert transmission task of secret information. Therefore, in this paper, an attack module is added to improve the robustness of the proposed image steganography scheme in the training process. The corresponding experimental results for robustness verification are shown in Fig. 6.

In this paper, the simulated attack is Gaussian noise whose mean value is 0 and variance is 0.1. As shown in Fig. 6,

TABLE II

EXPERIMENTAL COMPARISON RESULTS OF THE STEGO IMAGES AFTER BEING ATTACKED AND SUBJECTED TO DE-COLORIZATION AND COLORIZATION

Attack	16×16		32×32		64×64	
	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
Mean Filter	27.4167	0.9316	27.1216	0.9122	26.8311	0.8900
Gaussian Noise	20.9566	0.6344	20.3443	0.6195	19.8512	0.6214
Salt & pepper Noise	24.5455	0.9102	24.1228	0.8850	23.8634	0.8617
Edge Sharpening	20.6840	0.8333	20.4766	0.8128	20.0875	0.7977
Median Filter + Gaussian Noise	19.8465	0.6021	19.2908	0.5944	18.9522	0.5836
Salt & pepper Noise + Gaussian Filter	27.8521	0.9050	27.4479	0.8854	27.1050	0.8767
Edge Sharpening + Gaussian Noise	19.0577	0.5904	18.5455	0.5245	18.0126	0.5353
JPEG 70 + Salt & pepper Noise	24.3443	0.8508	24.0257	0.8167	23.4565	0.8204
JPEG 70 + Median Filter	27.9077	0.9369	26.3955	0.9208	26.1221	0.9125
JPEG 70 + Mean Filter	26.9612	0.9378	26.6707	0.9255	26.1475	0.8870
JPEG 70 + Gaussian Noise	19.4545	0.6123	19.0422	0.6064	18.9942	0.6109
JPEG 70 + Edge Sharpening	20.7745	0.8127	20.6032	0.7855	19.9100	0.7316
De-colorization and Colorization	**31.1312**	**0.9658**	**30.6054**	**0.9522**	**29.7125**	**0.9403**

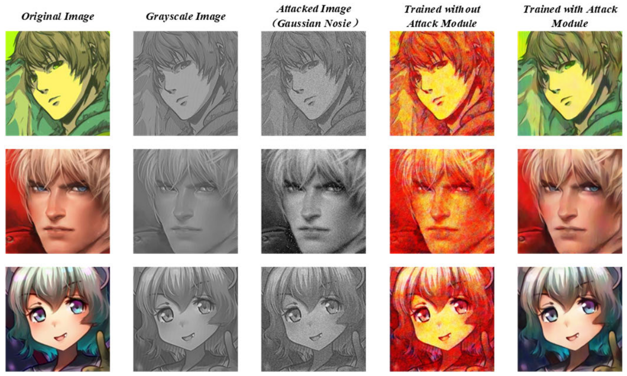

Original Image
Grayscale Image
Attacked Image (Gaussian Nosie)
Trained without Attack Module
Trained with Attack Module

Fig. 6. The experimental comparison results of reconstructed images obtained from the models trained with or without attack module.

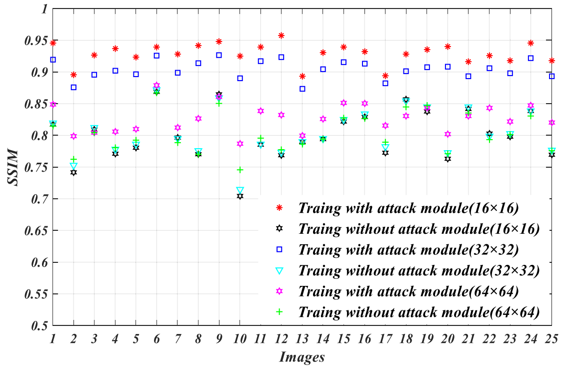


Fig. 7. The SSIM values of reconstructed stego images (embed secret images with different sizes (16 × 16, 32 × 32, and 64 × 64)) obtained from the models trained with or without the attack module.

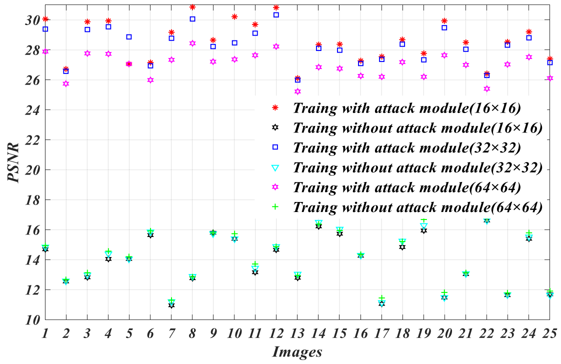


Fig. 8. The PSNR values of reconstructed stego images (embed secret images with different sizes (16 × 16, 32 × 32, and 64 × 64)) obtained from the models trained with or without the attack module.

the fourth column represents the reconstructed original images obtained from the trained model without attack module. From the experimental results, compared with original images, the reconstructed original images obtained from the trained model without attack module have obvious color distortion. the fifth column represents the reconstructed original images obtained from the trained de-colorized and colorized models with attack module. The experimental results in the fifth column show that the robust image steganography model can better reconstruct the color information of the original images.

To show the effect of the attack module more intuitively, we calculated the corresponding PSNR and SSIM (25 test images) of the reconstructed images with or without attack module. The experimental results are shown as follows:

By analyzing the experimental results in Fig. 7 and Fig. 8, it can be found that the visual quality (SSIM and PSNR) of steg images reconstructed from the de-colorized and colorized models with attack module is significantly higher than that of the model without attack module. This phenomenon also shows that the model with the attack module is more robust. From the detailed analysis of the above experimental results, it can be seen that the reconstruction effect of stego images with secret image of 16 × 16 is the best, which also indicates that the size of the secret information is inversely proportional to the robustness of the proposed image steganography scheme.

C. Security

Security is the key to the proposed image steganography in this paper. First, we compared the de-colorized image with other grayscale images, and the experimental results can be seen in Fig. 9. In-depth analysis, compared with other single-channel grayscale images, the obtained de-colorized grayscale image has better contrast and structure. For a color image, the corresponding grayscale version can be selected from a variety of options, such as a grayscale image with maximum value, grayscale image with average value, and single-channel grayscale image et al. Therefore, as long as the generated grayscale image is normal and meaningful,

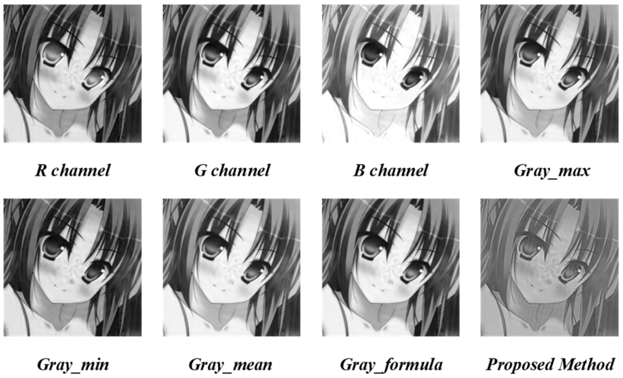


Fig. 9. The experimental comparison results for different grayscale images.

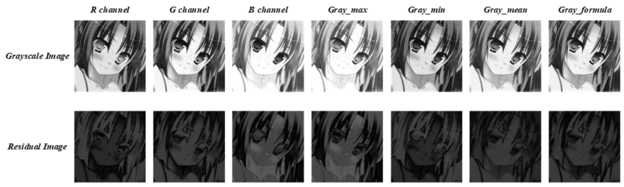


Fig. 10. The residual images between different grayscale cover images and corresponding de-colorized stego image.

it is difficult for the attacker to determine the existence of steganography behavior, because there are no fixed grayscale images to compare with the generated grayscale image.

For image steganography based on deep learning, there is not only image distortion, but even secret information leakage, such as deep steganography. Therefore, a similar experiment is conducted to verify whether the proposed image steganography has secret information leakage, and the corresponding experimental results are shown in Fig. 10. And from the experimental results in Fig. 10, we can see that there is only small amount of edge information of stego images in the residual image (linear processing between different grayscale cover images and corresponding de-colorized stego image) and no secret information leakage occurs. The above experimental results show that the proposed steganography scheme has high security.

To further investigate the security of the proposed steganography scheme, SRNet and YeNet are utilized to verify the anti-steganalysis ability. In terms of the security of this scheme, the main verification is which of the color image embedded with secret information and the generated grayscale image has stronger anti-steganalysis ability. Since the types of stego images include both color images and grayscale images in the verification process, the input dimensions of the steganalysis networks are correspondingly changed during the training process.

As shown in the experimental results in Table III, compared with the color images, the generated grayscale images obviously have greater advantages in anti-steganalysis. In addition, when the size of the secret image rises to 64×64, neither color nor grayscale images can avoid detection by SRNet and YeNet, which indicates the proposed image steganography method is more suitable for covert transmission of secret information with a small capacity. Therefore, it is not difficult to see that enhancing the transmitted capacity of the proposed image steganography method is also one of the main focuses of our future research.

TABLE III
The Detection Accuracy of SRNet and YeNet for the Proposed Scheme

	Color images			Grayscale images		
	16×16	32×32	64×64	16×16	32×32	64×64
SRNet	51.6%	89.3%	100%	28.3%	57.0%	100%
YeNet	57.3%	92.6%	100%	34.0%	65.3%	100%

D. Ablation Studies

To improve the feasibility of the proposed steganography scheme, four different image datasets are created for training the de-colorized and colorized models. In general, the training dataset to a large extent can determine the performance of the model. The four-training datasets include the original images without embedding secret image, the images embedded with the secret image with the size of 16×16, the images embedded with the secret image with the size of 32×32, and the images embedded with the secret image with the size of 64×64. The specific details about different training images can be seen in Fig. 4. For our experiments in this section, 20 test images are randomly selected to evaluate the optimal image dataset. In the experiments, we selected three representative evaluation indicators (PSNR, SSIM, and BER) to serve as base-lines for comparison, among which, PSNR and SSIM can evaluate the perceptual similarity between the original stego images and the reconstructed stego images, while BER can be utilized to estimate the impact of image dataset on the secret information extraction. The experimental results of SSIM, PSNR, and BER obtained by the proposed steganography using different training datasets are shown in Table IV.

For the setup of the experiments, the image datasets are determined as the only variable in the training process of the proposed image steganography. The following conclusions can be drawn from the experimental results in Table III. When training the image steganography model using the images embedded with the secret image with the size of 64×64, the visual quality of the reconstructed images and extracted ability of secret information can both reach optimal levels. Through in-depth analysis of the experimental results in Table IV, it is observed that the images embedded with the secret image with the size of 64×64 have severe distortion, and using it as the training dataset can enhance the generalization ability of the proposed image steganography model, thereby achieving the best experimental results. From the experimental results in Table IV, there is little difference in the reconstruction performance of the steganography models trained with different training image datasets, which indicates that the training process of the proposed method does not need to create a specific image dataset. Therefore, the experimental results in Table IV also verify from the side that the proposed scheme has strong generalization.

To further evaluate the feasibility of the proposed image steganography scheme, we selected some representative images and present the corresponding experimental results in Fig. 11. The first column shows the original stego image,

TABLE IV
THE EXPERIMENTAL RESULTS OF PSNR, SSIM AND BER BY USING DIFFERENT IMAGE DATASETS TO TRAIN THE MODEL

The training data	PSNR			SSIM			BER		
	16×16	32×32	64×64	16×16	32×32	64×64	16×16	32×32	64×64
The original images without embedding secret image	28.4346	28.4554	27.7491	0.9533	0.9578	0.9418	0.1382	0.1212	0.1576
The images embedded with the secret image with the size of 16×16	27.3677	28.1038	27.2953	0.9425	0.9545	0.9494	0.1372	0.1129	0.1495
The images embedded with the secret image with the size of 32×32	28.6133	28.7667	27.6900	0.9521	0.9623	0.9605	0.1306	0.1042	0.1555
The images embedded with the secret image with the size of 64×64	**28.8536**	**29.7535**	**28.2266**	**0.9623**	**0.9711**	**0.9541**	**0.1276**	**0.0813**	**0.1464**

TABLE V
EXPERIMENTAL COMPARISON RESULTS OBTAINED FROM DIFFERENT METHODS

Methods	PSNR			SSIM			BER		
	16×16	32×32	64×64	16×16	32×32	64×64	16×16	32×32	64×64
Gaussian noise	20.6167	20.6202	20.6036	0.6244	0.6398	0.6853	0.1758	0.1231	0.1985
Median filter + Gaussian noise	20.0435	20.0705	19.8679	0.5656	0.5779	0.5801	0.1836	0.1660	0.2664
JPEG compression + Gaussian noise	20.5969	20.5695	20.4264	0.6239	0.6363	0.6689	0.1797	0.1613	0.2026
Self-contained Stylization [39]	13.2770	12.8508	12.6007	0.5302	0.5008	0.4123	0.5234	0.4482	0.4551
Cycle-GAN [40]	18.3145	18.9263	17.5152	0.7314	0.7823	0.7078	0.2534	0.2177	0.2852
Style Removal [41]	26.1054	25.8804	24.3433	0.8332	0.8465	0.8297	0.1526	0.0906	0.1385
Proposed Method	**31.1312**	**0.9658**	**30.6054**	**0.9522**	**29.7125**	**0.9403**	**0.1086**	**0.0830**	**0.1395**

the second column displays the generated grayscale image, and the third column presents the reconstructed stego image. The fourth and fifth columns depict the original secret image and the extracted secret image, respectively. According to the experimental results, although the secret image cannot be extracted from the reconstructed stego image in a lossless manner, the key contents of the secret image can be perfectly reconstructed, indicating the high feasibility of the proposed scheme.

From the experimental results in Fig. 11, we can see that when the size of the secret image is 16 × 16, the BER of the reconstructed secret image is 0.1006. As the size of the secret image increases to 32 × 32, the BER decreases to 0.0686. However, when the size of the secret image is further increased to 64 × 64, the BER reaches its maximum value of 0.1516. It can be concluded that with the increase in the size of the secret image, the BER initially decreases and then increases. In fact, the phenomenon above is easily explainable. The smaller the size of the secret image, the less impact of its embedding operation has on the appearance and distribution of the stego image. Consequently, only when the detailed information of the stego image is reconstructed well, the secret image can be potentially extracted in its entirety. This also indicates that an appropriate embedding capacity can strike a balance between the extraction of the secret image and the reconstruction effect of the stego image.

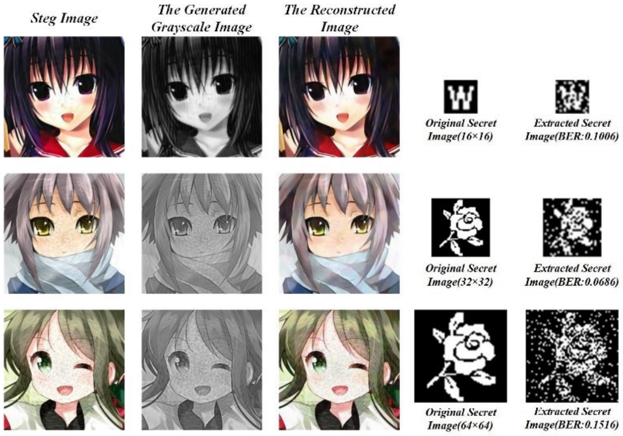

Steg Image
The Generated Grayscale Image
The Reconstructed Image
Original Secret Image(16×16)
Extracted Secret Image(BER:0.1006)
Original Secret Image(32×32)
Extracted Secret Image(BER:0.0686)
Original Secret Image(64×64)
Extracted Secret Image(BER:0.1516)

Fig. 11. The experimental results for grayscale images, reconstructed images, and extracted secret images (including different sizes and corresponding BERs).

E. The Comparison With the Different Methods

To evaluate the extraction ability and the feasibility of the proposed image steganography, we first compare the

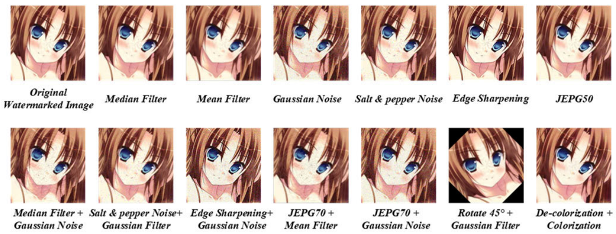

Original Watermarked Image
Median Filter
Mean Filter
Gaussian Noise
Salt & pepper Noise
Edge Sharpening
JEPG50
Median Filter + Gaussian Noise
Salt & pepper Noise+ Gaussian Filter
Edge Sharpening+ Gaussian Noise
JEPG70 + Mean Filter
JEPG70 + Gaussian Noise
Rotate 45° + Gaussian Filter
De-colorization + Colorization

Fig. 12. The experimental results obtained from different attacked stego images for visual quality.

reconstructed image obtained from the colorized model with several traditional attack methods, and the experimental results are shown in Fig. 12. As can be seen from the visual quality comparison results in Fig. 12, compared with the traditional attack methods, the stego image reconstructed from the colorized model proposed in this paper has obvious advantages in visual quality. In addition, to show the advantages of the proposed image steganography more intuitively, we conduct the experiments to calculate the PSNR, SSIM, and BERs obtained by traditional attacks and the mainstream deep learning-based methods and the comparison experimental results can be seen in Table V.

From the experimental results in Table V, we can conclude that: No matter whether compared with the traditional attack methods or current mainstream deep learning-based methods, the PSNR, SSIM, and BERs obtained by the method proposed in this paper have achieved considerable advantages, which means that the proposed image steganography can accomplish the covert transmission task of secret information safely.

V. Conclusion and Future Work

In this paper, we propose a robust image steganography method via color conversion. Unlike current image steganography schemes, our method modifies cover images twice to eliminate the embedding effect of secret information. Initially, the secret information is embedded into the cover image to obtain the stego image, as in traditional steganography. Subsequently, the stego image is de-colorized to produce a corresponding grayscale image, which can be transmitted over public channels. For the information receiver, a colorization network is designed to reconstruct the stego image, allowing the extraction of the secret image from the reconstructed stego image. Throughout the steganography process, the stego image undergoes de-colorization and colorization, which can be considered as an attack. Therefore, the secret information embedding algorithm is based on quaternion exponent moments, commonly used in robust watermarking. Additionally, an attack module employing Gaussian noise is designed to enhance the robustness of the proposed image steganography. Theoretically, since there is no fixed choice for grayscale images, the proposed method can fundamentally resist steganalysis detection. To our knowledge, this is the first time that deep-learning-based de-colorization and colorization technologies have been introduced to the field of image steganography.

The proposed steganography scheme currently still have issues such as limited options for secret information and restricted application scenarios. In future work, we will focus on improving the network architecture to enhance the visual quality of reconstructed stego images and further ensure the integrity of the extracted secret information.

References

[1] Y. Xian, X. Wang, and L. Teng, "Double parameters fractal sorting matrix and its application in image encryption," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 32, no. 6, pp. 4028–4037, Jun. 2022.

[2] S. Gao et al., "Asynchronous updating Boolean network encryption algorithm," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 33, no. 8, pp. 4388–4400, Aug. 2023.

[3] Z. Qian and X. Zhang, "Reversible data hiding in encrypted images with distributed source encoding," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 26, no. 4, pp. 636–646, Apr. 2016.

[4] X. Mao et al., "From covert hiding to visual editing: Robust generative video steganography," 2024, *arXiv:2401.00652*.

[5] X. Wang, X. Wang, B. Ma, Q. Li, and Y.-Q. Shi, "High precision error prediction algorithm based on ridge regression predictor for reversible data hiding," *IEEE Signal Process. Lett.*, vol. 28, pp. 1125–1129, 2021.

[6] T. Qiao, X. Luo, T. Wu, M. Xu, and Z. Qian, "Adaptive steganalysis based on statistical model of quantized DCT coefficients for JPEG images," *IEEE Trans. Dependable Secure Comput.*, vol. 18, no. 6, pp. 2736–2751, Nov. 2021.

[7] Z. Zhang, G. Fu, R. Ni, J. Liu, and X. Yang, "A generative method for steganography by cover synthesis with auxiliary semantics," *Tsinghua Sci. Technol.*, vol. 25, no. 4, pp. 516–527, Aug. 2020.

[8] P. Wei, S. Li, X. Zhang, G. Luo, X. Qian, and Q. Zhou, "Generative steganography network," in *Proc. 30th ACM Int. Conf. Multimedia*, vol. 2022, pp. 1621–1629.

[9] X. Hu, S. Li, Q. Ying, W. Peng, X. Zhang, and Z. Qian, "Establishing robust generative image steganography via popular stable diffusion," *IEEE Trans. Inf. Forensics Security*, vol. 19, pp. 8094–8108, 2024.

[10] T. Filler, J. Judas, and J. Fridrich, "Minimizing additive distortion in steganography using syndrome-trellis codes," *IEEE Trans. Inf. Forensics Security*, vol. 6, no. 3, pp. 920–935, Sep. 2011.

[11] T. Pevný, T. Filler, and P. Bas, "Using high-dimensional image models to perform highly undetectable steganography," in *Proc. Int. Conf. Inf. Hiding (IH)*, 2010, pp. 161–177.

[12] 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.

[13] V. Holub, J. Fridrich, and T. Denemark, "Universal distortion function for steganography in an arbitrary domain," *EURASIP J. Inf. Secur.*, vol. 2014, no. 1, p. 1, Dec. 2014.

[14] Q. Li, B. Ma, X. Wang, C. Wang, and S. Gao, "Image steganography in color conversion," *IEEE Trans. Circuits Syst. II, Exp. Briefs*, vol. 71, no. 1, pp. 106–110, Jan. 2024.

[15] 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.

[16] I. J. Goodfellow et al., "Generative adversarial networks," *Commun. ACM*, vol. 63, no. 11, pp. 139–144, 2020.

[17] Y. Sun, J. Liu, and R. Zhang, "Generative image steganography based on guidance feature distribution," *ACM Trans. Multimedia Comput., Commun., Appl.*, vol. 20, no. 11, pp. 1–18, Nov. 2024.

[18] T. Bui, S. Agarwal, N. Yu, and J. Collomosse, "RoSteALS: Robust steganography using autoencoder latent space," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2023, pp. 933–942.

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

[20] 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.

[21] M. Xia, X. Liu, and T.-T. Wong, "Invertible grayscale," *ACM Trans. Graph.*, vol. 37, no. 6, p. 246, Dec. 2018.

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

[23] X. Wang, X. Wang, B. Ma, Q. Li, C. Wang, and Y. Shi, "High-performance reversible data hiding based on ridge regression prediction algorithm," *Signal Process.*, vol. 204, Mar. 2023, Art. no. 108818.

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

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

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

[27] S. Liu, "Two decades of colorization and decolorization for images and videos," 2022, *arXiv:2204.13322*.

[28] K. Kim, C. Lee, and H.-J. Lee, "A sub-pixel gradient compression algorithm for text image display on a smart device," *IEEE Trans. Consum. Electron.*, vol. 64, no. 2, pp. 231–239, May 2018.

[29] R. Bala and R. Eschbach, "Spatial color-to-grayscale transform preserving chrominance edge information," in *Proc. Color Imag. Conf.*, 2004, pp. 82–86.

[30] K. Rasche, R. Geist, and J. Westall, "Detail preserving reproduction of color images for monochromats and dichromats," *IEEE Comput. Graph. Appl.*, vol. 25, no. 3, pp. 22–30, May/Jun. 2005.

[31] B. Duinkharjav, K. Chen, A. Tyagi, J. He, Y. Zhu, and Q. Sun, "Color-perception-guided display power reduction for virtual reality," *ACM Trans. Graph.*, vol. 41, no. 6, p. 210, Dec. 2022.

[32] K. Liu et al., "JPEG robust invertible grayscale," *IEEE Trans. Vis. Comput. Graphics*, vol. 28, no. 12, pp. 4403–4417, Dec. 2022.

[33] T. Zong, Y. Xiang, I. Natgunanathan, S. Guo, W. Zhou, and G. Beliakov, "Robust histogram shape-based method for image watermarking," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 25, no. 5, pp. 717–729, May 2015.

[34] X. Liu, G. Han, J. Wu, Z. Shao, G. Coatrieux, and H. Shu, "Fractional Krawtchouk transform with an application to image watermarking," *IEEE Trans. Signal Process.*, vol. 65, no. 7, pp. 1894–1908, Apr. 2017.

[35] K. Haribabu, G. R. K. S. Subrahmanyam, and D. Mishra, "A robust digital image watermarking technique using auto encoder based convolutional neural networks," in *Proc. IEEE Workshop Comput. Intell., Theories, Appl. Future Directions (WCI)*, Dec. 2015, pp. 1–6.

[36] K. Hao, G. Feng, and X. Zhang, "Robust image watermarking based on generative adversarial network," *China Commun.*, vol. 17, no. 11, pp. 131–140, Nov. 2020.

[37] M. Meng and Z. Ping, "Decompose and reconstruct images based on exponential-Fourier moments," *Nei Menggu Shifan Daxue Xuebao(Ziran Kexue Ban)*, vol. 40, no. 3 pp. 258–260, 2011.

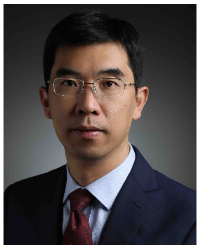
[38] (2018). *Seeprettyface.com, BUTY_GWY, Contributes The Dataset*. From www.seeprettyface.com/mydataset_page3.html

[39] H.-Y. Chen, I.-S. Fang, C.-M. Cheng, and W.-C. Chiu, "Self-contained stylization via steganography for reverse and serial style transfer," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Mar. 2020, pp. 2163–2171.

[40] 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.

[41] Q. Li et al., "Image steganography based on style transfer and quaternion exponent moments," *Appl. Soft Comput.*, vol. 110, Oct. 2021, Art. no. 107618.

Qi Li received the B.S. degree in computer science and technology and the M.S. degree in computer application from the Qilu University of Technology (Shandong Academy of Science) in 2016 and 2019, respectively, and the Ph.D. degree in computer science and technology from Dalian Maritime University in 2023. He is currently a Post-Doctoral Fellow with Shandong Academy of Science. His research interests include information hiding, computer vision, and machine learning.



Bin Ma received the M.S. and Ph.D. degrees from Shandong University in 2005 and 2008, respectively. From 2008 to 2013, he was an Associate Professor with the School of Information Science, Shandong University of Political Science and Law. He visited New Jersey Institute of Technology at Newark as a Visiting Scholar from 2013 to 2015. He is currently a Professor with the School of Cyber Security, Qilu University of Technology (Shandong Academy of Sciences), Shandong, China. His research interests include reversible data hiding, multimedia security, and image processing.

Xianping Fu received the Ph.D. degree in communication and information system from Dalian Maritime University, Dalian, in 2005. His major research interests include image processing for content recognition, multimedia technology, video processing for very low bit-rate compression, hierarchical storage management system for multimedia data, image processing for content retrievals, and driving and traffic environment perception.

Xiaoyu Wang received the B.S. degree in computer science and technology and the M.S. degree in computer application from Qilu University of Technology (Shandong Academy of Science) in 2016 and 2019, respectively, and the Ph.D. degree in computer science and technology from Dalian Maritime University in 2024. She is currently a Post-Doctoral Fellow with the School of Information Science and Technology, Dalian Maritime University, and also an adjunct Associate Professor with Qilu University of Technology. Her research interests include reversible data hiding, machine vision, and image processing.

Chunpeng Wang (Member, IEEE) received the B.E. degree in computer science and technology from Shandong Jiaotong University in 2010, the M.S. degree from the School of Computer and Information Technology, Liaoning Normal University, in 2013, and the Ph.D. degree from the School of Computer Science and Technology, Dalian University of Technology, in 2017. He is currently an Associate Professor with the School of Cyber Security, Qilu University of Technology (Shandong Academy of Sciences). His research interests include image processing and multimedia information security.

Xiaolong Li (Member, IEEE) received the B.S. degree from Peking University, China, in 1999, the M.S. degree from the École Polytechnique, France, in 2002, and the Ph.D. degree in mathematics from the ENS de Cachan, France, in 2006. He was a Post-Doctoral Fellow and a Researcher with Peking University from 2007 to 2016. He is currently a Professor with the Institute of Information Science, Beijing Jiaotong University, Beijing, China. His research interests include image processing and information hiding.