Image Steganography in Color Conversion

Qi Li, Bin Ma, *Member, IEEE*, Xiaoyu Wang, Chunpeng Wang, *Member, IEEE*, and Suo Gao

***Abstract*—For image steganography, no matter how ingenious the distortion function is designed, the embedding operation will lead to changes in the distribution of cover image, which causes security risks. In this brief, an image steganography based on de-colorization and colorization is proposed, which can accomplish the embedding and extraction task of secret information during the process of color translation. Unlike traditional image steganography, the proposed steganography counteracts the embedding influence of secret information through de-colorization network. In addition, a robust embedding algorithm is utilized to embed secret image into cover image by considering the de-colorization and colorization as an attack. Moreover, different training datasets are created to analysis their impact on the reconstruction performance of the colorization network. Finally, simulation results are given to verify the feasibility of the proposed steganography.**

***Index Terms*—Image steganography, color translation, de-colorization, colorization.**

I. Introduction

TO ENSURE data security, the most common methods are to use encryption technology to encrypt target images into meaningless cipher images [1], [2]. However, the encryption behavior exposes the existence of secret images, which is easy to attract the attention of the attacker and causes security problems. Therefore, the image steganography technology, which can not only protect content of secret information, but also can hide the embedded behavior of secret information, has received more and more attention [3].

Early image steganography is dominated by non-adaptive steganography. Among them, the Least Significant Bit (LSB) [4] replacement is the representative method in the non-adaptive steganography, which converts image pixel into binary, and replace the LSB of image with secret information to complete the embedding task. After LSB replacement, there will be a value pair effect on image pixels, which is easy to detect by steganalysis. To overcome this defect, an LSB matching algorithm [5] is proposed, but the differences in high-order statistical properties caused by this algorithm are easily captured. Afterwards, coding algorithms such as EMD coding [6] and ZZW coding [7] have been proposed. However, non-adaptive steganography assigns the same modification priority to image pixels in different regions, and the obtained stego images are easily detected by basic steganalysis algorithms such as chi-square test. In response to the problems of non-adaptive steganography, researchers have proposed the adaptive steganography, which assigns the different embedding distortion costs to the image pixels in different regions, and embeds secret information into the areas with relatively complex texture to improve security [8]. However, traditional steganography schemes modify the original cover image with artificially embedding strategies to hide the secret information. Due to the limitation of manual feature extraction, the traditional steganography schemes is seriously insecure.

The development of convolutional neural network has pushed the application of deep learning to a new level in the field of image processing, which also provides an opportunity for the combination of steganography and deep learning due to its powerful learning and modeling capabilities for complex data. In 2016, Volkhonskiy et al. [9] proposed a steganography model based on generative adversarial network (SGAN), by adding a steganalysis branch to the deep convolutional generative adversarial network, the generative cover image can resist the detection of steganalysis to a certain extent after embedding secret information. On the basis of SGAN, Shi et al. [10] proposed a security steganography based on generative adversarial network (SSGAN), which used WassersteinGAN to replace DCGAN in SGAN and utilized Gaussian-Neuron Convolutional Neural Network (GNCNN) to redesign the discriminative network and steganalysis network. Meng et al. [11] utilized Faster Objection Recognition Convolutional Neural Networks (FRCNN) to identify the texture region of cover image and designed an optimal adaptive steganography according to the texture complexity, which improve the security and imperceptibility of stego image. Hayes and Georg [12] applied the encoder-decoder network to the field of image steganography, and proposed a steganography model named SteGAN, which exploits the adversarial training to reduce the differences between cover image and stego image. The visual quality of stego image obtained by SteGAN is relatively poor, resulting in low security. Through the analysis of the development status for traditional image steganography and deep-learning-based steganography, although previous

Manuscript received 13 April 2023; accepted 25 July 2023. Date of publication 1 August 2023; date of current version 8 January 2024. This work was supported in part by the National Natural Science Foundation of China under Grant 61672124, Grant 61802212, and Grant 61872203; in part by the Password Theory Project of the 13th Five-Year Plan National Cryptography Development Fund under Grant MMJJ20170203; in part by the Liaoning Province Science and Technology Innovation Leading Talents Program Project under Grant XLYC1802013; in part by the Key Research and Development Projects of Liaoning Province under Grant 2019020105-JH2/103; in part by the Jinan City '20 universities' funding Projects Introducing Innovation Team Program under Grant 2019GXRC031; and in part by the Research Fund of Guangxi Key Laboratory of Multi-Source Information Mining and Security under Grant MIMS20-M-02. This brief was recommended by Associate Editor J. Meng. *(Corresponding author: Bin Ma.)*

Qi Li, Bin Ma, Xiaoyu Wang, and Chunpeng Wang are with the School of Cyber Security, Qilu University of Technology, Jinan 250353, China (e-mail: qluliqi@163.com; sddxmb@126.com; qluwxy@163.com; mpeng1122@163.com).

Suo Gao is with the School of Cyber Security, Harbin Institute of University, Harbin 250353, China.

Color versions of one or more figures in this article are available at https://doi.org/10.1109/TCSII.2023.3300330.

Digital Object Identifier 10.1109/TCSII.2023.3300330

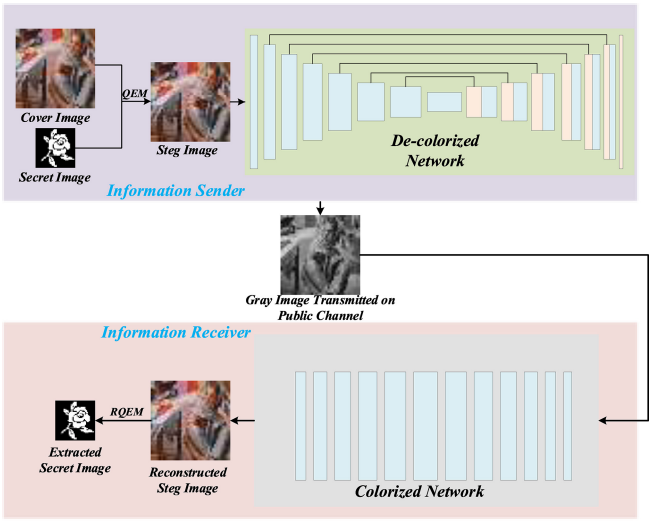

Fig. 1. The architecture for the proposed image steganography based on de-colorization and colorization. The overall architecture is simultaneously trained to work as a pair specifically [22]. Before transmitting the information, we assume that both parties (sender and receiver) already have the model that has been trained well.

research works have proposed many effective steganography schemes, the impact of embedding operation for secret information on the cover image has not been well resolved.

To solve the security risk existing in current image steganography, an image steganography based on de-colorization and colorization is proposed. The major contributions are concluded as follows:

1) The motivation behind the proposed scheme is to utilize color conversion of the steg images to mitigate the impact of embedding secret information on the cover images, such as color distortion and distribution changes, etc. The color conversion of the steg images can prevent attackers from determining whether there is steganographic behavior, as both color conversion and the embedding of secret information can lead to changes to the cover images.

2) Different from current image steganography, the proposed scheme applies image de-colorization and colorization techniques to the field of image steganography, which can accomplish the embedding and extraction task of secret information during the process of color translation.

3) The de-colorization network borrows from the U-Net network [13] architecture and the loss functions of IG (Invertible Grayscale) [14], which encodes the information of color image into the grayscale image for better accomplishing the extracting task of secret information.

II. The Proposed Image Steganography

Our proposed steganography scheme is overviewed in Fig. 1, which can describe as two stages: In the first stage, information sender uses robust embedding algorithm to embed secret image into the cover image to obtain the steg image. To improve security, the de-colorized network is utilized to de-colorize steg image for generating the grayscale image, which can be transmitted on the public channel. In the second stage, the corresponding colorized network is designed to restore the color information of steg images. Then information receiver can extract the secret image from the reconstructed steg image.

In the first stage, the embedding algorithm based on quaternion exponential moments (QEM) is utilized to embed the secret image I_s into the cover image I_c for obtaining the steg image I_t. This process can be described as follows:

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

Then the information receiver uses de-colorized network to conduct secondary operation on the steg image I_t, that is, de-colorizing the steg image to obtain the grayscale image I_g, which can be transmitted on the public channel. This process is expressed as follows:

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

where I_g represents the generated grayscale image and D represents the designed de-colorized network, respectively. The architecture of designed de-colorized network borrows from the form of U-Net and implicitly encodes the color information in the generated grayscale image by means of skip connections.

In the second stage, the corresponding colorized network C is utilized to restore the information of steg image for obtaining the reconstructed steg image I_r, this process can be described as follows:

$$I_r = C(I_g) \tag{3}$$

Then the information receiver can extract the transmitted secret image from the reconstructed steg image I_r, which can be expressed as follows:

$$I_{rs} = Extract_{RQEM}(I_r) \tag{4}$$

The loss functions in IG (Invertible Grayscale) are utilized to better train the de-colorized network and the corresponding colorized network. As can be seen from Fig. 1, to ensure no difference between the steg image I_t and the reconstructed steg image I_r in appearance, a MSE-based loss function is defined as follows:

$$\mathcal{L}_{MSE} = \mathbb{E}_{I_t \sim p(I_t), I_r \sim p(I_r)} ||I_t - I_r||_2 \tag{5}$$

where, $p(I_t)$ and $p(I_r)$ represent the distribution of the steg image and the reconstructed steg image, respectively. The MSE-based loss function can reduce the difference in appearance and distribution between the steg image and the reconstructed steg image.

To improve the security of proposed steganography, the generated grayscale image is transmitted on the public channel. Therefore, another loss function needs to be designed to ensure that the generative grayscale image is normal and meaningful. For this purpose, luminance loss, contrast loss and local structure loss are combined to train the networks. Firstly, the luminance loss is utilized to ensure the generated grayscale image is consistent with the steg image in terms of luminance. And the loss function can be expressed as follows:

$$\mathcal{L}_1 = \mathbb{E}|| \max |I_g - L(I_t)| - \theta, \mu||_1 \tag{6}$$

where L represents the luminance channel, and the values of θ and μ are 70 and 0, respectively. This function can

TABLE I
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%	100%
YeNet	57.3%	92.6%	100%	34.0%	65.3%	100%

Fig. 2. The comparison experimental results (visual quality, PSNR (Peak Signal to Noise Ratio) and SSIM (Structure Similarity Index Measure)) for the original images after being attacked and reconstruction (de-colorization and colorization operation).

reduce the difference between the luminance components of the steg image and the generated grayscale image. In addition, a contrast loss function is defined according to IG (Invertible Grayscale), which can be expressed as follows:

$$\mathcal{L}_2 = ||E_{vgg19}(I_g) - E_{vgg19}(I_t)||_1 \tag{7}$$

where E_{vgg19} represents the pre-trained VGG19 network, and the feature maps obtained from "conv4_4" layer in E_{vgg19} are selected as the representation of image contrast. Finally, a local structure loss is designed to preserve the local features of the images. The local features can be calculated as follows:

$$\mathcal{L}_3 = ||Var(I_g) - Var(I_t)||_1 \tag{8}$$

where the *Var* represents the mean values of local variation for an image. Therefore, the total grayscale loss function can be concluded as:

$$\mathcal{L}_{Gray} = \lambda_1\mathcal{L}_1 + \lambda_2\mathcal{L}_2 + \lambda_3\mathcal{L}_3 \tag{9}$$

where λ_1, λ_2 and λ_3 are the weight values that balance three losses. For grayscale images, contrast and content structure are more important than brightness. And inspired by Invertible Grayscale, λ_1, λ_2 and λ_3 are set to 1.0, 3.0 and 10, respectively.

In conclusion, two loss functions are simultaneously utilized to optimize the de-colorized network and colorized network to obtain the grayscale image and reconstructed image. The parameters of networks can be optimized by minimizing the overall loss function as follows:

$$\mathcal{L}_{total} = \mathcal{L}_{MSE} + \mathcal{L}_{Gray} \tag{10}$$

III. SIMULATION RESULTS

Extensive experiments are conducted to evaluate the feasibility of the proposed image steganography. In the process of experiments, the model is trained with GPU NVIDIA GeForce Tesla V100 32G in the environment of Pytorch 1.6 and python 3.6. The training images are cropped to the size of 256×256, which are randomly selected from the anime face dataset. And the size of training dataset is set to 3K. In the verification experiments, the size of secret image is set to the 16×16, 32×32 and 64×64, respectively.

To investigate the security of the proposed scheme, the SRNet [15] and YeNet [16] were utilized for verifying the anti-steganalysis ability of the scheme. Because the steg images include both color images and grayscale images, the input layer of steganalysis networks are modified in the training process. And the experimental results are shown in Table I.

From the experimental results of Table I, it can be seen that using generated grayscale images for transmission on public channel can significantly resist detection by steganalysis. However, when the size of the embedded secret image is 64×64, neither color nor grayscale images can avoid detection by steganalysis, due to the significant distortion of the cover image caused by the large amount of embedded secret image.

From the perspective of robust watermarking field, the operations of de-colorization and colorization on steg images can be regarded as an attack. Therefore, a variety of traditional methods are utilized to attack the steg images in the experiments, such as median filter, Gaussian noise, JEPG compression and cropping, etc. For the watermarking algorithm based on QEM, the secret image can be extracted from the attacked steg images with low BER. Therefore, from the perspective of image steganography, if the quality of reconstructed steg images is better than the attacked steg images for information receiver, the secret image can be successfully extracted from the reconstructed steg images. The experimental results can be seen from Fig. 2. From the analysis of experimental results in Fig. 2, in terms of visual quality, PSNR and SSIM, the reconstructed images are better than those of

TABLE II
THE CORRESPONDING EXPERIMENTAL RESULTS FOR PSNR, SSIM AND BERS

	PSNR			SSIM			BER		
	16×16	32×32	64×64	16×16	32×32	64×64	16×16	32×32	64×64
Training images without embedding secret image	32.3148	32.1404	31.1119	0.9780	0.9754	0.9628	0.1432	0.0900	0.1622
Training images with embedding secret image of 16×16	31.8863	31.8038	31.0935	0.9825	0.9815	0.9762	0.1572	0.1092	0.1595
Training images with embedding secret image of 32×32	32.4184	32.4679	31.8512	0.9818	0.9810	0.9763	0.1306	0.0921	0.1514
Training images with embedding secret image of 64×64	**32.5063**	**32.6685**	**32.6622**	**0.9832**	**0.9835**	**0.9827**	**0.1166**	**0.0898**	**0.1388**

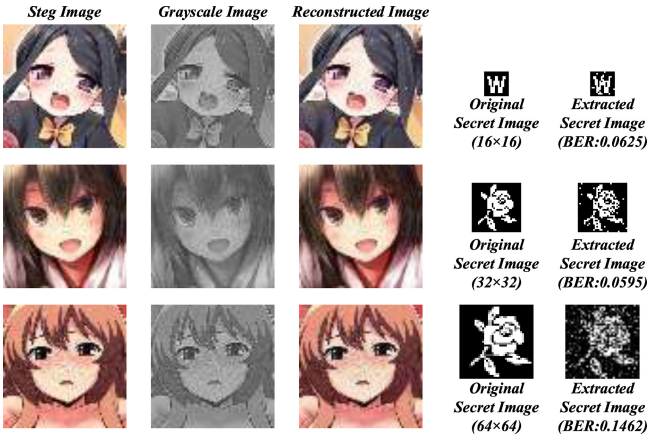

Fig. 3. The experimental results for grayscale images, reconstructed images and extracted secret images (including different size and corresponding BER (Bit Error Rate)).

attacked images. Since the secret image can be extracted from these attacked images with lower BERs, the secret image should be extracted from the reconstructed images with better quality. These experimental results also verify that the proposed image steganography is feasible.

For reversible data hiding, the secret information must be extracted in a lossless form, and the cover image can also be reconstructed losslessly. However, for image steganography, a certain degree of information loss can be acceptable, such as deep steganography [17], DeepMIH [18] and HiNet [19], which has a relatively high error rate.

To further evaluate the feasibility of the proposed image steganography, the steg image is firstly decolorized to obtain a grayscale image, which can be transmitted on the public channel. Then the receiver reconstructs the steg image from the grayscale image through a colorization network, and then extracts the secret image from the reconstructed steg image. The experimental results for grayscale images, reconstructed images and extracted secret images are shown in Fig. 3.

From the experimental results in Fig. 2, although the secret images cannot be completely extracted from reconstructed images in a lossless form, the contents of reconstructed secret images are complete, which indicates that the proposed image steganography is feasible. From the further analysis of experimental results, when the size of secret image is 16×16, the BER for reconstructed secret image can reach 0.0625; and when the size of secret image is 32×32, the BER drops to 0.0595. But when the size of secret image is up-scaled to 64×64, the BER is increased to 0.1462. This phenomenon is interpretable. The smaller the secret image is, the less influence its embedding operation has on the appearance and distribution of the cover image. that is, the better the details of steg image are reconstructed, the more complete the secret image can be extracted. Therefore, this can explain why the BER of extracted secret image with size of 32×32 is lower than the BER of extracted secret image with size of 16×16. Appropriate embedding capacity can compromise between the extraction of secret image and the reconstruction effect of steg image. With the increase of embedding capacity, the distribution and appearance of cover image are greatly affected, so the extraction of secret image largely depends on the reconstructed effect of steg image. Therefore, when the size of secret image is 64×64, the BER is the lowest. Nevertheless, the content information of the extracted secret image remains intact.

In addition, we created four different image datasets to train the decolorization-colorization model for obtaining the best extraction effect of secret image. The image datasets including: original images without embedding secret image, steg images with embedding secret image of 16×16, steg images with embedding secret image of 32×32 and steg images with embedding secret image of 64×64. The corresponding experimental results for PSNR, SSIM and BERs are shown in Table II.

The experimental results in Table II are obtained by calculating the mean values of PSRN, SSIM and BER for 20 randomly selected tested images. From the experimental results in Table II, when the size of secret image is 32×32, the BERs is the lowest. This phenomenon also verifies that the experimental results in Fig. 3 are convincing. When the size of secret image is 16×16 or 64×64, the BERs are relatively high. This indicates that the BER is proportional to the quality of reconstructed image. Another conclusion can be drawn from the experimental results in Table II, that is, the reconstructed ability of model is also proportional to the images of the training set. Through experiments, it can be also found that when the PSNR value is between 31-32 and the SSIM value is between 0.97-0.98, the integrity of extracted secret image is within the acceptable range. In addition, the experimental

TABLE III
THE CORRESPONDING EXPERIMENTAL RESULTS FOR BERS OF EXTRACTING SECRET IMAGE

Methods	BER		
	16×16	32×32	64×64
Gaussian noise	0.1758	0.1231	0.1985
Salt & pepper noise	0.0352	0.0450	0.1384
Median filter + Gaussian noise	0.1992	0.2129	0.2625
JPEG compression+ Salt & pepper noise	0.0391	0.0430	0.1426
JPEG compression + Gaussian noise	0.1484	0.1191	0.2026
Proposed Method	0.0742	0.0566	0.1357

TABLE IV
THE CORRESPONDING EXPERIMENTAL RESULTS FOR BERS OF EXTRACTING SECRET IMAGE

Methods	BER		
	16×16	32×32	64×64
Self-contained Stylization [20]	0.5078	0.4609	0.4981
CycleGAN [21]	0.2534	0.2177	0.2852
Style Removal [22]	0.1526	0.0906	0.1385
Proposed Method	0.0742	0.0566	0.1357

results indicate that the training process of the model dose not to need to create a specific image dataset, which shows the proposed steganography scheme has good generalization.

In order to verify the effectiveness and novelty of our proposed scheme, we compared the BREs of secret image extracted from the stego image obtained by various attacks and some state-of-the-art deep learning-based steganography methods. The comparison results are shown in Table III and Table IV. For image steganography, the successful transmission of secret information is the top priority. Therefore, the extraction effect of secret information can reflect the feasibility of steganography. From the experimental results in Table III and Table IV, the proposed scheme has obvious advantages in BERs compared with both traditional and deep learning-based steganography methods, which indicates that the proposed scheme is highly feasible.

IV. CONCLUSION

In this brief, an image steganography based on de-colorization and colorization is proposed. Different from the current steganography schemes, the transmission task of secret information can be accomplished during the process of color translation. The colorization and de-colorization technologies are introduced and applied to the field of image steganography. In the transmission process of secret information, colorization and de-colorization operations are regarded as attack behaviors, so the robust embedding algorithm based on quaternion exponential moments (QEM) is utilized to obtain the steg image. In addition, for obtaining the reconstructed images with better quality, the architecture and loss function of de-colorization network borrow from the U-Net and Invertible Grayscale, respectively. Extensive experiments are conducted to have verified the feasibility of the proposed method. In the future work, we focus on improving the architecture of the model for enhancing the quality of reconstructed images, thereby improving the generalization of the proposed steganography scheme. And we will strive to enhance the GPU device to enhance the ability to process the images with size of 512×512 in our steganography scheme.

REFERENCES

[1] B. Li, M. Wang, J. Huang, and X. Li, "A new cost function for spatial image steganography," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, 2014, pp. 4206–4210.
[2] V. Sedighi, R. Cogranne, and J. Fridrich, "Content-adaptive steganography by minimizing statistical detectability," *IEEE Trans. Inf. Forensics Security*, vol. 11, pp. 221–234, 2016.
[3] 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.
[4] C.-K. Chan, and L. M. Cheng, "Hiding data in images by simple LSB substitution," *Pattern Recognit.*, vol. 37, no. 3, pp. 469–474, 2004.
[5] J. Mielikainen, "LSB matching revisited," *IEEE Signal Process. Lett.*, vol. 13, no. 5, pp. 285–287, May 2006, doi: 10.1109/LSP.2006.870357.
[6] J. Fridrich and D. Soukal, "Matrix embedding for large payloads," *IEEE Trans. Inf. Forensics Security*, vol. 1, pp. 390–395, 2006, doi: 10.1109/TIFS.2006.879281.
[7] W. Zhang, X. Zhang, and S. Wang, "Maximizing steganographic embedding efficiency by combining Hamming codes and wet paper codes," in *Proc. Int. Workshop Inf. Hiding*, May 2008, pp. 60–71.
[8] J. Fridrich and J. Kodovsky, "Rich models for steganalysis of digital images," *IEEE Trans. Inf. Forensics Security*, vol. 7, pp. 868–882, 2012.
[9] D. Volkhonskiy, I. Nazarov, and E. Burnaev "Steganographic generative adversarial networks," in *Proc. 12th Int. Conf. Mach. Vis.*, 2020, pp. 1–15.
[10] H. Shi, J. Dong, and W. Wang, "SSGAN: Secure steganography based on generative adversarial networks," in *Proc. Pac. Rim Conf. Multimedia*, 2017, pp. 534–544.
[11] H. Meng, S. Rice, J. Wang, and X. Sun, "A fusion steganographic algorithm based on faster R-CNN," *Comput. Mater. Continua*, vol. 55, no. 1, pp. 1–16, 2018.
[12] J. Hayes and D. Georg, "Generating steganographic images via adversarial training," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, 2017, pp. 554–561.
[13] R. Olaf, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.*, 2015, pp. 234–241.
[14] M. Xia, X. Liu, and T. Wong, "Invertible grayscale," *ACM Trans. Graph.*, vol. 37, no. 6, pp. 1–10, 2018.
[15] J. Ye, J. Ni, and Y. Yi, "Deep learning hierarchical representations for image steganalysis," *IEEE Trans. Inf. Forensics Security*, vol. 12, pp. 2545–2557, 2017.
[16] M. Boroumand, M. Chen, and J. Fridrich, "Deep residual network for steganalysis of digital images," *IEEE Trans. Inf. Forensics Security*, vol. 14, pp. 1181–1193, 2019.
[17] S. Baluja, "Hiding images in plain sight: Deep steganography," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, 2017, pp. 1–11.
[18] 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.
[19] 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.*, 2021, pp. 4733–4742.
[20] 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/CVF Winter Conf. Appl. Comput. Vis.*, 2020, pp. 2152–2160.
[21] 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.*, 2017, pp. 2223–2232.
[22] Q. Li et al., "Image steganography based on style transfer and quaternion exponent moments," *Appl. Soft Comput.*, vol. 110, Oct. 2021, Art. no. 107618.