ARWGAN: Attention-Guided Robust Image Watermarking Model Based on GAN

Jiangtao Huang, Ting Luo, *Member, IEEE*, Li Li, Gaobo Yang, Haiyong Xu, *Member, IEEE*, and Chin-Chen Chang, *Fellow, IEEE*

***Abstract*— In the existing deep learning-based watermarking models, extracted image features for fusing with watermark are not abundant enough and more critically, and essential features are not highlighted to be learned with the purpose of robust watermarking, both of which limit the watermarking performance. To solve those two drawbacks, this article proposes an attention-guided robust image watermarking model based on a generative adversarial network (ARWGAN). To acquire a great deal of representational image features, a feature fusion module (FFM) is devised to learn shallow and deep features effectively for multilayer fusion with watermark, and meanwhile, reuse of those features by the dense connection enhances robustness. To alleviate image distortion caused by embedding watermark, an attention module (AM) is deployed to compute the attention mask by mining the global features of the original image. Specifically, with the guidance of the attention mask, image features representing inconspicuous regions and texture regions are enhanced for embedding the high strength of watermark, and simultaneously, other features are suppressed to improve the watermarking performance. Furthermore, the noise subnetwork is adopted for robustness enhancement by simulating various image attacks in iterative training. The discriminator is used to distinguish the encoded image from the original image for improving watermarking invisibility continuously. Experimental results demonstrate that the ARWGAN is superior to the existing state-of-the-art (SOTA) watermarking models, and ablation experiments prove the effectiveness of the FFM and the AM. The code is available at https://github.com/river-huang/ARWGAN.**

***Index Terms*— Attention mask, dense connection, feature fusion, generative adversarial network (GAN), robust watermarking.**

Manuscript received 2 February 2023; revised 19 April 2023; accepted 28 May 2023. Date of publication 19 June 2023; date of current version 28 June 2023. This work was supported in part by the National Natural Science Foundation of China under Grant 61971247, Grant 61501270, and Grant 62171243; in part by the Zhejiang Provincial Natural Science Foundation of China under Grant LY22F020020; and in part by the Natural Science Foundation of Ningbo under Grant 2021J134 and Grant 2022J066. The Associate Editor coordinating the review process was Dr. Eduardo Cabal-Yepez. *(Corresponding author: Ting Luo.)*

Jiangtao Huang and Haiyong Xu are with the Faculty of Information Science and Engineering, Ningbo University, Ningbo 315211, China (e-mail: 2201100032@nbu.edu.cn; xuhaiyong@nbu.edu.cn).

Ting Luo is with the College of Science and Technology, Ningbo University, Ningbo 315212, China (e-mail: luoting@nbu.edu.cn).

Li Li is with the School of Computer Science and Technology, Hangzhou Dianzi University, Hangzhou 310018, China (e-mail: lili2008@hdu.edu.cn).

Gaobo Yang is with the School of Information Science and Engineering, Hunan University, Changsha 410082, China (e-mail: yanggaobo@hnu.edu.cn).

Chin-Chen Chang is with the Department of Information Engineering and Computer Science, Feng Chia University, Taichung 40724, Taiwan (e-mail: alan3c@gmail.com).

Digital Object Identifier 10.1109/TIM.2023.3285981

I. Introduction

DIGITAL watermarking is an old yet still active research topic in the community of information security, which has attracted wide interest from numerous researchers in both academia and industry. It can be used for a wide range of applications, such as copyright protection, source tracking, and integrity authentication. Image watermarking is the most popular due to the easy availability and the widespread of digital images [1], [2], [3], [4], [5], [6], [7], [8].

The existing image watermarking works can be divided into two categories, namely, fragile watermarking and robust watermarking. Fragile watermarking is mainly used for tampering detection and integrity verification [9], [10], whereas robust watermarking is used for copyright protection and tracing against various image processing attacks [11], [12]. The traditional robust watermarking works usually embed watermark into various transform domains for resisting image attacks [13], [14], [15], [16], [17], [18]. Ko et al. [19] presented a robust image watermarking by modifying discrete cosine transform (DCT) coefficients while exploiting interblock coefficient correlation. Hu et al. [20] investigated the use of singular value decomposition (SVD) with mixed modulation incorporated for robust color image watermarking. Liu et al. [21] presented an optimized image watermarking method based on Hessenberg decomposition (HD) and SVD in the discrete wavelet transform (DWT) domain. To further improve watermarking performance, machining learning techniques, which include support vector machine (SVM), support vector regression (SVR), and basic function neural network (BFNN), have also been employed for image watermarking [22], [23], [24]. Niu et al. [25] presented a robust color image watermarking method against geometric distortions, in which watermark is recovered by using the SVR technique. Chen and Zhang [26] presented to improve watermarking robustness against geometrical attacks, where the trained SVM is used to roughly correct the geometric parameters. Liu and Jiang [24] proposed an image watermarking algorithm by exploiting BFNN to simulate human visual specialty so as to determine the watermark embedding intensity endured by DCT coefficients. However, these works usually require a lot of prior experience and a series of complex operations, such as preprocessing and postprocessing. Furthermore, relevant image features to define embedding positions and strategies

are extracted manually, and inaccurate features will affect watermarking performance.

In recent years, deep learning has achieved very impressive successes in computer vision tasks, such as image classification and image segmentation, due to its strong representation capabilities [27], [28]. Compared with machining learning, deep learning can learn deeper image features automatically and achieve better generalization capability by training on large-scale datasets to cope with diverse scenes. Although vanishing and exploding gradients may reduce the performance of networks, some optimization techniques, such as the skip connection, the dense connection, and batch Normalization, solve these problems satisfactorily [29], [31]. Deep learning has also been introduced into image watermarking, and the existing deep network models for image watermarking can be divided into two categories, namely, convolutional neural network (CNN) and generative adversarial network (GAN) [32]. Haribabu et al. [33] presented the pioneer CNN-based robust image watermarking model, where the autoencoder-based CNN is used to learn image features for fusing with watermark. Mun et al. [34] also proposed a CNN-based robust image watermarking model, where an iterative learning framework is presented to ensure watermarking robustness. In essence, the learned network is an extension of the frequency domain that is widely used in the existing watermarking schemes. Ahmadi et al. [35] proposed a deep end-to-end diffusion watermarking framework (ReDMark), which can learn a new watermarking model in any desired transform space. The framework is composed of two full CNNs with residual structure, which simulates various attacks as a differentiable network layer to facilitate end-to-end training. The watermark data are diffused in a relatively wide area of the cover image to enhance security and robustness. Moreover, Luo et al. [36] presented a new framework for distortion-agnostic deep watermarking (DA), which does not require the explicit modeling of the image distortion during training. Instead, its robustness comes from two sources: adversarial training and channel coding. Compared with training on a fixed set of distortions and noise levels, it achieves comparable or better results on distortions available during training and better performance on unknown distortions.

Besides CNN, GAN provides an alternative for designing deep image watermarking. As we know, GAN is made up of a generator and a discriminator, which play an adversarial game [37]. This makes GAN inherently suitable for robust image watermarking by exploiting the adversarial relationship of the generator and the discriminator to compromise capacity, invisibility, and robustness. In recent years, there are a few GAN-based image watermarking works. Zhu et al. [38] presented a pioneer work, namely, HiDDeN, by exploiting the adversarial relationship between the generator and the discriminator for robust image watermarking. Specifically, it uses the encoder–noise–decoder structure and jointly trains the encoder and the decoder, where the encoder produces a visually indistinguishable encoded image, whereas the decoder recovers well the original watermark. Hao et al. [39] also proposed a GAN-based robust image watermarking approach. The proposed model includes two modules, namely, generator and adversary. The generator is mainly used to produce watermarked images and decode the image damaged by noise to obtain watermark. The adversary is used to discriminate whether the image is embedded with watermark and damages the image by noise. It is actually an improved HiDDeN, which adds a high-pass filter in front of the discriminator, making the watermark tend to be embedded in the mid-frequency region of the image. Liu et al. proposed a two-stage separable deep learning framework (TSDL) for practical blind watermarking [40]. Specifically, the framework is composed of noise-free end-to-end adversary training (FEAT) and noise-aware decoder-only training (ADOT). A redundant multilayer feature encoding network is developed in FEAT to obtain the encoder, while ADOT is used to get the decoder, which is robust and practical enough to accept any type of noise. Jia et al. [41] introduced a watermarking framework using a mini-batch of real and simulated JPEG compression (MBRS) to enhance robustness. In particular, real JPEG and simulated JPEG noises are trained alternatively to improve robustness of JPEG compression in the real world. Fernandez et al. [42] proposed a self-supervised watermarking model to embed watermark into latent spaces (SSLW). In SSLW, a self-supervised encoder utilizes data augmentation and image preprocessing to extract image features for fusing with watermark to obtain robustness. However, in the above watermarking models, extracted image features for fusing with watermark are still not abundant and robust enough, and consequently, the ability to resist image noises is affected. More importantly, they neglect to highlight critical image features for learning, which degrades the watermarking performance.

To obtain high invisibility and robustness, we propose an attention-guided robust image watermarking model based on GAN (ARWGAN). ARWGAN employs the encoder–noise–decoder structure for end-to-end training. In the encoder, to overcome the drawback of extracting limited image features, the feature fusion module (FFM) is deployed to learn shallow and deep image features with multiple layers for fusing with watermark. Watermark fused with deep image features has the capability of resisting image noises, yet the reuse of those shallow and deep features by the dense connection further promotes robustness. To decrease the image distortion caused by watermark fusion, the attention module (AM) is designed for obtaining the attention mask of the original image by mining global image features, which determines watermarking strengths for different regions. Specifically, the flat and sensitive areas are suppressed, and oppositely, inconspicuous, and textured areas are enhanced for high embedding strength. In addition, to increase the image quality of the encoded image, the adversarial relationship between the encoder and the discriminator is established, and meanwhile, the discriminator is utilized to distinguish the encoded image from the original image in the iterative training with the help of the AM. To increase the capability of resisting various image attacks, different types of image attacks are randomly selected in the noise subnetwork for training. Under jointly learning all parts

of ARWGAN, high watermarking invisibility and robustness are achieved. Experimental results show that the AM and the FFM are essential to the proposed watermarking model, which is superior to the state-of-the-art (SOTA) watermarking models in terms of the watermarked image quality and resistance to different attacks. The main contributions of this work are summarized as follows.

1) To overcome the drawback of extracting limited image features in the existing deep learning-based watermarking model, we deploy the FFM to learn shallow and deep image features with multiple layers for fusing with watermark. Meanwhile, those abundant image features are reused by the dense connection to be learned for improving watermarking robustness. This kind of multilayer fusion obtains the ability to resist different image attacks.
2) To enhance image features with the purpose of high watermarking invisibility and robustness, we present the AM to obtain the attention mask by computing probability distribution among feature channels of the original image. The attention mask is used to guide watermark strength toward different image regions adaptively. In particular, inconspicuous and texture regions are emphasized to embed watermark with high embedding strength and oppositely, and other regions are suppressed with low strength.
3) Combined with the FFM and the AM, we present ARWGAN for robust watermarking with the adversarial relationship between the encoder and the discriminator. In FFM, a great deal of multilayer features is enhanced and learned to be filtered with the help of the AM for increasing robustness without image degradation. Various comparative experiments validate the watermarking performance of ARWGAN, and ablation studies are also conducted to explore the effectiveness of the two ideas.

The rest of this article is organized as follows. Section II discusses the related work and background of the dense network (Dense-Net) and the attention mechanism. Section III introduces the proposed watermarking model. Section IV gives experimental results and discussions, and Section V provides conclusions.

II. Related Work

In this section, the Dense-Net [31] is introduced at first, which inspires our work. Then, the attention mechanism is described for different applications.

A. Dense-Net

In recent years, the deep network has high performance in varieties of computer vision tasks [43], [44]. It seems that more convolution layers concatenated can improve the representational capacity of the deep network, yet it is still challenging to prevent error accumulation when convolution layers are concatenated via the successive connection [45], which results in some problems, such as vanishing/exploding gradients and model degradation [46], [47]. He et al. proposed the residual network (Res-Net), which adopts the skip connection to solve the above problems [31]. The skip connection reuses preceding features and achieves great success. Inspired by Res-Net, Huang et al. [31] proposed a Dense-Net to promote the performance of the deep network.

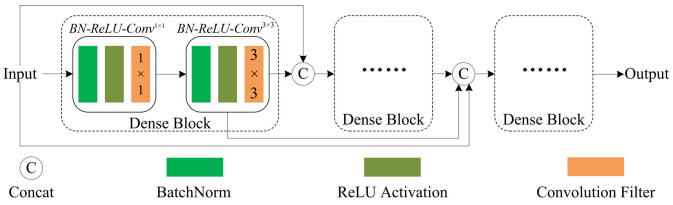


Fig. 1. Structure of Dense-Net.

In Dense-Net, the dense connection is designed to collect the output of all previous layers as the input of the next layer, which encourages feature reuse and alleviates the vanishing gradient problem. Due to the good performance of Dense-Net, it has a wide application in the field of image processing. For instance, Wu et al. [48] used Dense-Net for medical image segmentation. Zhang et al. [49] utilized Dense-Net for improving the accuracy of image classification. Zhang et al. [50] employed Dense-Net for image steganography to increase the embedding capacity. As illustrated in Fig. 1, Dense-Net contains several dense blocks, and each dense block is composed of a BN–ReLU–Conv$^{1\times1}$ and a BN–ReLU–Conv$^{3\times3}$, wherein BN–ReLU–Conv$^{j\times j}$ is consisting of a BatchNorm function, a rectified linear unit (ReLU) activation, and a convolution with the kernel size of 3×3. The output of the ith dense block is calculated as follows:

$$X_i = Y_i(\text{Concat}(X_1, X_2, X_3, \ldots, X_{i-1})) \tag{1}$$

where Concat$(\cdot)$ is the concatenation operation and $Y_i(\cdot)$ represents a series of nonlinear transformations of the dense block.

These dense blocks of Dense-Net are concatenated by the dense connection, which contributes to improving the corresponding performance. The dense connection preserves all features to ensure each dense block learns all previous ones. Not only the utilization rate of image features is improved to boost the efficiency of training, but also relearning of previous features significantly suppresses the problems, such as vanishing/exploding gradients and model degradation. Moreover, multiple features are extracted by these dense blocks, and the reuse of these features is beneficial to the representational capacity of the network model. Drawing on the advantages of the dense connection, the reuse of multilayer features is incorporated into the watermarking model for robustness.

B. Attention Mechanism Based on Deep Learning

For each second, the human eyes receive a great deal of information, and the significant processing of the human attention mechanism can distinguish important data from redundant information [51]. Inspired by the human attention mechanism, the key principle behind the attention mechanism based on deep learning involves finding the most important target for solving some tasks. The deep learning attention has been successfully applied in various computer vision

tasks, such as image classification, object detection, and so on [52], [53], [54], [55], [56], [57], [58], [59]. Wang et al. [60] came up with the residual attention network, which generates attention-aware features adaptively for image classification. Hu et al. [61] exploited the interchannel relationship for computing the channel attention to enhance the representational power of the network. Woo et al. [62] proposed convolutional block AM (CBAM), which combines channel attention and spatial attention to improve the performance of the network. Chen et al. [63] used CBAM to improve the accuracy of tamper detection.

Meanwhile, the attention mechanism is added to the watermarking model based on deep learning for improving watermarking performance. Considering different tasks, the attention mechanism plays different roles in the watermarking model. For the improvement of watermarking invisibility, Yu [64] employed the Res-Net to design the AM, which is used to find inconspicuous areas for watermark embedding, but this watermarking model lacks robustness. In order to obtain robustness, Zhang et al. [65] paid attention to finding robust locations by computing the inverse gradient attention mask for extracting watermark. Zhang et al. [66] computed an attention mask based on the image content to guide watermark embedding or extracting, but robustness is not high, because when the image is distorted, the attention mask computed in the process of watermark extraction is much different from that of the embedding process. Thus, for high-intensity noises, the accuracy of watermark extraction is decreased much.

Above watermarking models adopt the attention mechanism for improvement of watermarking invisibility or robustness, whereas it is still a challenge to obtain high watermarking invisibility and robustness at the same time. Differing from previous works, our work utilizes global features to compute the attention mask to find inconspicuous and rich texture areas for improving watermarking performance.

III. Proposed Watermarking Network Model

In this article, we propose an attention-guided robust image watermarking model based on a GAN (ARWGAN). In this section, the general structure of ARWGAN is introduced first, and then, all parts of ARWGAN are described in detail.

A. Overview

The robust watermarking model must satisfy two properties: 1) different image attacks on the encoded image should have minimal impact on embedded watermark and 2) watermark is embedded by applying invisible perturbations on the original image. Considering these two requirements, robust features should be extracted to be fused with watermark, and meanwhile, watermark is distributed over inconspicuous areas for invisibility.

As illustrated in Fig. 2, ARWGAN is mainly composed of the encoder E_θ, the decoder D_φ, the noised layers N, and the discriminator A_γ, where θ, φ, and γ donate trained parameters of the encoder, the decoder, and the discriminator, respectively. These parameters will be updated constantly during iterative

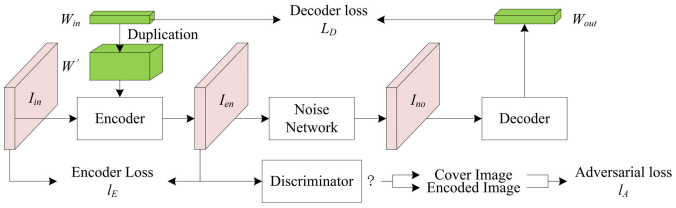


Fig. 2. Architecture of ARWGAN.

training to arrive at high watermarking invisibility and robustness. Given an original image I_{in} with the size of $H \times W \times C$ and the original binary watermark W_{in} with the length of L, when I_{in} is input to E_θ, the encoded image I_{en} is generated, wherein E_θ includes the AM and the FFM. The purpose of the AM is to compute an attention mask for guiding the generation of I_{en} by calculating the global features of the original image. The goal of the FFM is to extract shallow and deep features in multiple layers based on the dense connection for fusing with watermark.

To increase the quality of the encoded image I_{en}, the discriminator A_γ computes the possibility that I_{en} belongs to the original image according to the similarity between I_{in} and I_{en}, and the feedback of A_γ will help E_θ to generate I_{en}. To obtain watermarking robustness, different types of noises are added to N. At the same time, the adversarial training is built that D_φ is utilized to extract decoded watermark W_{out} from I_{no}, and W_{out} is tried to be as same as W_{in}. Moreover, before encoding, the original watermark W_{in} is copied and preprocessed as follows:

$$W' = \begin{bmatrix} W_{\text{in}} & \cdots & W_{\text{in}} \\ \vdots & \ddots & \vdots \\ W_{\text{in}} & \cdots & W_{\text{in}} \end{bmatrix}. \tag{2}$$

It ensures each point of W' contains W_{in}, which can help the encoder embed W_{in} at any spatial location of the encoded image.

B. Encoder

In E_θ, watermark is embedded into I_{in} by fusing image features, and the attention mechanism is added to help distribute watermark over different areas for watermarking invisibility and robustness. Moreover, the residual structure is employed by adopting the global residual skip connection for boosting training. I_{en} is generated by

$$I_{\text{en}} = E\left(I_{\text{in}}, W'\right) \tag{3}$$

where $E(\cdot)$ represents the encoding process, as illustrated in Fig. 3. In the following, the AM and the FFM are depicted, respectively.

In general, embedding watermark into different image regions causes distinct image distortion and robustness. Thus, a suitable attention mask is necessary to guide watermark to distribute over different image regions. For instance, inconspicuous regions holding watermark can improve the visual image quality, and on the other hand, embedding watermark into texture regions can increase robustness. Therefore, the

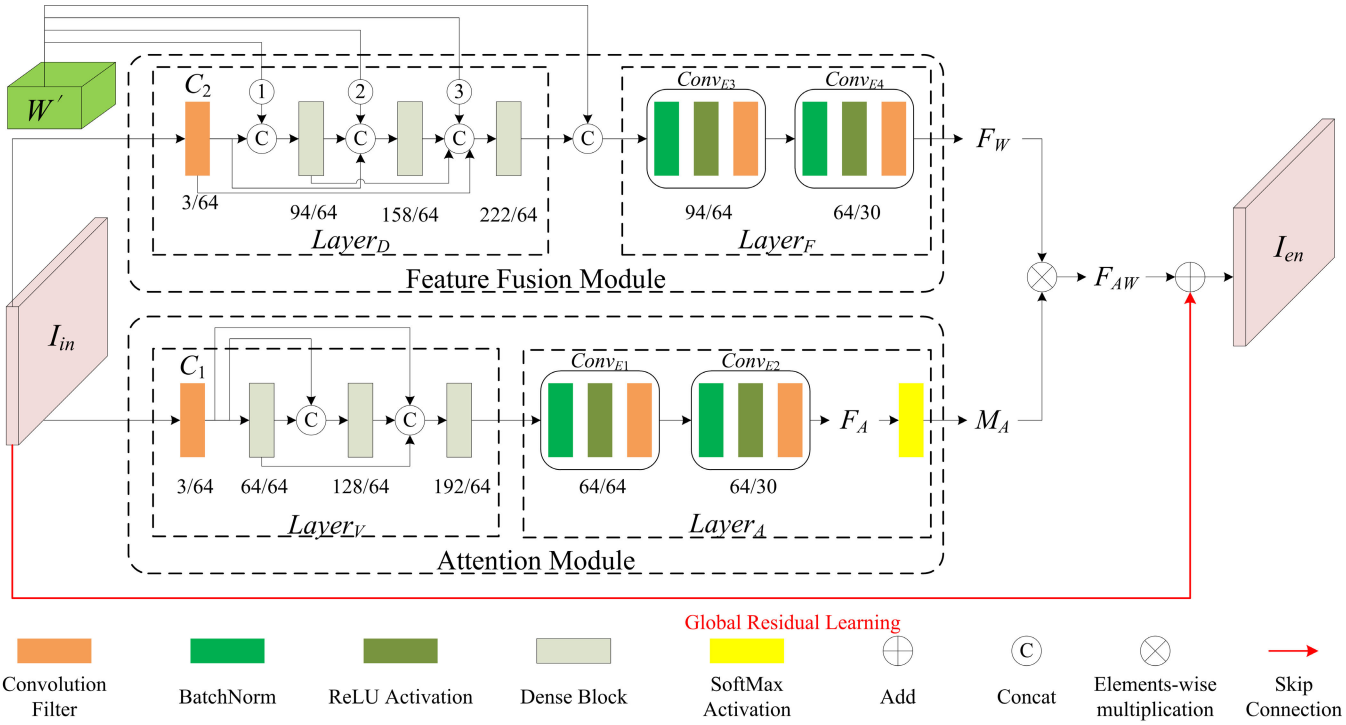


Fig. 3. Structure of encoder, and x and y represent the input dimensions and output dimensions, respectively.

AM is deployed to find inconspicuous and rich texture areas for watermarking considering invisibility and robustness at the same time. To be concrete, according to global image features of I_{in}, the AM generates an attention mask M_A to help the generation of I_{en}. The main reason is that M_A represents global information of I_{in}, which adjusts features of I_{en} effectively to limit the image distortion caused by watermark embedding for decreasing visibility perturbations. On the other hand, M_A finds inconspicuous and rich texture areas for high embedding strength to improve robustness according to the probability feature distribution of the original image. M_A is computed by

$$M_A = E_{\text{AM}}(I_{\text{in}}) \tag{4}$$

where $E_{\text{AM}}(\cdot)$ represents the process of the AM, which is consisting of Layer_V and Layer_A. Specifically, Layer_V contains a convolution filter C_1 with the 3×3 kernel and three dense blocks, which are concatenated by using the dense connection. Layer_A consists of Conv_{E1}, Conv_{E2}, and SoftMax activate function, wherein the output of Conv_{E2} is F_A, and sent to the SoftMax activate function for computing the probability distribution among channels of F_A

$$M_A(i,j,k) = \frac{e^{F_A(i,j,k)}}{\sum_{u=1}^{C} e^{F_A(i,j,u)}} \times C \tag{5}$$

where C represents the channel number of F_A and M_A is the final attention mask for E_θ. In essence, Layer_V is utilized to extract the deep features of I_{in}, while, based on those features, Layer_A is employed to generate M_A, which is utilized to readjust watermarked features for improving the watermarking performance.

To resist image attacks, robust image features are expected to be extracted for fusing with watermark. Thus, to search proper image features, all representational image features should be mined as the candidate for learning. The FFM is utilized to extract shallow and deep features with multiple layers, and then, those image features are reused by the dense connection for fusing with watermark. The main purpose of this kind of multilayer watermark fusion is to increase watermarking robustness. The image feature F_W from this module is extracted as follows:

$$F_W = E_{\text{FFM}}(I_{\text{en}}, W') \tag{6}$$

where $E_{\text{FFM}}(\cdot)$ represents the process of the FFM, which is composed of Layer_D and Layer_F. To be specific, Layer_D is consisting of a convolution filter C_2 with the 3×3 kernel and three dense blocks, which are concatenated by the dense connection as well. In general, the output of the first convolution filter is defined as the shallow features, and the output of dense blocks is called the deep features. Shallow and deep features in multiple layers are extracted for fusing with watermark, and consequently, the output features of Layer_D contain watermark. More importantly, the dense connection reuses all previous features, including shallow and deep features, which are relearned to make the encoder learn different watermarking patterns for increasing robustness. Moreover, Layer_F is consisting of two convolution layers Conv_{E3} and Conv_{E4}, where Conv_{E3} is used to improve robustness by tuning watermarked features again, and Conv_{E4} reduces the dimension.

After obtaining F_W, to increase watermarking invisibility and robustness, the attention mask M_A is utilized to adjust the

distribution of F_W by

$$F_{AW} = F_W \times M_A. \tag{7}$$

For boosting training of the watermarking model, E_θ adopts the global residual skip connection to generate I_{en} by

$$I_{\text{en}} = I_{\text{in}} + F_{AW}. \tag{8}$$

The objective of encoder training is to generate the encoded image I_{en} with high invisibility, and the encoding loss function l_E is designed to minimize the distance between I_{in} and I_{en} by updating the parameter θ. l_E is composed of the image reconstruction loss l_{E1} and the visual loss l_{E2}

$$\begin{aligned} l_E &= \lambda_1 l_{E1} + \lambda_2 l_{E2} \\ &= \lambda_1 \text{MSE}(I_{\text{in}}, I_{\text{en}}) + \lambda_2 \text{SSIM}(I_{\text{in}}, I_{\text{en}}) \end{aligned} \tag{9}$$

where MSE$(\cdot)$ is the mean-square-error function and SSIM$(\cdot)$ is the structural similarity index metric. λ_1 and λ_2 represent the weights of image reconstruction loss and visual loss, respectively.

C. Noise Subnetwork

In general, watermark can be extracted when it is hidden in regions without serious distortion or corresponding features, including watermark, that can tolerate image noises. Therefore, the distribution of differentiable noises can be learned to avoid watermark embedded into serious distorted areas, since these noises support backpropagation. To improve watermarking robustness, the noise subnetwork N is designed for simulating various attacks as some differentiable network layers in iterative training. On the one hand, keeping noises in the training loop learns the robust watermarking pattern to resist various image attacks. On the other hand, noise training can obtain robustness on other untrained noises, which have similar natures of trained noises.

In N, the noised image I_{no} is generated by using different types of noises

$$I_{no} = N_{\text{net}}(I_{\text{en}}, N_t) \tag{10}$$

where $N_{\text{net}}(\cdot)$ represents the noise adding and N_t is a trained noise.

As illustrated in Fig. 4, the noise subnetwork consists of five noise layers, and each layer provides a corresponding noise in the training loop. Dropout (P), Gaussian blur (δ), JPEG (Q), resize (R), and random crop (P) are provided for each noise layer, respectively. Dropout (P) represents that some pixels of I_{en} are randomly selected by probability P to be retained, and the rest is filled by I_{in}. Gaussian blur (δ) means the Gaussian filter with a 3×3 convolution kernel and the variance of δ. JPEG (Q) represents JPEG compression with the quality factor Q. Resize (R) denotes that the image is enlarged or reduced according to the scaling factor R. Random crop (P) is randomly clipped with the probability P. For improving watermarking robustness, one noise layer is randomly selected for each training loop, as illustrated in Fig. 4.

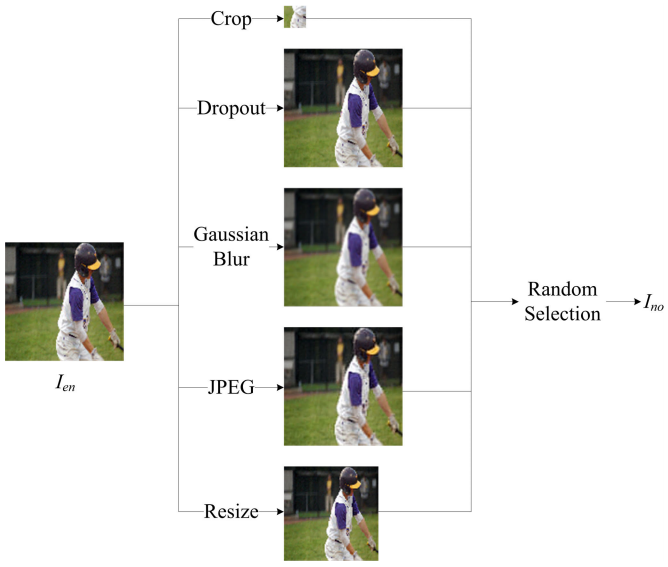


Fig. 4. Structure of noise subnetwork.

D. Decoder for Watermark Extraction

The decoder D_φ is the inverse process of E_θ without the AM, and D_φ is used to extract W_{out} from I_{no}

$$W_{\text{out}} = D(I_{no}) \tag{11}$$

where $D(\cdot)$ represents the decoding process. D_φ consists of a feature extraction layer Layer_E, a watermark generation layer Layer_C, and a threshold function *Hard_Threshold*. Layer_E is composed of four convolution sublayers, which adopt the dense connection to extract the deep feature F_N of I_{no}, and Layer_C includes a convolution sublayer, an adaptive average pooling, and a linear function.

As illustrated in Fig. 5, these convolution sublayers of Layer_E and Layer_C employ the structure of $\text{Conv}^{3\times3}$-BN-LR, wherein $\text{Conv}^{3\times3}$–BN–LR consists of a convolution filter with the 3×3 kernel, BatchNorm, and a leaky ReLU (LeakyReLU) activation. The convolution sublayer of Layer_C is utilized to reduce the dimension of F_N for generating the feature F_C with the same dimension as L. The adaptive average pooling computes the average values of each dimension in F_C, and the linear function obtains a sequence S for watermark extraction. W_{out} is generated with a threshold function by

$$W_{\text{out}} = \textit{Hard\_Threshold}(S, 0.5) \tag{12}$$

where $\textit{Hard\_Threshold}(a, b)$ represents the value of a greater than b is 1 and the value of a less than b is 0.

To obtain watermarking robustness, the objective of decoder training is to minimize the difference between W_{out} and W_{in} by updating parameter φ, and the decoding loss l_D is expressed by

$$l_D = \frac{\| W_{\text{in}} - W_{\text{out}} \|_2^2}{L}. \tag{13}$$

As illustrated in Fig. 2, the decoder and the noise subnetwork adopt cotraining to obtain watermarking robustness, and the purpose of cotraining ensures that (13) keeps fit during

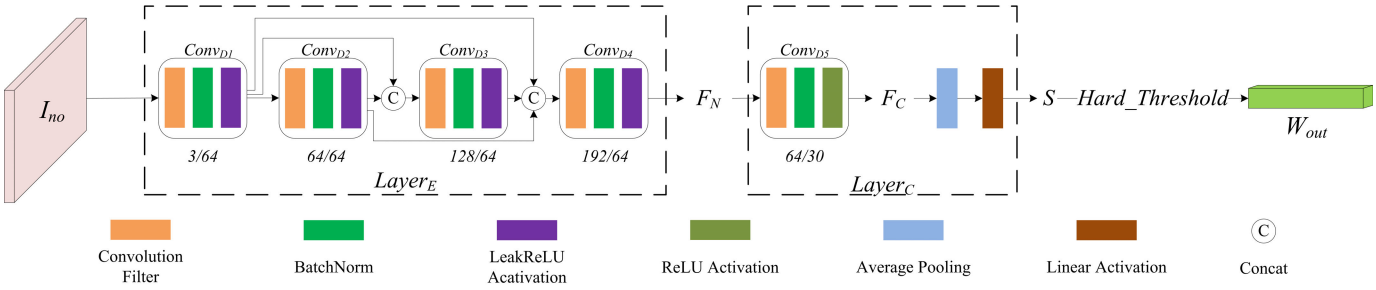


Fig. 5. Structure of decoder.

the training. Although I_{en} contains a complete watermark, N provides different noises to distort watermark. Since this kind of distortion is not uniform on each image region, robust positions are adjusted to preserve more information of watermark during training. D_φ learns the noise distribution, and meanwhile, the feedback of D_φ forces E_θ to embed watermark into robust regions. Therefore, I_{en} is distorted to some extent, but watermark can be extracted correctly. In addition, the preprocessing of W_{in} ensures that the decoder could extract watermark, while any location of I_{no} is visited, and watermark copies of all locations are considered to form the final watermark.

As above stated, watermark extraction is the inverse process of embedding, and the decoder adopts a similar structure as that of the FFM to extract the image features of the encoded image for recovering watermark. In the decoding process, the dense connection is still employed to extract deep features, which are hoped to be similar as features extracted from the encoder. Then, the adaptive average pooling is used to averages extracted features across each dimension, and the hard threshold function is utilized to obtain final decoded watermark by (12). In order to increase the watermark extraction accuracy, the adaptive average pooling and the hard threshold function supply error tolerance to the decoder via noise training. For example, if watermark is distorted by noises to some extent, that is, the extracted features from the decoder slightly differ from encoding features, and watermark can be still extracted accurately. However, when image distortion exceeds maximum error tolerance of the decoder, watermark cannot be extracted, and the decoding ability of the decoder is suppressed. In order to avoid embedding watermark into fragile regions, the feedback of the decoder forces the encoder to adjust image features during noise training. Meanwhile, the encoded image is also tuned as well for accurate decoding. Thus, with the help of the decoding loss, the similar structure and iterative learning ensure that watermarked features can be captured effectively for watermark extraction. Without the original image, watermark can be reconstructed blindly in the decoder.

E. Discriminator

The discriminator A_γ tries to discriminate the encoded image I_{en} from the original image I_{in} by using the distribution of I_{in}, yet in the challenge, the encoder will deceive A_γ by generating I_{en} similar to I_{in}, so that A_γ cannot make a correct judgment. In essence, A_γ judges whether the encoded image contains watermark. In iterative training, there is an adversarial relationship for improving the encoded image quality constantly until Nash equilibrium is reached. As illustrated in Fig. 6, A_γ uses the similar structure as that of the decoder; that is, Conv_{A1}, Conv_{A2}, Conv_{A3}, Conv_{A4}, and Conv_{A5} use the structure of $\text{Conv}^{3\times3}$–BN–LR. In particular, the output of the discriminator is a binary classification rather than the binary watermark of the decoder for judging.

To increase invisibility of the encoded image, the objective of the discriminator is to minimize the possibility that the discriminator makes an incorrect judgment by updating parameter γ, and the adversarial loss l_A is expressed by

$$l_A = \mathbb{E}_{p\sim I_{\text{in}}}\left[\log(1 - A(I_{\text{en}}))\right] \tag{14}$$

where $A(\cdot)$ represents the probability that I_{en} contains W_{in}.

In order to obtain high watermarking invisibility and robustness, the objective of model training is to minimize the above loss function with different weights, and the total loss is expressed by

$$l_t = \lambda_1 \times l_{E1} + \lambda_2 \times l_{E2} + \lambda_3 \times l_D + \lambda_4 \times l_A \tag{15}$$

where λ_1–λ_4 represent the weights of l_{E1}, l_{E2}, l_D, and l_A, respectively.

ARWGAN adopts end-to-end training to achieve the objective by reducing the total loss, which includes the encoding loss, the decoding loss, and the adversarial loss. Specifically, the parameters of the encoder and the decoder are updated by minimizing the encoding loss and the decoding loss, respectively. In particular, to embed watermark into imperceptible and robustness areas, it is important to ensure that the updated parameters of the encoder are not only controlled by the encoding loss, but also indirectly constrained by the decoding loss. For instance, when image features for fusing watermark are not robust enough to resist trained noises, the feedback of the decoder forces the encoder to adjust image features. Meanwhile, the encoded image is also tuned for accurate decoding. In addition, the discriminator utilizes the adversarial loss to update its parameters for accurately evaluating the effect of embedding on image quality, which gives feedback to the encoder for further boosting watermarking invisibility during the iterative training. Thus, the encoder, the decoder, and the discriminator are jointly trained for obtaining the satisfactory performance.

IV. Experimental Results and Discussions

The proposed watermarking model is implemented by PyTorch and executed on NVIDIA GeForce RTX 2080 Ti. The Microsoft Common Objects in Context (COCO) [67] dataset including 80 different scenarios is used for training and testing, wherein more than 100 000 images are selected for training,

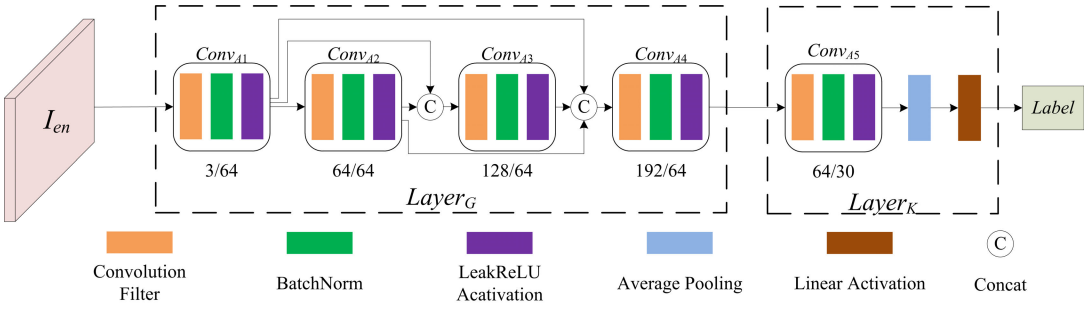


Fig. 6. Structure of discriminator.

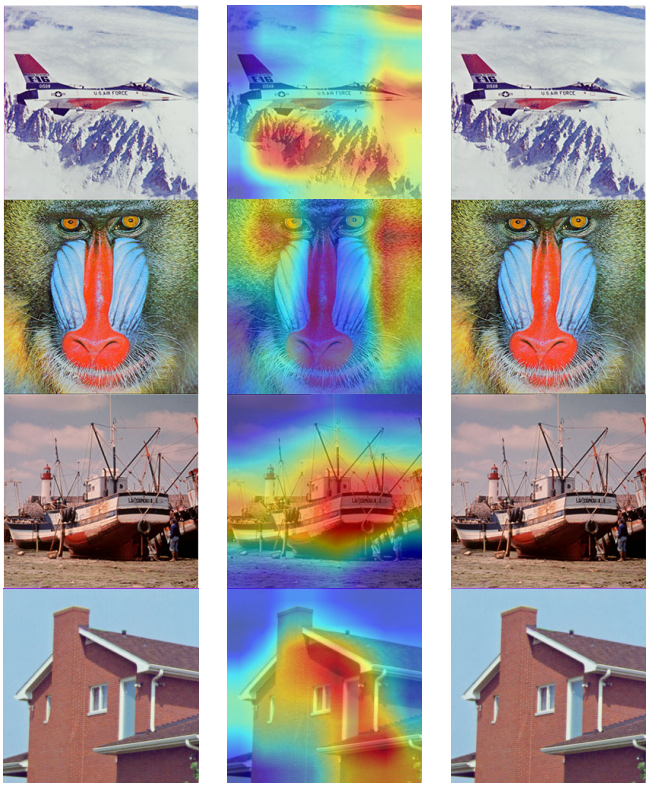
and without overlapping with the trained image, 3000 images of COCO are randomly chosen to test the performance of the proposed watermarking model. Besides COCO, 1000 images from pattern analysis, statistical modelling and computational learning visual object classes 2012 (PASCAL VOC2012) [68], 3000 images from ImageNet [69], and 600 images from the underwater image dataset enhancing underwater visual perception (EUVP) [70] are used for testing as well to assess the cross-dataset generalization ability of ARWGAN. PASCAL VOC2012 supplies much bigger objects in different scenarios compared with COCO, and ImageNet has 1000 different scenarios, most of which are not included in COCO, such as bathtub, carriage, and tank. In addition, EUVP is the specific dataset for providing different underwater scenarios. All these images are resized to $128 \times 128 \times 3$. Peak signal-to-noise ratio (PSNR) and similarity index metric (SSIM) are used to evaluate watermarking invisibility, while bit accuracy (BA) is used to validate watermarking robustness. The weights of loss functions are $\lambda_1 = 0.7$, $\lambda_2 = 0.1$, $\lambda_3 = 1.5$, and $\lambda_4 = 10e^{-3}$, and for Adam's gradient descent, the learning rate is $l_r = 10e^{-3}$.

In order to explore the effectiveness of noise training, two versions of ARWGAN are trained. The first one is trained without noise, which is named ARWGAN-NN. The second one is named ARWGAN-CN and trained with five types of noises, which are dropout (30%), Gaussian blur (2.0), JPEG (50), resize (0.8), and random crop (3.5%), respectively. For a comprehensive comparison, we also train two versions of [38] as the baseline. One trained without noise is named identity, and the other trained with the same noises as those of ARWGAN-CN is named HiDDeN. In addition, five SOTA watermarking models are used for comparison as well, which are ReDMark [35], DA [36], TSDL [40], MBRS [41], and SSLW [42], respectively. For ReDMark, DA, and TSDL, we try to conduct their experiments, but the performance is not completely reproduced. Therefore, to respect their experimental results, we directly use the results published in these articles to be compared. In addition, MBRS and SSLW provide pretrained models for testing, and the test results are used for comparison. For a fair comparison, the length of binary watermark is $L = 30$ for all comparative watermarking models.

In the following, at first, watermarking invisibility and robustness of ARWGAN are discussed. Second, the superiority of the proposed watermarking model is demonstrated by comparisons among SOTA watermarking models. Furthermore, PASCAL VOC2012, ImageNet, and EUVP are also employed to validate the cross-dataset generalization ability of ARWGAN-CN. Third, ablation experiments are given to prove the effectiveness of the AM and the FFM. Finally, the computational cost of the proposed watermarking model is analyzed.

TABLE I
INVISIBILITY OF DIFFERENT WATERMARKING MODELS

Model	Invisibility	
	PSNR[.dB]	SSIM
Identity [38]	36.72	0.9727
ARWGAN-NN	**41.48**	**0.9900**
HiDDeN [38]	32.20	0.9227
ARWGAN-CN	**35.87**	**0.9688**



(a) (b) (c)

Fig. 7. Watermarking performance of ARWGAN. (a) Original image. (b) Attention mask. (c) Encoded image.

A. Watermarking Invisibility Evaluation

Fig. 7 shows the subjective invisibility of ARWGAN-CN, which contains the original image, visualization of the attention mask, and the encoded image. Obviously, the original

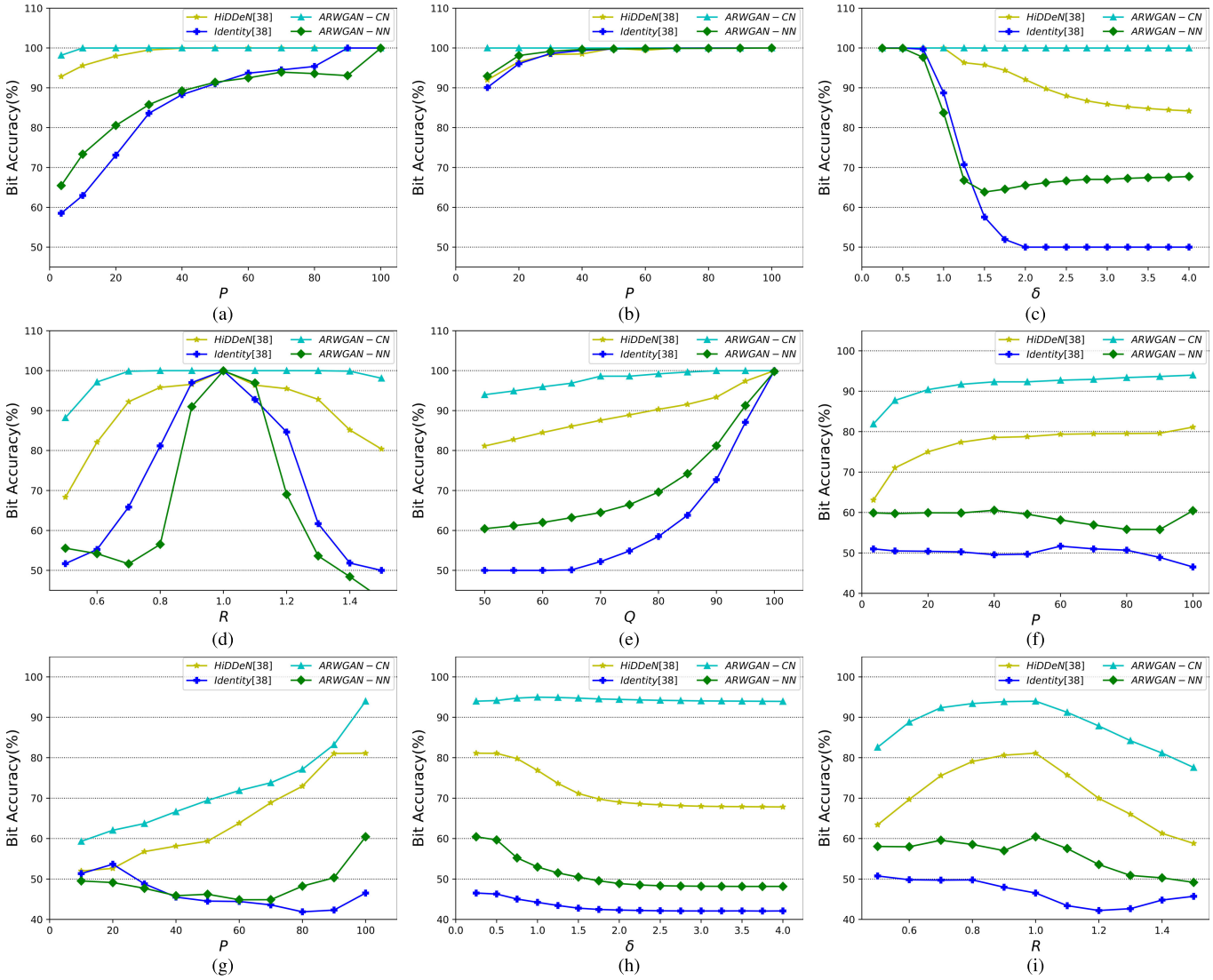


Fig. 8. Robustness on trained noises. (a) Crop. (b) Dropout. (c) Gaussian blur. (d) Resize. (e) JPEG. (f) JPEG (50) + crop. (g) JPEG (50) + dropout. (h) JPEG (50) + Gaussian blur. (i) JPEG (50) + resize.

image and the encoded image are visually indistinguishable, as illustrated in Fig. 7(a) and (c), which demonstrates watermark has been hidden in imperceptible areas. Fig. 7(b) shows the attention masks for different images, where different colors are used to represent different attention degrees. The color similar to red represents inconspicuous and rich texture areas, which are used for the high strength of watermark, while other areas are used to embed the low strength of watermark.

Besides the subjective quality validation, the objective quality of the proposed model is also evaluated, as shown in Table I, wherein PSNRs and SSIMs of those watermarking models are calculated by averaging 3000 encoded images of COCO, respectively. It is obvious that PSNRs and SSIMs of identity and ARWGAN-NN are higher than those of other models with noise training. It is mainly because without noise constraints, network training pays much attention to increasing the image quality, while watermark can be extracted completely under the decoding loss. PSNR and SSIM of ARWGAN-NN are much higher than those of identity, although they are trained without noises, which denotes the AM helps watermark distribution to increase the image quality. For two noise-trained models, ARWGAN-CN is still better than HiDDeN considering PSNR and SSIM. More importantly, PSNR and SSIM of ARWGAN-CN are only 0.85 dB and 0.0039 lower than those of identity, respectively. The main reason is that the AM reduces the embedding influence on the image and has a significant improvement of the image quality.

To evaluate watermarking robustness of the proposed watermarking model, BA is computed by averaging all tested images. At first, different strengths of trained noises are tested by comparing with identity and HiDDeN, as illustrated in Fig. 8. In Fig. 8(a)–(e), when one single noise is attacked on the image, most BAs of ARWGAN-CN are higher than those of the other three models. Especially for Gaussian blur, resize, and JPEG, robustness is much more obvious, and it can be concluded that ARWGAN-CN has higher robustness than the other three models. Moreover, although the image quality of identity and ARWGAN-NN is better than those of HiDDeN and ARWGAN-CN, respectively, BAs of HiDDeN and ARWGAN-CN are much higher, which denotes that the noise subnetwork of the watermarking model is essential to resist image attacks effectively. Many BAs of identity are lower than 75%, such as Gaussian blur with $\delta \geq 1.25$ and JPEG compression with $Q \leq 90$, which means that it cannot

TABLE II
ROBUSTNESS ON UNTRAINED NOISES (%)

Noises	*Identity* [38]	*ARWGAN-NN*	*HiDDeN* [38]	*ARWGAN-CN*
Gaussian Noise (0.06)	84.85	71.01	92.38	**92.84**
Gaussian Noise (0.08)	79.14	66.19	**90.39**	89.60
Gaussian Noise (0.10)	74.85	63.42	**87.89**	86.61
Salt&Pepper (0.05)	94.00	96.42	93.57	**99.41**
Salt&Pepper (0.10)	76.61	85.33	83.02	**96.34**
Salt&Pepper (0.15)	65.16	73.37	73.21	**90.73**
Median Filter (3×3)	45.27	57.74	84.66	**99.71**
Median Filter (5×5)	41.83	57.10	70.23	**91.76**
Median Filter (7×7)	39.54	58.12	63.70	**86.47**
Adjust Brightness(1.1)	91.33	89.82	93.97	**98.28**
Adjust Brightness(1.2)	88.68	86.53	93.00	**94.41**
Adjust Brightness(1.3)	85.39	82.98	91.58	**94.28**
Adjust Constrast(1.0)	81.19	83.93	90.02	**94.97**
Adjust Constrast(1.5)	73.10	77.49	87.77	**93.09**
Adjust Constrast(2.0)	67.38	72.49	85.04	**90.95**
Adjust Hue(0.2)	88.96	93.41	90.69	**93.89**
Adjust Hue(0.3)	83.28	86.30	82.77	**84.94**
Adjust Hue(0.4)	79.59	79.02	**79.91**	78.13
Adjust Saturation (15)	77.85	78.21	85.25	**95.41**
Adjust Saturation (20)	76.86	77.74	84.98	**93.82**
Adjust Saturation (25)	76.23	77.51	83.77	**93.41**
Grid Crop(50%)	96.32	94.07	94.25	**99.82**
Grid Crop(60%)	90.58	86.59	90.39	**99.34**
Grid Crop(70%)	78.55	76.63	81.87	**97.60**
Cropout(30%)	77.63	86.67	93.82	**99.98**
Cropout(20%)	69.23	79.22	89.57	**98.08**
Cropout(10%)	58.40	67.31	76.44	**88.13**
Gaussian Noise(0.08) +Salt&Pepper(0.10)	61.55	61.75	76.96	**82.48**
Adjust Brightness(1.2) + Adjust Contrast(1.5)	73.98	83.40	86.75	**90.99**
Cropout(20%) + Grid Crop (60%)	53.63	59.79	54.94	**73.49**
Grid Crop(50%)+Adjust Brightness(1.1)+Median Filter(3×3)	52.97	54.94	68.53	**77.62**
Average	73.68	76.27	83.76	**91.82**

protect image copyright to some extent. In some cases, BAs of ARWGAN-NN are higher than those of HiDDeN, such as crop and dropout, which shows that the proposed watermarking model without the noise subnetwork still has robustness. It proves that deep and shallow image features are extracted to be fused with watermark for robustness.

Then, to show watermarking robustness of the proposed watermarking model again, combined noises are used for evaluation as well. Fig. 8(f)–(i) shows that ARWGAN-CN outperforms the other three models, since most BAs are the highest. It is clearly seen that the proposed watermarking model can resist combined noises validly. However, in con-

TABLE III
COMPARISON OF DIFFERENT MODELS ON COCO

Model	Invisibility		Robustness [.%]					
	PSNR[.dB]	SSIM	JPEG (50)	Cropout (30%)	Dropout (30%)	Crop (3.5%)	Gaussian (2.0)	Average
ReDMark [35]	**35.93**	0.9660	74.60	92.50	92.00	**100.00**	50.00	81.82
DA [36]	33.70	-	81.70	-	97.90	93.50	60.00	83.28
TSDL [40]	33.50	-	76.20	97.30	97.40	89.00	98.60	91.70
MBRS [41]	35.84	0.8899	91.97	**99.98**	99.96	92.68	**100.00**	96.92
SSLW[42]	34.00	0.8725	83.01	79.66	88.11	50.73	98.96	80.09
ARWGAN-CN	35.87	**0.9688**	**93.98**	99.82	**100.00**	98.17	99.99	**98.39**

TABLE IV
COMPARISON OF DIFFERENT MODELS ON PASCAL VOC2012

Model	Invisibility		Robustness[.%]					
	PSNR[.dB]	SSIM	JPEG (50)	Cropout (30%)	Dropout (30%)	Crop (3.5%)	Gaussian (2.0)	Average
HiDDeN [38]	33.37	0.9333	77.58	96.05	95.76	96.02	93.40	91.76
MBRS [41]	36.64	0.8938	**97.99**	**100.00**	**100.00**	93.21	**100.00**	98.24
SSLW[42]	33.90	0.8613	77.33	77.65	87.88	50.65	99.40	78.90
ARWGAN-CN	**36.83**	**0.9698**	94.45	**100.00**	**100.00**	**99.99**	**100.00**	**98.89**

trast to ARWGAN-CN and HiDDeN, it is an irregular trend that BAs of ARWGAN-NN and identity are not increased when corresponding noises are decreased, as illustrated in Fig. 8(g) and (i). This irregular trend indicates that the noise subnetwork is necessary for the robust watermarking model again. In all, the above experiments prove that the proposed ARWGAN-CN performs well on the trained noise, including the combined noise.

A practical watermarking system must be robust to a wide range of image noises, not just in the range of trained noises. Nine types of untrained noises are used for testing the generalization of resisting various image attacks, i.e., Gaussian noise (δ), salt and pepper (P), median filter (W), adjust contrast (R), adjust brightness (R), adjust hue (R), adjust saturation (R), grid crop (P), and cropout (P), wherein W means the filter window size. Meanwhile, combined noises are utilized to evaluate watermarking robustness, and these noises represent heavier distortions than a single noise. These combinations include two or three noises, such as adjust brightness (1.2) + adjust contrast (1.5) and median filter (3×3) + adjust brightness (1.1) + grid crop (50%). In Table II, BAs of different watermarking models for those untrained noises are compared. When the average value of BA is computed, results lower than 50% are truncated to 50%, since they are no better than random chance. We can see that some BAs are lower than 70%, such as median filter (5×5) and median filter (7×7) for identity, and it denotes that the corresponding watermarking models do not have the capability of resisting those noises. From Table II, we can see that most BAs of ARWGAN-CN for resisting untrained noises are higher than 75%, which proves the capability of copyright protection. For most of the single untrained noises, ARWGAN-CN outperforms the other three models, and only for a few noises, such as Gaussian noise (0.10) and adjust contrast (2.0), BAs of ARWGAN-CN are a little lower than those of HiDDeN. For mixed noises, ARWGAN-CN has stronger robustness than other models, and especially for grid crop (50%) + adjust brightness (1.1) + median filter (3×3), BA of ARWGAN-CN is at least 11% higher than those of other models. Even if, BA of ARWGAN-CN is lower than 75%, is 18% higher than that of HiDDeN for grid crop (60%) + cropout (20%). Totally, the average BA is 8% higher than those of the other three models at least, which indicates ARWGAN-CN is also robust to untrained noises besides resisting trained noises, and has the generalization ability to resist various noises. Furthermore, it is interesting to note that ARWGAN-NN still has high robustness on these untrained noises, and BAs of ARWGAN-NN are similar to those of HiDDeN. The main reason is that multiple features with different layers, which are reused by the dense connection, can represent main image energies, and more importantly, those features fusing with watermark are learned iteratively for improving robustness. It also proves that multilayer watermark fusion can increase robustness much. From experiments on resisting trained and untrained noises, it is clearly seen that ARWGAN has a strong ability to protect the image copyright. Except above tested noises, ARWGAN-CN may fail to resist some untrained image attacks, since their distributions of image distortions are much different from those of trained ones. Thus, if a special noise cannot be handled by the proposed watermarking model, and

TABLE V
COMPARISON OF DIFFERENT MODELS ON IMAGENET

Model	Invisibility		Robustness[.%]					
	PSNR[.dB]	SSIM	JPEG (50)	Cropout (30%)	Dropout (30%)	Crop (3.5%)	Gaussian (2.0)	Average
HiDDeN [38]	33.26	0.9322	77.30	95.98	95.67	96.02	92.89	91.57
MBRS [41]	36.44	0.8919	92.64	99.99	99.98	92.93	**100.00**	97.11
SSLW[42]	33.50	0.8412	77.48	80.22	86.88	50.52	98.91	79.13
ARWGAN-CN	**36.66**	**0.9685**	**94.08**	**99.99**	**100.00**	**99.95**	99.98	**98.80**

TABLE VI
COMPARISON OF DIFFERENT MODELS ON EUVP

Model	Invisibility		Robustness[.%]					
	PSNR[.dB]	SSIM	JPEG (50)	Cropout (30%)	Dropout (30%)	Crop (3.5%)	Gaussian (2.0)	Average
HiDDeN [38]	33.53	0.9272	79.98	95.49	95.34	95.07	93.76	91.93
MBRS [41]	36.19	0.9009	91.88	**99.98**	**99.96**	93.11	**100.00**	96.99
SSLW[42]	34.20	0.8531	83.62	77.69	88.22	50.13	99.27	79.79
ARWGAN-CN	**36.39**	**0.9623**	**93.63**	**99.99**	**100.00**	**99.97**	99.97	**98.72**

in the real application, we will add it to the noise subnetwork for training, and consequently, the ability to resist that noise can be obtained.

B. Comparison With SOTA Models

To further prove the effectiveness of the proposed model, the existing SOTA watermarking models with noise training are used to be compared. Table III shows invisibility and robustness comparisons on COCO. From Table III, we can discover that ARWGAN-CN achieves the best performance among all SOTA models. The image quality and robustness of ARWGAN-CN are all better compared with DA, TSDL, and SSLW. Especially, since DA cannot resist Gaussian noise, and SSLW performs badly on crop, their average BAs for different noises are lower, which indicates those two watermarking models do not have the capability of resisting varieties of image attacks. For the average BA, ARWGAN-CN nearly obtains a 17% gain over RedMark, although PSNR of ARWGAN-CN is only 0.07 dB lower than that of RedMark. Moreover, the average BA of ARWGAN-CN is still 1.47% higher than that of MBRS, which is the second best watermarking model in terms of robustness. In particular, considering SSIM, ARWGAN-CN has 0.0789 higher than MBRS, which demonstrates that ARWGAN-CN can generate encoded images with the greater quality.

To show the cross-dataset generalization ability, PASCAL VOC2012, ImageNet, and EUVP are also utilized for evaluation. HiDDeN, MBRS, SSLW, and ARWGAN-CN are compared, as shown in Tables IV–VI. At first, from Table IV on PASCAL VOC2012, we can see that ARWGAN-CN is much better than HiDDeN considering large gains of PSNR. In contrast to SSLW, ARWGAN-CN has stronger robustness on each noise and, meanwhile, has better image quality. Compared with MBRS, only for JPEG (50), BA of the proposed watermarking model is lower, but BAs for other noises are higher. Moreover, PSNR and SSIM are higher, and especially, SSIM gain is 0.076, which indicates the visual quality is better. Second, although ImageNet has much more scenarios compared with COCO and PASCAL VOC2012, ARWGAN-CN is still superior to the other two watermarking models, as shown in Table V. To be concrete, PSNR/SSIM is 3.4 dB/0.0363 and 0.22 dB/0.0766 higher than those of HiDDeN and MBRS, respectively, and for average BA, ARWGAN-CN has 7.23% and 1.69% gains over HiDDeN and MBRS, respectively. Third, EUVP is specific and differs from the above three datasets, and because in contrast with natural images, underwater images have color deviations, low contrast and brightness, blurred details, and so on. Notably, similar to the results of Tables IV and V, Table VI shows that ARWGAN-CN is much better than HiDDeN, MBRS, and SSLW as well, and MBRS is also the second. Moreover, we also discover that the average BAs of MBRS for different datasets vary from 98.24% to 96.99%, yet the average BAs of the proposed watermarking are changed a little, which means the proposed watermark model performs stable on different datasets.

We can conclude that ARWGAN-CN obtains a significant advantage over SOTA watermarking models on four different datasets, which demonstrates the superiority and also proves the cross-dataset generalization ability of the proposed watermarking model. The main reason for the optimal watermarking performance is that the FFM employs multilayer fusion to obtain robustness, and the reuse of image features by the dense connection increases robustness further. More importantly, the AM is utilized to compute the attention mask for guiding watermark distribution over different image regions to decrease image distortion caused by embedding watermark, so that image quality is also ensured. In the following, the effectiveness of the AM and the FFM is also demonstrated via the ablation experiment.

TABLE VII
COMPARISON WITH ARWGAN-NA

Model	λ_2	Invisibility		Robustness[.%]					
		PSNR[.dB]	SSIM	Crop (3.5%)	Dropout (10%)	Resize (0.5)	JPEG (50)	Gaussian Blur (4.0)	Average
ARWGAN-CN	0.10	35.87	0.9688	98.17	**99.98**	**88.20**	93.98	**99.98**	96.06
	0.05	33.80	0.9497	97.24	99.56	85.85	92.09	99.96	94.74
	0	30.45	0.8859	**98.65**	99.79	87.06	**96.85**	99.08	**96.28**
ARWGAN-NA	0.10	**36.41**	**0.9689**	96.53	98.77	83.01	84.95	99.96	92.64
	0.05	33.30	0.9385	93.92	94.21	83.18	89.55	95.18	91.21
	0	28.76	0.8348	97.89	99.17	81.85	93.14	99.82	94.37

TABLE VIII
COMPARISON OF AM AND CBAM

Position	Invisibility		Robustness[.%]					
	PSNR[.dB]	SSIM	Crop (3.5%)	Dropout (10%)	Resize (0.5)	JPEG (50)	Gaussian Blur (4.0)	Average
ARWGAN-CBAM	**35.98**	0.9661	96.87	99.95	79.40	76.41	99.96	90.52
ARWGAN-CN	35.87	**0.9688**	**98.17**	**99.98**	**88.20**	**93.98**	**99.98**	**96.06**

C. Ablation Experiments and Analysis

In this section, a series of ablation experiments are designed to precisely illustrate the effectiveness of the AM and the FFM.

1) Influence of AM: As a crucial component of the encoder, the AM is designed for guiding image generation. In order to evaluate the effectiveness of the AM, two watermarking models are designed for comparisons. The first model excludes the AM from ARWGAN, namely, ARWGAN-NA. Since the weight of the visual loss function λ_2 is related to the image quality, to quantitatively analyze the influence of the AM, ARWGAN-CN and ARWGAN-NA are trained with $\lambda_2 = 0.10$, $\lambda_2 = 0.05$, and $\lambda_2 = 0$, respectively. Table VII shows the comparison between ARWGAN-CN and ARWGAN-NA. We can see that when λ_2 is decreased, the image quality is reduced as well. For instance, as λ_2 is set from 0.1 to 0.05, PSNR and SSIM decreases of ARWGAN-CN are 2.07 dB and 0.0197, respectively, and PSNR and SSIM declines of ARWGAN-NA are 3.11 dB and 0.0304, respectively. Thus, in contrast to ARWGAN-NA, the image quality decrease of ARWGAN-CN is smaller. Moreover, it is clearly seen that both ARWGAN-CN ($\lambda_2 = 0$) and ARWGAN-NA ($\lambda_2 = 0$) have low invisibility due to training without the visual loss, whereas it is worth mentioning that ARWGAN-CN ($\lambda_2 = 0$) has higher invisibility compared with ARWGAN-NA ($\lambda_2 = 0$). Although PSNR of ARWGAN-CN ($\lambda_2 = 1.0$) is lower than that of ARWGAN-NA ($\lambda_2 = 1.0$), their SSIMs are very close, which indicates their visual qualities have no difference. More importantly, in terms of average BA, ARWGAN-CN ($\lambda_2 = 1.0$) achieves a 3.42% gain over ARWGAN-NA ($\lambda_2 = 1.0$). Thus, combining robustness and invisibility, ARWGAN-CN is superior to ARWGAN-NA. For robustness, it is clearly seen that when λ_2 is set from 0 to 0.05, the average BA of ARWGAN-NA is reduced from 94.37% to 91.21%, while the average BA of ARWGAN-CN is only decreased from 96.06% to 94.74%. It is interesting that the average BA of ARWGAN-CN ($\lambda_2 = 0.10$) is 0.23% lower than that of ARWGAN-CN ($\lambda_2 = 0$), while the average BA of ARWGAN-NA ($\lambda_2 = 0.10$) is 1.73% lower than that of ARWGAN-NA ($\lambda_2 = 0$). From the above analysis, it is easily known that although invisibility of ARWGAN-CN is changed when λ_2 is set to different values, its robustness is stable. Thus, it also proves ARWGAN-CN is much better than ARWGAN-NA and reveals that the AM has the ability to maintain strong robustness with invisibility improvement. The main reason for the effectiveness of the AM is the attention mask, which forces the encoded image to follow the original image features for reducing the distortion caused by embedding watermark to some extent. Besides the inconspicuous areas for the image quality, the AM finds the texture areas to embed high embedding strength for robustness.

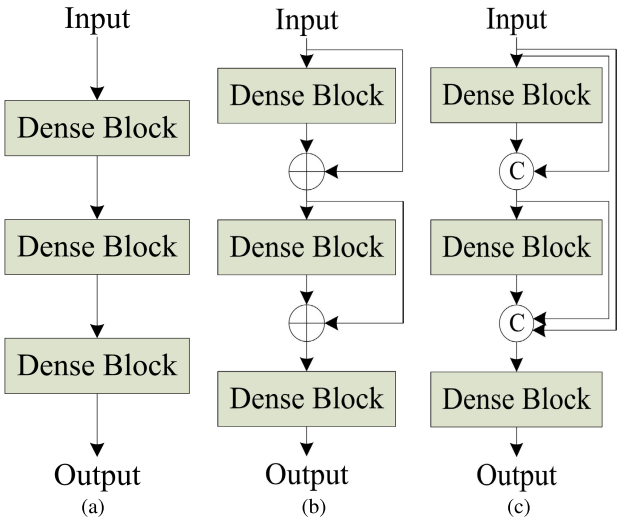


Fig. 9. Three different connections. (a) Successive connection. (b) Skip connection. (c) Dense connection.

The second model uses the CBAM [62] as the attention mechanism in ARWGAN instead of the AM, namely, ARWGAN-CBAM, since the CBAM is very effective in many vision tasks [71], [72]. As shown in Table VIII, despite the fact that PSNR of ARWGAN-CN is only 0.11 dB lower than that of ARWGAN-CBAM, SSIM is 0.0027 higher, which means their visual qualities are similar. In terms of robustness comparisons, BA of ARWGAN-CN is higher than that of ARWGAN-CBAM for each noise. Especially for resize (0.5) and JPEG (50), BAs of ARWGAN-CN are 8.8% and 17.57% higher than those of ARWGAN-CBAM, respectively. Obviously, ARWGAN-CN is superior to ARWGAN-CBAM. It is mainly because the computed attention mask of the CBAM is not proper for guiding to distribute watermark

TABLE IX
COMPARISON OF DIFFERENT CONNECTIONS

Position	Invisibility		Robustness[.%]					
	PSNR[.dB]	SSIM	Crop (3.5%)	Dropout (10%)	Resize (0.5)	JPEG (50)	Gaussian Blur (4.0)	Average
ARWGAN-SeC	**36.12**	0.9676	98.92	99.07	84.10	86.97	98.98	93.61
ARWGAN-SkC	35.21	0.9592	**99.51**	99.75	85.73	89.27	99.91	94.83
ARWGAN-CN	35.87	**0.9688**	98.17	**99.98**	**88.20**	**93.98**	**99.98**	**96.06**

TABLE X
INFLUENCE ON DIFFERENT POSITIONS FOR WATERMARK FUSION

Position	Invisibility		Robustness[.%]					
	PSNR[.dB]	SSIM	Crop (3.5%)	Dropout (10%)	Resize(0.5)	JPEG (50)	Gaussian Blur (4.0)	Average
①	36.82	0.9737	46.97	47.17	52.10	57.97	46.80	50.20
②	36.78	0.9737	70.97	71.11	67.76	78.44	73.66	72.39
③	**37.07**	**0.9758**	78.17	78.64	68.71	73.07	70.54	73.83
①②	36.73	0.9733	75.64	77.26	71.17	81.02	76.96	76.41
①③	37.04	0.9756	83.51	84.35	71.56	77.13	78.40	78.99
②③	36.08	0.9703	97.96	98.99	83.69	90.92	96.93	93.70
①②③	35.87	0.9688	**98.17**	**99.98**	**88.20**	**93.98**	**99.98**	**96.06**

due to its different structures with that of the FFM. On the contrary, the AM and the FFM have the similar structure, and hence, the AM can learn similar features as those of the FFM. Meanwhile, probability distribution among those feature channels is computed to capture the attention mask for adjusting watermarked features in the FFM. In contrast to the CBAM, the AM of ARWGAN-CN simulates the embedding processes to extract image features for computation of the attention mask, so that more robust features are learned to improve robustness for ARWGAN-CN. Based on comparison with ARWGAN-CBAM, we prove the effectiveness of the AM again and also show that not all existing attention techniques are adaptive to the proposed watermarking model. In conclusion, the AM is an essential part of ARWGAN.

2) Influence on Different Positions for Watermark Fusion: The FFM is designed to extract shallow and deep image features, and meanwhile, those features are reused by the dense connection for obtaining robustness. Thus, in order to validate the dense connection of the FFM, two models used for comparisons replace the dense connection of ARWGAN-CN with the successive connection and the skip connection, namely, ARWGAN-SuC and ARWGAN-SkC, respectively. Three connections are illustrated in Fig. 9. The training environments of the three watermarking models are the same, and their corresponding results are shown in Table IX.

As shown in Table IX, PSNR and SSIM of ARWGAN-SkC are 0.91 dB and 0.0084 lower than those of ARWGAN-SeC, respectively, but the average BA is 1.22% higher. It is mainly because compared with the successive connection, the skip connection of ARWGAN-SkC has the ability to reuse image features for increasing robustness. However, only the forward features of the previous dense block are reused in ARWGAN-SkC. Notably, the dense connection of ARWGAN-CN reuses more forward features, and robustness and invisibility are enhanced much. Compared with ARWGAN-SkC, the average BA of AGWGAN-CN is 1.6% higher, and PSNR/SSIM is 0.66 dB/0.0096 higher. The principle reason is that besides the deep features, each dense block in ARWGAN-CN relearns shallow features for fusing with watermark to increase robustness, and simultaneously, shallow features for holding watermark can reduce image distortion. However, in ARWGAN-SkC, only one dense block reuses shallow features for fusing with watermark, and thus, more deep features are used to embed watermark for robustness, which degrades the image quality. Compared with ARWGAN-SeC, despite PSNR of ARWGAN-CN being only 0.25 dB lower, SSIM is a little higher. More critically, the average BA is 2.79% higher than that of ARWGAN-SeC. The main cause is that ARWGAN-SeC only uses deep features of the dense block without reusing, which degrades robustness greatly.

As above stated, compared with the successive connection and the skip connection, the dense connection has more bypasses to preserve the preceding features, which indicates that shallow and deep features are relearned in each dense block, which improves the watermarking performance. As a result, it can be easily concluded that the dense connection is an effective part of the FFM to improve the watermarking performance.

Besides the dense connection, multilayer fusion also increases robustness in the FFM. In the FFM, multiple positions of Layer_D are utilized for feature fusion with watermark and are discussed to prove the rationality of the designed FFM. Table X displays invisibility and robustness of watermark fusion in different positions, as illustrated in Fig. 3. In Table X, "①" denotes that only position ① in the FFM is used for fusing with watermark, and "①②" means watermark fusion exists in both positions ① and ②.

As shown in Table X, watermark fusion in different positions has a limited impact on invisibility, but a huge impact on robustness. Compared with position ①, robustness of positions ② and ③ is increased by 22.19% and 23.63%, respectively. It is mainly because image features of position ① belong to shallow features, which are more fragile than deep features, whereas features of positions ② and ③ are deep and robust to be easily unchanged under noises. In addition, it can be easily discovered that position ②③ has higher robustness than positions ①② and ①③. The main reason is that shallow features of position ① are more fragile in contrast to that of positions ② and ③. However, compared with position ①②③, robustness of position ②③ is still less, and thus, position ① is still necessary for fusing with watermark to

TABLE XI
COMPARISON OF COMPUTATIONAL COST

Model	Param. (M)	FLOPs (G)	Speed (im/s)
ARWGAN-CN	1.5M	21.14G	47.25
ARWGAN-NA	0.9M	11.29G	51.08
ARWGAN-SuC	1.3M	19.17G	40.94
ARWGAN-SkC	1.3M	19.17G	37.71
HiDDeN[38]	**0.4M**	**3.52G**	**57.39**
MBRS[41]	5.8M	13.36G	17.67

increase robustness. The reason lies in the reuse of shallow features by the dense connection.

Based on the above experiments on different positions for watermark fusion, we can conclude that fusing deep features with watermark obtains robustness, and at the same time, relearning shallow and deep features by the dense connection enhances the capability of resisting different noises. Multilayer fusion with three different positions is necessary to be utilized for improving watermarking robustness. In a word, the design of the FFM is proven to be effective for the proposed ARWGAN.

D. Discussion of Computational Cost

To evaluate computational costs of different watermarking models, the model size, the time complexity, and the throughput are used for comparisons, as shown in Table XI, where the model size is represented by the number of parameters, the time complexity is provided by the number of floating-point operations per second (FLOPs), and the throughput is donated by the number of images processed per second.

Table XI shows that the parameter number and FLOPs of ARWGAN-CN are greater than those of ARWGAN-NA, while the speed is lower. It is mainly because the AM of ARWGAN-CN increases the computational cost. Compared with ARWGAN-SuC and ARWGAN-SkC, the parameter number and FLOPs of ARWGAN-CN are much more, and the speed of ARWGAN-CN is lower. The critical reason is that the dense connection reuses more features than the successive connection and the skip connection do. Despite HiDDeN having the fastest processing speed and the smallest model size among all watermark models due to using simple convolution operations, its watermarking performance is not good as above stated. MBRS is the second best in terms of robustness and invisibility, yet has much more parameters than ARWGAN-CN. The main cause is that MBRS needs to preprocess watermark in a more complex way and to process a great deal of squeeze-and-excitation and convolution operations [41]. Although FLOPs of ARWGAN-CN are larger than that of MBRS, the computational speed is faster. Thus, ARWGAN-CN is still superior to MBRS considering invisibility, robustness, and computational speed.

To obtain high invisibility and robustness, the AM and the FFM are designed in ARWGAN-CN. It is evident that the AM and the FFM significantly improve the performance much via the ablation experiments, but they increase the computational complexity, since the computation of the attention mask and the multilayer fusion with watermark increase the corresponding parameters. Thus, we have to admit that despite ARWGAN-CN obtaining the good performance in terms of watermarking invisibility and robustness, the high computational cost of ARWGAN-CN limits its application due to the requirements of real time in practical watermarking systems. Meanwhile, the generality of the proposed watermarking model is limited as well due to the high computational cost.

V. CONCLUSION

In this article, we have presented an ARWGAN. The proposed model, which is trained with the end-to-end framework, consists of the encoder, the decoder, the noise subnetwork, and the discriminator. In the existing deep learning-based watermarking models, not all image features are not learned enough to be fused with watermark, and meanwhile, some important feature are not enhanced, so that the watermarking performance is degraded. In order to solve those shortcomings, the FFM and the AM in the encoder are deployed. The FFM adopts the dense connection to extract the shallow and deep features, which are reused to be fused with watermark for boosting watermarking robustness. To reduce image distortion caused by embedding watermark, the AM is incorporated into ARWGAN for allocating watermark into inconspicuous regions by computing probability distribution among original image feature channels. Besides improving the image quality, robustness is still increased by finding rich texture regions to embed watermark. In addition, the noise subnetwork and the discriminator are utilized to improve robustness and invisibility during the training loop, respectively. Experimental results denote ARWGAN not only obtains the high image quality, but also resists different trained and untrained noises effectively. Compared with the SOTA watermark models, ARWGAN is better in terms of watermarking invisibility and robustness for different datasets. Ablation experiments also prove the effectiveness of the AM and the FFM. However, ARWGAN has a heavy computation cost due to incorporating the AM and the FFM. Furthermore, the embedding capacity is not large enough. In the future work, in order to gain the watermarking generality, we will mine effective lightweight network modules to decrease the computation time for real time application, and meanwhile, watermarking invisibility and robustness are still hoped to be improved.

REFERENCES

[1] L. Xiong, X. Han, C. Yang, and Y. Shi, "Robust reversible watermarking in encrypted image with secure multi-party based on lightweight cryptography," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 32, no. 1, pp. 75–91, Jan. 2022.

[2] I. C. Dragoi and D. Coltuc, "On the security of reversible data hiding in encrypted images by MSB prediction," *IEEE Trans. Inf. Forensics Security*, vol. 16, pp. 187–189, 2021.

[3] A. Kamili, N. N. Hurrah, S. A. Parah, G. M. Bhat, and K. Muhammad, "DWFCAT: Dual watermarking framework for industrial image authentication and tamper localization," *IEEE Trans. Ind. Informat.*, vol. 17, no. 7, pp. 5108–5117, Jul. 2021.

[4] M. Sadeghi, R. Toosi, and M. A. Akhaee, "Blind gain invariant image watermarking using random projection approach," *Signal Process.*, vol. 163, pp. 213–224, Oct. 2019.

[5] Y. Huang, B. Niu, H. Guan, and S. Zhang, "Enhancing image watermarking with adaptive embedding parameter and PSNR guarantee," *IEEE Trans. Multimedia*, vol. 21, no. 10, pp. 2447–2460, Oct. 2019.

[6] V. Amanipour and S. Ghaemmaghami, "Video-tampering detection and content reconstruction via self-embedding," *IEEE Trans. Instrum. Meas.*, vol. 67, no. 3, pp. 505–515, Mar. 2018.

[7] K. Wang, L. Li, T. Luo, and C. C. Chang, "Deep neural network watermarking based on texture analysis," in *Proc. Int. Conf. Artif. Intell. Secur.* Singapore: Springer, Jul. 2020, pp. 558–569.

[8] L. Kuo and C. Tai, "Robust image-based water-level estimation using single-camera monitoring," *IEEE Trans. Instrum. Meas.*, vol. 71, pp. 1–11, 2022.

[9] J. Molina-Garcia, B. P. Garcia-Salgado, V. Ponomaryov, R. Reyes-Reyes, S. Sadovnychiy, and C. Cruz-Ramos, "An effective fragile watermarking scheme for color image tampering detection and self-recovery," *Signal Process., Image Commun.*, vol. 81, Feb. 2020, Art. no. 115725.

[10] P. Lefèvre, P. Carré, C. Fontaine, P. Gaborit, and J. Huang, "Efficient image tampering localization using semi-fragile watermarking and error control codes," *Signal Process.*, vol. 190, Jan. 2022, Art. no. 108342.

[11] F. Peng, Z. Lin, X. Zhang, and M. Long, "A semi-fragile reversible watermarking for authenticating 2D engineering graphics based on improved region nesting," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 31, no. 1, pp. 411–424, Jan. 2021.

[12] X. Wang, X. Li, and Q. Pei, "Independent embedding domain based two-stage robust reversible watermarking," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 30, no. 8, pp. 2406–2417, Aug. 2020.

[13] J. A. Cortés-Osorio, J. B. Gómez-Mendoza, and J. C. Riaño-Rojas, "Velocity estimation from a single linear motion blurred image using discrete cosine transform," *IEEE Trans. Instrum. Meas.*, vol. 68, no. 10, pp. 4038–4050, Oct. 2019.

[14] Y. Zhang and G. Sun, "A watermark algorithm based on space-domain and transform-domain," in *Proc. IEEE 9th Int. Conf. Electron. Inf. Emergency Commun. (ICEIEC)*, Jul. 2019, pp. 41–44.

[15] T. Jin and W. Zhang, "A novel interpolated DFT synchrophasor estimation algorithm with an optimized combined cosine self-convolution window," *IEEE Trans. Instrum. Meas.*, vol. 70, pp. 1–10, 2021.

[16] H. Sadreazami and M. Amini, "A robust image watermarking scheme using local statistical distribution in the contourlet domain," *IEEE Trans. Circuits Syst. II, Exp. Briefs*, vol. 66, no. 1, pp. 151–155, Jan. 2019.

[17] S. Shi, T. Luo, J. Huang, and M. Du, "A novel HDR image zero-watermarking based on shift-invariant shearlet transform," *Secur. Commun. Netw.*, vol. 2021, pp. 1–12, Mar. 2021.

[18] M. Refiyanti, G. Budiman, L. Novamizanti, and M. A. Y. Pratama, "Medical image watermarking using spread spectrum and compressive sensing techniques," in *Proc. 4th Int. Conf. Comput. Informat. Eng. (ICIE)*, Sep. 2021, pp. 406–411.

[19] H.-J. Ko, C.-T. Huang, G. Horng, and S.-J. Wang, "Robust and blind image watermarking in DCT domain using inter-block coefficient correlation," *Inf. Sci.*, vol. 517, pp. 128–147, May 2020.

[20] H.-T. Hu, L.-Y. Hsu, and H.-H. Chou, "An improved SVD-based blind color image watermarking algorithm with mixed modulation incorporated," *Inf. Sci.*, vol. 519, pp. 161–182, May 2020.

[21] J. Liu et al., "An optimized image watermarking method based on HD and SVD in DWT domain," *IEEE Access*, vol. 7, pp. 80849–80860, 2019.

[22] C. Zhou et al., "A new model transfer strategy among spectrometers based on SVR parameter calibrating," *IEEE Trans. Instrum. Meas.*, vol. 70, pp. 1–13, 2021.

[23] C. Yang, P. An, and L. Shen, "Blind image quality measurement via data-driven transform-based feature enhancement," *IEEE Trans. Instrum. Meas.*, vol. 71, pp. 1–12, 2022.

[24] Q. Liu and X. Jiang, "Design and realization of a meaningful digital watermarking algorithm based on RBF neural network," in *Proc. 6th World Congr. Intell. Control Autom.*, 2006, pp. 214–218.

[25] P.-P. Niu, X.-Y. Wang, Y.-P. Yang, and M.-Y. Lu, "A novel color image watermarking scheme in nonsampled contourlet-domain," *Exp. Syst. Appl.*, vol. 38, no. 3, pp. 2081–2098, Mar. 2011.

[26] H. Chen and L. Zhang, "A novel digital watermarking approach under the DCT and SVM," in *Proc. IEEE 4th Int. Conf. Autom., Electron. Electr. Eng. (AUTEEE)*, Nov. 2021, pp. 154–156.

[27] X. Wei, W. Li, M. Zhang, and Q. Li, "Medical hyperspectral image classification based on end-to-end fusion deep neural network," *IEEE Trans. Instrum. Meas.*, vol. 68, no. 11, pp. 4481–4492, Nov. 2019.

[28] A. Lin, B. Chen, J. Xu, Z. Zhang, G. Lu, and D. Zhang, "DS-TransUNet: Dual Swin transformer U-Net for medical image segmentation," *IEEE Trans. Instrum. Meas.*, vol. 71, pp. 1–15, 2022.

[29] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *Proc. Int. Conf. Mach. Learn.*, 2015, pp. 448–456.

[30] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Las Vegas, NV, USA, 2016, pp. 770–778.

[31] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 2261–2269.

[32] B. Chen, Z. Zhang, Y. Li, G. Lu, and D. Zhang, "Multi-label chest X-ray image classification via semantic similarity graph embedding," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 32, no. 4, pp. 2455–2468, Apr. 2022.

[33] 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*, Dec. 2015, pp. 1–6.

[34] S.-M. Mun, S.-H. Nam, H.-U. Jang, D. Kim, and H.-K. Lee, "A robust blind watermarking using convolutional neural network," 2017, *arXiv:1704.03248*.

[35] M. Ahmadi, A. Norouzi, N. Karimi, S. Samavi, and A. Emami, "ReDMark: Framework for residual diffusion watermarking based on deep networks," *Exp. Syst. Appl.*, vol. 146, May 2020, Art. no. 113157.

[36] X. Luo, R. Zhan, H. Chang, F. Yang, and P. Milanfar, "Distortion agnostic deep watermarking," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 13545–13554.

[37] Z. Huang, J. Zhang, Y. Zhang, and H. Shan, "DU-GAN: Generative adversarial networks with dual-domain U-Net-based discriminators for low-dose CT denoising," *IEEE Trans. Instrum. Meas.*, vol. 71, pp. 1–12, 2022.

[38] J. Zhu, R. Kaplan, J. Johnson, and L. Fei-Fei, "HiDDeN: Hiding data with deep network," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 657–672.

[39] 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.

[40] 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.

[41] Z. Jia, H. Fang, and W. Zhang, "MBRS: Enhancing robustness of DNN-based watermarking by mini-batch of real and simulated JPEG compression," in *Proc. 29th ACM Int. Conf. Multimedia*, Oct. 2021, pp. 41–49.

[42] P. Fernandez, A. Sablayrolles, T. Furon, H. Jégou, and M. Douze, "Watermarking images in self-supervised latent spaces," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2022, pp. 3054–3058.

[43] Y. Yang, J. Liu, S. Huang, W. Wan, W. Wen, and J. Guan, "Infrared and visible image fusion via texture conditional generative adversarial network," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 31, no. 12, pp. 4771–4783, Dec. 2021.

[44] T. Geng, X. Liu, X. Wang, and G. Sun, "Deep shearlet residual learning network for single image super-resolution," *IEEE Trans. Image Process.*, vol. 30, pp. 4129–4142, 2021.

[45] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, *arXiv:1409.1556*.

[46] H. H. Tan and K. H. Lim, "Vanishing gradient mitigation with deep learning neural network optimization," in *Proc. 7th Int. Conf. Smart Comput. Commun. (ICSCC)*, Jun. 2019, pp. 1–4.

[47] G. Philipp, D. Song, and J. G. Carbonell, "The exploding gradient problem demystified–definition, prevalence, impact, origin, tradeoffs, and solutions," 2017, *arXiv:1712.05577*.

[48] Z. Wu, J. Hai, L. Zhang, J. Chen, G. Cheng, and B. Yan, "Cascaded fully convolutional DenseNet for automatic kidney segmentation in ultrasound images," in *Proc. 2nd Int. Conf. Artif. Intell. Big Data (ICAIBD)*, May 2019, pp. 384–388.

[49] K. Zhang, Y. Guo, X. Wang, J. Yuan, and Q. Ding, "Multiple feature reweight DenseNet for image classification," *IEEE Access*, vol. 7, pp. 9872–9880, 2019.

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

[51] S. Rebuffi, R. Fong, X. Ji, and A. Vedaldi, "There and back again: Revisiting backpropagation saliency methods," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 8836–8845.

[52] C. F. R. Chen, Q. Fan, and R. Panda, "CrossViT: Cross-attention multi-scale vision transformer for image classification," in *Proc. IEEE/CVF Comput. Vis. Pattern Recognit.*, Aug. 2021, pp. 357–366.

[53] C. Yan et al., "Task-adaptive attention for image captioning," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 32, no. 1, pp. 43–51, Jan. 2022.
[54] Y. Mei, Y. Fan, and Y. Zhou, "Image super-resolution with non-local sparse attention," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 3516–3525.
[55] X. Ying et al., "Multi-attention object detection model in remote sensing images based on multi-scale," *IEEE Access*, vol. 7, pp. 94508–94519, 2019.
[56] Q. Yan et al., "Attention-guided network for ghost-free high dynamic range imaging," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 1751–1760.
[57] Y. Yan, W. Ren, X. Hu, K. Li, H. Shen, and X. Cao, "SRGAT: Single image super-resolution with graph attention network," *IEEE Trans. Image Process.*, vol. 30, pp. 4905–4918, 2021.
[58] J. Liu, X. Fan, J. Jiang, R. Liu, and Z. Luo, "Learning a deep multi-scale feature ensemble and an edge-attention guidance for image fusion," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 32, no. 1, pp. 105–119, Jan. 2022.
[59] S. W. Zamir, A. Arora, S. Khan, M. Hayat, F. S. Khan, and M. Yang, "Restormer: Efficient transformer for high-resolution image restoration," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 5718–5729.
[60] F. Wang et al., "Residual attention network for image classification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 6450–6458.
[61] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 7132–7141.
[62] S. Woo, J. Park, J. Y. Lee, and I. S. Kweon, "CBAM: Convolutional block attention module," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 3–19.
[63] B. Chen, W. Tan, G. Coatrieux, Y. Zheng, and Y. Shi, "A serial image copy-move forgery localization scheme with source/target distinguishment," *IEEE Trans. Multimedia*, vol. 23, pp. 3506–3517, 2021.
[64] C. Yu, "Attention based data hiding with generative adversarial networks," in *Proc. AAAI Conf. Artif. Intell.*, vol. 34, no. 1, 2020, pp. 1120–1128.
[65] H. Zhang, H. Wang, Y. Cao, C. Shen, and Y. Li, "Robust data hiding using inverse gradient attention," 2020, *arXiv:2011.10850*.
[66] K. Alex Zhang, L. Xu, A. Cuesta-Infante, and K. Veeramachaneni, "Robust invisible video watermarking with attention," 2019, *arXiv:1909.01285*.
[67] T. Lin et al., "Microsoft COCO: Common objects in context," in *Proc. 13th Eur. Conf. Comput. Vis. (ECCV)*. Cham, Switzerland: Springer, Sep. 2014, pp. 740–755.
[68] H. Ibrahem, A. D. A. Salem, and H. Kang, "Real-time weakly supervised object detection using center-of-features localization," *IEEE Access*, vol. 9, pp. 38742–38756, 2021.
[69] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2009, pp. 248–255.
[70] S. Liu, H. Fan, S. Lin, Q. Wang, N. Ding, and Y. Tang, "Adaptive learning attention network for underwater image enhancement," *IEEE Robot. Autom. Lett.*, vol. 7, no. 2, pp. 5326–5333, Apr. 2022.
[71] B. Chen, X. Liu, Y. Zheng, G. Zhao, and Y. Shi, "A robust GAN-generated face detection method based on dual-color spaces and an improved xception," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 32, no. 6, pp. 3527–3538, Jun. 2022.
[72] W. Wang, X. Tan, P. Zhang, and X. Wang, "A CBAM based multiscale transformer fusion approach for remote sensing image change detection," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 15, pp. 6817–6825, 2022.

Jiangtao Huang received the M.S. degree from the Faculty of Information Science and Engineering, Ningbo University, Ningbo, China, in 2022, where he is currently pursuing the Ph.D. degree.

His current research interests include image watermarking.

Ting Luo (Member, IEEE) received the Ph.D. degree from the Faculty of Information Science and Engineering, Ningbo University, Ningbo, China, in 2016.

He is currently a Professor with the College of Science and Technology, Ningbo University. His research interests include multimedia security, image processing, data hiding, and pattern recognition.

Li Li received the B.S. and M.S. degrees in mathematics and the Ph.D. degree in computer science from Zhejiang University, Hangzhou, China, in 1994, 1997, and 2004, respectively.

She is currently with Hangzhou Dianzi University, Hangzhou, where she was a Professor from 2002 to 2005. Her research interests include digital image watermarking and computer animation. Her current research interests include image/video/3-D mesh watermarking, QR code, and image/video processing.

Gaobo Yang received the Ph.D. degree in communication and information system from Shanghai University, Shanghai, China, in 2004.

He is currently a Professor with Hunan University, Changsha, China, where he is also a Principle Investigator of several projects, including the Natural Science Foundation of China (NSFC), the Special Pro-Phase Project on National Basic Research Program of China (973), and the Program for New Century Excellent Talents (NCET). His current research interests include the area of digital media forensics, and image and video signal processing.

Haiyong Xu (Member, IEEE) received the B.S. degree in applied mathematics from Jilin University, Changchun, China, in 2003, the M.S. degree in applied mathematics from Sun Yat-sen University, Guangzhou, China, in 2005, and the Ph.D. degree from Ningbo University, Ningbo, China, in 2020.

He is currently an Associate Professor with the School of Mathematics and Statistics, Ningbo University. His research interests include underwater image processing and deep learning.

Chin-Chen Chang (Fellow, IEEE) received the Ph.D. degree in computer engineering from National Chiao Tung University, Hsinchu, Taiwan, in 1987.

On numerous occasions, he was invited to serve as a visiting professor, the chair professor, an honorary professor, an honorary director, an honorary Chairperson, a distinguished alumnus, a distinguished researcher, and a research fellow by universities and research institutes. He has been the Chair Professor with the Department of Information Engineering and Computer Science, Feng Chia University, Taichung, Taiwan, since February 2005. His current research interests include database design, computer cryptography, image compression, and data structures.