Robust Blind Watermarking Framework for Hybrid Networks Combining CNN and Transformer

Baowei Wang* wbw.first@163.com
Ziwei Song 202212200022@nuist.edu.cn
Yufeng Wu 20211220033@nuist.edu.cn
Nanjing University of Information Science and Technology, Nanjing, China

Editors: Berrin Yanıkoğlu and Wray Buntine

Abstract

As an essential means of copyright protection, the deep learning-based robust watermarking method is being studied extensively. Its framework consists of three main parts: the encoder, the noise layer and the decoder. But practically all of the schemes are directed at the encoder rather than the decoder. And the whole network is structured by shallow Convolutional Neural Networks (CNNs) for primary feature extraction, while CNNs capture local information and do not model non-local information in watermarked images well. To solve this problem, we consider the use of Transformer networks with a spatially self-attention mechanism. We propose to construct a novel decoder network by combining Transformer and CNNs, which can not only enriches local feature information but also enhances the ability to explore global representations. Meanwhile, to embed secret messages more perfectly, we design a multi-scale attentional feature fusion module to achieve an efficient aggregation of cover image features and secret message features, resulting in the encoded images with rich hybrid features. In addition, perceptual loss is introduced to better evaluate the visual quality of the watermarked images. Extensive experimental results show that our proposed method achieves better results in terms of imperceptibility and robustness compared with existing State-Of-The-Art (SOTA) methods.

Keywords: Robust blind watermarking; Hybrid network of CNN-Transformer; Attentional feature fusion; Perceptual loss

1. Introduction

Digital watermarking technology Van Schyndel et al. (1994). has become a significant study direction in the field of multimedia information security as multimedia technology has advanced. For image blind watermarking Hamidi et al. (2018); Ko et al. (2020), as its security depends on the content of the image, secret information must be embedded in the cover image in an invisible way to obtain the secret image. Even if the secret image is attacked, the secret information can be extracted from the secret image, which requires that the image must have a certain degree of robustness and imperceptibility. With the rise of deep learning, Convolutional Neural Networks (CNNs) have achieved great success in computer vision tasks, and some researchers use CNNs in the field of watermarking Kandi et al. (2017); Mun et al. (2017, 2019). However, in recent years, researchers have tended to focus on improving the encoder while ignoring the decoder's potential for advancement Zhu

[*] Corresponding author

et al. (2018); Liu et al. (2019); Ahmadi et al. (2020); Jia et al. (2021); Fang et al. (2022); Lu et al. (2022); Ma et al. (2022). For example, the first end-to-end trainable deep network HiDDeN was proposed in Zhu et al. (2018), where the encoder of HiDDeN merges the secret message replicated on space with the image features of the cover processed by convolution, ensuring that the complete secret message is obtained at any spatial location at the next layer of convolutional processing. A novel Two-stage Separable Deep Learning (TSDL) watermarking framework is designed in Liu et al. (2019), which can obtain a powerful encoder. The encoder uses a redundant multi-level feature encoding network as a framework to obtain robust watermarking. Ahmadi et al. (2020) proposed a framework to support operation in different domains such as the DCT domain and used a fully convolutional neural network with residual structure as an encoder to improve network performance. In MBRS Jia et al. (2021), a secret message processor is proposed to improve the performance of the encoder, and a significant number of SE blocks are introduced to the encoder to increase feature extraction and make it easier for the network to comprehend the preprocessed features. However, the decoders in these works Zhu et al. (2018); Liu et al. (2019); Ahmadi et al. (2020); Jia et al. (2021) only adopt the shallow design based on CNNs, ignoring the contribution of decoders to the model. In the decoder, the CNN-based architecture has a good ability to process two-dimensional feature information. However, the inherent characteristics of convolution operation, namely the independence of the local acceptance domain and input information, hinder the ability of the model to model non-local information. Moreover, CNNs cannot obtain the overall information of the image at the primary layer of the network and thus cannot learn the two-dimensional distribution of images better. So how can enhance the ability of the model to handle global information?

Recently, many researchers have introduced the Transformer Vaswani et al. (2017) into vision tasks. The difference between CNNs and Transformer is that traditional CNNs can only capture local information when processing sequential data, while the self-attention mechanism in Transformer can calculate global dependencies to better capture global representations. One problem with using Transformer directly for visual tasks, however, is that Transformer is not good at solving pixel-level tasks. Therefore, to make Transformer more suitable for visual tasks, many researchers have improved Transformer architecture to be suitable for image classification Dosovitskiy et al. (2021); Liu et al. (2021), segmentation Xie et al. (2021); Lee et al. (2022); Cao et al. (2022), image restoration Li et al. (2023); Liang et al. (2021), multi-task learning Ye and Xu (2022); Qu et al. (2022), target detection Zhu et al. (2019); Carion et al. (2020), image generation Jiang et al. (2021), image fusion Zhao et al. (2023), and other fields. However, to our knowledge, the Transformer has not been applied to the deep learning-based image watermarking technology at present. Almost all watermarking schemes use CNNs as the backbone network for feature extraction, but CNNs pay too much attention to local information and cannot capture long-distance feature dependency in watermarked images.

Therefore, to investigate a better watermarking model, we attempt to construct a novel decoder by combining Transformer and CNNs to complement each other's strengths, enhance the extraction of global representation and local information, and improve the robustness and imperceptibility of watermarking. Inspired by the work of Dai et al. (2021), we consider that linear methods like addition and concatenation neither fuse features well nor are context-conscious. Therefore, we designed a Multi-scale Attention Feature Fusion

Module (MA-FFN) and combined it with the iterative architecture iAFF Dai et al. (2021) for the final fusion of the cover image and secret messages, to obtain encoded images with rich context information. To be specific, combined with Transformer and CNNs as deep feature extraction modules, self-attention mechanism and convolution operation are used to enhance hybrid representation learning, a cascade multi-scale attentional fusion module is adopted to further improve feature aggregation ability, and perceptual loss is introduced to improve the visual quality of watermarked images. We conducted comprehensive experiments, and our model achieved better results compared to other deep learning-based models.

In summary, the main contributions of this paper are as follows:

- We propose a multi-branch decoder network combining CNNs and Transformer with stronger decoding and feature analysis capabilities by exploiting the inductive biasing capability of CNNs and the global feature processing capability of Transformer.
- We propose a Multi-scale Attentional Feature Fusion Module (MA-FFN) that aggregates multi-scale feature contexts, allowing local and non-local pixels to interact effectively and greatly enhancing the robustness of encoded images.
- We improved the low-level loss function and added additional high-level perceptual loss to enhance the visual quality of the images. And we conducted a preliminary study on perceptual loss.
- To the best of our knowledge, our model is the first to combine CNNs and Transformer in the deep learning-based image watermarking works.

2. Related Works

2.1. Vision Transformer

The great potential of Transformer Vaswani et al. (2017) in Natural Language Processing (NLP) has led researchers to explore its use in computer vision. ViT Dosovitskiy et al. (2021) is the pioneering work of Transformer in the field of computer vision and has achieved better results than the existing CNNs algorithm in image classification. But ViT directly used the standard Transformer used in NLP, so it is not suitable for pixel-level visual tasks. To address this problem, Liu et al. (2021) developed a layered Shift Windows (Swin) Transformer, which can better process images by applying a layered structure similar to that of CNNs. Since then, the Swin Transformer has been used as a general-purpose vision backbone network architecture, setting performance records on a variety of vision tasks Cao et al. (2022); Liang et al. (2021); Fan et al. (2022); Bhattacharjee et al. (2022). Cao et al. (2022) proposed a U-shaped network, Swin-Unet, for image segmentation tasks. Bhattacharjee et al. (2022) proposed a multi-task learning framework based on the Swin Transformer. Fan et al. (2022) proposed SUNet, an image denoising model using the Swin Transformer layer as the basic block. All these Swin Transformer-based methods have demonstrated excellent performance, even surpassing CNN-based methods. Motivated by the success of Swin Transformer, we investigated how to design a powerful watermarking model by Swin Transformer to further improve the performance of image watermarking tasks.

2.2. Hybrid model combining CNNs and Transformer

In CNNs, the convolution operation is good at extracting local features but has limitations in capturing global feature representation. In contrast, the self-attention mechanism and multilayer perceptron architecture in Transformer can reflect complex spatial transformations and long-distance feature dependencies to achieve a global feature representation while ignoring the details of local features. Therefore, some researchers have proposed combining CNNs with Transformer to improve the learning capability of the network for features. The combination methods are mainly divided into structural concatenation and feature fusion. Structure concatenation refers to forming a new network structure through a reasonable combination of CNNs and Transformer modules. Mehta and Rastegari (2021) embedded Transformer as a convolutional layer into the convolutional neural network, enabling the interaction of local and global information. Yuan et al. (2021) combined the Feed-Forward Network (FFN) with the convolutional module to extract local information. Guo et al. (2022) combined the traditional convolution and the Transformer to form a CMT module for hierarchical extraction of local and global features. Feature fusion refers to the integration of CNNs features and Transformer features at the feature level. Some works Chen et al. (2022); Peng et al. (2021) employed a parallel architecture of CNNs-branches and Transformer-branches, and use bridging modules to fuse local features and global representations interactively. All of these hybrid structures obtained a performance comparable to that of CNNs.

3. Proposed Method

In this section, we first describe the overall pipeline and the hybrid structure used for image watermarking. Then, we present the details of the multi-branch decoder network, which consists of three main key branches: CNNs branch, Transformer branch, and Identity branch. Next, we describe the specific architecture of the proposed multi-scale attentional feature fusion module. Finally, we introduce the perceptual loss function Zhang et al. (2018).

3.1. Overall Pipeline

In this paper, based on the research of MBRS Jia et al. (2021), a new watermarking model is designed that is more excellent and secure. The overall architecture of the model is shown in Figure 1. For the message processor and noise layer, we adopt the same network structure as in MBRS.

The encoder first receives the RGB cover image I_{co} of size $3 \times H \times W$ and performs a feature extraction operation on it through one 3×3 convolutional layer and four Squeeze-and-Excitation (SE) blocks Hu et al. (2018) to obtain an intermediate feature vector $I_{inter} \in \mathbb{R}^{C \times H \times W}$. Then, the M_{en} obtained after the message processor is fed together with I_{inter} into the iterative attentional feature fusion module (iAFF-Pro) to integrate the features. After that, the tensor obtained after iAFF-Pro and the cover image are concatenated into a new tensor by skip connection and fed into a 1×1 convolutional layer to recover to the number of original channels, finally generating an encoded image I_{en} of shape $3 \times H \times W$. The decoder recovers the embedded secret message $M^{'}$ of length L from the noised image I_{co}

through three-branch architecture. Finally, for the feature downsampling and upsampling operations, we apply pixel-unshuffle and pixel-shuffle to prevent checkerboard artifacts in the watermarked image.

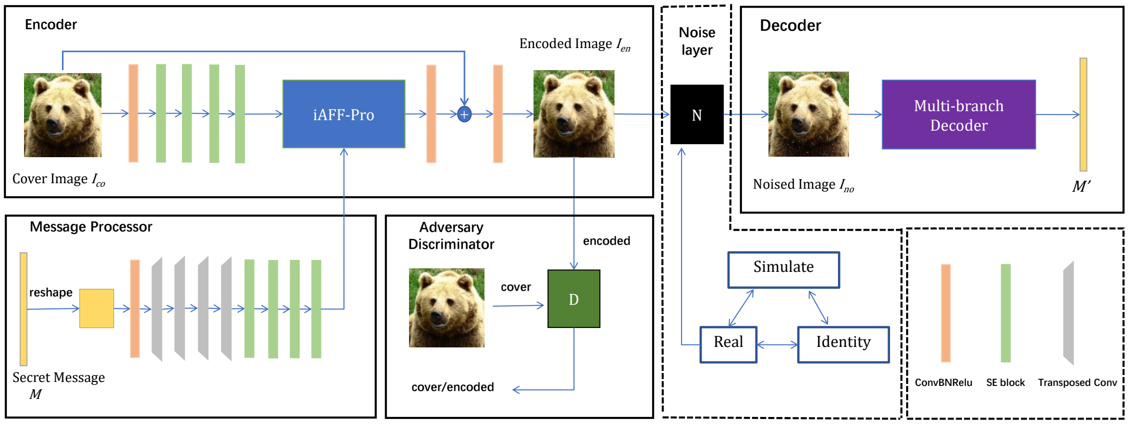

Figure 1: Overall model architecture.The encoder integrates the cover image features and the message features through the iAFF-Pro module. The noise layer changes the type of noise according to MBRS method. The decoder recovers the secret message from the noised image through the multi-branch architecture. The adversary discriminator is trained to distinguish whether an image is encoded or not.

3.2. Multi-branch decoder

In MBRS, the authors designed a decoder network with CNNs combined with the attention mechanism as the main architecture to perform better learning of image features in the extraction phase. However, the network focuses more on local features of the image and fails to capture the overall attributes of the image, such as color features, texture features, and shape features, which results in the watermark information in the image under attack not being easily extracted in its entirety and is not conducive to the protection of images.

To solve this problem, we designed a multi-level decoder network combining CNNs and Transformer, as shown in Figure 2, which adopts a three-branch architecture: CNNs branch, Transformer branch, and Identity branch.

CNNs Branch: The CNNs branch uses SE blocks with the attention mechanism to learn the data in the channel dimension, to better assign the features in a ranked manner, and to obtain the importance of each channel in the feature map. The CNNs branch consists of two convolutional layers and five SE blocks; the structure of the SE block is shown in Figure 2(a), which consists of two main parts, Squeeze and Excitation, respectively. The F_{sq} in the figure is the Squeeze operation, which globally pools the input feature map to obtain a single channel feature map. The F_{ex} in the figure corresponds to the Excitation operation, which is the computation of two fully connected layers on the output of the Squeeze operation to obtain a weight vector with the same number of channels as the

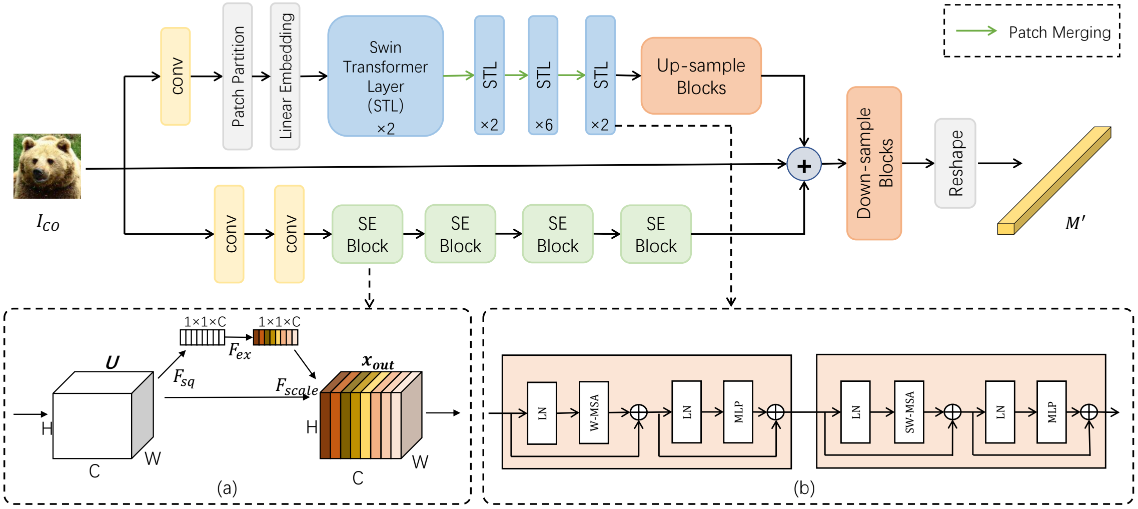

Figure 2: The architecture of proposed multi-branch decoder network. (a) The architecture of SE block; (b) two successive Swin Transformer Blocks. The up-sampling and down-sampling blocks consist of several pixel-shuffle and pixel-unshuffle operations respectively.

original input. Finally, through F_{scale} the original input is multiplied by the weight vector to obtain the enhanced feature map.

Transformer Branch: First, the initial feature extraction is performed by a feature extraction block consisting of convolutional layers to retain some inductive bias properties of the convolution. Then, the feature is partitioned into non-overlapping patches through the patch partition module, and further linear embedding layers are applied to project them to arbitrary dimensions. Finally, it is sent into the Transformer block for global feature collation. To improve performance, we use Swin Transformer Liu et al. (2021) as the main Transformer block, which uses a hierarchical Transformer computed with shifted windows. By limiting the computation of self-attention to non-overlapping local windows and allowing connections between non-adjacent windows, it is allowed to make watermarked images retain richer information. The specific architecture of the Swin Transformer Layer is shown in Figure 2(b). Each Swin Transformer Layer consists of two consecutive Swin Transformer blocks. Since after the feature transformation by applying the Swin Transformer block, the feature size will be halved and the channel doubled by each patch merging operation, several upsampling operations are adopted to restore the original feature size for fusion with other branches.

Identity Branch: The Identity branch is similar to the residual structure of ResNet He et al. (2016), which adds skip connections to the network so that the network can learn the residual function and better fit the data.

After the three-branch network, we fuse the resulting feature maps and then send them to the downsampling network. Finally, through the reshaping operation, we get the finally decoded secret message M'.

3.3. Multi-scale attentional feature fusion module

In order to better embed watermarks into the cover image, we propose a Multi-scale Attentional Feature Fusion Module (MA-FFM). To be specific, we design the MA-FFM through two main paths: the local path and the global path. As shown in Figure 3(a), given an input tensor $X \in \mathbb{R}^{C\times H\times W}$, it is fed into the two paths to fuse local features and global representations at different scales in an interactive manner.

For the local path, we choose deep-wise convolution to encode information on spatially close pixel locations, using 3×3 and 5×5 depth-wise convolution to enhance the extraction of multi-scale local information, respectively. After the first deep convolution layer, the initial contextual feature information is cross-aggregated. Then, the 3×3 and 5×5 depth-wise convolution layers are fed again respectively for a second feature fusion in the channel dimension to enrich the contextual feature information. The contextual features $L(X)$ of the local path are computed by the following structure:

$$X_l' = f_{1\times1}^{conv}(X) \tag{1}$$

$$X_l^{c3} = \omega(\beta(f_{3\times3}^{dwc}(X_l'))); X_l^{c5} = \omega(\beta(f_{5\times5}^{dwc}(X_l'))) \tag{2}$$

$$X_l^{c33} = \omega(\beta(f_{3\times3}^{dwc}[X_l^{c3}, X_l^{c5}])); X_l^{c55} = \omega(\beta(f_{5\times5}^{dwc}[X_l^{c3}, X_l^{c5}])) \tag{3}$$

$$L(X) = f_{1\times1}^{conv}[X_l^{c33}, X_l^{c55}] \tag{4}$$

where $f_{1\times1}^{conv}$ represents 1×1 convolution, $f_{3\times3}^{dwc}$ and $f_{5\times5}^{dwc}$ represents 3×3 and 5×5 depth-wise convolution, $\omega(\cdot)$ is a ReLU activation, and $[\cdot]$ is the channel-wise concatenation.

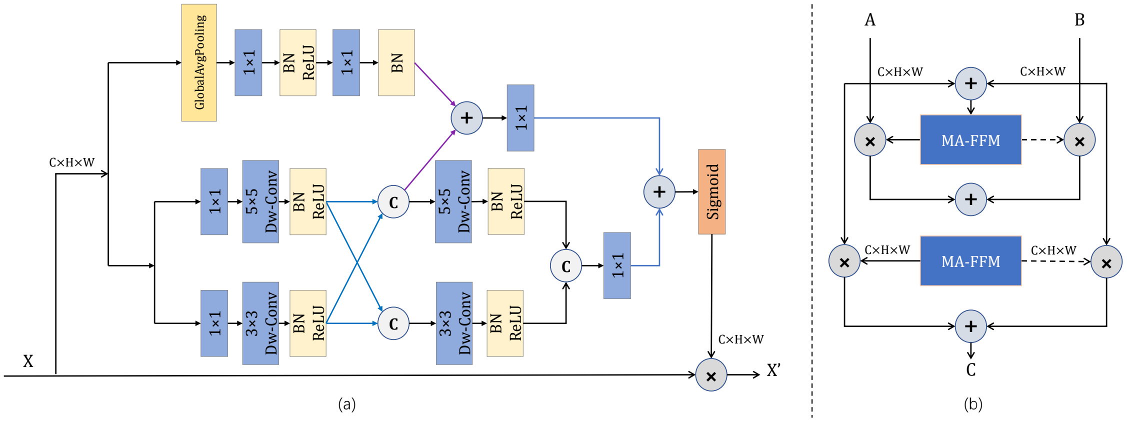

Figure 3: (a)The architecture of MA-FFM; (b)The architecture of iAFF-Pro.

For the global path, we first compress the spatial information by taking the input feature map through global averaging pooling to obtain a feature map that summarizes the rich spatial information of the image. The global averaging pooling operation forces the correspondence between the cover image and the secret message, which makes the encoded image more robust against geometric attacks. And then we use two 1×1 convolution layers. One is to process the characteristic channel information to strengthen the correlation between channels. The other is used to extend the channels to add them to the feature map on the left after cross-fusion. This step is the highlight of this module, aiming to integrate local details and global information of different levels for the first time, improve feature reuse rates, and effectively enrich features of different levels. Finally, 1×1 convolution is used to extend the channel to the original input dimension for the final aggregation between the local information and the global information.

After getting the local channel context $L(X)$ and the global channel context $G(X)$, the refined features $X^{'}\in\mathbb{R}^{C\times H\times W}$ obtained by MA-FFM are denoted as follows:

$$X^{'} = X \otimes \sigma(L(X) \oplus G(X)) \tag{5}$$

where σ is the Sigmoid function, $\oplus$ represents the broadcasting addition and $\otimes$ represents the element-wise multiplication.

Dai et al. (2021) demonstrated that the initial integration problem of feature mapping can be alleviated by adding another level of attention. We address the initial integration problem by using a combination of MA-FFM and their iterative architecture iAFF. We call it iAFF-Pro. The structure of the iAFF-Pro is shown in Figure 3(b).

In general, MA-FFM controls the flow of information at each level of the pipeline, so that each level can focus on details that complement the other levels. By fusing these different levels of feature information it is possible to better capture information about multi-layered features from the original image and to interact across channels. The feature fusion operation further enhances the representational power of the CNN and thus embeds the secret message more seamlessly into the cover image.

3.4. Loss function

In the field of deep learning-based watermarking, researchers usually use Mean Square Error (MSE) as a loss function to measure the quality of watermarked images, but MSE only reflects the degree of difference at the pixel level of an image and cannot explain many nuances of human perception, much less the difference between image structures. What we really want is a ”perceptual distance” that can measure the similarity between the encoded image and the cover image. By reducing this distance so that the human eye cannot distinguish the differences between two images. Thus, we add the perceptual loss LPIPS Zhang et al. (2018) to the total loss $\mathcal{L}_m$ of MBRS to improve the perceptual quality and make the watermarked images look more natural and visually better.

The perceptual loss $\mathcal{L}_L$ is calculated as:

$$\mathcal{L}_L = d(x, x_0) = \sum_{l} \frac{1}{H_l W_l} \sum_{h,w} \|w_l \odot (\hat{y}^l_{hw} - \hat{y}^l_{0hw})\|^2_2 \tag{6}$$

where x is the cover image, x_0 is the encoded image, $\hat{y}^l_{hw}$ and $\hat{y}^l_{0hw}$ are the features of the cover and encoded images at layer l, w_l is the cosine distance, and H_l and W_l are the height and width of the feature map at layer l.

The total loss function is:

$$\mathcal{L}_{Total} = \mathcal{L}_m + \lambda_L \mathcal{L}_L \tag{7}$$

where λ_L is weight factor. For the choice of λ_L, please see the Results and Analysis section for details.

4. Experiments

4.1. Implementation Details

Our proposed model is implemented by PyTorch. In the training phase, we use a mini-batch of size 4, with 100 epochs trained for each noise. We use the Adam Kingma and Ba (2014) optimizer with a learning rate of 1e-3. The strength factor S is all set to 1 during training and is given different values during testing. For the parameters of the loss function, we choose λ_L= 1.2, the same parameters in $\mathcal{L}_m$ as in MBRS Jia et al. (2021).

We randomly selected 10,000 images from the COCO dataset Lin et al. (2014) as the training set. And 5000 images were randomly selected from the validation set as the validation set and 5000 images as the test set. Each image was processed to a size of 128 × 128 before training. The secret message is composed of a random 0 and 1 of length 64 bits.

4.2. Metrics

We use PSNR and SSIM to measure the quality of the encoded image. In general, the higher the PSNR, the better the image quality; the closer the value of SSIM is to 1, the smaller the difference between the encoded image and the cover image. Robustness is measured using the Bit Error Rate (BER), which indicates the probability that the number of wrong bits of information in the secret message recovered from the decoder accounts for the total number of bits; the lower the BER, the higher the extraction accuracy.

4.3. Baselines

Our baseline for comparison is HiDDeN Zhu et al. (2018), TSDL Liu et al. (2019) and MBRS Jia et al. (2021). The reason is that all of these works are watermarking algorithms with CNN as the main architecture and MBRS as the SOTA method. The strength factor S is a parameter to adjust robustness and image quality. For a fair comparison, we set the length of the embedded secret messages to 64 and the strength factor S to 1 during training and assigned different values during testing.

5. Results and analysis

5.1. Imperceptibility and robustness

To evaluate the imperceptibility and robustness of the watermarked images, we trained with seven different noises, including differentiable noise: Crop (p = 0.3), Dropout (p = 0.3), Gaussian Filtering (GF, $\sigma = 2$), Gaussian Noise (GN, $\sigma^2 = 0.005$), Salt and Pepper noise (SP, $D = 0.01$), Median Filtering (MF, $w = 3$) and non-differentiable noise: JPEG compression ($Q = 50$). We use PSNR and SSIM to measure the similarity between the

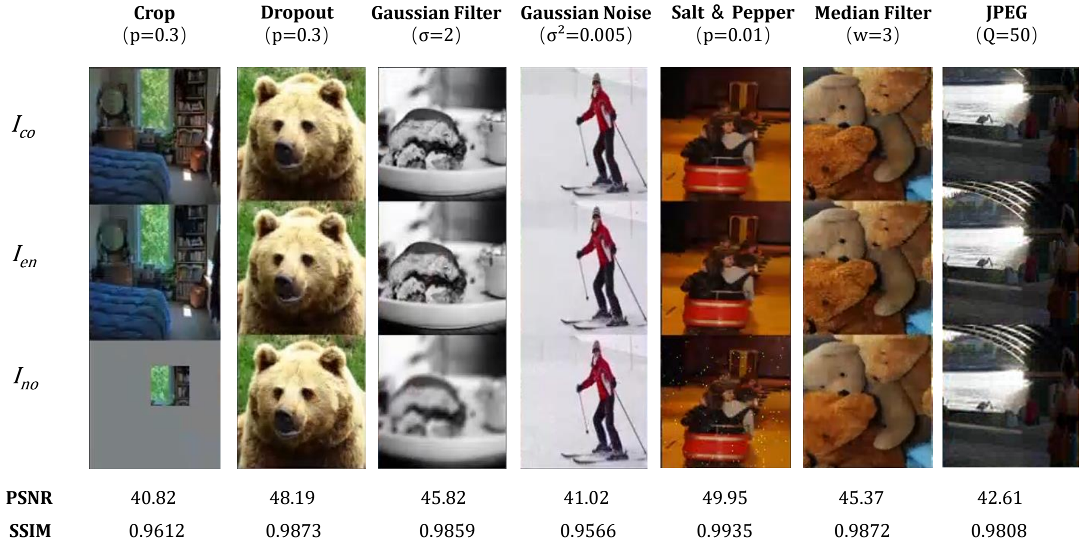

Figure 4: Image quality for our models trained with the seven types of noise. Top: cover image I_{co}; Middle: encoded image I_{en}; Bottom: noisy image I_{no}. PSNR and SSIM reflect the similarity between the original cover image and the encoded image.

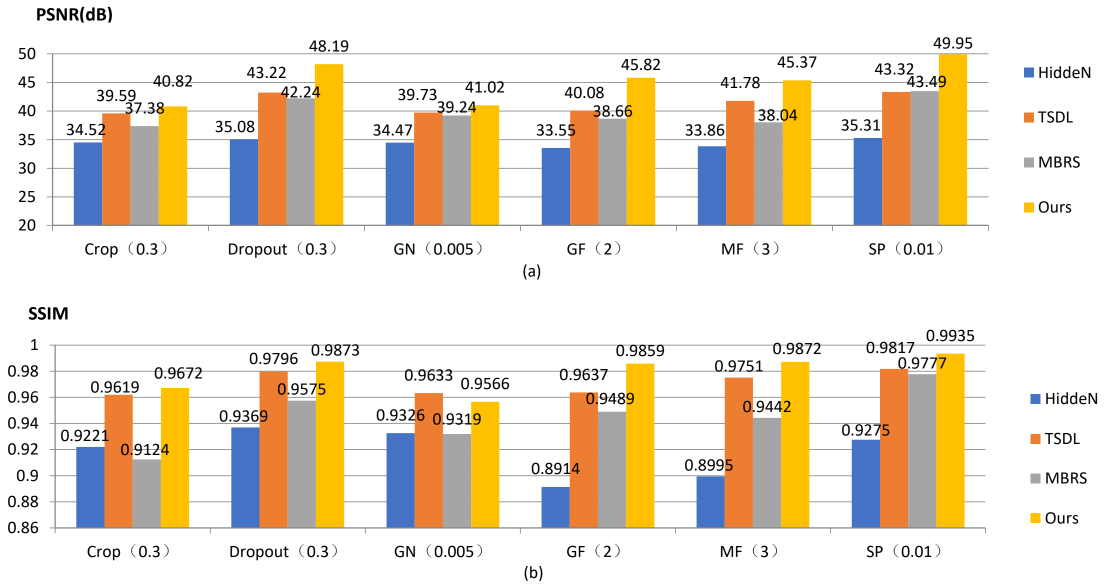

Figure 5: Imperceptibility against differentiable noise compared to other methods. For PSNR and SSIM, higher is better.

original cover image and the watermarked image. The PSNR and SSIM values of seven different watermarked images are shown in Figure 4. We can find that the encoded images obtained by our model are visually indistinguishable from the cover images and have certain imperceptibility.

We compared the results of our scheme in this paper with HiDDeN Zhu et al. (2018), TSDL Liu et al. (2019) and MBRS Jia et al. (2021). The experimental results for differentiable noise are shown in Figure 5, PSNR and SSIM can be over 40 dB and 0.95 respectively, indicating high image quality. Therefore, the perceptual quality of our scheme for watermarked images is significantly superior to that of other methods Zhu et al. (2018); Liu et al. (2019); Jia et al. (2021). In addition, we calculate the BER of the extracted watermark under different noises conditions to prove the robustness of the proposed method. The data in Table 1 show that the BER of our method can be reduced by an order of magnitude under Dropout, GN, and SP compared with the SOTA method Jia et al. (2021). As for JPEG compression, we compared it with the SOTA method Jia et al. (2021) under the condition that PSNR = 33.5 dB, and the results in Table 2 show that our model outperforms Jia et al. (2021) when Q > 50. This indicates that our method has some resistance to JPEG compression. In summary, the scheme in this paper performs better robustness in general and does not degrade the image quality.

Table 1: Robustness comparison with other methods. It shows that our model performs the best of the four methods.

Noise	Crop (p=0.3)	Dropout (p=0.3)	GN (σ^2=0.005)	GF (σ=2)	MF (w=3)	SP (D=0.01)
HiddeN	30.16%	39.77%	27.90%	4.00%	23.00%	27.30%
TSDL	32.58%	5.91%	22.43%	7.20%	22.64%	10.15%
MBRS	21.15%	0.00188%	0.00141%	0.00281%	0.00266%	0.00203%
Ours	**20.95%**	**0.00078%**	**0.00047%**	**0.00188%**	**0.00219%**	**0.00078%**

Table 2: Robustness against non-differentiable noise compared to Jia et al. (2021). We made a fair comparison under PSNR = 33.5 for different Q values in JPEG compression.

JPEG	Q = 10	Q = 30	Q = 50	Q = 70	Q = 90
MBRS	5.79%	0.011%	4.14%	0.00003%	0.94%
Ours	5.84%	0.0017%	**0.00%**	**0.000003%**	**0.87%**

5.2. Ablation studies

In this section, we validate the effectiveness of the proposed method through ablation experiments. Specifically, the effects of the Transformer branch, the iAFF-Pro module and the perceptual loss are discussed below.

Table 3: Ablation experiment results of the proposed method. Here, we only show the experimental results of MF(w=3) for better observation.

Transformer branch	iAFF-Pro module	LPIPS	PSNR	SSIM	BER
			38.59	0.9486	0.31380%
✓			40.94	0.9729	0.00239%
	✓		39.61	0.9575	0.00281%
		✓	43.73	0.9816	0.00287%
✓	✓		41.21	0.9733	0.00218%
✓		✓	43.86	0.9825	0.00226%
	✓	✓	42.12	0.9694	0.00234%
✓	**✓**	**✓**	**45.37**	**0.9872**	**0.00189%**

Unlike the existing CNN-based watermarking model, we introduce Transformer blocks rationally based on CNN and combine them into a new decoder. We also design a multi-scale attentional feature fusion module (MA-FFM) to achieve efficient fusion of image features and message features. For better observation, we conducted experiments using MF (w=3). As can be seen in Table 3, the PSNR value increases by at least 2 dB and the SSIM value rises by 0.02 after adding the Transformer branch; then the PSNR value increases by at least 3 dB and the SSIM value rises by 0.03 after adding the iAFF-Pro module, and the BER decreases by two orders of magnitude. In addition, we improved the low-level loss function and added an additional high-level perceptual loss. As shown in Table 3, the PSNR and SSIM of the images are significantly improved with the addition of perceptual loss, with the PSNR value increasing by 5 dB and the SSIM value increasing by 0.04. To conclude, the combination of the three modules achieves the best performance.

5.3. Research on λ_L

In order to obtain visually lossless encoded images, we conducted a preliminary study and analysis of the parameter λ_L of the perceptual loss $\mathcal{L}_L$. Specifically, the effect of the parameter λ_L on the BER, PSNR value and SSIM value are discussed below.

We use different noises for training and set different values of λ_L. After extensive experiments, we found that the performance of the model was extremely unstable as λ_L increased. Therefore, we chose the value of λ_L to range from 0 to 2 for easy comparison. For better observation, we only show the experimental results for Dropout (p=0.3). The corresponding BER, PSNR values and SSIM values are shown in Figure 6. It can be seen from Figure 6 that either too large or too small a value of λ_L has some effect on the results, with the lowest BER and highest PSNR and SSIM values when the λ_L value is around 1.2. We conjecture that the reason for this is that adding the appropriate perceptual metrics leads to the best performance of the model, while the opposite is counterproductive. Therefore, we finally chose $\lambda_L = 1.2$ for the subsequent experiments.

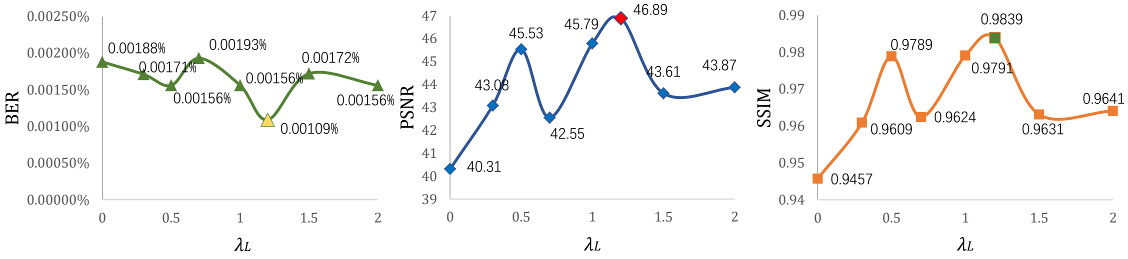

Figure 6: The influence of the perceptual loss parameter λ_L on experimental results.

6. Conclusion

In this paper, we design a multi-branch decoder network structure combining CNN and Transformer for secret message extraction. Compared with the decoder of pure CNN architecture, our model has more powerful decoding and feature analysis capabilities. To better embed secret messages, we developed a multi-scale attentional feature fusion module to aggregate image features and message features. In addition, we add perceptual loss to the loss function to enhance the deeper understanding of the cover image and the encoded image by the network, resulting in better perceptual capabilities. Experimental results show that our approach achieves the best performance in most types of noise compared to state-of-the-art algorithms. To our knowledge, our model is the first method to use the Transformer in the watermarking field, which further demonstrates the feasibility of the Transformer in the field of computer vision. In the future, we will explore Transformer models that are more suitable for use in the field of image watermarking.

References

Mahdi Ahmadi, Alireza Norouzi, Nader Karimi, Shadrokh Samavi, and Ali Emami. Redmark: Framework for residual diffusion watermarking based on deep networks. *Expert Systems with Applications*, 146:113157, 2020.

Deblina Bhattacharjee, Tong Zhang, Sabine Süsstrunk, and Mathieu Salzmann. Mult: an end-to-end multitask learning transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12031–12041, 2022.

Hu Cao, Yueyue Wang, Joy Chen, Dongsheng Jiang, Xiaopeng Zhang, Qi Tian, and Manning Wang. Swin-unet: Unet-like pure transformer for medical image segmentation. In *European Conference on Computer Vision*, pages 205–218. Springer, 2022.

Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part I 16*, pages 213–229. Springer, 2020.

Yinpeng Chen, Xiyang Dai, Dongdong Chen, Mengchen Liu, Xiaoyi Dong, Lu Yuan, and Zicheng Liu. Mobile-former: Bridging mobilenet and transformer. In *Proceedings of the*

IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 5270–5279, 2022.

Yimian Dai, Fabian Gieseke, Stefan Oehmcke, Yiquan Wu, and Kobus Barnard. Attentional feature fusion. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 3560–3569, 2021.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, and Sylvain and Gelly. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021.

Chi-Mao Fan, Tsung-Jung Liu, and Kuan-Hsien Liu. Sunet: swin transformer unet for image denoising. In *2022 IEEE International Symposium on Circuits and Systems (ISCAS)*, pages 2333–2337. IEEE, 2022.

Han Fang, Zhaoyang Jia, Hang Zhou, Zehua Ma, and Weiming Zhang. Encoded feature enhancement in watermarking network for distortion in real scenes. *IEEE Transactions on Multimedia*, 2022.

Jianyuan Guo, Kai Han, Han Wu, Yehui Tang, Xinghao Chen, Yunhe Wang, and Chang Xu. Cmt: Convolutional neural networks meet vision transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12175–12185, 2022.

Mohamed Hamidi, Mohamed El Haziti, Hocine Cherifi, and Mohammed El Hassouni. Hybrid blind robust image watermarking technique based on dft-dct and arnold transform. *Multimedia Tools and Applications*, 77:27181–27214, 2018.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7132–7141, 2018.

Zhaoyang Jia, Han Fang, and Weiming Zhang. Mbrs: Enhancing robustness of dnn-based watermarking by mini-batch of real and simulated jpeg compression. In *Proceedings of the 29th ACM international conference on multimedia*, pages 41–49, 2021.

Yifan Jiang, Shiyu Chang, and Zhangyang Wang. Transgan: Two pure transformers can make one strong gan, and that can scale up. *Advances in Neural Information Processing Systems*, 34:14745–14758, 2021.

Haribabu Kandi, Deepak Mishra, and Subrahmanyam RK Sai Gorthi. Exploring the learning capabilities of convolutional neural networks for robust image watermarking. *Computers & Security*, 65:247–268, 2017.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Hung-Jui Ko, Cheng-Ta Huang, Gwoboa Horng, and WANG Shiuh-Jeng. Robust and blind image watermarking in dct domain using inter-block coefficient correlation. *Information Sciences*, 517:128–147, 2020.

Youngwan Lee, Jonghee Kim, Jeffrey Willette, and Sung Ju Hwang. Mpvit: Multi-path vision transformer for dense prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7287–7296, 2022.

Yawei Li, Yuchen Fan, Xiaoyu Xiang, Denis Demandolx, Rakesh Ranjan, Radu Timofte, and Luc Van Gool. Efficient and explicit modelling of image hierarchies for image restoration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18278–18289, 2023.

Jingyun Liang, Jiezhang Cao, Guolei Sun, Kai Zhang, Luc Van Gool, and Radu Timofte. Swinir: Image restoration using swin transformer. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1833–1844, 2021.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer, 2014.

Yang Liu, Mengxi Guo, Jian Zhang, Yuesheng Zhu, and Xiaodong Xie. A novel two-stage separable deep learning framework for practical blind watermarking. In *Proceedings of the 27th ACM International conference on multimedia*, pages 1509–1517, 2019.

Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10012–10022, 2021.

Junxiong Lu, Jiangqun Ni, Wenkang Su, and Hao Xie. Wavelet-based cnn for robust and high-capacity image watermarking. In *2022 IEEE International Conference on Multimedia and Expo (ICME)*, pages 1–6. IEEE, 2022.

Rui Ma, Mengxi Guo, Yi Hou, Fan Yang, Yuan Li, Huizhu Jia, and Xiaodong Xie. Towards blind watermarking: Combining invertible and non-invertible mechanisms. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 1532–1542, 2022.

Sachin Mehta and Mohammad Rastegari. Mobilevit: light-weight, general-purpose, and mobile-friendly vision transformer. *arXiv preprint arXiv:2110.02178*, 2021.

Seung-Min Mun, Seung-Hun Nam, Han-Ul Jang, Dongkyu Kim, and Heung-Kyu Lee. A robust blind watermarking using convolutional neural network. *arXiv preprint arXiv:1704.03248*, 2017.

Seung-Min Mun, Seung-Hun Nam, Haneol Jang, Dongkyu Kim, and Heung-Kyu Lee. Finding robust domain from attacks: A learning framework for blind watermarking. *Neurocomputing*, 337:191–202, 2019.

Zhiliang Peng, Wei Huang, Shanzhi Gu, Lingxi Xie, Yaowei Wang, Jianbin Jiao, and Qixiang Ye. Conformer: Local features coupling global representations for visual recognition. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 367–376, 2021.

Linhao Qu, Shaolei Liu, Manning Wang, and Zhijian Song. Transmef: A transformer-based multi-exposure image fusion framework using self-supervised multi-task learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 2126–2134, 2022.

Ron G Van Schyndel, Andrew Z Tirkel, and Charles F Osborne. A digital watermark. In *Proceedings of 1st international conference on image processing*, volume 2, pages 86–90. IEEE, 1994.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Enze Xie, Wenhai Wang, Zhiding Yu, Anima Anandkumar, Jose M Alvarez, and Ping Luo. Segformer: Simple and efficient design for semantic segmentation with transformers. *Advances in Neural Information Processing Systems*, 34:12077–12090, 2021.

Hanrong Ye and Dan Xu. Inverted pyramid multi-task transformer for dense scene understanding. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXVII*, pages 514–530. Springer, 2022.

Kun Yuan, Shaopeng Guo, Ziwei Liu, Aojun Zhou, Fengwei Yu, and Wei Wu. Incorporating convolution designs into visual transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 579–588, 2021.

Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 586–595, 2018.

Zixiang Zhao, Haowen Bai, Jiangshe Zhang, Yulun Zhang, Shuang Xu, Zudi Lin, Radu Timofte, and Luc Van Gool. Cddfuse: Correlation-driven dual-branch feature decomposition for multi-modality image fusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5906–5916, 2023.

Jiren Zhu, Russell Kaplan, Justin Johnson, and Li Fei-Fei. Hidden: Hiding data with deep networks. In *Proceedings of the European conference on computer vision (ECCV)*, pages 657–672, 2018.

Xizhou Zhu, Han Hu, Stephen Lin, and Jifeng Dai. Deformable convnets v2: More deformable, better results. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9308–9316, 2019.