Chinese Journal of Electronics
vol. 33, no. 4, pp. 965–978, July 2024
https://doi.org/10.23919/cje.2022.00.452

Special Focus on Explainability, Robustness, and Security in AI Systems

RESEARCH ARTICLE

Lightweight Steganography Detection Method Based on Multiple Residual Structures and Transformer

Hao LI[1], Yi ZHANG[1], Jinwei WANG[2], Weiming ZHANG[3], and Xiangyang LUO[1]

1. *Key Laboratory of Cyberspace Situation Awareness of Henan Province, Zhengzhou 450001, China*
2. *Nanjing University of Information Science & Technology, Nanjing 210044, China*
3. *University of Science and Technology of China, Hefei 230052, China*

Corresponding author: Xiangyang LUO, Email: luoxy_ieu@sina.com
Manuscript Received December 28, 2022; Accepted January 12, 2024

Abstract — Existing deep learning-based steganography detection methods utilize convolution to automatically capture and learn steganographic features, yielding higher detection efficiency compared to manually designed steganography detection methods. Detection methods based on convolutional neural network frameworks can extract global features by increasing the network's depth and width. These frameworks are not highly sensitive to global features and can lead to significant resource consumption. This manuscript proposes a lightweight steganography detection method based on multiple residual structures and Transformer (ResFormer). A multi-residuals block based on channel rearrangement is designed in the preprocessing layer. Multiple residuals are used to enrich the residual features and channel shuffle is used to enhance the feature representation capability. A lightweight convolutional and Transformer feature extraction backbone is constructed, which reduces the computational and parameter complexity of the network by employing depth-wise separable convolutions. This backbone integrates local and global image features through the fusion of convolutional layers and Transformer, enhancing the network's ability to learn global features and effectively enriching feature diversity. An effective weighted loss function is introduced for learning both local and global features, BiasLoss loss function is used to give full play to the role of feature diversity in classification, and cross-entropy loss function and contrast loss function are organically combined to enhance the expression ability of features. Based on BossBase-1.01, BOWS2 and ALASKA#2, extensive experiments are conducted on the stego images generated by spatial and JPEG domain adaptive steganographic algorithms, employing both classical and state-of-the-art steganalysis techniques. The experimental results demonstrate that compared to the SRM, SRNet, SiaStegNet, CSANet, LWENet, and SiaIRNet methods, the proposed ResFormer method achieves the highest reduction in the parameter, up to 91.82%. It achieves the highest improvement in detection accuracy, up to 5.10%. Compared to the SRNet and EWNet methods, the proposed ResFormer method achieves an improvement in detection accuracy for the J-UNIWARD algorithm by 5.78% and 6.24%, respectively.

Keywords — Steganalysis, Multiple residual blocks, Transformer, Channel shuffle.

Citation — Hao LI, Yi ZHANG, Jinwei WANG, *et al.*, "Lightweight Steganography Detection Method Based on Multiple Residual Structures and Transformer," *Chinese Journal of Electronics*, vol. 33, no. 4, pp. 965–978, 2024. doi: 10.23919/cje.2022.00.452.

I. Introduction

With the development of Internet technology, there has been an explosive growth of multimedia data such as images and audio. People can express and transmit information through digital multimedia. Steganography, which utilizes the redundancy in multimedia data to hide secret information, has become an important means and method for secure covert communication [1], [2]. The development of steganographic algorithms, especially adaptive steganographic algorithms, has further enhanced the security of covert communication. However, the application of steganography also provides the potential for se-

Guest Editor: Prof. Chao SHEN, Xi'an Jiaotong University.

cret information transmission by malicious actors. As a countermeasure to steganography, steganalysis can detect whether a carrier contains hidden information. Therefore, conducting research on steganalysis methods for detecting and uncovering illicit steganographic communication behaviors is of great significance for safeguarding social and national security.

Steganalysis methods distinguish between cover and stego images by statistically analyzing the differences between their features. Steganalysis methods have evolved from specialized detection techniques targeting known steganographic algorithms [3], [4] to more general approaches for detecting unknown steganographic algorithms [5], [6], resulting in improved applicability and detection performance. However, with the development of adaptive steganographic algorithms, the aforementioned detection algorithms may struggle to effectively analyze feature differences in complex textured regions, making detection more challenging. To address this, some researchers have focused on enhancing feature diversity and leveraging machine learning techniques to analyze the differences between features, leading to the proposal of high-dimensional and rich-model steganalysis methods. One of the most representative methods in this category is the high-dimensional spatial rich model (SRM) proposed by Fridrich *et al.* in 2011 [7]. Subsequently, the same team proposed a series of high-dimensional steganalysis techniques [8], [9]. In 2016, Tang *et al.* [10] improved upon the rich model approach and introduced the adaptive spatial rich model (adaptive SRM) detection method. The high-dimensional rich model methods effectively enhance the detection rate of adaptive steganographic algorithms, but they come with high feature dimensions and significant resource consumption. In 2019, Ma *et al.* [11] addressed this issue by using rough sets to eliminate redundant features among high-dimensional features. This approach significantly reduces computational resources while maintaining detection effectiveness. Through optimization and dimension reduction, the efficiency and performance of high-dimensional steganalysis methods have been improved. However, existing high-dimensional steganalysis methods are based on feature extraction and machine learning (classification) approaches. The limited effective steganographic features learned through these methods hinder the improvement of detection accuracy.

Steganalysis is generally regarded as a binary classification problem. The powerful automatic learning and classification capabilities of deep learning networks have yielded many notable research achievements in addressing both binary and multi-class classification problems [12]–[14]. Several researchers have integrated deep learning techniques into steganalysis methods to address the limitations of existing approaches. Here, we will introduce existing deep learning-based steganalysis methods in chronological order. In 2015, Qian *et al.* [15] enhanced the non-linearity of features by using high-pass filtering and increased the depth of the convolutional layers to improve network expressiveness. Their detection performance was comparable to the high-dimensional SRM method. In 2016, Xu *et al.* [16] used high-pass filtering in the preprocessing layer to obtain steganographic features and utilized the ABS (absolute) activation function to capture the symmetry of residual features. This combined approach effectively captured steganographic features and achieved detection performance similar to the high-dimensional SRM method. In the next year, Xu [17] designed a 20-layer fully convolutional network by incorporating residual structures based on the foundation of [16]. This network improved the detection performance for the J-UNIWARD steganographic algorithm. Also in 2017, Chen *et al.* [18] proposed a steganalysis network based on JPEG phase information. Their preprocessing utilized four different high-pass filters and a JPEG phase-aware module, enhancing the network's ability to extract frequency domain steganographic features and effectively improving detection accuracy for frequency domain images. In the same year, Ye *et al.* [19] enhanced steganographic signals using SRM (spatial rich model) high-pass filtering and designed a new activation function to remove the impact of outliers on feature calculations, achieving good detection performance. In 2018, YedroudjNet [20] combined convolution and multiple activation functions to further enhance the detection performance. Also in the same year, Tsang *et al.* [21] combined high-pass filtering, truncation functions, and feature statistical moments to propose a steganalysis method suitable for multi-scale images, achieving good detection performance on a small multi-scale dataset. In 2019, Boroumand *et al.* [22] applied residual structures to steganalysis methods, designing multiple residual blocks and proposing a SRNet method suitable for spatial and JPEG domains. This method significantly improved the detection performance in both spatial and JPEG domains and became a milestone research achievement. In the next year, Zhang *et al.* [23] utilized different-sized convolutional operations to preprocess images, combined with residual structures to enhance the capture of residual features, achieving good detection performance. In 2021, You *et al.* [24] combined the siamese framework with residual structures and proposed the SiaStegNet method, which enhanced the capture of residual features and improved the detection performance for large-scale multi-scale images. In the same year, Su *et al.* [25] combined residual structures with deconvolution to develop a fully convolutional method suitable for multi-scale frequency domain images, effectively improving the detection performance. In 2022, Liu *et al.* [26] introduced the CSANet method, which utilized channel attention mechanisms, pyramid pooling, and residual blocks to construct an efficient steganalysis method, achieving good detection accuracy. Also in the same year, Weng *et al.* [27] proposed the LWENet method, which reduced the parameter count and computational complexity while en-

suring detection accuracy by using bottleneck structures and depth-wise separable networks. In 2023, Li *et al.* [28] further enhanced the detection performance for large-scale multi-scale datasets by combining inverted residual structures with depth-wise separable convolution to enrich residual features.

The above-mentioned method utilizes convolutions to extract residual features and increases the depth and width of the network to capture global features. However, the captured global features are limited, which restricts the improvement of detection performance. Moreover, increasing the depth and width of the network also leads to significant computational resource consumption. Inspired by the design of deep learning-based steganalysis networks, some current classification schemes have proposed effective methods for domain classification [29]–[32]. Therefore, this paper designs a multi-residual structure and proposes a Transformer structure capable of extracting both local and global features. It constructs a spatial and JPEG domain image steganalysis framework based on the combination of residual and Transformer structures. The aim is to effectively reduce the computational resource consumption and parameter quantity of the network while ensuring the detection accuracy.

The main contributions of this paper are as follows:

1) The paper proposes a novel preprocessing layer called the multi-residual layer, which is based on channel shuffle. This layer enhances hidden signals by employing high-pass filtering and channel attention mechanisms. The multi-residual blocks are then used to capture rich residual features, and the expressive power of these features is further enhanced through channel shuffle.

2) The paper presents a lightweight hybrid feature extraction layer. By using depth-wise separable convolution, the resource consumption of the network is reduced. Simultaneously, several convolutional blocks are employed to extract local features, while the attention mechanism in the Transformer captures global features. The combination of these two modules increases the diversity of features.

3) The paper constructs an effective weighted loss function for learning feature diversity. The cross-entropy loss is utilized to learn steganographic features, while the contrastive loss function enhances the expressive power of the features. The BiasLoss is employed to learn both local and global features of the network. The integration of these three components results in a weighted loss function that is well-suited for hybrid networks.

The remainder of this paper is organized as follows. Section II presents the related work of typical steganalysis methods. Section III introduces the ResFormer method in detail. Section IV gives the experimental results and analysis. Section V summarizes the whole manuscript and gives the next research.

II. Related Work

Current deep learning-based steganalysis methods have effectively improved the detection accuracy by improving the network structure and incorporating prior knowledge of steganography. The improvement of network structure has been inspired by some structures from the computer vision field, which have effectively enhanced the detection performance. In 2019, Boroumand *et al.* [22] introduced residual structures into steganalysis methods. In 2021, You *et al.* [24] borrowed this idea and proposed SiaStegNet, an effective method for multi-size image steganalysis using multiple residual blocks. In the same year, Su *et al.* [25] combined multiple residual structures with fully convolutional structures and proposed the EWNet method. In 2022, Liu *et al.* and Weng *et al.* [26], [27] respectively designed CSANet and LWENet methods using residual structures. In 2023, Li *et al.* [28] proposed the SiaIRNet method based on SiaStegNet. CSANet and LWENet were designed with reference to the structure of SRNet, while SiaStegNet and EWNet were effective for multi-size spatial and frequency domain steganalysis, respectively. Therefore, in order to better understand the network structures, the SRNet, SiaStegNet, and EWNetare taken as examples to introduce the types of residual structures used in these three methods, as well as their advantages and limitations.

The steganography detection method based on a convolutional neural network can further extract rich residual features by increasing the depth of the network, which is helpful to improve the detection efficiency. To this end, Boroumand *et al.* [22] proposed a 26-layer SRNet method by designing a variety of residual blocks, which included four different convolution blocks, two of which were residual blocks, T2 and T3, as shown in Figure 1. On the one hand, the residual block increases the depth of the network and enhances the ability of the model to extract residual features. On the other hand, the residual block can prevent overfitting caused by the depth of the network. Compared with the previous methods, the detection effect of SRNet method has been significantly improved. However, the increase in network depth also brings great resource consumption, and the global features captured by stacking convolution are limited, which also limits the further improvement of the detection effect of SRNet method.

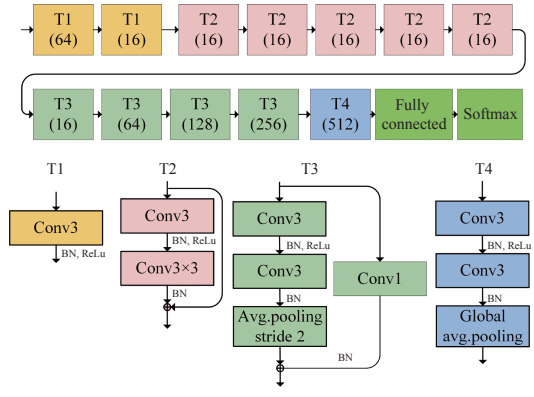

Figure 1 Architecture of SRNet method.

You *et al.* [24] proposed an 18-layer SiaSteNet method by taking images of different sizes in real scenes as training objects. The residual structures, BlockA and BlockB, used by the SiaSteNet method are shown in Figure 2. The method extracts richer subgraph features by using a twin framework and increases the depth of the network using two residual structures. The SiaSteNet method has good detection results on both fixed-size images and large multi-size images. Although the twin structure used by the SiaSteNet method uses less number of parameters, it is computationally intensive due to the increase of the network.

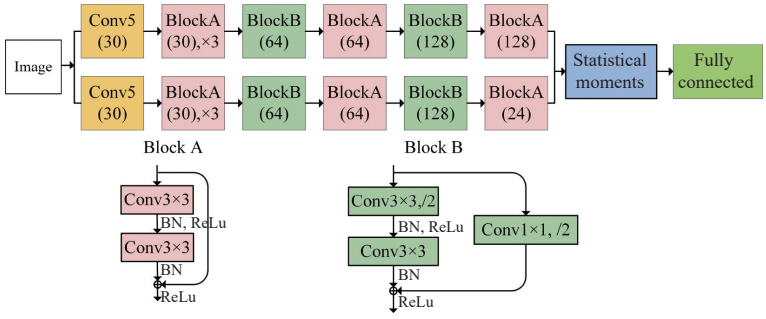

Figure 2 Architecture of SiaStegNet method.

The SiaSteNet method is aimed at the detection of fixed-size and multi-size spatial domain images. Currently, there is a lack of the corresponding detection of JPEG fixed-size and multi-size images. Therefore, Su *et al.* [25] proposed the EWNet method using the full convolutional network. In this method, two kinds of residual structures are designed and combined with deconvolution to increase the depth and width of the network, which enhances the ability of the network to extract residual features and effectively improve the detection performance of fixed-size and multi-size JPEG images. The residual structure is shown in Figure 3. The EWNet method uses two residual structures to extract more abundant local features, but the global features extracted are limited, and this method also brings great resource consumption.

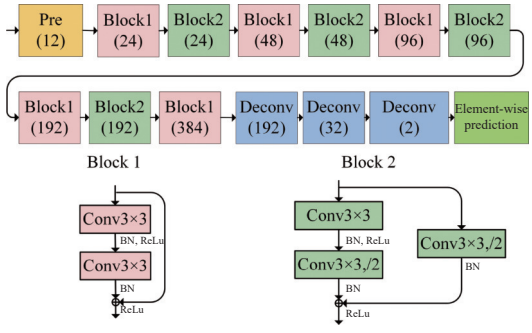

Figure 3 Architecture of EWNet method.

The steganography detection method discussed above is based on a convolutional neural network, which extracts both local and global features from images using the residual structures. However, the global features captured through stacked convolution are limited. This limitation not only hampers the improvement of the detection accuracy of the model but also increases its computational effort. To address these issues, this manuscript proposes a lightweight steganography detection method that combines convolutional and Transformer architectures. This method aims to extract rich local and global features while reducing resource consumption and enhancing the detection effectiveness.

III. Proposed Method

The existing steganography detection methods based on deep learning primarily enhance the effectiveness of steganography detection by increasing the network depth, which, in turn, leads to higher resource consumption. This manuscript proposes a multi-residual structure that utilizes depth-separable convolution to reduce the computational overhead and the number of parameters in the model. Additionally, we introduce a convolutional Transformer feature extraction layer that enhances the diversity of features, thereby allowing for a reduction in network depth. The goal of this method is to reduce the computational cost of the model while ensuring accuracy.

1. Basic principles and main steps

The proposed ResFormer method first trains the lightweight model and then uses the model to detect. In the training phase, first, rich and effective steganography features are extracted by using high-pass filtering and multiple residual structures. Then, convolution and Transformer are used to extract local features and global features to increase the diversity of features. Finally, the full connection layer is used to generate the training model. For image detection, the training model is initialized and classification is performed using the Softmax function. The ResFormer method is depicted in Figure 4.

1) The preprocessing layer includes high pass filtering and multiple residual structures (C-Block1).

a) High pass filtering: SRM (spatial rich model) high pass filters are used for high pass filtering, which can make the model pay more attention to steganography features.

b) C-Block1: Multiple residual structures. SRM helps to identify steganography information. C-Block1 helps to

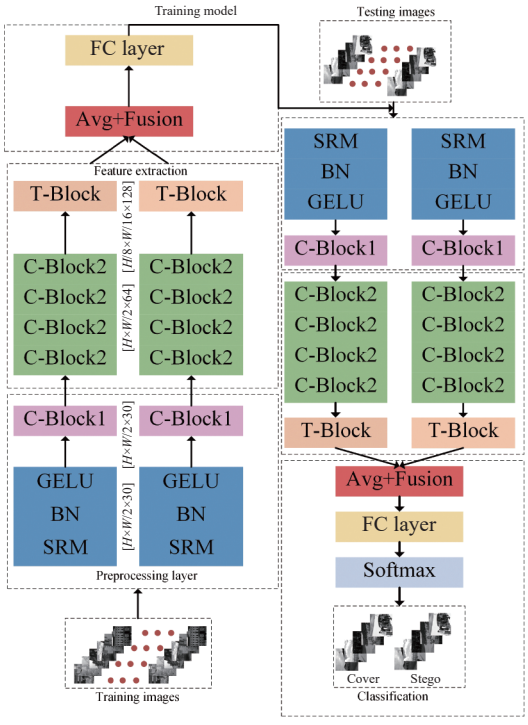

Figure 4 Overview of the proposed ResFormer method architecture.

obtain rich and effective steganography features by using multiple residual blocks. Detailed content will be provided in multiple residual structures and performance analysis of ResFormer method.

2) The feature extraction layer includes four layers of convolution blocks (C-Block2) and one layer of Transformer blocks (T-Block).

C-Block2 and T-Block: Four C-Block2 blocks are used to extract local features. T-Block contains the module of the Transformer structure to extract global features. Detailed content will be given in performance analysis of ResFormer method and convolution and Transformer structure backbone.

3) The generation model includes a global-level pooling layer, a feature fusion layer, and a full connection layer.

a) Global average pooling layer and feature fusion layer: In order to reduce dimension, global average pooling is used. We integrate features to increase feature diversity using the method of [24].

b) Full connection layer: To distinguish the cover and stego images and build the training model, the full connection layer with classification function is used.

4) Classification stage: By loading the training model, multiple residual blocks and Transformer structure are used to extract the features of the image to be detected, and Softmax is used to detect the image.

To provide a better description and understanding of the proposed network architecture, we present an algorithmic description of the training process shown in Algorithm 1.

Algorithm 1 Training process of the proposed method

Input: $M_i = (m_1, m_2, m_3, \cdots, m_n)$, where m represents a grayscale image of size $W \times H \times C$ (W and H denote the width and height of the image, respectively, C represents the number of channels), and K represents the filter.

Output: Model($m_{\text{cover}}, m_{\text{stego}}$).

1: Data processing:

The image m is evenly divided into two parts, m_i and m_j, along the vertical direction. Here, we use m_i as an example to illustrate the training process of the image.

2: Initialization of the preprocessing layer:

$$\boldsymbol{R}_i = \text{C-Block1}(m_i * K)$$

where * denotes convolution and $\boldsymbol{R}_i$ denotes the residual feature having a size of $W/2 \times H \times 30$.

3: Feature extraction layer:

First, pass through four C-Block2 convolutions:

$$\boldsymbol{R}_i = \text{C-Block2}(\boldsymbol{R}_i)$$

where $\boldsymbol{R}_i$ has a size of $W/2 \times H \times 64$.

Then, pass through T-Block and global average pooling:

$$\boldsymbol{R}_i = \text{T-Block}(\boldsymbol{R}_i)$$

where $\boldsymbol{R}_i$ has a size of $W/8 \times H/16 \times 128$.

$$\boldsymbol{R}_i = \text{Avg}(\boldsymbol{R}_i)$$

Repeat the same operations for the other half of the image to obtain the residual feature map $\boldsymbol{R}_j$, Concatenate the obtained features to get $\boldsymbol{R} = \text{Stack}(\boldsymbol{R}_i, \boldsymbol{R}_j)$.

4: Train the model:

First, compute the maximum, minimum, mean, and variance of the residual feature map:

$$\boldsymbol{R} = [\boldsymbol{R}_{\max}, \boldsymbol{R}_{\min}, \boldsymbol{R}_{\text{mean}}, \boldsymbol{R}_{\text{var}}]$$

Next, use fully connected layers and a loss function to obtain the trained model:

$$\boldsymbol{R} = \text{FC}(\boldsymbol{R})$$

$$\text{Model}(m_{\text{cover}}, m_{\text{stego}}) = \text{Loss}(\boldsymbol{R})$$

5: **Return** Model($m_{\text{cover}}, m_{\text{stego}}$).

2. Multiple residual structures

To enhance the ability of preprocessing layer to extract rich residual features, this manuscript designs multiple residual structures, which can enhance the ability of the network to capture residual characteristics, and boost its capability to identify steganography characteristics. The designed residual structure is shown in Figure 5. The steps of the multiple residual structures are as follows. For the input features in the left branch of the multiple residual structures, initially, the residual features are extracted through the application of 3×3 convolution and 1×1 convolution, followed by the utiliza-

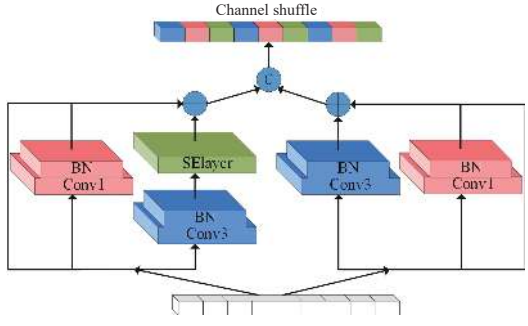

Figure 5 Multiple residual structure (C-Block1).

tion of the channel attention mechanism to grasp the inter-channel correlation of residuals. Then, the extraction features of channel attention, the residual features obtained from 1×1 convolution extraction are combined with the input features to enhance the residual characteristics. In the right branch of the multiple residual structures, the residual features are first extracted by 3×3 convolution and 1×1 convolution, and then the residual features are fused with the input features. After the residual features are extracted from the left and right branches of the multiple residual structures, the residual feature needs to be concatenated, and the channel shuffle is used to enhance the expression ability of the features. To reduce parameters and Flops, the convolution used by multiple residual structures is depthwise separable convolution [14], that is, 3×3 is depthwise convolution and 1×1 is pointwise convolution.

To verify the effect of the multiple residual structures (the method without the multiple residual structures is denoted by ResFormer-R in the following experiments). Three adaptive steganography algorithms [33]–[35] are detected at two embedding rates on the BossBows dataset, and the specific experimental results are shown in Table 1.

Table 1 The detection results of ResFormer-R and ResFormer

Steganography algorithm	Detector	0.2 bpp		0.4 bpp	
		Accuracy (%)	AUC (%)	Accuracy (%)	AUC (%)
WOW	ResFormer-R	87.01	95.89	90.98	97.92
	ResFormer	**88.36**	**96.51**	**92.13**	**98.35**
S-UNIWARD	ResFormer-R	77.97	88.87	88.10	96.49
	ResFormer	**79.44**	**89.67**	**89.64**	**97.28**
HILL	ResFormer-R	74.48	85.26	83.86	93.73
	ResFormer	**75.17**	**85.89**	**84.55**	**94.14**

Table 1 gives the performance of the ResFormer-R method and the ResFormer method for the three steganography algorithms at two embedding rates. It seems that in Table 1, the performance of the ResFormer method is substantially improved. For example, when the embedding rate is 0.2 bpp, compared to ResFormer-R. The detection effect of the ResFormer method has been improved by 1.35% and 1.47%, respectively; when the embedding rate is set at 0.4, the detection efficiency of the ResFormer method is 1.18% and 1.54% higher than that of ResFormer-R method. In summary, it can be seen that the multiple residual structures improve the detection accuracy.

3. Based on convolution and transformer structure backbone

Steganography detection based on deep learning reduces the depth of the model, which is helpful to reduce parameters and Flops, but the detection effect of the model will also decrease. Enriching the diversity of features helps to improve the detection effect of the model. Transformer structure can extract global features. Combining convolution with a Transformer structure can extract local features and global features, which helps to increase the diversity of features. Therefore, this paper designs a Transformer structure (T-Block) and a convolution block C-Block2, as shown in Figure 6. The main steps of the C-Block2 structure are as follows: C-Block2 first bisects the channels, then it uses multiple convolution blocks to extract features from the feature map after bisecting. The SElayers are to enhance the correlation between residual errors in the model extraction channels. Finally, the two features are concatenated and the channel is shuffled. To reduce computation and parameter complexity, C-Block2 utilizes depthwise separable convolution. The main steps of the T-Block structure are as follows: 1×1 convolution is used for channel fusion, and two 3×3 convolutions are used to extract local residual features. The ReAttention module and FFN (feedforward neural) module extract the global features and fuse the local features.

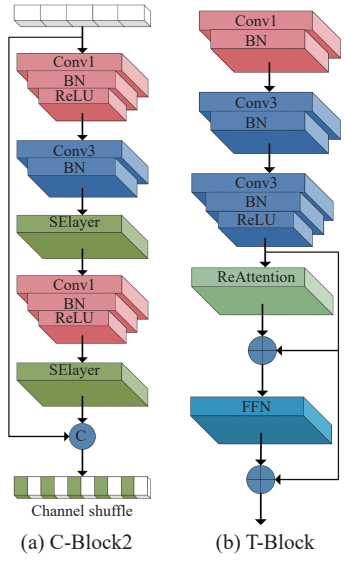

Figure 6 Convolutional block and Transformer block.

To further verify the detection effect of Transformer structure (ResFormer-T below denotes the method without Transformer structure), The BossBows dataset is used to test steganography algorithms with two different pay-

Table 2 The detection results of ResFormer-T and ResFormer

Steganography algorithm	Detector	0.2 bpp		0.4 bpp	
		Accuracy (%)	AUC (%)	Accuracy (%)	AUC (%)
WOW	ResFormer-T	87.35	96.06	91.09	97.93
	ResFormer	**88.36**	**96.51**	**92.13**	**98.35**
S-UNIWARD	ResFormer-T	78.58	89.16	88.73	96.83
	ResFormer	**79.44**	**89.67**	**89.64**	**97.28**
HILL	ResFormer-T	73.22	83.91	82.48	92.88
	ResFormer	**75.17**	**85.89**	**84.55**	**94.14**

loads. The experimental findings can be found in Table 2.

Table 2 gives the detection results of the methods ResFormer-T and ResFormer methods at payloads 0.2 bpp and 0.4 bpp for the three adaptive steganography algorithms, WOW, S-UNIWARD, and HILL. From Table 2, it can be seen that the detection accuracy of ResFormer for the three steganography algorithms has been improved. When the embedding rate is set to 0.2 bpp, in comparison to ResFormer-T. The detection effect of the ResFormer method for two steganography algorithms exhibits enhancements of 1.01% and 1.95%, respectively; when the embedding rate is 0.4 bpp, in comparison to ResFormer-T, the detection efficiency of the ResFormer method for three algorithms exhibits enhancements of 1.99% and 2.07%, respectively. In summary, it can be seen that the Transformer structure can improve the detection accuracy for the adaptive steganography algorithms.

4. Loss function design

The loss function is to allow the network to learn the difference between the cover image and the stego images. In this manuscript, convolution and Transformer blocks are used to extract local and global features to increase feature diversity. The loss function commonly used in steganalysis is cross entropy loss function, but this loss function uses the features after pooling. In order to better utilize the fusion features, BiasLoss [36] is used. BiasLoss uses variance as a simple measure of diversity. The larger the variance, the higher the opportunity to obtain diversity features.

In order to calculate the variance, we assume that the feature before pooling is X, $X \in \mathbb{R}^{B\times C\times H\times W}$, where B represents Batchsize and 30 is taken in this paper, C represents the number of channels, and H and W represent the height and width of the feature graph, respectively. The feature map of the nth data in batch processing can be expressed as

$$\sigma_n = \frac{1}{N-1}\sum_{i=1}^{N}\left(f_i - \frac{\sum_{i=1}^{N} f_i}{N}\right)^2 \tag{1}$$

where X is expanded into a two-dimensional array f, $f \in \mathbb{R}^{B\times N}$, $N = C \times H \times W$. In order to make the variance easier to use for the loss function, the variance is scaled at [0, 1].

$$\sigma_n = \frac{\sigma_n - G_{\min}}{G_{\max} - G_{\min}} \tag{2}$$

$G_{\min}$ and $G_{\max}$ represent the minimum and maximum values activated in a batch of feature graphs in each training iteration, avoiding the instability of the model caused by the too large loss function. BiasLoss can be expressed as

$$\begin{aligned} L_1 = &-[p\cdot\log(q) + (1-p)\cdot\log(1-q)] \\ &* [e(\sigma_n * \alpha_1) - \alpha_2] \end{aligned} \tag{3}$$

where p is the cover image (label 0) or the stego image (label 1), q is the probability of label 1, α_1 and α_2 are constants and 0.3 is taken in this manuscript, and * denotes convolution. The cross-entropy loss function can be expressed as

$$L_2 = -[p\cdot\log(q) + (1-p)\cdot\log(1-q)] \tag{4}$$

To better describe the features of the double-branch, the contrast loss function [19] is used here to learn the features of cover and stego images in the double-branch:

$$L_3 = \frac{p[\max(0, s - ||F_1 - F_2||_2)]^2 + (1-p)(||F_1 - F_2||_2^2)}{2} \tag{5}$$

where F_1 and F_2 represent double branching features. The loss function used by the proposed ResFormer method can be expressed as

$$L_{\text{Proposed}} = L_1 + \frac{1}{10}L_2 + \lambda L_3 \tag{6}$$

where λ is a constant, which is taken as 1 in this manuscript. In order to verify the effect of λ value on the detection effect, the detection effect of the proposed ResFormer is given here when λ takes different values, as shown in Table 3.

5. Performance analysis of ResFormer method

1) Resource consumption of ResFormer method

In deep learning method, the resource consumption of the model is usually measured from the two aspects of

Table 3 The detection results of λ

λ =0.1		λ =0.2		λ =0.3		λ =0.4		λ =0.5	
Accuracy (%)	AUC (%)	Accuracy (%)	AUC (%)	Accuracy (%)	AUC (%)	Accuracy (%)	AUC (%)	Accuracy (%)	AUC (%)
91.73	98.24	91.42	98.13	91.45	98.07	91.33	98.09	91.58	98.19
λ =0.6		λ =0.7		λ =0.8		λ =0.9		λ =1.0	
Accuracy (%)	AUC (%)	Accuracy (%)	AUC (%)	Accuracy (%)	AUC (%)	Accuracy (%)	AUC (%)	Accuracy (%)	AUC (%)
91.61	98.21	91.57	98.16	91.48	98.14	91.67	98.21	**92.13**	**98.35**

the parameters and Flops. The number of parameters can be expressed as the multiplication of input channels, output channels, and the convolution size, that is, parameters = input channels × output channels × the convolution size. Flops can usually be expressed as the multiplication of parameter and output feature size, that is, calculation amount = parameter × output feature size. It can be seen from the parameter and Flops that input channels, output channels and the convolution size will all affect the size of parameter number. In order to reduce parameter and Flops as much as possible, depth separable convolution is adopted in this manuscript. Depth separable convolution includes 3×3 depthwise convolution and 1×1 pointwise convolution, whose parameter can be expressed as $C \times 3 \times 3 + C \times 1 \times 1 \times C'$. The parameters of 3×3 standard convolution can be expressed as $C \times 3 \times 3 \times C'$, where C and C' are input channels and output channels, respectively, so compared with standard convolution, the Flops and parameter of depth-separable convolution are reduced by $\frac{1}{C'} + \frac{1}{9}$ times.

Table 4 gives the parameters and Flops of the proposed ResFormer method, ResFormer-R, and ResFormer-T structures, respectively. From Table 4, when the proposed method does not use the Transformer structure, the number of parameters is reduced substantially and the computation is also reduced, but the detection effectiveness decreases as shown in Table 2; when the Transformer structure is used without the multiple residual structure, the number of parameters and the computation remain unchanged. The multiple residual structure enriches the residual features through residual connections, which does not significantly impact computation or the number of parameters. However, the multiple residual structure has a significant impact on the detection performance, as presented in Table 1. To provide a comprehensive understanding of the ResFormer framework, the detailed structure of the proposed method is presented in Table 5.

Table 4 The parameters and Flops of multiple residual structure and Transformer structure (M stands for 10^6 and G for 10^9)

Detector	Image size	Parameters	Flops
ResFormer	256 × 256	0.39M	1.80G
ResFormer-T		0.03M	1.38G
ResFormer-R		0.39M	1.80G

Table 5 The detailed framework of the proposed ResFormer method

Layer	Output size	Repeat	Channel
Image	256 × 128	1	1
SRM	252 × 126	1	30
C-Block1	252 × 126	1	30
C-Block2	252 × 126	4	64
T-Block	32 × 16	3	128
FC	–	–	2
Flops	–	–	1.80G
Parameters	–	–	0.39M

The detailed framework of the proposed ResFormer method is given in Table 5. Since the proposed ResFormer method uses twin models, the framework of the model is presented here with one of them as an example. As can be seen from Table 5, the table has five columns, the initial column is the number of layers of the model, followed by the feature map output, the convolution kernel size is indicated in the third column, the number of times the layer is repeated is shown in the fourth column, and the output channels are specified in the last column. For a 256×128 image, firstly, it goes through SRM filtering and C-Block1 preprocessing, and the output feature map is 252×126 with 30 channels. Then it goes through C-Block2 and T-Block feature extraction layers and the output feature is 32×16 with 128 channels. Finally, it goes through a fully connected (FC) layer. The number of parameters and computation of the whole network are 0.39M and 1.80G, respectively.

2) Analysis of the role of the backbone structure

The proposed ResFormer method utilizes a multi-residual structure to capture abundant residual characteristics, which enhance the diversity of local and global features in the backbone. The effectiveness of the multi-residual structure and the Transformer structure are verified in Sections III.2 and III.3, respectively. In order to further demonstrate the impact of these two structures on the classification performance of steganalysis, we visualize the distribution of steganographic noise in the image and use the Grad-CAM method [37] to visualize the influence of the two structures, as shown in Figure 7.

From Figure 7(a), we give the cover images and the distribution of steganography signals of the stego images. The adaptive steganography algorithm pays more attention to complex texture regions when embedding information, which can be clearly seen from the distribution

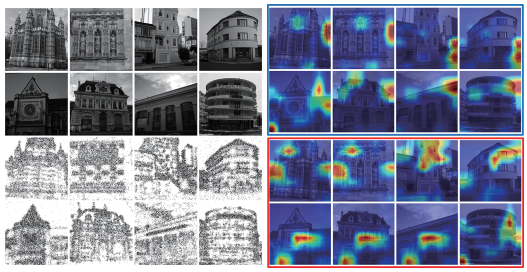

Figure 7 Visualization of the ResFormer. (a) Cover and the noise distribution of stego. (b) The heat map before (the blue box) and after (the red box) using multiple residual structure and Transformer block.

of steganography signals in stego images. Figure 7(b) shows the heat maps before and after using the two structures in this manuscript. The Grad-CAM method can reflect the importance of a certain location in an image to that category. In the blue box of Figure 7(b), the proposed ResFormer pays less attention to the area of complex textured areas, and some focus on non-complex textured areas. In the red box, after using two structure, it can be clearly seen that the ResFormer method proposed focuses on more complex textured areas and fewer non-complex textured areas, which can help the model learn more effective and rich steganography features. In summary, combined with Figures 7(a) and (b), it can be seen that the two structures proposed in this manuscript can effectively learn steganography features and help to enhance the model performance.

IV. Experimental Results and Analysis

Three public databases, BossBase-1.01 [38], BOWS2 (http://bows2.ec-lille.fr), and ALASKA #2 [39], are used in our experiments. ALASKA #2 is a large heterogeneous dataset containing a variety of datasets. Here we use the ALASKA512 dataset and randomly select 10k images ($1k = 10^3$) from it to construct the ALASKA512 dataset. BossBase-1.01, BOWS2, and ALASKA512 are composed of 10k grayscale images of size 512×512. The Boss256, Bows256, and ALASKA256 datasets are composed of 10k grayscale images of size 256×256, which are generated by scaling the above three datasets. The BossBow dataset contains 20k images, which is composed of Boss256 and Bows256. The Boss256_75 and Boss256_85 datasets are compressed with Boss256 dataset at quality factors 75 and 85, respectively. The spatial image databases used for steganography detection is BossBows, the steganography algorithms are WOW, S-UNIWARD and HILL, and the payloads are 0.1–0.4 bpp. The JPEG datasets used for steganography detection are Boss256_75 and Boss256_85 datasets. The steganography algorithm is J-UNIWARD, and the payloads are 0.1–0.4 bpnzac. The training set, verification set, and test set in the spatial domain and JPEG domain are randomly selected at a ratio of 6:1:3. All the experimental results are obtained on a single NVIDIA TelsaV 100 GPU.

1. Experimental parameter setting

The proposed ResFormer method uses the Admax optimizer [40], in which the parameters are set by default, β_1=0.9, β_2=0.99, and eps=1E−7. The Batchsize is set to 30. The initial learning rate is set to 0.001, and when Epoch = 400, the learning rate is adjusted to 0.0001. Accuracy and ROC (receiver operating characteristic) [22], [24] are used to evaluate the performance of the proposed ResFormer method. Figure 8 shows the accuracy and loss of the proposed ResFormer method for the WOW steganography algorithm at payload 0.4 bpp on BossBows dataset. It can be seen from Figure 8(a) and (b) that when Epoch $\leq$ 300, the accuracy of the training set and the validation set will increase with the increase of the Epoch, and the loss function will decrease. When Epoch $\geq$ 300, the accuracy will further increase with the adjustment of the learning rate, and the loss function will further decrease. When Epoch $\geq$ 400, the model loss tends to be constant, and the proposed ResForme method tends to converge. Figure 9 shows the AUC value for the three steganography algorithms WOW, S-UNIWARD, and HILL at payload 0.4 bpp on BossBows dataset.

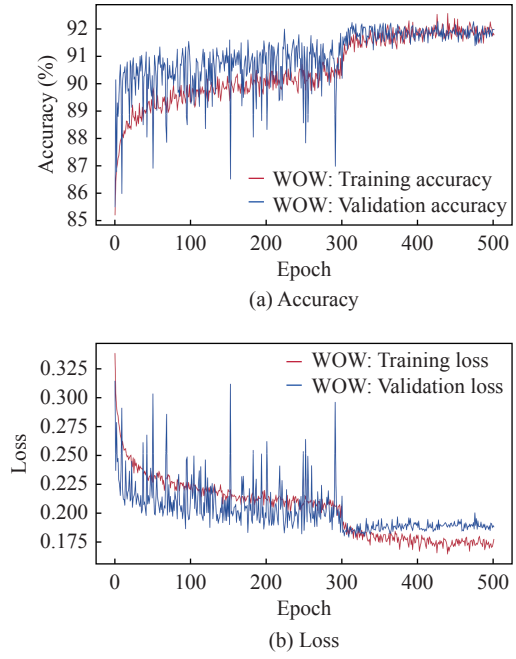

Figure 8 Accuracy and loss of proposed ResFormer for WOW algorithm.

2. Parameters and Flops

In deep learning, the metrics to measure whether a network is lightweight or not usually use parameters and Flops. In Section III.5, the calculation of parameters and Flops are given in this manuscript, and the parameters and Flops of the proposed ResFormer method is given here and compared with the existing steganography detection methods based on deep learning, as shown in

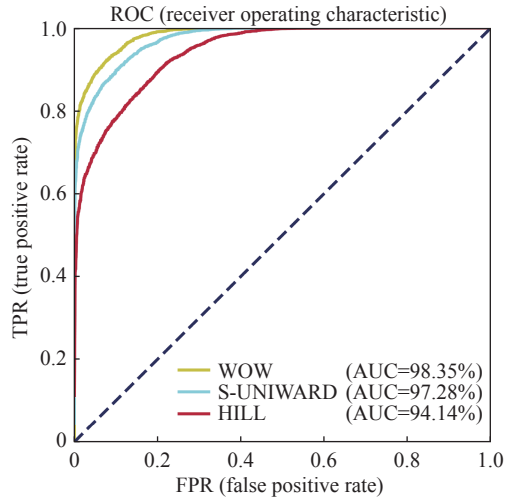

Figure 9 ROC of the proposed ResFormer for three algorithms.

Table 6.

Table 6 shows the number of parameters and Flops of the proposed ResFormer method and the comparison method. As can be seen from Table 6, both the number of parameters and Flops of the proposed ResFormer method has been greatly reduced. Specifically, compared with the SRNet method, the number of parameters and Flops of the proposed ResFormer method has been reduced by 91.82% and 69.75%, respectively. Compared with the SiaStegNet method, the number of parameters and Flops of the ResFormer method is reduced by 45.07% and 75.27%, respectively. Compared to the CSANet method, the proposed ResFormer method reduces the parameters by 17.02% and the Flops by 78%. Compared to the LWENet method, the proposed ResFormer method has a similar parameter quantity but reduces the Flops by 62.66%. Compared to the SiaIRNet method, the proposed ResFormer method reduces the parameters by 45.83% and the Flops by 77.19%. Compared with the EWNet method, the number of parameters and Flops of the ResFormer method is reduced by 89.79% and 11.33%, respectively. EWNet method is the PyTorch version that we use when replicating.

Table 6 The parameters and Flops of SRNet [22], SiaStegNet [24], EWNet [25], CSANet [26], LWENet [27], SiaIRNet [28], and ResFormer

Detector	Image size	Parameters	Flops
SRNet [22]	256 × 256	4.77M	5.95G
SiaStegNet [24]		0.71M	7.28G
EWNet [25]		3.82M	2.03G
CSANet [26]		0.47M	8.18G
LWENet [27]		0.38M	4.82G
SiaIRNet [28]		0.72M	7.89G
ResFormer		0.39M	1.80G

3. Spatial domain images detection

To verify the detection efficiency of the proposed method against a variety of adaptive steganography algorithms, SRM (Spatial Rich Model) [7], SRNet [22], SiaStegNet [24], CSANet [26],LWENet [27], SiaIRNet [28], and the ResFormer method are used to detect three adaptive algorithms, and steganography methods based on deep learning all adopt the same training method. Table 7 gives the specific detection results.

Table 7 shows the detection effect of the six comparative methods and ResFormer method for three spatial adaptive algorithms. As can be seen from Table 7, in comparison to SRM detector, the performance of ResFormer for adaptive steganography algorithms has been greatly improved. Compared with the classical SRNet method and the latest SiaSteNet method, with the payloads increasing, the detection efficiency for adaptive algorithms increases. Specifically, when the embedding rate is set to 0.1 bpp, compared to SRNet, SiaSteg-Net, CSANet, LWENet, and SiaIRNet methods, the detection efficiency of the ResFormer method is exhibits enhancements of 4.39%, 5.10%, 2.60%, 0.44%, and 2.57%, respectively. the detection efficiency of the ResFormer method exhibits enhancements of 0.68%, 1.84%, 1.99%, 0.78%, and 1.81%, respectively. At an embedding rate of 0.3 bpp, the ResFormer method shows improved detection efficiency compared to SRNet, SiaStegNet, CSANet, LWENet, and SiaIRNet methods, with enhancements of 1.39%, 2.78%, 1.40%, 0.44%, and 1.60%, respectively. In comparison to SRNet, SiaStegNet, CSANet, LWENet and SiaIRNet method, the detection efficiency of the ResFormer method exhibits enhancements of 0.39%, 2.91%, 2.02%, 0.63%, and 2.40%, respectively. Overall, the performance of ResFormer method for the adaptive steganography algorithms has advantages.

4. JPEG domain images detection

In addition to spatial domain images, there are many JPEG images in the actual scene. To assess the performance of the ResFormer method on JPEG images, SRNet [22] and EWNet [25] are employed as benchmarking techniques. The J-UNIWARD [34] algorithm at four embedding rates is detected.

Table 8 shows the detection effects of SRNet [22], EWNet [25], and ResFormer for J-UNIWARD steganography algorithm at 0.2 bpnzac and 0.4 pnzac. With the payloads and quality factor decreasing, the detection accuracy decreases as shown in Table 8. That is, the higher the quality factor and the lower the embedding rate, the worse the detection effect; on the contrary, the smaller the quality factor and the higher the embedding rate, the better the detection effect. Specifically, when the quality factor is 75, in comparison to both SRNet and EWNet methods, the detection efficiency of the ResFormer method at an embedding rate of 0.3 bpnzac exhibits enhancements of 5.78% and 6.24%, respectively. When the quality factor is 85, in comparison to the SRNet method and EWNet method, the detection efficien-

Table 7 The detection results of SRM [7], SRNet [22], SiaStegNet [24], CSANet [26], LWENet [27], SiaIRNet [28], and ResFormer

Steganography algorithm	Detector	0.1 bpp		0.2 bpp		0.3 bpp		0.4 bpp	
		Accuracy (%)	AUC (%)	Accuracy (%)	AUC (%)	Accuracy (%)	AUC (%)	Accuracy (%)	AUC (%)
WOW	SRM [7]	55.20	61.30	67.26	78.70	70.12	83.19	72.26	85.55
	SRNet[22]	70.58	78.06	86.98	95.46	88.88	96.65	90.85	97.87
	SiaStegNet [24]	69.87	79.93	86.17	94.14	87.49	96.12	91.04	97.85
	CSANet [26]	72.37	82.93	87.73	96.25	88.87	96.97	91.16	97.98
	LWENet [27]	74.53	85.25	88.25	96.59	89.83	97.32	91.39	98.15
	SiaIRNet [28]	72.40	83.25	87.2	95.4	88.67	96.70	**92.19**	**98.41**
	ResFormer	**74.97**	**85.70**	**88.36**	**96.51**	**90.27**	**97.55**	92.13	98.35
S-UNIWARD	SRM [7]	54.47	59.17	60.79	70.86	67.22	78.66	73.03	85.88
	SRNet [22]	67.80	77.20	79.23	89.76	85.18	94.86	89.10	97.45
	SiaStegNet [24]	66.64	75.44	78.63	88.60	85.13	94.11	88.89	96.68
	CSANet [26]	66.49	76.11	79.06	90.07	84.64	94.36	89.81	97.28
	LWENet [27]	67.70	77.51	78.64	89.58	84.27	94.09	89.28	96.99
	SiaIRNet [28]	66.67	76.29	**80.17**	**90.63**	84.05	93.66	89.51	97.10
	ResFormer	**68.48**	**78.15**	79.44	89.67	**85.96**	**95.01**	**89.64**	**97.28**
HILL	SRM [7]	53.13	56.27	57.38	64.22	62.32	71.61	67.44	78.10
	SRNet [22]	65.77	74.68	74.30	85.76	81.25	91.86	83.75	94.88
	SiaStegNet [24]	64.22	72.45	74.20	83.95	78.73	88.89	83.35	93.23
	CSANet [26]	66.93	75.75	**75.68**	**86.59**	79.62	90.13	**84.95**	**94.55**
	LWENet [27]	67.19	76.71	75.48	86.46	81.03	91.40	84.58	94.25
	SiaIRNet [28]	64.12	72.75	74.45	84.48	79.24	89.55	84.94	93.81
	ResFormer	**66.91**	**76.05**	75.17	85.89	**81.64**	**91.81**	84.55	94.14

cy of the ResFormer method at an embedding rate of 0.3 bpnzac exhibits enhancements of 2.88% and 5.85%, respectively. On the whole, although the proposed ResFormer method has a slightly lower image detection effect than EWNet on QF85 at payloads 0.4 and 0.2 bpnzac, compared to the SRNet and EWNet methods. The ResFormer method has obvious advantages at embedding rates of 0.1 bpnzac and 0.3 bpnzac.

Table 8 The detection results of SRNet [22], EWNet [25], and ResFormer on Boss256 75 and Boss256 85 datasets

Quality factor	Detector	0.1 bpnzac		0.2 bpnzac		0.3 bpnzac		0.4 bpnzac	
		Accuracy (%)	AUC (%)	Accuracy (%)	AUC (%)	Accuracy (%)	AUC (%)	Accuracy (%)	AUC (%)
75	SRNet [22]	56.50	61.01	75.10	83.10	79.04	90.37	89.95	96.26
	EWNet [25]	64.35	68.47	**77.37**	83.44	78.58	90.23	89.67	95.87
	ResFormer	**64.98**	**70.37**	75.88	**84.34**	**84.82**	**92.72**	**90.43**	**96.84**
85	SRNet [22]	55.55	58.83	71.65	80.96	78.05	89.63	87.60	95.71
	EWNet [25]	61.52	65.74	**73.37**	78.86	75.05	87.10	**88.85**	94.33
	ResFormer	**61.86**	**66.40**	72.97	**82.02**	**80.88**	**89.89**	87.48	**95.13**

5. Cover mismatch

In steganography detection methods, the training and testing sets often do not belong to the same dataset, that is, carrier mismatch, and verifying the detection effect of carrier mismatch is of high reference value for practical applications. In this section, Boss256, Bows256, and ALASK256 are used. The payload of the training set, validation set, and test set is 0.4 bpp. The steganography algorithm used is WOW. Table 9 gives the specific detection results.

Table 9 shows the steganography detection effects of SRNet [22], SiaStegNet [24], and ResFormer on three datasets. As can be seen from Table 9, comparing the detection accuracy of the three datasets, the Boss256 dataset is the easiest to detect, while the ALASKA256 dataset is the most difficult to detect because it is generated by large heterogeneous datasets. Specifically, when utilizing the Bows256 training dataset and the Boss256 testing dataset, the detection accuracy of the proposed ResFormer method is 5.19% and 6.22% higher than that of SiaStegNet and SRNet, respectively. When the testing dataset is ALASKA256, ResFormer outperforms SiaStegNet and SRNet by 4.16% and 1.92%, respectively,

Table 9 The detection results of SRNet [22], SiaStegNet [24], and ResFormer on three datasets

Training set	Detector	Testing set					
		Boss256		Bows256		ALASKA256	
		Accuracy (%)	AUC (%)	Accuracy (%)	AUC (%)	Accuracy (%)	AUC (%)
Boss256	SRNet [22]	91.19	98.01	85.85	95.09	**68.10**	76.09
	SiaStegNet [24]	90.53	97.44	83.73	92.59	65.95	75.01
	ResFormer	**92.04**	**98.16**	**85.94**	**94.59**	66.71	**76.72**
Bows256	SRNet [22]	86.85	95.43	87.74	95.95	66.55	75.13
	SiaStegNet [24]	87.88	96.23	84.41	93.64	64.31	73.13
	ResFormer	**93.07**	**98.62**	**90.11**	**97.39**	**68.47**	**78.85**
ALASKA256	SRNet [22]	82.00	94.86	76.05	91.32	68.20	77.18
	SiaStegNet [24]	83.10	93.39	75.72	88.57	67.07	77.22
	ResFormer	**91.52**	**98.15**	**88.12**	**96.42**	**73.22**	**84.56**

in terms of accuracy. The ResFormer method demonstrates detection accuracy improvements of 8.42% and 9.52% over the SiaStegNet method and the SRNet method, respectively, when ALASK256 serves as the training dataset and Boss256 as the testing dataset. Switching to Bows256 as the testing dataset, the ResFormer method showcases detection enhancements of 12.40% and 12.07%, compared to the SiaStegNet method and the SRNet method, respectively. On the whole, the ResFormer method exhibits strong detection performance on the cover mismatch dataset.

V. Conclusions

In response to the issue of high computational resource consumption in deep steganalysis models, caused by the extraction of both local and global features, this manuscript introduces a steganalysis method that utilizes a multi-residual structure and Transformer. The proposed method utilizes a designed multi-residual structure to extract rich residual features, constructs a diverse set of features using convolutional neural network (CNN) and Transformer as the backbone, and enhances the learning capability of the network by incorporating a weighted loss function. The paper extensively analyzes the performance of the proposed approach in terms of CNN and Transformer structures, as well as the number of parameters and computational resources. A series of experiments are conducted, comparing the proposed ResFormer method with SRNet, SiaStegNet, SiaIRNet, and CSANet. The findings indicate that the ResFormer method decreases the parameter count by 91.82%, 45.07%, 17.02%, and 45.83%, respectively, compared to the aforementioned methods. Additionally, the computational resources are reduced by 69.75%, 75.27%, 78.00%, and 77.19%. In terms of detection performance, when the spatial image sets embedding rate to 0.1 bpp, the performance of the ResFormer method exhibits enhancements of 4.39%, 5.10%, 2.60%, 0.44%, and 2.57%, respectively, compared to SRNet, SiaStegNet, CSANet, LWENet, and SiaIRNet. The ResFormer method also improves the detection accuracy of the S-UNIWARD method by 0.68%, 1.84%, 1.99%, 0.78%, and 1.81%. Furthermore, when the quality factor is 75, compared to SRNet and EWNet methods, the proposed ResFormer method improves the detection accuracy of images with 0.3 bpnzac payload by 5.78% and 6.24%, respectively. Overall, the proposed method effectively improves the detection performance of spatial and JPEG images while significantly reducing the computational resource consumption of the model.

From the experimental results in Section III.3 of this work, it can be observed that the detection accuracy of the proposed ResFormer method, which incorporates the Transformer structure, is significantly improved. However, it is accompanied by increased computational and parameter requirements. The experimental comparison of parameter quantities is presented in Table 4 of this work. Currently, in the field of computer vision, there are lightweight transformer-based network architectures that can improve efficiency while effectively reducing resource consumption. Based on this, our next step will be to investigate steganalysis methods based on lightweight transformers. We aim to further reduce resource consumption while improving detection performance.

Acknowledgements

This work was supported by the National Natural Science Foundation of China (Grant Nos. 62172435, 62202495, and U2336206), the National Key Research and Development Program of China (Grant No. 2022 YFB3102900), the Zhongyuan Science and Technology Innovation Leading Talent Project, China (Grant No. 214200510019), and the Key Research and Development Project of Henan Province (Grant No. 221111321200).

References

[1] J. Wen, Y. Q. Deng, W. L. Peng, *et al.*, "Linguistic steganalysis via fusing multi-granularity attentional text features," *Chinese Journal of Electronics*, vol. 32, no. 1, pp. 76–84, 2023.

[2] J. W. Luo, M. J. Yu, X. L. Yin, *et al.*, "Binary image steganalysis based on symmetrical local residual patterns," *Chinese Journal of Electronics*, vol. 31, no. 4, pp. 752–763, 2022.

[3] A. Westfeld and A. Pfitzmann, "Attacks on steganographic

systems," in *Proceedings of 3rd International Workshop on Information Hiding*, Dresden, Germany, pp. 61–76, 2000.

[4] S. Dumitrescu, X. L. Wu, and Z. Wang, "Detection of LSB steganography via sample pair analysis," *IEEE Transactions on Signal Processing*, vol. 51, no. 7, pp. 1995–2007, 2023.

[5] K. Sullivan, U. Madhow, S. Chandrasekaran, *et al.*, "Steganalysis for Markov cover data with applications to images," *IEEE Transactions on Information Forensics and Security*, vol. 1, no. 2, pp. 275–287, 2006.

[6] Q. Z. Liu, A. H. Sung, and M. Y. Qiao, "Neighboring joint density-based jpeg steganalysis," *ACM Transactions on Intelligent Systems and Technology*, vol. 2, no. 2, article no. 16, 2011.

[7] J. Fridrich and J. Kodovsky, "Rich models for steganalysis of digital images," *IEEE Transactions on Information Forensics and Security*, vol. 7, no. 3, pp. 868–882, 2012.

[8] T. Denemark, V. Sedighi, V. Holub, *et al.*, "Selection-channel-aware rich model for steganalysis of digital images," in *2014 IEEE International Workshop on Information Forensics and Security*, Atlanta, GA, USA, pp. 48–53, 2014.

[9] T. Denemark, J. Fridrich, and P. Comesaña-Alfaro, "Improving selection-channel-aware steganalysis features," *Electronic Imaging*, vol. 28, no. 8, article no. art00016, 2016.

[10] W. X. Tang, H. D. Li, W. Q. Luo, *et al.*, "Adaptive steganalysis based on embedding probabilities of pixels," *IEEE Transactions on Information Forensics and Security*, vol. 11, no. 4, pp. 734–745, 2016.

[11] Y. Y. Ma, X. Y. Luo, X. L. Li, *et al.*, "Selection of rich model steganalysis features based on decision rough set α-positive region reduction," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 29, no. 2, pp. 336–350, 2019.

[12] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint*, arXiv: 1409.1556, 2015.

[13] K. M. He, X. Y. Zhang, S. Q. Ren, *et al.*, "Deep residual learning for image recognition," in *Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition*, Las Vegas, NV, USA, pp. 770–778, 2016.

[14] M. X. Tan and Q. Le, "Efficientnetv2: Smaller models and faster training," in *Proceedings of the 38th International Conference on Machine Learning*, Virtual Event, pp. 10096–10106, 2021.

[15] Y. L. Qian, J. Dong, W. Wang, *et al.*, "Learning and transferring representations for image steganalysis using convolutional neural network," in *Proceedings of the 2016 IEEE International Conference on Image Processing*, Phoenix, AZ, USA, pp. 2752–2756, 2016.

[16] G. S. Xu, H. Z. Wu, and Y. Q. Shi, "Structural design of convolutional neural networks for steganalysis," *IEEE Signal Processing Letters*, vol. 23, no. 5, pp. 708–712, 2016.

[17] G. S. Xu, "Deep convolutional neural network to detect J-UNIWARD," in *Proceedings of the 5th ACM Workshop on Information Hiding and Multimedia Security*, Philadelphia, PA, USA, pp. 67–73, 2017.

[18] M. Chen, V. Sedighi, M. Boroumand, *et al.*, "JPEG-phase-aware convolutional neural network for steganalysis of JPEG images," in *Proceedings of the 5th ACM Workshop on Information Hiding and Multimedia Security*, Philadelphia, PA, USA, pp. 75–84, 2017.

[19] J. Ye, J. Q. Ni, and Y. Yi, "Deep learning hierarchical representations for image steganalysis," *IEEE Transactions on Information Forensics and Security*, vol. 12, no. 11, pp. 2545–2557, 2017.

[20] M. Yedroudj, F. Comby, and M. Chaumont, "Yedroudj-Net: An efficient CNN for spatial steganalysis," in *Proceedings of the 2018 IEEE International Conference on Acoustics, Speech and Signal Processing*, Calgary, AB, Canada, pp. 2092–2096, 2018.

[21] C. F. Tsang and J. Fridrich, "Steganalyzing images of arbitrary size with CNNs," *Electronic Imaging*, vol. 30, no. 7, article no. art00005, 2018.

[22] M. Boroumand, M. Chen, and J. Fridrich, "Deep residual network for steganalysis of digital images," *IEEE Transactions on Information Forensics and Security*, vol. 14, no. 5, pp. 1181–1193, 2019.

[23] R. Zhang, F. Zhu, J. Y. Liu, *et al.*, "Depth-wise separable convolutions and multi-level pooling for an efficient spatial CNN-based steganalysis," *IEEE Transactions on Information Forensics and Security*, vol. 15, pp. 1138–1150, 2020.

[24] W. K. You, H. Zhang, and X. F. Zhao, "A Siamese CNN for image steganalysis," *IEEE Transactions on Information Forensics and Security*, vol. 16, pp. 291–306, 2021.

[25] A. T. Su, X. F. Zhao, and X. L. He, "Arbitrary-sized JPEG steganalysis based on fully convolutional network," in *Proceedings of 20th International Workshop on Digital Watermarking*, Beijing, China, pp. 197–211, 2021.

[26] Q. L. Liu, J. Q. Ni, and M. X. Jian, "Effective JPEG steganalysis using non-linear pre-processing and residual channel-spatial attention," in *Proceedings of the 2022 IEEE International Conference on Multimedia and Expo*, Taipei, China, pp. 1–6, 2022.

[27] S. W. Weng, M. F. Chen, L. F. Yu, *et al.*, "Lightweight and effective deep image steganalysis network," *IEEE Signal Processing Letters*, vol. 29, pp. 1888–1892, 2022.

[28] H. Li, J. W. Wang, N. Xiong, *et al.*, "A Siamese inverted residuals network image steganalysis scheme based on deep learning," *ACM Transactions on Multimedia Computing, Communications and Applications*, vol. 19, no. 6, pp. 1–23, 2023.

[29] Y. T. Chen, R. L. Xia, K. Zou, *et al.*, "FFTI: Image inpainting algorithm via features fusion and two-steps inpainting," *Journal of Visual Communication and Image Representation*, vol. 91, article no. 103776, 2023.

[30] Y. T. Chen, R. L. Xia, K. Yang, *et al.*, "MFFN: Image super-resolution via multi-level features fusion network," *The Visual Computer*, vol. 10, no. 2, pp. 489–504, 2024.

[31] Y. T. Chen, R. L. Xia, K. Yang, *et al.*, "DGCA: High resolution image inpainting via DR-GAN and contextual attention," *Multimedia Tools and Applications*, vol. 82, no. 30, pp. 47751–47771, 2023.

[32] Y. T. Chen, R. L. Xia, K. Yang, *et al.*, "DARGS: Image inpainting algorithm via deep attention residuals group and semantics," *Journal of King Saud University-Computer and Information Sciences*, vol. 35, no. 6, article no. 101567, 2023.

[33] V. Holub and J. Fridrich, "Designing steganographic distortion using directional filters," in *Proceedings of the 2012 IEEE International Workshop on Information Forensics and Security*, Costa Adeje, Spain, pp. 234–239, 2012.

[34] V. Holub and J. Fridrich, "Digital image steganography using universal distortion," in *Proceedings of the First ACM Workshop on Information Hiding and Multimedia Security*, Montpellier, France, pp. 59–68, 2013.

[35] B. Li, M. Wang, J. W. Huang, *et al.*, "A new cost function for spatial image steganography," in *Proceedings of the 2014 IEEE International Conference on Image Processing*, Paris, France, pp. 4206–4210, 2014.

[36] L. Abrahamyan, V. Ziatchin, Y. M. Chen, *et al.*, "Bias loss for mobile neural networks," in *Proceedings of the 2021*

IEEE/CVF International Conference on Computer Vision, Montreal, QC, Canada, pp. 6536–6546, 2021.

[37] R. R. Selvaraju, M. Cogswell, A. Das, *et al.*, "Grad-CAM: Visual explanations from deep networks via gradient-based localization," in *Proceedings of the IEEE International Conference on Computer Vision*, Venice, Italy, pp. 618–626, 2017.

[38] P. Bas, T. Filler, and T. Pevný, ""Break our steganographic system": The ins and outs of organizing boss," in *Proceedings of 13th International Workshop on Information Hiding*, Prague, Czech Republic, pp. 59–70, 2011.

[39] R. Cogranne, Q. Giboulot, and P. Bas, "The ALASKA steganalysis challenge: A first step towards steganalysis," in *Proceedings of the ACM Workshop on Information Hiding and Multimedia Security*, Paris, France, pp. 125–137, 2019.

[40] D. P. Kingma and J. L. Ba, "Adam: A method for stochastic optimization," in *Proceedings of 3rd International Conference on Learning Representations*, San Diego, CA, USA, pp. 1–15, 2015.

Hao LI received the B.S. degree in computer science and technology from Zhengzhou University, Zhengzhou, China, in 2015 and the M.S. degree from Central South University of Forestry and Technology, Changsha, China, in 2019. He is currently pursuing the Ph.D. degree with Key Laboratory of Cyberspace Situation Awareness of Henan Province, Zhengzhou, China. His research interests are image steganography and steganalysis technique.
(Email: li15575963101hao@163.com)

Yi ZHANG reveived the B.S., M.S., and Ph.D. degrees from the Key Laboratory of Cyberspace Situation Awareness of Henan Province, Zhengzhou, China, in 2014, 2017, and 2021, respectively. Her current research interests include image robust steganography and steganalysis technique.
(Email: tzyy4001@sina.com)

Jinwei WANG received the B.S. degree in automatic control from Inner Mongolia University of Technology, Hohhot, China, in 2000, the Ph.D. degree in information security from Nanjing University of Science & Technology, Nanjing, China, in 2007. He is a Professor with Nanjing University of Information Science & Technology, Nanjing, China. He was a visiting scholar with Service Anticipation Multimedia Innovation (SAMI) Lab of France Telecom R&D Center (Beijing) in 2006. He worked as a Senior Engineer at the 28th Research Institute, CETC from 2007 to 2010, and as a visiting scholar at New Jersey Institute of Technology, Newark, USA, from 2014 to 2015. His research interests include multimedia copyright protection, multimedia forensics, multimedia encryption, and data authentication. He has published more than 50 papers, hosted and participated in more than 10 projects.
(Email: wjwei_2004@163.com)

Weiming ZHANG received the M.S. and Ph.D. degrees from Zhengzhou Information Science and Technology Institute, Zhengzhou, China, in 2002 and 2005 respectively. Currently, he is a Professor with the School of Information Science and Technology, University of Science and Technology of China, Hefei, China. His research interests include information hiding and multimedia security.
(Email: zhangwm@ustc.edu.cn)

Xiangyang LUO received the B.S., M.S., and Ph.D. degrees from the State Key Laboratory of Mathematical Engineering and Advanced Computing, University of Information Engineering, Zhengzhou, China, in 2001, 2004, and 2010, respectively. He is the author or co-author of more than 200 refereed international journal and conference papers. He is currently a Full Professor with Key Laboratory of Cyberspace Situation Awareness of Henan Province, Zhengzhou, China. His research interests include network and information security.
(Email: luoxy_ieu@sina.com)