Journal of Research in Science and Engineering (JRSE)
ISSN: 1656-1996 Volume-7, Issue-11, November 2025
DOI: 10.53469/jrse.2025.07(11).10

Image Steganalysis Model with Residual Connections and Pyramid Scene Parsing

Ningning Gong[1,*], Xiaoli Zhong[2], Guohui Niu[3]

[1,2,3] College of Intelligent Science and Engineering, Qinghai Minzu University, Xining, 810000, China.
[2] Public Computer Experimental Teaching Demonstration Center, Qinghai Minzu University, Xining, 810000, China.
Correspondence Author

Abstract: ***To address the detection challenges posed by adaptive steganographic algorithms such as HUGO, HILL, and WOW, this paper proposes an improved deep learning model based on the ZhuNet architecture. The model introduces deep residual blocks with learnable scaling factors, replacing standard convolutional blocks to effectively mitigate the vanishing gradient problem in deep networks. Furthermore, the Spatial Pyramid Pooling (SPP) module is superseded by a Pyramid Scene Parsing (PSP) module to enhance multi-scale feature extraction capabilities. Experimental results demonstrate that at a 0.4 bpp embedding rate, the proposed model achieves a detection accuracy of 79.12%, marking a significant improvement over the original ZhuNet (71.58%) and the variant employing only the PSP module (74.38%). Additionally, the improved model exhibits more stable convergence behavior and faster performance improvement during training, validating the effectiveness of the proposed enhancements for steganalysis tasks.***

Keywords: Steganalysis, Deep Learning, ZhuNet, Residual, Pyramid Scene Parsing**.**

1. Introduction

With the rapid advancement of digital multimedia technology, steganography, as a highly covert information hiding technique, poses increasingly severe challenges to network security [1]. Attackers can embed malicious data or scripts into carriers such as digital images and web advertisements. These carriers appear visually normal to ordinary users, allowing the concealed harmful content to execute undetected, potentially leading to system damage, data breaches, and other security risks [2], or even being utilized to conspire illegal activities on public social platforms [3].

To address these challenges, steganalysis techniques have continuously evolved. Content-adaptive steganographic algorithms, represented by HUGO [4], HILL [5], and WOW [6], have become among the most covert steganographic methods currently available. By precisely modeling local image features and concentrating modification operations in complex texture regions, these algorithms effectively evade detection by traditional feature-domain steganalysis methods like SRM [7], presenting a significant challenge to existing detection frameworks.

In recent years, deep learning-based steganalysis has achieved remarkable progress. Specifically, YeNet [8] employed end-to-end learning of SRM filters, achieving performance comparable to traditional handcrafted features. SRNet [9] utilized a residual architecture for direct pixel-domain processing, enabling automatic learning of feature representations suitable for steganalysis and demonstrating excellent detection accuracy. ZhuNet [10] integrated the domain knowledge of SRM filters with deep convolutional neural networks, significantly reducing model complexity and computational overhead while maintaining high detection accuracy.

However, these advanced models exhibit respective limitations. YeNet's relatively simple architecture, despite its enhanced feature selection capability via an attention mechanism, suffers from increased training difficulty due to its complex multi-branch design and shows limited performance improvement under low embedding rates. Although SRNet achieves the highest detection accuracy, its deep residual structure lacks dynamic regulation mechanisms, leading to convergence oscillations during training. Furthermore, its complex architecture and large parameter count result in significantly constrained training and inference efficiency. In comparison, ZhuNet strikes a better balance between efficiency and accuracy, featuring a lightweight structure and commendable detection performance.

Nevertheless, ZhuNet still has room for improvement. Firstly, its standard convolutional blocks are prone to the vanishing gradient problem during training, restricting further network deepening. Secondly, the Spatial Pyramid Pooling (SPP) module [11] lacks an effective feature refinement mechanism during fusion, potentially causing loss of critical spatial contextual information and compromising multi-scale feature extraction efficiency.

To overcome these limitations, this paper proposes an enhanced ZhuNet architecture. We design deep residual blocks [12] with learnable scaling factors to replace the original standard convolutional blocks, effectively mitigating the vanishing gradient problem and enabling the construction of deeper networks. A Pyramid Scene Parsing (PSP) module [13] is introduced to substitute the conventional SPP module, enhancing the model's multi-scale perception of steganographic noise through integrated multi-scale pooling and feature reconstruction. Additionally, we optimize the network normalization strategy and classification head design, further improving model stability and generalization capability. Experimental results demonstrate that the proposed model achieves a peak accuracy of 79.12% in detecting the HUGO algorithm, representing a 7.54% improvement over the original ZhuNet, alongside faster convergence speed and superior training stability.

2. Improved Steganalysis Model Design

2.1 Overall Architecture

To address the trade-off between performance and efficiency in existing deep steganalysis models, this paper proposes an improved ZhuNet architecture, whose overall structure is illustrated in Figure 1. The proposed model retains the advantageous SRM pre-processing of the original ZhuNet while achieving performance breakthroughs through the integration of residual connections and a Pyramid Scene Parsing (PSP) module. Specifically, during the feature extraction stage, standard convolutional layers are replaced with residual blocks incorporating learnable scaling factors to mitigate the vanishing gradient problem. In the feature aggregation stage, the Spatial Pyramid Pooling (SPP) module is superseded by a PSP module, enabling precise capture of multi-scale contextual features. Furthermore, the network incorporates Layer Normalization and a redesigned classification head at its terminal stage, which collectively reduce overfitting risks and enhance the model's generalization capability on unseen samples. The model accepts 256×256 pixel grayscale images as input. These inputs are first processed by SRM filters, then sequentially passed through a separable convolutional enhancement module, a four-stage residual feature extraction backbone, and the PSP multi-scale fusion module, before the lightweight classification head ultimately produces the steganalysis result.

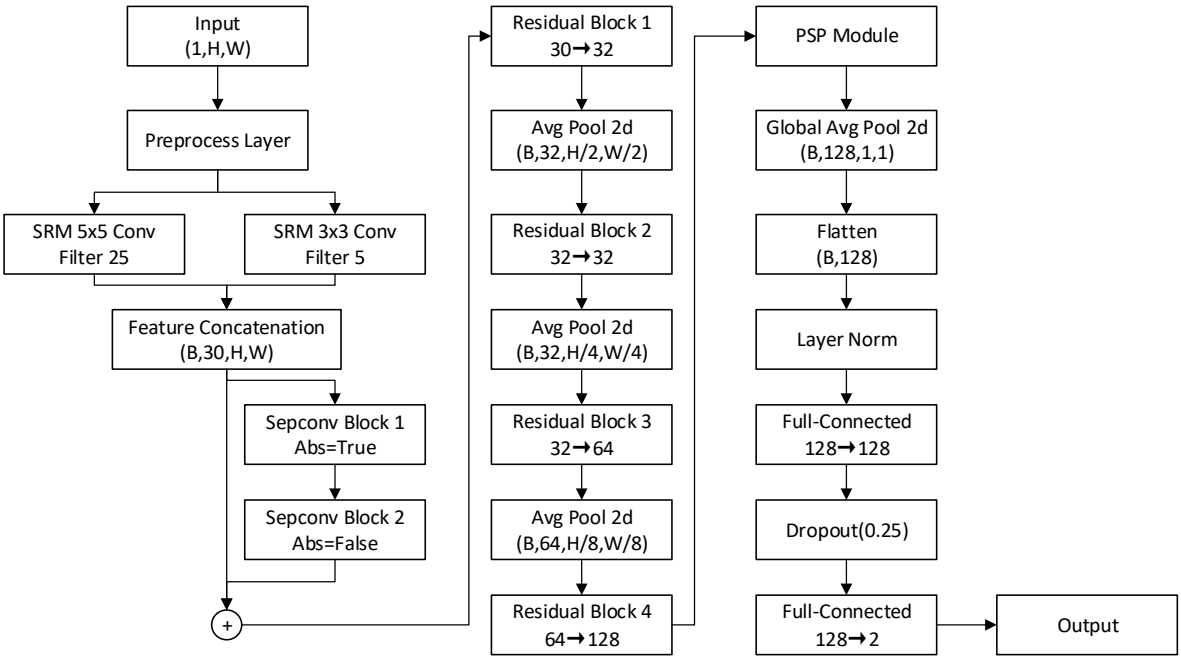

Figure 1: Overall Architecture of the Proposed Model

3. Technical Improvements

3.1 Learnable Residual Scaling Mechanism

To address the vanishing gradient problem in the original ZhuNet caused by standard convolutional blocks, this paper introduces a residual block incorporating a learnable scaling factor. The core idea is to adaptively balance the contribution between the residual branch and the identity mapping. The specific implementation of this mechanism is described as follows:

Let the input feature be $x \in \mathbb{R}^{C\times H\times W}$, where C、H、W denote the number of channels, height, and width, respectively. The primary forward propagation process within the residual block is defined by the following formula:

$$z = ReLU\left(BN\left(Conv\left(W_2, ReLU\left(BN\left(Conv(W_1, x)\right)\right)\right)\right)\right)$$

where $W_1, W_2 \in \mathbb{R}^{C\times C\times 3\times 3}$ are the weight parameters of the two 3×3 convolutional layers. $Conv$ denotes the convolution operation, BN represents batch normalization, and $ReLU$ is the activation function.

The shortcut connection $x_{shortcut}$ is defined as:

$$x_{shortcut} = \begin{cases} x & if\ stride - 1\ and\ C_{in} = C_{out} \\ BN(W_8 * x), & otherwise \end{cases}$$

where C_{in} and C_{out} represent the input and output channel dimensions of the residual block, $W_8 \in \mathbb{R}^{C_{out}\times C_{in}\times 1\times 1}$ denotes the weights of the optional 1×1 convolutional projection, and stride refers to the stride of the convolution within the block.

The final output y is computed as:

$$y = ReLU(Scale(z, \alpha) + x_{shortcut})$$

where $\alpha \in \mathbb{R}$ is the learnable residual scaling factor, and $Scale$ operation multiplies each element of z by α.

Compared to the traditional ResNet residual block, this work introduces the scaling parameter αα at the terminus of the residual path as a trainable variable. Initialized to 1 at the beginning of training, it is subsequently optimized automatically via gradient descent:

$$\alpha = \alpha - \eta \frac{\partial L}{\partial \alpha}$$

where η is the learning rate and L is the loss function.

This design enables the network to dynamically adjust the

contribution of features from the residual path relative to the identity path, thereby adaptively balancing information flow, mitigating the vanishing gradient problem, accelerating convergence during training, and enhancing the network's capacity to capture subtle noise patterns characteristic of steganographic features in deeper layers.

3.2.2 Pyramid Scene Parsing Module

n steganalysis tasks, steganographic noise is often distributed across different scale spaces within an image. To effectively capture these multi-scale features, this paper employs a Pyramid Scene Parsing (PSP) module to replace the original Spatial Pyramid Pooling (SPP) module. Through systematic multi-scale pooling and feature reconstruction, the PSP module significantly enhances the model's perception of contextual information. Its core operations are formally described as follows:

Given an input feature map $F_{in} \in \mathbb{R}^{C\times H\times W}$, the module first captures contextual information at different scales through four parallel adaptive average pooling branches:

$$P_k = Upsample\big(Conv_{1\times1}(Pool_{s_k}(F_{in}))\big),\ k = 1,2,3,4$$

where, $s_k \in \{(1,1),(2,2),(3,3),(6,6)\}$ represents the four different pooling sizes, $Conv_{1\times1}$ reduces the channel dimension of each branch, and $Upsample$ restores the feature map to the original spatial dimensions via bilinear interpolation.

The features from all branches are then concatenated:

$$F_{concat} = [F_{in}, P_1, P_2, P_3, P_4] \in \mathbb{R}^{2C\times H\times W}$$

Where F_{concat} denotes the result of concatenating the multi-scale branch outputs with the original input feature along the channel dimension.

Finally, a bottleneck layer performs feature fusion and dimensionality reduction:

$$F_{out} = ReLU(BN(Conv_{1\times1}(F_{concat}))) \in \mathbb{R}^{C\times H\times W}$$

Compared to the original SPP module, the proposed PSP module introduces a $1\times11\times1$ convolution after each pooling branch for feature transformation and dimensionality reduction, enhancing nonlinear representation capacity and strengthening feature expression. It replaces the vector unfolding operation in SPP with bilinear interpolation for upsampling, preserving the spatial structural integrity of the feature maps, which is beneficial for subsequent pixel-level analysis tasks. Furthermore, it controls parameter growth while integrating multi-scale features, improving computational efficiency and steganalysis accuracy. These characteristics enable the model to capture both local subtle noise anomalies and global statistical feature distributions, which is crucial for detecting the complex noise patterns generated by adaptive steganographic algorithms like HUGO and HILL.

3.2.3 Stability Optimization Strategy

In order to further enhance the training stability and generalization capability of the model, and to reduce the occurrence of gradient explosion, this paper introduces several stability optimization strategies in the classifier head design. After the global average pooling layer, Layer Normalization is introduced to stabilize the feature distribution. Its calculation process is as follows:

Given the input feature $x \in \mathbb{R}^{B\times C}$, where B is the batch size and C is the feature dimension, the output of layer normalization is calculated as:

$$y = \gamma\frac{x-\mu}{\sqrt{\sigma^2+\epsilon}} + \beta$$

where $\mu = \frac{1}{C}\sum_{i=1}^{C} x_i$ is the mean of the input features, $\sigma^2 = \frac{1}{C}\sum_{i=1}^{C}(x_i-\mu)^2$ is the variance, γ and β are learnable scale and shift parameters that reduce the model's sensitivity to different data distributions, and ϵ is a small constant to prevent division by zero, ensuring numerical stability.

3.2.4 Classifier Head Optimization Strategy

This paper redesigns the original classifier head, adopting a lightweight architecture that combines global adaptive average pooling with fully connected layers. Through this optimization strategy, the model's parameter count and computational complexity are significantly reduced, while convergence speed is accelerated. The specific operations are as follows:

First, the feature map output from the PSP module is converted into a feature vector through a global average pooling layer:

$$f = Pool(F_{psp})$$

Then, the feature vector is normalized using Layer Normalization:

$$f_{norm} = LayerNorm(f)$$

Finally, a two-layer classifier implements the final classification through nonlinear transformation, significantly reducing the number of model parameters while maintaining performance:

$$h = ReLU(W_1 f_{norm} + b_1)$$

$$h_{drop} = Dropout(h, 0.25)$$

$$o = W_2 h_{drop} + b_2$$

where $W_1 \in \mathbb{R}^{128\times128}$ and $W_2 \in \mathbb{R}^{2\times128}$ are the weight matrices of the fully connected layers, and the Dropout layer randomly discards neurons with a probability of 0.25 to prevent overfitting.

4. Experiments and Results Analysis

To comprehensively evaluate the effectiveness of the improved ZhuNet model, this paper designs a systematic experimental scheme. Comparative analyses with traditional methods and mainstream deep learning models are conducted on standard datasets.

4.1 Datasets

The experiments utilized the BOSSBase 1.01 and ALASKA2 datasets, from which 10,000 images were randomly selected.

All images were resized to 256×256 pixels using the resize function from the Python PIL library and were subsequently randomly divided into a training set (7,000 images) and a test set (3,000 images) in a 7:3 ratio. Three representative adaptive steganographic algorithms—HUGO, HILL, and WOW—were employed to generate stego images for subsequent training and testing by embedding secret information into cover images at a payload of 0.4 bpp.

4.2 Experimental Setup

All deep learning models were trained using the Adam optimizer with an initial learning rate of 0.001, a batch size of 32, and for 100 epochs. The experiments were conducted in a PyTorch 2.9 framework, utilizing a single NVIDIA RTX 5060 GPU.

4.3 Evaluation Metrics

The performance of the steganalysis models was assessed using Accuracy (A) and the Area Under the ROC Curve (AUC). Accuracy measures the overall detection performance of the model, with a higher value indicating better performance. In this binary steganalysis task, samples are categorized into two classes: cover images and stego images (containing secret information). Defining stego images as Positive (P) samples and cover images as Negative (N) samples, we denote:

- **TP (True Positives)**: The number of correctly classified stego images.
- **TN (True Negatives)**: The number of correctly classified cover images.
- **FP (False Positives)**: The number of cover images misclassified as stego.
- **FN (False Negatives)**: The number of stego images misclassified as cover.

The formula for calculating Accuracy is as follows:

$$A = \frac{TP+TN}{TP+TN+FP+FN}$$

AUC (Area Under the ROC Curve) represents the area under the Receiver Operating Characteristic curve and is used to evaluate the model's discriminative ability. A higher AUC value indicates superior classification performance. The curve is constructed by plotting the True Positive Rate (TPR) against the False Positive Rate (FPR) at various classification thresholds.

The TPR, also referred to as sensitivity, quantifies the proportion of stego images correctly identified. A higher TPR is desirable and is calculated as:

$$TPR = \frac{TP}{TP+FN}$$

The **FPR**, in contrast, measures the proportion of cover images incorrectly classified as stego. A lower FPR is preferred, and its calculation is given by:

$$FPR = \frac{FP}{FP+TN}$$

4.4 Overall Performance Comparison

Table 1 presents a comparative analysis of the detection accuracy achieved by various methods against three different steganographic algorithms.

Table 1: Detection Accuracy (%) of Different Methods at 0.4 bpp Embedding Rate

Method	HUGO	HILL	WOW	Average Accuracy
CovNet [14]	68.85	74.22	90.63	77.90
SiastegNet [15]	73.51	77.58	90.61	80.57
YeNet [8]	60.12	66.87	75.76	67.58
SRNet [9]	71.98	78.42	91.37	80.59
ZhuNet [10]	71.58	76.89	88.74	79.07
PSP-ZhuNet	74.38	78.60	89.53	80.84
Proposed Method	79.12	82.95	93.72	85.26

As evidenced by the results in Table 1, the improved ZhuNet model proposed in this paper achieves superior performance across all three steganographic algorithms, attaining an average accuracy of 85.26%. This represents a significant improvement of 6.19 percentage points over the original ZhuNet. Notably, the proposed method demonstrates a substantial advantage in detecting the HUGO algorithm, achieving a leading accuracy of 79.12%.

4.3 Ablation Study

To validate the individual contributions of each proposed improvement module, a systematic ablation study was conducted. The results are summarized in Table 2.

Table 2: Results of the Ablation Study (Accuracy: %)

Residual Block	PSP Module	Stability Optimization	HUGO	HILL	WOW	Average
×	×	×	71.58	76.89	88.74	79.07
×	√	×	74.38	78.60	89.53	80.84
√	×	×	72.43	75.43	89.68	79.18
√	√	×	77.31	80.84	91.63	83.26
√	√	√	79.12	82.95	93.72	85.26

As shown in Table 2, the introduction of the PSP module alone, while keeping other components unchanged, increased the average accuracy from 79.07% to 80.84%, a relative gain of 1.77 percentage points. This improvement was most pronounced for the HUGO algorithm, demonstrating that the multi-scale feature fusion mechanism of the PSP module effectively enhances the model's ability to perceive complex steganographic noise. In contrast, employing the residual block alone provided only a marginal performance gain. This observation suggests that the advantages of the residual block cannot be fully realized without the support of sufficient multi-scale features provided by the PSP module.

When the residual block was combined with the PSP module, a significant synergistic effect was observed, boosting the average accuracy substantially to 83.26%. The rich, multi-scale features provided by the PSP module establish a foundation for effective deep feature learning within the residual blocks, which in turn ensure these features are utilized efficiently.

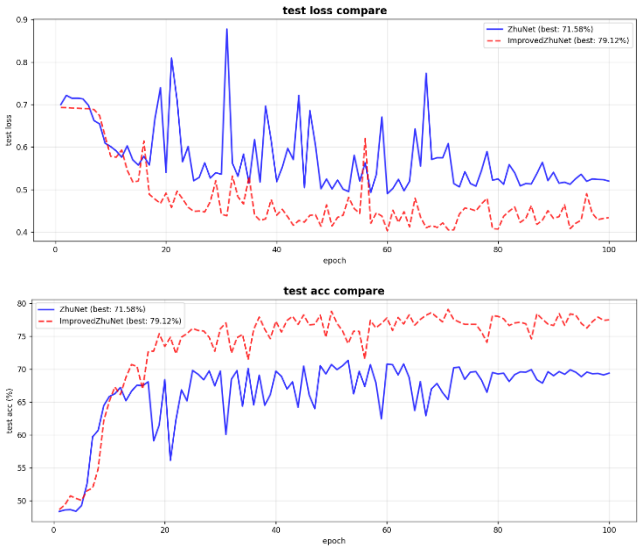

Figure 2: Model Comparison

Figure 2 compares the loss and accuracy curves of the original ZhuNet and the improved ZhuNet on the test set. In conjunction with the data from Table 2, it can be observed that the training process of the original model was characterized by significant oscillations in both loss and accuracy, which hindered further performance enhancement. However, after integrating the stability optimization strategy on the foundation of the residual block and PSP module, the model achieved superior performance, attaining an average accuracy of 85.26%. This strategy effectively improved the training stability of the model and contributed to a notably faster convergence speed.

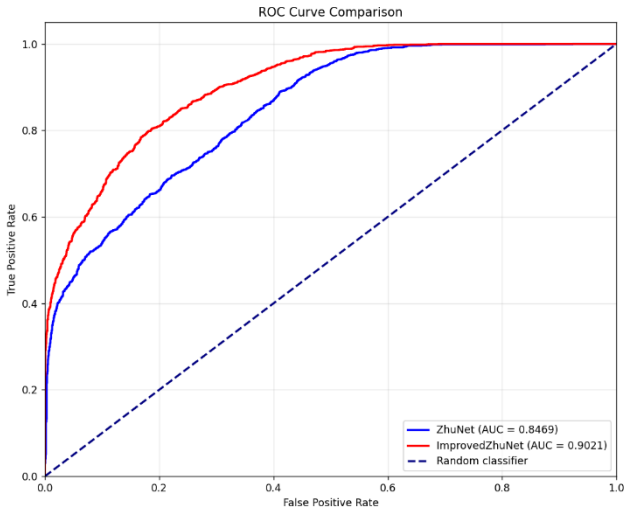

Figure 3: ROC Curve Comparison

Figure 3 presents a comparison of the ROC curves between the original ZhuNet and the improved ZhuNet. The proposed model demonstrates a larger AUC, indicating its superior capability in discriminating between cover and stego images and confirming its advantageous detection performance.

5. Conclusion

To address the challenges faced by current deep learning steganalysis models in detecting adaptive steganographic algorithms, this paper presents an enhanced model based on the ZhuNet architecture. Extensive experimental results demonstrate that the proposed learnable residual scaling mechanism effectively mitigates the vanishing gradient problem in deep network training, enabling the construction of deeper architectures without performance degradation. The replacement of the SPP module with the PSP module significantly strengthens the model's multi-scale perception of steganographic noise. Furthermore, the integration of Layer Normalization and the redesigned classification head substantially improve training stability. Comprehensive experimental results confirm that the proposed improvements lead to a remarkable enhancement in the detection performance of the original model, providing an effective technical solution for practical applications.

References

[1] Michaylov, K. D., & Sarmah, D. K. (2024). Steganography and steganalysis for digital image enhanced Forensic analysis and recommendations. Journal of Cyber Security Technology, 9(1), 1–27. https://doi.org/10.1080/23742917.2024.2304441

[2] Mittal, P., Kaur, R., & Dalal, M. (2026). State-of-the-art image and video-based steganalysis techniques: A comprehensive review, challenges and future recommendations for digital forensic experts. *Computer Science Review*, 59, 100852. https://doi.org/10.1016/j.cosrev.2025.100852

[3] BASHIR B, MIR R N, QURESHI S. Unveiling the evolving landscape of image steganalysis: A comprehensive survey, challenges, and emerging trends [EB/OL]. SSRN, 2025. (2025-10-08) [2024-01-01]. https://ssrn.com/abstract=5579870. DOI:10.2139/ssrn.5579870.

[4] Pevný, T., Filler, T., & Bas, P. (2010). Using High-Dimensional Image Models to Perform Highly Undetectable Steganography. In Information Hiding (pp. 161-177). Springer. DOI: 10.1007/978-3-642-16435-4_13

[5] Li B , Wang M , Huang J , et al. A new cost function for spatialimage steganography [C]// 2014 IEEE International Conference on Image Processing (ICIP). IEEE, 2015.

[6] Holub V,Fridrich J J.Designing steganographic distortion using directional filters. [C]// IEEE International Workshop on Information Forensics & Security. IEEE, 2012.

[7] Fridrich, J., & Kodovský, J. (2012). *Rich Models for Steganalysis of Digital Images*. IEEE TIFS, 7(3), 868-882.DOI: 10.1109/TIFS.2012.2190067

[8] Ye J, Ni J Q, Yi Y. Deep learning hierarchical representations for image steganalysis [J]. IEEE Transactions on Information Forensics and Security, 2017, 12(11): 2545-2557

[9] Boroumand M, Chen M, Fridrich J. Deep residual network for steganalysis of digital images [J]. IEEE Trans actions on Information Forensics and Security, 2019, 14(5): 1181-1193

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

[11] HE K, ZHANG X, REN S, et al. Spatial pyramid pooling in deep convolutional networks for visual recognition [J]. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2015, 37(9): 1904-1916.
[12] TIAN S, YU Z. Improve Generalization Ability of Deep Wide Residual Network with A Suitable Scaling Factor [EB/OL]. *arXiv preprint arXiv:2403.04545*, 2024. (2024-03-07) [2024-01-01]. https://arxiv.org/abs/2403.04545
[13] ZHAO H, SHI J, QI X, et al. Pyramid scene parsing network [C/OL]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2017: 2881-2890. arXiv: 1612.01105 [cs.CV]. (2016-12-04) [2024-01-01]. https://arxiv.org/abs/1612.01105
[14] Deng X, Chen B, Luo W, et al. Fast and effective global covariance pooling network for image steganalysis [C]//Proceedings of the ACM Workshop on Information Hiding and Multimedia Security, 2019: 230-234.
[15] You W, Zhang H, Zhao X. A Siamese CNN for image steganalysis [J]. IEEE Transactions on Information Forensics and Security, 2020, 16: 291-306.