electronics

MDPI

Article

A Novel Technique for Image Steganalysis Based on Separable Convolution and Adversarial Mechanism

Yuwei Ge [1], Tao Zhang [2,*], Haihua Liang [2], Qingfeng Jiang [2] and Dan Wang [3]

[1] School of Computer Science and Technology, Soochow University, Suzhou 215006, China; 20195227092@stu.suda.edu.cn
[2] School of Computer Science and Engineering, Changshu Institute of Technology, Changshu 215500, China; llh@cslg.edu.cn (H.L.); qingfeng_jiang@cslg.edu.cn (Q.J.)
[3] National Key Laboratory of Science and Technology on Blind Signal Processing, Chengdu 610041, China; danwang@fudan.edu.cn
* Correspondence: tzhang@cslg.edu.cn

Abstract: Image steganalysis is a technique for detecting the presence of hidden information in images, which has profound significance for maintaining cyberspace security. In recent years, various deep steganalysis networks have been proposed in academia, and have achieved good detection performance. Although convolutional neural networks (CNNs) can effectively extract the features describing the image content, the difficulty lies in extracting the subtle features that describe the existence of hidden information. Considering this concern, this paper introduces separable convolution and adversarial mechanism, and proposes a new network structure that effectively solves the problem. The separable convolution maximizes the residual information by utilizing its channel correlation. The adversarial mechanism makes the generator extract more content features to mislead the discriminator, thus separating more steganographic features. We conducted experiments on BOSSBase1.01 and BOWS2 to detect various adaptive steganography algorithms. The experimental results demonstrate that our method extracts the steganographic features effectively. The separable convolution increases the signal-to-noise ratio, maximizes the channel correlation of residuals, and improves efficiency. The adversarial mechanism can separate more steganographic features, effectively improving the performance. Compared with the traditional steganalysis methods based on deep learning, our method shows obvious improvements in both detection performance and training efficiency.

Keywords: image steganalysis; deep learning; convolutional neural networks; adversarial training

Citation: Ge, Y.; Zhang, T.; Liang, H.; Jiang, Q.; Wang, D. A Novel Technique for Image Steganalysis Based on Separable Convolution and Adversarial Mechanism. *Electronics* **2021**, *10*, 2742. https://doi.org/10.3390/electronics10222742

Academic Editor: Stefanos Kollias

Received: 23 October 2021
Accepted: 8 November 2021
Published: 10 November 2021

1. Introduction

Image steganography is a technology that hides secret information in images. Due to its simplicity, variability, and difficulty of detection and extraction [1,2], it can be easily used by illegal organizations to engage in activities that will endanger both national and public security. This situation makes steganalysis—an attack technology against steganography—a research hotspot in the field of cyberspace security.

Traditional steganalysis methods include two categories: specific steganalysis, and universal steganalysis. Specific steganalysis is an effective detection method for specific steganography algorithms; its advantage is that its false alarm rate is low, and can accurately reflect the steganographic facts, but it has the problem of small application scope in practical use. Classical specific steganalysis algorithms include regular-sigular (RS) analysis [3], based on the correlation between neighboring pixels, raw quick pair (RQP) analysis [4] to observe changes in test statistics through active steganography, and blockiness analysis on OutGuess [5]. Universal steganalysis regards steganographic detection as a classification problem, extracting high-dimensional features for classification based on machine learning. Classical methods include subtractive pixel adjacency matrix (SPAM) [6] feature analysis

for steganography corrupting the correlation between neighboring pixels, steganalysis of JPEG images based on Markov features [7], spatial rich model (SRM) [8] features extracted by multiple submodels, and several of its variants [9–11]. These methods significantly improve detection performance, but inevitably increase the training time due to the use of high-dimensional features. Feature design is the core element in steganalysis. The features involved in the model are often obtained by manual design. On the one hand, the features require a substantial amount of manual intervention and professional knowledge; on the other hand, the performance of the model directly depends on the quality of the manually defined features.

In recent years, deep learning has flourished in various fields. Some researchers have applied it to steganalysis, with remarkable achievements. Classical methods include Gaussian-neuron convolutional neural network (GNCNN), based on convolutional neural networks and a Gaussian activation function, as proposed by Qian et al. [12]. Xu et al. [13] proposed a CNN structure called Xu-Net containing five convolutional layers, and its detection performance exceeded the spatial rich model for the first time. Ye et al. [14] designed a new truncated linear unit (TLU) as the activation function, based on which the TLU–CNN was proposed. Fridrich et al. [15] designed SRNet on the basis of residual networks, and You et al. [16] designed Ke-Net on the basis of Siamese network. Deep neural networks automatically obtain the feature representations for steganographic detection through sample training, avoiding the dependence on manually defined features. The core problem shifts to the structural design of deep neural networks. In the spatial image steganalysis tasks, what needs to be extracted is the very subtle steganography information features hidden behind the image content and texture, which is significantly different from the traditional computer vision task. Therefore, increasing the signal-to-noise ratio and maximizing the residual information are usually necessary in order to improve the steganographic detection performance.

In this paper, we propose a new end-to-end network to improve the performance of steganalysis tasks, which balances the accuracy and efficiency of steganographic detection. Given that steganographic detection utilizes weak signals hidden in the image content, most previous approaches have introduced high-pass filters to enhance the signal-to-noise ratio. In this paper, separable convolution and an adversarial mechanism are introduced to separate the steganographic signal from the content signal in the spatial image, thus enabling better extraction of steganographic embedding features and improving the performance of image steganographic detection, without the interference of image content. The following steps were taken in the design of the network in order to improve its performance:

A separable convolution module was introduced into the network, which not only can enable it to obtain higher accuracy, but also makes the network converge quickly, and improves efficiency. The module divides the normal convolution into two parts—pointwise convolution, and depthwise convolution—separating the spatial feature learning from the channel feature learning, maximizing the channel correlation of residuals, and effectively enhancing the signal-to-noise ratio.

We introduced an adversarial mechanism into the network structure to suppress image content information and highlight steganographic information as much as possible. In the process of adversarial training, the generator extracts more image content features to mislead the classifier and, thus, isolates the required steganographic features. The introduction of a gradient reversal layer (GRL) allows the network to better extract the steganographic embedding features and improve the performance of steganographic detection, without the interference of image content.

To better train and evaluate the proposed method, we detected a variety of adaptive steganography algorithms on BOSSBase1.01 and BOWS2. The experimental results demonstrate that the separable convolution and the adversarial mechanism have better effects on the extraction of the existence features of hidden information. The introduced separable convolution improves the signal-to-noise ratio, maximizes the channel correla-

tion of the residuals, reduces the number of training parameters, and improves efficiency. More steganographic embedding features can be separated via the adversarial mechanism, which effectively improves the performance of steganalysis.

The rest of the paper is organized as follows: In Section 2, we briefly review classical steganalysis network architectures. Section 3 focuses on the network model on the basis of the separable convolution and the adversarial mechanism proposed in this paper. Section 4 provides and analyzes the experimental results on BOSSBase1.01 and BOWS2. Finally, Section 5 presents the concluding remarks.

2. Related Works

The earliest use of deep learning for steganalysis can be traced back to 2014, when Tan et al. [17] used a stacked convolutional autoencoder for steganographic detection; they found that the network usually failed to converge after directly applying a randomly initialized CNN to the steganalysis task, and that using a KV kernel to initialize the weights of the first layer of the network could effectively improve the accuracy.

Qian et al. [12] proposed a customized GNCNN for steganographic detection; the network structure contains three parts: a preprocessing layer with high-pass filters, a convolutional layer for feature extraction, and a fully connected layer for classification. This method is the first to apply CNNs to the task of steganalysis, achieving results that are comparable to traditional methods using hand-crafted features.

Xu et al. [13] proposed a CNN structure with five convolutional layers, introducing batch normalization (BN) and global average pooling, which are commonly used in image classification tasks. The network uses various activation functions—including absolute (ABS) activation, hyperbolic tangent (TanH) activation function, and rectified linear unit (ReLU)—to improve the experimental results; its performance exceeds the SRM scheme [8], and the improved Xu-Net achieves better results for steganalysis in the JPEG domain [18].

Ye et al. [14] proposed a new method in 2017, which uses a set of high-pass filters in SRM to detect the steganographic signal in the image. The method of initializing the preprocessing layer parameters is significantly better than random initialization. The method uses TLU for the first time. On this basis, TLU–CNN was designed. The idea of selection-channel-aware steganalysis was introduced, and a selection-channel-aware TLU–CNN network was proposed. Experimental results show that the detection performance of this network has obvious advantages over the traditional rich-model method.

Yedroudj et al. [19] proposed Yedroudj-Net in 2018. This method borrows excellent results from Xu-Net and Ye-Net, uses 30 filters from SRM [8] as initialization values of the preprocessing layer, and then adds batch normalization layers and truncating linear units. The method still achieves good performance without the use of selection-channel awareness.

Li et al. [20] designed a CNN network with a parallel subnet structure by using linear and nonlinear filters, which further improved the performance of detection. Boroumand et al. [15] proposed SRNet, which does not use high-pass filters in the traditional sense, but maximizes the noise residuals introduced by the steganography algorithms, and is one of the current methods that can achieve high accuracy. Zhu et al. [21] proposed a CNN on the basis of separable convolution, multilevel pooling, and spatial pyramid pooling for steganalysis, which achieved good performance in detecting arbitrary-sized images.

The main idea of the above approaches is to regard the image steganographic detection task as an image binary classification problem, and then use the classical image classification framework based on the CNN. Nevertheless, a significant difference clearly exists between the steganalysis task and the image classification task. Image classification relies on content information, whereas steganographic detection requires subtle noise signals hidden under the image content. Consequently, directly adopting the CNN framework is difficult for steganalysis tasks. The existing methods are generally solved by adding high-pass filters in the preprocessing layer, but the manually defined filters are not always optimal, and may suppress part of the steganographic signal. In view of the fact that the performance of

image steganographic detection depends heavily on the signal-to-noise ratio, this paper introduces separable convolution and adversarial mechanism to enhance the signal-to-noise ratio and improve detection performance.

Introduction of separable convolution module: Separable convolution splits normal convolution into pointwise convolution and depthwise convolution. The module first separates the channels and performs independent spatial convolution for each channel; it then concatenates the output channels via pointwise convolution to perform spatial feature learning and channel feature learning, thereby maximizing the channel correlation of the noise residuals to improve the signal-to-noise ratio in order to detect the subtle differences between the cover signal and the steganographic signal.

Introduction of adversarial training: The image contains content information reflecting the visual perception of the image, along with steganographic information reflecting the embedding of steganographic messages. In this paper, we use the idea of transfer learning for reference, and introduce adversarial training [22] to suppress content information and highlight steganographic information as much as possible; doing so can better extract steganographic embedding features and improve the detection performance of the network, without the interference of content information.

By introducing the above two modules, the proposed network significantly improves the accuracy of steganalysis.

3. Proposed Method

In this section, we elaborate on the proposed network structure, key modules, and implementation details, and provide the training process of the network and parameter settings.

3.1. Architecture

3.1.1. General Structure

Figure 1 shows the structure of the CNN-based steganalysis network proposed in this paper. The network is an end-to-end network where the input is an image of size 256 × 256 and the outputs are two class labels: cover image, and steganographic image. The network consists of one image preprocessing layer, two separable convolutional (sepconv) modules, one gradient reversal layer (GRL), four base convolutional modules, and multiple fully connected layers.

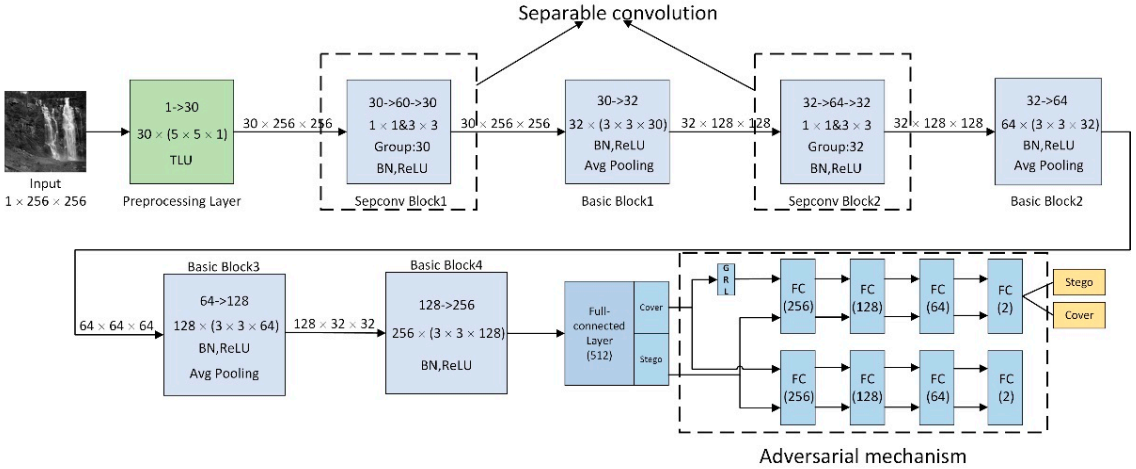


Figure 1. Architecture of the proposed CNN: For each block, $x_1 \rightarrow x_2$; $x_2 \times (a \times a \times x_1)$ denotes the block with the kernel size $a \times a$ for x_1 input feature maps and x_2 output feature maps. Batch normalization is abbreviated as BN. Fully connected layer is abbreviated as FC.

The function of the preprocessing layer is to calculate the noise residuals. In recent years, most methods perform residual calculation before feature extraction, and generally use high-pass filters to filter out content information in order to enhance steganographic

information. In this paper, the preprocessing layer is initialized using the 30 high-pass filters in SRM [8], and these weights are not optimized during the training.

The separable convolution module and adversarial mechanism are introduced to the network, enabling it to more effectively extract image steganographic embedding features, and improving its detection performance. The next section provides detailed descriptions of the separable convolution module and the generative adversarial module.

3.1.2. Separable Convolution Module

Separable convolution has recently made great progress in computer vision tasks, such as Inception [23] and Xception [24]. Xception can be seen as the extreme version of Inception, which reduces storage space and enhances the expressiveness of the model. The general steganalysis approach is to extract attributes via direct normal convolution in 3D space, which ignores the correlation between multiple channels of the image itself, and does not fully utilize the channel correlation of the residual information. Based on the Xception module, a separable convolution module is designed to maximize the channel correlation of residual information, in order to better extract the steganographic embedding features.

Separable convolution includes pointwise convolution and depthwise convolution to separate spatial feature learning from channel feature learning. As shown in Figure 2a, each separable convolution module contains 1×1 pointwise convolution and 3×3 depthwise convolution, where the 1×1 pointwise convolution extracts the channel correlation of the residuals, and the 3×3 depthwise convolution extracts the spatial correlation. The spatial correlation and channel correlation of residuals are independent. Thus, the introduction of the separable convolution allows the attainment of channel correlations of the residuals, and improves the performance of the network. The separable convolution module has the following advantages:

1. It separates the normal convolution into pointwise convolution and depthwise convolution, maximizing the channel correlation of the residual information;
2. It increases the signal-to-noise ratio and extracts more steganographic embedding features, without the interference of image content information;
3. It adjusts the size and number of convolution kernels in order to reduce the number of parameters and the computation involved, shortening the training time and improving efficiency.

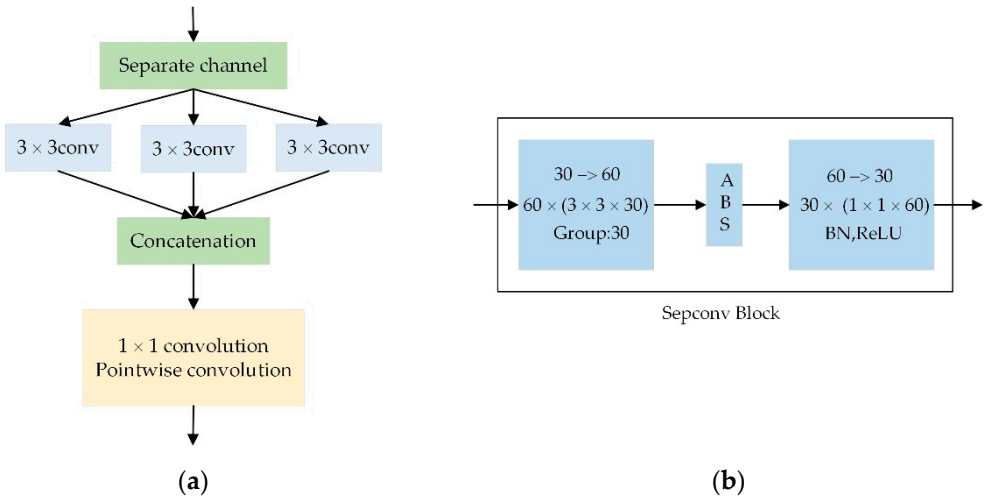


Figure 2. (**a**) Separable convolution structure; (**b**) separable convolution module.

3.1.3. Generative Adversarial Mechanism

In recent years, generative adversarial networks (GANs) [25] have been widely used in various fields. A GAN is essentially a generative model where the training process is in a state of confrontation and its main structure consists of a generator and a discriminator.

The generator generates samples that mislead the discriminator as much as possible, whereas the discriminator needs to classify as accurately as possible. Adversarial training is essentially a zero-sum game between the generator and the discriminator, which can usually be implemented as a complex min–max problem. We simplify this problem to a simple minimization problem by introducing GRL [22,26], which has no parameters associated with it. During the forward propagation, GRL acts as an identity transformation. During the backpropagation, GRL takes the gradient from the subsequent level, and then changes its sign before passing it to the preceding layer.

In the application of image steganographic detection, considering that the image contains both content information reflecting the visual perception of the image and steganographic information reflecting the embedding of secret messages, the steganographic information is hidden in the content information. Thus, extracting the steganographic embedding features that describe the existence of the steganographic information is extremely difficult. This paper draws on the ideas of transfer learning [22] to introduce adversarial training in steganographic detection, in order to suppress the content information of the image as much as possible and highlight the steganographic information to extract the steganographic embedding features more effectively. As shown in Figure 3, the adversarial mechanism can be divided into three parts:

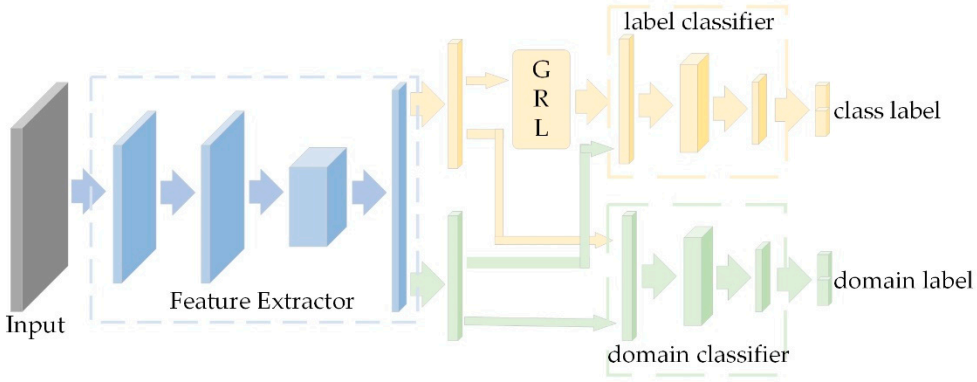


Figure 3. Architecture of adversarial mechanism; GRL: gradient reversal layer.

Feature extractor: This part extracts features and decomposes them into image content features and steganographic embedding features. For steganalysis tasks, image content features are interference signals, and should be suppressed as much as possible, while steganographic embedding features are useful features required by subsequent classification tasks.

Label classifier: This part introduces GRL [22,26], which simplifies the adversarial training between the feature extractor and the classifier; it will optimize the loss function in the direction of negative gradient, and extract more image content features in order to mislead the discriminator's classification; thus, it is deemed more conducive to the detection of the existence of steganographic information. Through adversarial training, more steganographic embedding features can be isolated, and the accuracy of the network can be improved.

Domain classifier: The domain classifier classifies the decomposed features into two classes, distinguishing as much as possible between image content features and steganographic embedding features.

The advantage of introducing adversarial mechanism into the network is that it can fully suppress image content information, highlight useful steganographic information, and extract more steganographic features from it, improving the detection performance and generalization ability of the network.

The introduced separable convolution and adversarial mechanism are interrelated in the network. The separable convolution module separates spatial feature learning from channel feature learning, and further processes the results of the preprocessing layer in or-

der to maximize the channel correlation of the noise residuals. The feature maps generated by the separable convolution module retain less image content information, so that separable convolution blocks facilitate subsequent basic convolution blocks and the adversarial module. The feature maps passed to the adversarial module ensure better extraction of the steganographic embedding features, improving the accuracy of the steganalysis. The separable convolution module and the adversarial mechanism work together to ensure that the network can fully extract the steganographic embedding features and complete the steganalysis task, without the interference of image content information.

3.2. Implementation Details

The training process of the network is as follows: an input image of 256 × 256 pixels, after passing through high-pass filters in the preprocessing layer, is sent to the subsequent separable convolution module and the basic convolution module for feature extraction; after a fully connected layer, the feature decomposition is used to split the feature into two parts, one of which needs to pass through the GRL [22,26], and is then sent to the subsequent fully connected layer for classification in order to complete the task of image steganalysis.

The network can be divided into three parts: image preprocessing, feature extraction, and classification. The implementation details of each part are described below.

The starting part of the network is the image preprocessing layer, and the parameters of this layer are initialized using the 30 high-pass filters in the SRM [8]. The method proposed in this paper resizes all of the original convolution kernels to 5 × 5, and pads the inconsistently sized convolution kernels with zeros. The noise residuals of the input are calculated using the optimized convolution kernels and stacked together as the input of the subsequent feature extraction layer.

The feature extraction section consists of three types of modules: separable convolution modules, basic convolution modules, and generative adversarial modules.

3.2.1. Separable Convolution Module

As shown in Figure 2b, the separable convolution module includes 1 × 1 convolution and 3 × 3 convolution, and the input is the noise residual calculated by the preprocessing layer. The module separates the multichannel feature maps from the previous layer into the feature map of a single channel, performs independent spatial convolution for each channel and, finally, concatenates the output channels by 1 × 1 pointwise convolution. All convolutions in the module do not use bias, use ReLU activation function, and introduce the absolute value function (ABS) to ensure sign symmetry.

3.2.2. Basic Convolution Module

Each basic convolution module consists of the following steps:

1. Convolution layer: In the basic convolution module, a 3 × 3 convolution kernel is used. The optimized convolution kernel allows the network to have fewer parameters, reducing the time required for training and increasing efficiency. The padding and step size are set to 1, and no bias is used;
2. Batch normalization (BN) layer: Batch normalization can normalize the distribution to a zero-mean and a unit variance during the training. The benefit of using a BN layer is that it can recalibrate the value sent into the nonlinear activation function to the appropriate position, and reduce the sensitivity of parameters in the initialization process. Experiments show that adding a batch processing layer can effectively accelerate training and improve the accuracy of detection;
3. Nonlinear activation function: The nonlinear activation functions commonly used in spatial image steganalysis tasks are TanH, ReLU, and TLU. ReLU and TLU are

used in the network structure proposed in this paper, and TLU is also used in the preprocessing layer. The equation for TLU is expressed as shown below:

$$Trunc(x) = \begin{cases} -T, & x < -T \\ x, & -T \le x \le T \\ T, & x > T \end{cases} \quad (1)$$

where the threshold T is set to 3. The basic convolution module uses ReLU as a nonlinear activation function;

4. Average pooling layer: Average pooling is used in the basic convolution module, which downsamples the feature maps and reduces the complexity of the features by adjusting the size of the feature maps. Considering the ability to abstract image features, the average pooling can meet the needs of the network for generalization capability. The last basic convolution module uses global average pooling, allowing it to be fed into the subsequent fully connected layers.

3.2.3. Generative Adversarial Module

The generative adversarial module includes a GRL [22,26] and several fully connected layers. The introduction of the GRL simplifies the min–max problem of adversarial training to a minimization problem, which is equivalent to a constant change in values when the network is forward-propagated, whereas in backpropagation the layer changes the sign of the acquired gradient and passes it to the preceding layer.

The fully connected layer in the generative adversarial module can be divided into two parts: one as a label classifier, and the other as a domain classifier. The function of the label classifier is to complete the image steganographic detection task and perform accurate classification on the basis of the input features; the domain classifier is to classify the image content information and steganographic information, keeping the two types of features separated as much as possible in order to obtain better steganographic detection features and improve the accuracy of steganalysis.

Given a training dataset $\{x_i, y_i\}_{i=1}^{N}$, where x_i is an input, y_i is the corresponding category vector. The feature extractor E, the label classifier C, and the domain classifier D are parameterized by θ_e, θ_c, *and* θ_d, respectively. First, x_i is fed into the feature extractor, and the extracted feature vector is decomposed into the image content feature vector c_i and the steganographic embedding feature vector s_i, whose domain labels are represented by the vectors b_i and q_i, respectively. Then, c_i and s_i are simultaneously fed into the subsequent label and domain classifiers. The label classifier obtains the output vectors $\hat{h}_i$ and $\hat{y}_i$. The domain classifier obtains the output vectors $\hat{b}_i$ and $\hat{q}_i$. The network is trained to reduce the loss function as follows:

$$Loss = L_s + L_c + L_d \quad (2)$$

The loss function consists of three components:

Loss obtained by feeding the steganographic embedding feature vector into the label classifier L_s. This section uses the cross-entropy (CE) loss function to calculate the sum of the cross-entropy between y_i and $\hat{y}_i$, where $\hat{y}_i = C(s_i)$. L_s can be expressed as:

$$L_s(\theta_e, \theta_c) = \sum_{i=1}^{N} CE(y_i, \hat{y}_i) \quad (3)$$

Loss obtained by feeding the image content feature vector into the label classifier L_c. Generating adversarial training is usually implemented by a complex min–max problem, as follows:

$$\min_{\theta_c} \max_{\theta_e} \sum_{i=1}^{N} CE(y_i, \hat{h}_i) \quad (4)$$

where $\hat{h}_i = C(c_i)$. This paper introduces a GRL [22,26] into the steganalysis task in order to simplify this complex problem into a minimization problem. Mathematically, we treat the GRL as a pseudo-function R defined by two equations describing its forward propagation and backpropagation behaviors:

$$R(x) = x \tag{5}$$

$$\frac{dR}{dx} = -I \tag{6}$$

Therefore, adversarial training can be implemented by the minimization of the L_c:

$$L_c(\theta_e, \theta_c) = \sum_{i=1}^{N} CE(y_i, \hat{h}_i) \tag{7}$$

where $\hat{h}_i = C[R(s_i)]$, and R stands for the function implemented by the GRL layer, which implements the constant transform in forward propagation and acquires the gradient from the subsequent layer in backward propagation, changing the sign and then passing it to the preceding layer.

Loss of domain classifier L_d. The function of the domain classifier is to distinguish as much as possible between image content features and steganographic embedding features, and this partial loss is defined as:

$$L_d(\theta_e, \theta_d) = \frac{1}{2}\left[\sum_{i=1}^{N} CE(b_i, \hat{b}_i) + \sum_{i=1}^{N} CE(q_i, \hat{q}_i)\right] \tag{8}$$

where $\hat{b}_i = D(c_i), \hat{q}_i = D(s_i)$. By minimizing the loss function, the two types of features can be further distinguished in order to separate other steganographic features and improve the recognition performance of the network.

4. Experiments

4.1. Dataset and Software Platform

To obtain fair comparison results, all experiments used the same dataset, and the two standard datasets were as follows:

BOSSBase1.01 [27]: This dataset contains 10,000 uncompressed grayscale images with a size of 512 × 512 pixels, which are derived from 7 different brands of cameras.

BOWS2 [28]: This dataset also contains 10,000 uncompressed gray images with a size of 512 × 512 pixels, and the image distribution in the dataset is very similar to that in BOSSBase1.01.

Experiments on the steganographic detection of the spatial adaptive steganography algorithms spatial-universal wavelet relative distortion (S-UNIWARD) [29], high-pass, low-pass, low-pass (HILL) [30] and wavelet obtained weights (WOW) [31] were performed on two image databases as above, using MATLAB for two steganographic embeddings of the cover images at 0.2 bpp and 0.4 bpp, using a random embedding key during the steganography process. The network was trained, validated, and tested in a PyTorch environment. The method was compared with Ye-Net [14], SRNet [15], Yedroudj-Net [19], and Zhu-Net [21].

4.2. Training, Validation, and Testing

Due to the limited GPU computing power, training the network using the original 512 × 512 images would be time consuming. Accordingly, we used MATLAB to change the original images to 256 × 256 pixels, and all subsequent experiments were conducted on the basis of the images of 256 × 256 pixels.

The designed experiment was divided into three parts:

The first part of the experiment focused on the effectiveness of the separable convolution and the adversarial mechanism. The experiment used 10,000 modified images on

the basis of BOSSBase1.01, with each cover image having its own corresponding steganographic image, for a total of 20,000 images. The training set contained 6000 pairs of images, the validation set contained 2000 pairs of images, and the remaining 2000 pairs of images were used as the test set, with no overlay in the three-part image set. This part of the experiment verified the effectiveness of the separable convolution and the adversarial mechanism by removing the separable convolution module and the GRL, respectively, from the network structure.

The second part of the experiment compared our network with other steganalysis methods based on CNNs. The size of the original cover images in the BOSSBase1.01 dataset was modified, and then multiple adaptive steganography algorithms were performed to obtain 10,000 pairs of images as the dataset. Similarly, 6000 pairs were used as the training set, 2000 pairs as the validation set, and 2000 pairs as the test set. This part of the experiment compared the detection performance of the method proposed in this paper with various CNN-based steganalysis methods at 0.2 bpp and 0.4 bpp.

The third part of the experiment considered the impact of data expansion on network performance. Considering that a larger training set is effective in avoiding overfitting for experiments based on CNNs, 10,000 pairs of images from the BOWS2 dataset were added to this part of the experiment; together with 6000 pairs of images from BOSSBase1.01, the training set totaled 16,000 pairs of images; 20% of BOSSBase1.01 was used as the validation set, and the remaining part was used as the test set for the experiments.

According to the above experimental design, the proposed method was trained and tested with the same hyperparameters and settings as the previous method, and the test results were taken as the final performance of the model.

4.3. Hyperparameters

The method proposed in this paper applies a mini-batch stochastic gradient descent (SGD) to train the CNN network, with the momentum set to 0.9 and the weight decay set to 0.0005. Due to the limited computing power, the batch size in the training was set to 16 (8 cover/stego pairs). All convolutional layers in this network structure were initialized using the Kaiming method [32], and all linear layers were initialized by random numbers generated from zero-mean Gaussian distribution with a standard deviation of 0.01. In this paper, the parameters of the preprocessing layer were initialized using the values of the high-pass filter in the SRM, and the threshold T of the TLU in this layer was set to 3. The experiments used a cross-entropy loss function, and the cross-entropy loss decreased continuously in the process of network training. The initial learning rate was 0.01, and the number of epochs was set to 200. As the training process progressed, the learning rate was changed to one-fifth of the original rate after a certain number of steps. The reduction in learning rate ensured that the loss was still effectively reduced rather than repeatedly oscillating in the later stage of training, thus further improving the accuracy.

4.4. Results

4.4.1. Verification of the Effectiveness of Separable Convolution and the Adversarial Mechanism

To investigate whether the introduced separable convolution and adversarial training can retain less information about the image content in the extracted features, we removed the separable convolution module and the GRL from the network structure in order to verify the performance of the network separately. We compared the networks without the introduction of separable convolution (labelled as Our method/wosep) and with the introduction of separable convolution (labelled as Our method/wisep); Table 1 shows the experimental results.

Table 1. Detection accuracy (%) for steganography at 0.2 bpp and 0.4 bpp on BOSSBase1.01 using S-UNIWARD for networks without and with the introduction of the separable convolution module.

Payload	Our Method/Wosep	Our Method/Wisep
0.2 bpp	71.4	76.2
0.4 bpp	84.3	88.7

We compared the networks without the introduction of the adversarial mechanism (labelled as Our method/woadv) and with the introduction of the adversarial mechanism (labelled as Our method/wiadv) on the same dataset and with the same hyperparameters; Table 2 shows the experimental results.

Table 2. Detection accuracy (%) for steganography at 0.2 bpp and 0.4 bpp on BOSSBase1.01 using S-UNIWARD for networks without and with the introduction of the adversarial mechanism.

Payload	Our Method/Woadv	Our Method/Wiadv
0.2 bpp	72.7	76.2
0.4 bpp	85.9	88.7

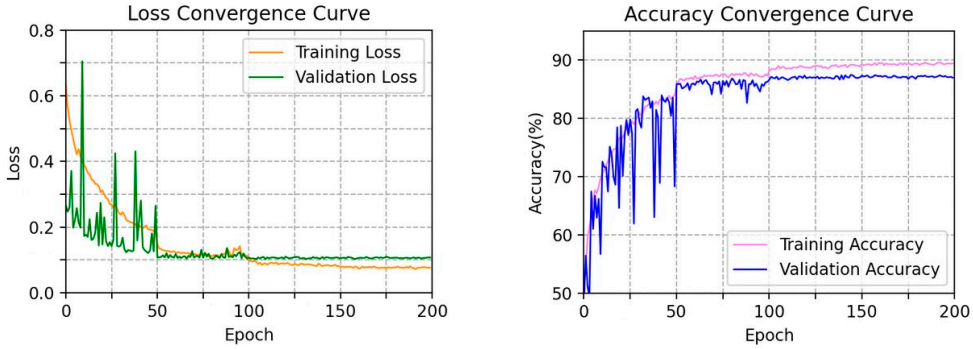
For this subsection, we experimentally verified the effectiveness of the introduced separable convolution and adversarial mechanism. Table 1 shows the performance comparison between the networks without the introduction of separable convolution and with the introduction of separable convolution. By observing the data in Table 1, the network with the introduction of separable convolution can obtain higher accuracy in steganographic detection at different payloads. Owing to the introduction of separable convolution, the accuracy of the network improves by 4.8% and 4.4% for S-UNIWARD at 0.2 bpp and 0.4 bpp, respectively. This indicates that separable convolution can maximize the residual information and extract more steganographic embedding features, thus improving the accuracy.

In addition, we also compared the results achieved by the networks without the introduction of the adversarial mechanism and with the introduction of the adversarial mechanism. As can be observed in Table 2, the network with the introduction of the adversarial mechanism outperforms that without it, improving the accuracy by 3.5% and 2.8% for S-UNIWARD at 0.2 bpp and 0.4 bpp, respectively. The above experimental results verify the effectiveness of introducing separable convolution and adversarial mechanism into the network structure.

4.4.2. Performance Comparison between this Method and other CNN-Based Steganalysis Methods

The experimental results reported in this section can be divided into two parts: The first part visualizes the training process of the proposed method via an accuracy and loss epoch chart. The second part compares the performance of the method proposed in this paper with other popular steganalysis methods. All of the experimental results are from the final iteration. When training and validating 256×256 images sourced from BOSSBase1.01 for S-UNIWARD at 0.4 bpp, our proposed network is capable of fast convergence; the detailed data are shown in Figure 4.

We trained the network on BOSSBase1.01 for 200 epochs—a process which took ~7 h. From the chart, we can observe that the loss and accuracy tended to stabilize around the 100th epoch. To prevent the network from overfitting, we stopped training at the 200th epoch. The loss curve drops obviously at the 50th epoch, which we believe is due to the learning rate decay strategy, which effectively reduces the loss and improves accuracy.



(**a**) (**b**)

Figure 4. (**a**) Loss convergence curve; (**b**) accuracy convergence curve.

The proposed method was compared with several common steganalysis networks, such as Ye-Net, Yedroudj-Net, SRNet, and Zhu-Net. Table 3 shows the experimental results. The proposed method achieves good results regardless of the embedding method and payload. Given that the network further introduces separable convolution and adversarial mechanism based on the foundation of the high-pass filter, it can better extract the steganographic embedding features and, thus, improve the accuracy of steganographic detection.

Table 3. Detection accuracy (%) of multiple CNN-based steganalysis methods against HILL, S-UNIWARD, and WOW at 0.2 bpp and 0.4 bpp.

Algorithms	HILL		S-UNIWARD		WOW	
	0.2 bpp	0.4 bpp	0.2 bpp	0.4 bpp	0.2 bpp	0.4 bpp
Ye-Net	58.5	66.9	59.3	68.4	67.1	76.9
Yedroudj-Net	61.8	72.2	63.7	77.2	72.6	85.8
SRNet	67.3	78.1	69.8	82.5	75.3	86.9
Zhu-Net	69.6	80.4	72.2	84.3	77.4	88.7
Our method	72.5	82.7	76.2	88.7	80.6	89.2

In Table 3, we further illustrate the detection accuracy of three common steganography methods—HILL, S-UNIWARD, and WOW—at payloads of 0.2 bpp and 0.4 bpp. Based on the data in Table 3, the network proposed in this paper obviously outperforms several other CNN-based steganalysis methods; it is 12.3–20.3% better than YeNet, 3.4–12.5% better than Yedroudj-Net, 2.3–6.4% better than SRNet, 0.5–4.4% better than Zhu-Net. For the WOW algorithm, the proposed method achieves an accuracy of 89.2% at 0.4bpp.

Briefly, these experimental results demonstrate well that the method proposed in this paper can extract the steganographic features more effectively, achieving higher accuracy than other networks. According to the results of the first part of the experiment, it is believed that the introduction of separable convolution and adversarial mechanism to the network contribute greatly to the superior performance of CNN-based steganalyzers over the other approaches. Note that the above experiments were operated without using the knowledge of channel awareness, a larger database, or a virtual augmentation of the database.

4.4.3. Impact of Data Expansion on Network Performance

In deep learning, it is significant to use a larger database to ensure a good performance, but also to avoid overtraining. Academics are prone to using large datasets to improve the performance of networks and to prevent overfitting. This part of the experiment expanded the dataset by adding 10,000 pairs of images from BOWS2 to BOSSBase1.01, for a total of 16,000 pairs of training set images containing 6000 pairs of images from BOSSBase1.01 and 10,000 pairs of images from BOWS2. The remaining images in BOSSBase1.01 were used as

the validation set and test set, and the enhanced dataset was noted as the extended BOSS. The network was trained using the above dataset to verify whether the expansion of the dataset could improve the accuracy of detecting steganographic images.

Table 4 shows the comparisons of Ye-Net, Yedroudj-Net, Zhu-Net, and our method trained on the original BOSSBase1.01 and the extended BOSS, against the steganography algorithm S-UNIWARD at payloads of 0.2 bpp and 0.4 bpp.

Table 4. Accuracy (%) of Ye-Net, Yedroudj-Net, Zhu-Net, and our proposed method for S-UNIWARD at 0.2 bpp and 0.4 bpp on the original BOSS dataset and the extended BOSS dataset.

Algorithms	BOSS	Extended BOSS	BOSS	Extended BOSS
	0.2 bpp		0.4 bpp	
Ye-Net	59.3	64.2	68.4	71.2
Yedroudj-Net	63.7	66.8	77.2	79.6
Zhu-Net	72.2	76.4	84.3	86.5
Our method	76.2	80.5	88.7	89.8

From the data in Table 4, we can observe that the detection performance of the network gradually improves as the training set is incremented. For all of the steganalysis algorithms involved in the experiments, better results were achieved using the extended BOSS compared to training with only BOSSBase1.01. Ye-Net, Yedroudj-Net, Zhu-Net, and our method improved accuracy by up to 4.9%, 3.1%, 4.2%, and 4.3%, respectively. Especially for S-UNIWARD at 0.4 bpp, using the extended dataset for training, our method achieved the best results in all of the experimental replicas, reaching 89.8%. Similarly, when attempting to detect steganography at a lower payload, the network trained with the extended BOSS also achieved the best performance. This prompted us to use larger datasets for training the network; as opposed to using the BOSS training set only, the extended BOSS can significantly improve the detection accuracy. During the experiments, we also found that using a larger training set was effective in mitigating overfitting.

5. Conclusions

Benefiting from the application of CNNs in the field of image steganalysis, traditional manually-defined features are slowly being replaced by features extracted automatically by CNNs. In this paper, we introduced separable convolution and adversarial mechanism into the traditional CNN structure, and proposed a new method for spatial image steganalysis, which can detect steganographic images well. The algorithm shows significant improvement over the current CNN-based methods. We attribute the improved performance of steganographic detection to the following factors: a set of high-pass filters in the preprocessing layer, a separable convolution module, and the introduction of adversarial mechanism. The separable convolution module eliminates image content information from the features and increases the signal-to-noise ratio; the introduced adversarial mechanism forces the feature extractor to extract more content information features, and isolates more useful steganographic embedding features. These mechanisms can extract more steganographic embedding features and improve the accuracy of steganographic detection. We also experimentally demonstrated that the network performance can be further improved by data expansion. Extensive experiments demonstrate that the method proposed in this paper significantly improves the detection accuracy compared with other steganalysis networks.

We hope that our method can provide some inspiration for future research in image steganalysis. Our future work will focus on utilizing the current foundation in conjunction with the backbone of more advanced networks, which will extract more valuable steganographic features for the steganalysis of color images.

Author Contributions: Conceptualization, Y.G. and T.Z.; methodology, Y.G.; software, Y.G.; validation, Y.G.; formal analysis, Y.G. and T.Z.; investigation, Y.G., T.Z. and H.L.; resources, Y.G. and T.Z.; data curation, Y.G.; writing—original draft preparation, Y.G.; writing—review and editing, T.Z., H.L., Q.J. and D.W.; visualization, Y.G.; supervision, T.Z., H.L. and Q.J.; project administration, T.Z.; funding acquisition, T.Z. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by (1) the National Natural Science Foundation of China, under grant number 62072057; (2) the Humanity and Social Science Youth Foundation of the Ministry of Education of China, under grant number 18YJCZH068; and (3) the Natural Science Foundation of the Jiangsu Higher Education Institutions of China, under grant number 18KJB520002.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Abd-El-Atty, B.; Iliyasu, A.M.; Alaskar, H.; El-Latif, A.; Ahmed, A. A robust quasi-quantum walks-based steganography protocol for secure transmission of images on cloud-based E-healthcare platforms. *Sensors* **2020**, *20*, 3108. [CrossRef] [PubMed]
2. Abd El-Latif, A.A.; Abd-El-Atty, B.; Elseuofi, S.; Khalifa, H.S.; Alghamdi, A.S.; Polat, K.; Amin, M. Secret images transfer in cloud system based on investigating quantum walks in steganography approaches. *Phys. A Stat. Mech. Appl.* **2020**, *541*, 123687. [CrossRef]
3. Fridrich, J.; Goljan, M.; Du, R. Reliable detection of LSB steganography in color and grayscale images. In Proceedings of the 2001 Workshop on Multimedia and Security: New Challenges, Ottawa, ON, Canada, 5 October 2001; pp. 27–30.
4. Fridrich, J.; Long, M. Steganalysis of LSB encoding in color images. In Proceedings of the IEEE International Conference on Multimedia and Expo, Virtual Conference, 5–9 July 2000; pp. 1279–1282.
5. Fridrich, J.; Goljan, M.; Hogea, D. Attacking the outguess. In Proceedings of the ACM Workshop on Multimedia and Security, Princeton, NJ, USA, 7–8 September 2002.
6. Pevny, T.; Bas, P.; Fridrich, J. Steganalysis by subtractive pixel adjacency matrix. *IEEE Trans. Inf. Forensics Secur.* **2010**, *5*, 215–224. [CrossRef]
7. Shi, Y.Q.; Chen, C.; Chen, W. A Markov process based approach to effective attacking JPEG steganography. In Proceedings of the International Workshop on Information Hiding, Alexandria, VA, USA, 10–12 July 2006; pp. 249–264.
8. Fridrich, J.; Kodovsky, J. Rich models for steganalysis of digital images. *IEEE Trans. Inf. Forensics Secur.* **2012**, *7*, 868–882. [CrossRef]
9. Chen, L.; Shi, Y.Q.; Sutthiwan, P.; Niu, X. A novel mapping scheme for steganalysis. In Proceedings of the International Workshop on Digital Watermarking, Shanghai, China, 31 October–3 November 2012; pp. 19–33.
10. Denemark, T.D.; Boroumand, M.; Fridrich, J. Steganalysis features for content-adaptive JPEG steganography. *IEEE Trans. Inf. Forensics Secur.* **2016**, *11*, 1736–1746. [CrossRef]
11. Denemark, T.; Sedighi, V.; Holub, V.; Cogranne, R.; Fridrich, J. Selection-channel-aware rich model for steganalysis of digital images. In Proceedings of the 2014 IEEE International Workshop on Information Forensics and Security (WIFS), Atlanta, GA, USA, 3–5 December 2014; pp. 48–53.
12. Qian, Y.; Dong, J.; Wang, W.; Tan, T. Feature learning for steganalysis using convolutional neural networks. *Multimed. Tools Appl.* **2018**, *77*, 19633–19657. [CrossRef]
13. Xu, G.; Wu, H.Z.; Shi, Y.Q. Structural design of convolutional neural networks for steganalysis. *IEEE Signal Process. Lett.* **2016**, *23*, 708–712. [CrossRef]
14. Ye, J.; Ni, J.; Yi, Y. Deep learning hierarchical representations for image steganalysis. *IEEE Trans. Inf. Forensics Secur.* **2017**, *12*, 2545–2557. [CrossRef]
15. Boroumand, M.; Chen, M.; Fridrich, J. Deep residual network for steganalysis of digital images. *IEEE Trans. Inf. Forensics Secur.* **2018**, *14*, 1181–1193. [CrossRef]
16. You, W.; Zhang, H.; Zhao, X. A Siamese CNN for image steganalysis. *IEEE Trans. Inf. Forensics Secur.* **2020**, *16*, 291–306. [CrossRef]
17. Tan, S.; Li, B. Stacked convolutional auto-encoders for steganalysis of digital images. In Proceedings of the Signal and information processing association annual summit and conference (APSIPA), 2014 Asia-Pacific, Chiang Mai, Thailand, 9–12 December 2014; pp. 1–4.
18. Xu, G. Deep convolutional neural network to detect J-UNIWARD. In Proceedings of the 5th ACM Workshop on Information Hiding and Multimedia Security, Philadelphia, PA, USA, 20–22 June 2017; pp. 67–73.
19. Yedroudj, M.; Comby, F.; Chaumont, M. Yedroudj-net: An efficient CNN for spatial steganalysis. In Proceedings of the 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Calgary, AB, Canada, 15–20 April 2018; pp. 2092–2096.
20. Li, B.; Wei, W.; Ferreira, A.; Tan, S. ReST-Net: Diverse activation modules and parallel subnets-based CNN for spatial image steganalysis. *IEEE Signal Process. Lett.* **2018**, *25*, 650–654. [CrossRef]
21. Zhang, R.; Zhu, F.; Liu, J.; Liu, G. Efficient feature learning and multi-size image steganalysis based on CNN. *arXiv* **2018**, arXiv:1807.11428.

22. Ganin, Y.; Ustinova, E.; Ajakan, H.; Germain, P.; Larochelle, H.; Laviolette, F.; Marchand, M.; Lempitsky, V. Domain-adversarial training of neural networks. *J. Mach. Learn. Res.* **2016**, *17*, 2096-2030.
23. Szegedy, C.; Liu, W.; Jia, Y.; Sermanet, P.; Reed, S.; Anguelov, D.; Erhan, D.; Vanhoucke, V.; Rabinovich, A. Going deeper with convolutions. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 1–9.
24. Chollet, F. Xception: Deep learning with depthwise separable convolutions. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 1251–1258.
25. Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; Bengio, Y. Generative adversarial nets. In Proceedings of the Advances in Neural Information Processing Systems, Montreal, QC, Canada, 8–13 December 2014; pp. 2672–2680.
26. Chen, P.; Guo, Y.; Li, G.; Wang, L.; Wan, J. Discriminative adversarial networks for specific emitter identification. *Electron. Lett.* **2020**, *56*, 438–441. [CrossRef]
27. Bas, P.; Filler, T.; Pevný, T. "Break our steganographic system": The ins and outs of organizing BOSS. In Proceedings of the International Workshop on Information Hiding, Prague, Czech Republic, 18–20 May 2011; pp. 59–70.
28. Bas, P.; Furon, T. BOWS-2 Contest (Break Our Watermarking System). Available online: http://bows2.ec-lille.fr/ (accessed on 6 June 2021).
29. Holub, V.; Fridrich, J.; Denemark, T. Universal distortion function for steganography in an arbitrary domain. *EURASIP J. Inf. Secur.* **2014**, *2014*, 1–13. [CrossRef]
30. Li, B.; Wang, M.; Huang, J.; Li, X. A new cost function for spatial image steganography. In Proceedings of the 2014 IEEE International Conference on Image Processing (ICIP), Paris, France, 27–30 October 2014; pp. 4206–4210.
31. Holub, V.; Fridrich, J. Designing steganographic distortion using directional filters. In Proceedings of the 2012 IEEE International workshop on information forensics and security (WIFS), Tenerife, Spain, 2–5 December 2012; pp. 234–239.
32. He, K.; Zhang, X.; Ren, S.; Sun, J. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In Proceedings of the IEEE International Conference on Computer Vision, Las Condens, Chile, 11–18 December 2015; pp. 1026–1034.