applied sciences

MDPI

Article

Deep Fake Image Detection Based on Pairwise Learning

Chih-Chung Hsu [1], Yi-Xiu Zhuang [1] and Chia-Yen Lee [2,*]

[1] Department of Management Information System, National Pingtung University of Science and Technology, 1, Shuefu Road, Neipu, Pingtung 91201, Taiwan; cchsu@mail.npust.edu.tw (C.-C.H.); shaun85528@gmail.com (Y.-X.Z.)

[2] Department of Electrical Engineering, National United University, 2, Lienda, Miaoli 36063, Taiwan

* Correspondence: olivelee@nuu.edu.tw

Received: 1 December 2019; Accepted: 30 December 2019; Published: 3 January 2020

Abstract: Generative adversarial networks (GANs) can be used to generate a photo-realistic image from a low-dimension random noise. Such a synthesized (fake) image with inappropriate content can be used on social media networks, which can cause severe problems. With the aim to successfully detect fake images, an effective and efficient image forgery detector is necessary. However, conventional image forgery detectors fail to recognize fake images generated by the GAN-based generator since these images are generated and manipulated from the source image. Therefore, in this paper, we propose a deep learning-based approach for detecting the fake images by using the contrastive loss. First, several state-of-the-art GANs are employed to generate the fake–real image pairs. Next, the reduced DenseNet is developed to a two-streamed network structure to allow pairwise information as the input. Then, the proposed common fake feature network is trained using the pairwise learning to distinguish the features between the fake and real images. Finally, a classification layer is concatenated to the proposed common fake feature network to detect whether the input image is fake or real. The experimental results demonstrated that the proposed method significantly outperformed other state-of-the-art fake image detectors.

Keywords: forgery detection; GAN; contrastive loss; deep learning; pairwise learning

1. Introduction

Recently, deep learning-based generative models, such as variational autoencoders and generative adversarial networks (GANs), have been widely used to synthesize the photo-realistic partial or whole content of an image or a video. Furthermore, recent modifications of the GANs, such as progressive growth of GANs (PGGAN) [1] and BigGAN [2], have been used to synthesize a highly photo-realistic image or video, which is impossible recognize as a fake by humans in a limited time. In general, the generative applications perform image translation tasks [3], which can cause serious problems if a fake image is improperly used on social media networks. For instance, the cycleGAN can be used to synthesize the fake face image in a pornography video [4]. Furthermore, the GANs can create a speech video with the synthesized facial content of any famous politician, causing severe problems to the society, political, and commercial activities. Therefore, an effective fake face image detection technique is urgently needed. In this paper, our previous study [5] is extended to recognize generated fake images effectively and efficiently.

In the traditional image forgery detection approaches, two types of forensics schemes are commonly used, active schemes and passive schemes. In the active schemes, an externally additive signal (i.e., watermark) is embedded in the source image without visual artifacts. To determine if an image is a tampered image, the watermark extraction process is performed on the target image to

restore the watermark [6]. The extracted watermark image can be used to detect tampered regions in the target image. However, there is no source image for the images generated by the GANs so that the active image forgery detector cannot extract the watermark image. On the other hand, the passive image forgery detectors use the statistical information on the source image that is high consistency between different images. As a result, the intrinsic statistical information can be used to detect the fake regions in the image [7,8]. The passive image forgery detectors cannot be used to identify fake images generated by the GANs because they are synthesized from the low-dimensional random vector. Specifically, the fake images generated by the GANs are not modified from their original images.

Since deep neural networks have been widely used in various recognition tasks, we can also adopt a deep neural network to detect fake images generated by the GANs. Recently, the deep learning-based approached for fake image detection using supervised learning has been studied. In other words, fake image detection has been treated as a binary classification problem (i.e., fake or real image). For instance, the convolution neural network (CNN) network was used to develop the fake image detector [9,10]. In [11], the performance of the fake face image detection was further improved by adopting the most advanced CNN–Xception network [12]. In [13], a manipulated face detection algorithm was proposed based on a hybrid ensemble learning approach. However, none of these studies has investigated the fully generated image, but instead, they have been focused only on partial manipulation of face images; thus, they cannot be used to detect the fully generated fake images.

Many GANs have been proposed in recent years. Some of the recently proposed GANs [1–3,14–18] have been used to produce photo-realistic images. To develop a fake image detector, it is necessary to collect all of the GAN's images as the training set for deep neural networks to achieve the promising performance. However, it is difficult and very time-consuming to collect the training samples generated by all the GANs. In addition, such a supervised learning strategy [9–11] tends to learn the discriminative features of fake images generated by all the GANs, and as a result, the learned (trained) detector may not have a good generalization ability. In other words, the learned detector will be unable to recognize the fake images generated by the GANs that were not included in the detector training process.

To meet the current requirement for the GANs-based generator of fake image detection, we propose a modified network structure, including a pairwise learning approach, called the common fake feature network (CFFN). By using the pairwise learning, the proposed structure overcomes the shortcomings of the supervised learning-based CNNs, such as those presented in [9,11]. To verify the effectiveness of the proposed method, the proposed deep fake detector (DeepFD) is applied to identify the fake face and generic images. The main contributions of this work are as follows.

- A fake face image detector based on the novel CFFN, consisting of an improved DenseNet backbone network and Siamese network architecture, is proposed.
- The cross-layer features are investigated by the proposed CFFN, which can be used to improve the performance.
- The pairwise learning approach is used to improve the generalization property of the proposed DeepFD.

The rest of this paper is organized as follows. Sections 2 and 3 introduce the proposed CFFN for fake image detection with the pairwise learning intended for the face and general images, respectively. Section 4 presents obtained experimental results of the fake face and general images. Finally, Section 4 gives the conclusions.

2. Fake Face Image Detection

The most serious challenge in the image and video forgery detection field is the fake face image detection. Fake face images can be used to create fake identities on social media networks, thus stealing personal information illegally. For instance, the fake image generator can be used to produce

images of celebrities with inappropriate content, which has hazardous consequences. In this section, the proposed deep learning framework with the pairwise learning strategy is introduced in detail.

The proposed two-step learning method that combines the CFF based on pairwise learning strategy and the classifier learning is presented in Figure 1. Introducing the supervised learning strategy in the fake face image detection the problems related to both difficult collection of training samples generated by all possible GANs and the need to retrain the fake face detector to obtain an effective model for the fake face images generated by a new GAN, are addressed. Specifically, to overcome these problems, the fake and real images are paired and follow by using the pairwise information to construct the contrastive loss to learn the discriminative common fake feature (CFF) by the proposed CFFN. Once the discriminative CFF is learned, the classification network captures the discriminative CFF to identify whether the image is real or fake. The details of the proposed method are described in the following.

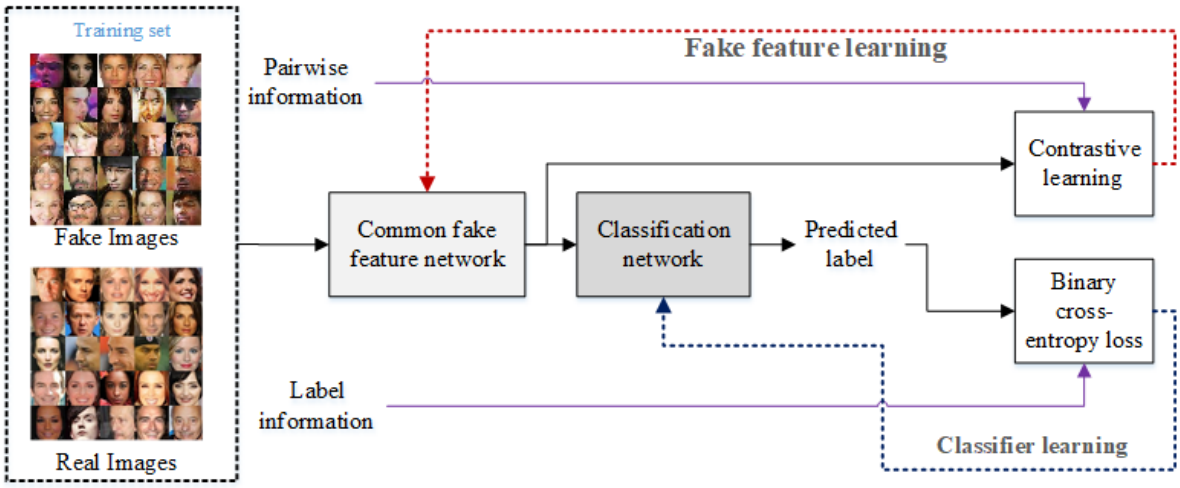


Figure 1. The flowchart of the proposed fake face detector based on the proposed common fake feature network with the two-step learning approach.

Let the set of the collected training images generated by M GANs be defined as: $\mathbf{X}_{fake} = [\mathbf{x}_{i=1}^{k=1}, \mathbf{x}_{i=2}^{k=1}, ..., \mathbf{x}_{i=N_1}^{k=1}, \mathbf{x}_{i=N_M}^{k=M}]$, where each GAN generates N_k training images. Let the training set consisted of real images be denoted as $\mathbf{X}_{real} = [\mathbf{x}_{i=1}, \mathbf{x}_{i=2}, ..., \mathbf{x}_{i=N_r}]$, containing N_r training images. Therefore, the total number of training images, including both real and fake images, will be $N_T = N_r + N_f = N_r + \sum_{k=1}^{M} N_k$. The label information set denoted as $\mathbf{Y} = [\mathbf{y}_1, \mathbf{y}_2, ..., \mathbf{y}_{N_t}]$ indicates whether an image is fake $(y = 0)$ or real $(y = 1)$. As stated previously, the pairwise information is necessary for the training stage so that the CFFN can learn the discriminative CFFs well. Toward this end, the pairwise information can be generated from the training set $\mathbf{X}$ and its corresponding label set $\mathbf{Y}$ by the permutation combination. Therefore, there are $C(N_T, 2)$ pairs $P = [p_{i=0,j=0}, p_{i=0,j=1}, ..., p_{i=0,j=N_r}, ..., p_{i=N_f,j=N_r}]$ generated from the training samples. In this paper, we set the total number of pairwise samples to $N_r = 2,000,000$.

2.1. Common Fake Feature Network

Many advanced CNN can be used to learn the fake features from the training set. Xception Network was used in [11] to capture the powerful feature from the training images in a purely supervised way. Other advanced CNNs, such as DenseNet [19], ResNet [20], Xception [12], can also be applied to the fake face detector training. However, most of these advanced CNNs are trained in a supervised way, so the classification performance depends on the training set. Rather than learn the fake features from all the GANs' images, we seek the CFF over different GANs. In this way, a suitable backbone network is needed for learning CFFs. However, the traditional CNNs (e.g., the DenseNet [19]) are not designed to learn the discriminative CFF. To overcome this shortcoming, we propose integrating

the Siamese network with the DenseNet [19], developing the CFFN to achieve the discriminative CFF learning.

A dense block is a basic component in the DenseNet [19], which is one of the state-of-the-art CNN models for image recognition. However, it is trained by the supervised learning strategy, while the proposed pairwise learning strategy for the CFFs denotes a semi-supervised learning strategy. The proposed CFFN is a two-streamed network designed to allow the pairwise input for CFF learning. On the other hand, the traditional CNNs, which are single-streamed networks, are unable to receive the paired information; thus, the common features can be difficultly learned by the traditional CNNs. In the proposed CFFN, the backbone network can be any of the advanced CNNs, such as ResNet [20], Xception [12], or DenseNet [19]. Once the backbone network is trained to have the best feature representation ability, the performance of the fake image recognition can be improved as well. To this end, DenseNet is selected as a backbone network of the proposed CFFN.

Moreover, it is well known that CNNs capture the hierarchical feature representation from a low level to a high level. In other words, the CNNs use only on high-level feature representation to identify whether the image is fake or not. However, the CFFs of fake face images may not exist only in the high-level representation but also in the middle-level feature representation. Inspired by [21], in this work, the cross-layer features are integrated into the classification layer to improve the fake image recognition performance, as shown in Figure 2.

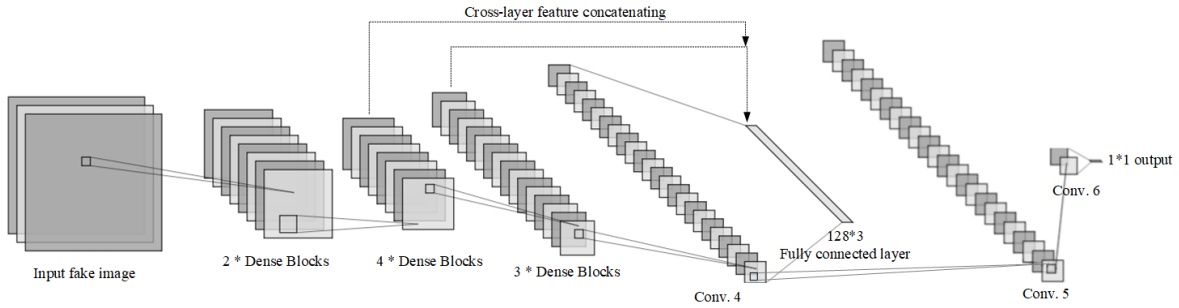


Figure 2. The structure of the proposed common fake feature network.

The proposed CFFN consists of three dense units that include two, four, and three dense blocks, respectively, and the number of channels in the three dense units are 48, 60, 78, and 126, respectively. The parameter θ in the transition layer is 0.5 and the growth rate is 24. Then, a convolution layer with 128 channels and 3×3 kernel size is concatenated to the output layer of the last dense unit. Finally, the fully connected layer is added to obtain the discriminative feature representation. To obtain the cross-layer feature representation, we also reshape the last layers of the first and second dense units to aggregate the cross-layer features into the fully connected layer. Therefore, in the final feature representation, there are $128 \times 3 = 384$ neurons.

In general, the classification of fake image can be performed by a different classification learning model, such as random forest, SVM, or Bayes classifier. However, the discriminative feature may be further improved by applying the back-propagation algorithm to the end-to-end structure. Therefore, in this work, the convolution and fully connected layers are concatenated into the last convolution layer of the proposed CFFN to obtain the final decision result. The details of the proposed CFFN are given in Table 1.

Table 1. Structure Parameters of the Proposed Common Fake Feature Network for Fake Face Image Detection.

Layer Number	Feature Learning	Classification
1	conv. layer, kernel = 7 × 7, stride = 4, #channels = 48	Conv. layer, kernel = 3 × 3, #channels = 2
2	Dense block × 2, #channels = 48	Global average pooling
3	Dense block × 3, #channels = 60	Fully connected layer, neurons = 2
4	Dense block × 4, #channels = 78	
5	Dense block × 2, #channels = 126	
6	Fully connected layer, neurons = 128	SoftMax layer

2.2. Discriminative Feature Learning

The main drawback of supervised learning is that it is hard to identify the subject that is excluded from the training phase. To enhance the performance of the proposed method, we introduce contrastive loss to learn the CFFs by pairwise learning. Therefore, the Siamese network structure [22] is used for allowing the pairwise inputs, as illustrated in Figure 3.

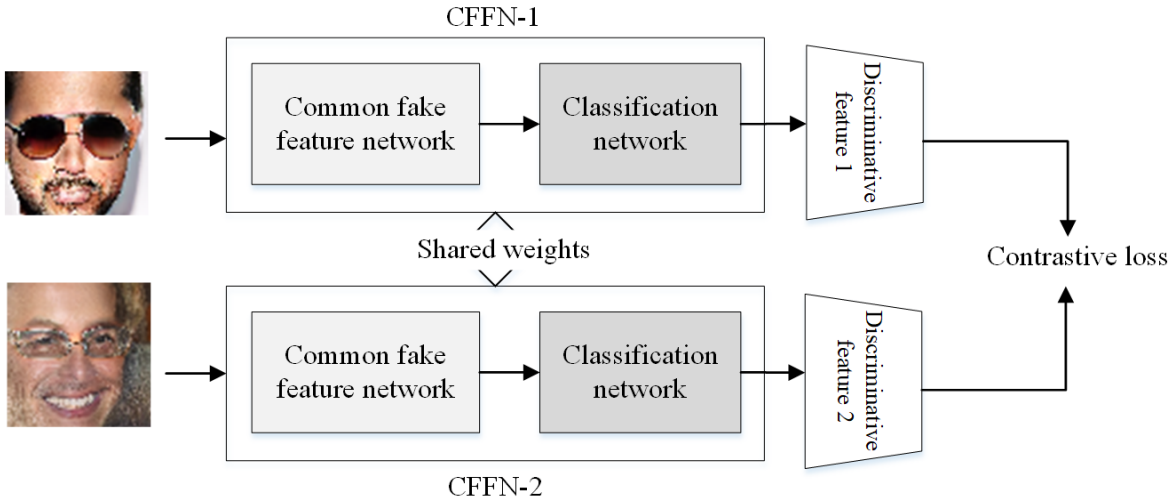


Figure 3. The proposed pairwise learning based on the Siamese network and contrastive loss.

With the aim to make the proposed CFFN learn the discriminative features during the training process, the contrastive loss term is incorporated into the energy function of the traditional loss function for supervised learning (i.e., the cross-entropy loss). Afterward, given the face image pair $(\mathbf{x}_1, \mathbf{x}_2)$ and the pairwise label $\mathbf{y}$, where $y = 0$ indicates an impostor pair, and $y = 1$ indicates a genuine pair, the energy function between two images is defined as:

$$E_W(\mathbf{x}_1, \mathbf{x}_2) = ||f_{CFFN}(\mathbf{x}_1) - f_{CFFN}(\mathbf{x}_2)||_2^2. \quad (1)$$

The most intuitive way to learn the discriminative features is to minimize the energy function E_W given by (1). Specifically, direct computation of $E_W(\mathbf{x}_1, \mathbf{x}_2)$ by calculating l_2 norm distance in the feature domain leads to a constant mapping, and the constant mapping makes any input to a constant vector such that the energy function E_W can be minimized. For instance, the learned mapping function can be $f_{CFFN}(\mathbf{x}_1) = f_{CFFN}(\mathbf{x}_2) = [1, 1, \ldots, 1]$. Thus, the constant mapping leads to useless feature representation. Therefore, to overcome this problem, the contrastive loss is introduced to learn the discriminative feature representation as well as to avoid constant mapping, which can be expressed as:

$$L(W, (P, \mathbf{x}_1, \mathbf{x}_2)) = 0.5 \times (y_{ij} E_w^2) + (1 - y_{ij}) \times max(0, (m - E_w)_2^2), \quad (2)$$

where m denotes the predefined threshold value. When the input is the genuine pair $y_{ij} = 1$, the cost function tends to minimize the energy (defined by the feature distance) E_W between two images. When the input is an impostor pair, the contrastive loss will minimize the function $max(0, (m - E_w)$. In other words, the energy E_W will be maximized if the feature distance between the impostor pair is smaller than the predefined threshold value m. In this way, it is possible to learn the common characteristics of the fake images generated by different GANs. When the contrastive loss is used, the feature representation $f_{CFFN}(\mathbf{x}_i)$ will tend to become similar to $f_{CFFN}(\mathbf{x}_j)$ at $y_{ij} = 1$ (i.e., for a fake-fake or real-real pair). By iteratively train the network f_{CFFN} using the contrastive loss, the CFFs of the collected GANs can be learned well.

2.3. Classification Learning

As stated previously, there are multiple existing classifiers for fake image detection. To improve the performance of the fake face image detection, we adopt a sub-network as a classifier. Thus, the classification learning can be quickly learned by the cross-entropy loss function, which is given by:

$$L_c(\mathbf{x}_i, p_i) = -\sum_{i}^{N_T}(f_{CLS}(f_{CFFN}(\mathbf{x}_i)) \log p_i), \tag{3}$$

where $p_i = 0$ indicates the real image, f_{CLS} denotes the classification sub-network consisting of a convolution layer with two channels, and a fully connected layer with two neurons. The classifier can be easily trained by the back-propagation algorithm [23]. One way to learn both the CFFs and classifier is the joint learning strategy incorporating the contrastive loss and cross-entropy loss into the total energy function. In another way, the CFFN is first trained by the proposed contrastive loss and follows by training the classifier based on cross-entropy loss. When the first strategy is applied, it is difficult to observe the impact of both contrastive and cross-entropy loss functions on the performance of the fake image detection tasks. Therefore, we adopt the second strategy to ensure the best performance of the proposed method. However, the first learning strategy is used as a baseline in the comparison, which will be presented in one of the following sections.

2.4. Two-Step Learning Policy

There are two loss functions, including contrastive loss and cross-entropy loss for the proposed CFFN and classifier learning in the proposed method, respectively. A joint learning policy can be adapted to optimize the proposed CFFN and classifier network based on two loss functions simultaneously. However, it is hard to determine the weighting values for two loss functions. In general, the weighting is determined empirically. Since the primary purpose of the proposed CFFN is the discriminative features learning, allowing that the CFF can be learned by minimizing the contrastive loss first. Afterward, any classifier can be used to recognize the fake face image based on the trained CFF. In this study, we adopt a small neural network as the classifier, enabling the capability of the end-to-end training. Moreover, it is well known that a classifier can be easily trained based on a better feature representation. Therefore, the CFFN is first trained based on contrastive loss, and then the classifier network is optimized by minimizing the cross-entropy loss. To verify whether the two-step learning policy is valid or not, we also conduct an experiment to compare the performance between the joint learning (i.e., Baseline-I) and the two-step learning policy in the experimental Section.

3. Fake General Image Detection

In contrast to the fake face image detection, the fake general image is more difficult to detect because the content of a general image varies significantly. Moreover, the fake feature of a general image is more complicated than that of a face image. Therefore, in this case, the more effective backbone network is required to be able to capture the CFFs of a general image, compared to the backbone network used in the fake face detection task. To this end, we increase the number of channels

in the proposed CFFN. As given in Table 2, the total number of the dense blocks in each dense unit is increased. The number of channels in each dense block is also increased to achieve better capturing of fake features of general images. Similarly, the contrastive loss and the classification sub-network presented in Section 2 are employed to detect whether an image is fake or real.

Table 2. Structure Parameters of the Proposed Common Fake Feature Network for Fake General Image Detection.

Layer Number	Feature Learning	Classification
1	conv. layer, kernel = 7 × 7, stride = 4, #channels = 48	Conv. layer, kernel = 3 × 3, #channels = 2
2	Dense block × 3, #channels = 60	Global average pooling
3	Dense block × 4, #channels = 78	Fully connected layer, neurons = 2
4	Dense block × 5, #channels = 99	
5	Dense block × 3, #channels = 171	
6	Fully connected layer, neurons = 128	SoftMax layer

4. Experimental Results

4.1. Fake Face Image Detection

4.1.1. Data Collection

The dataset used in the experiments was extracted from CelebA [24]. The images from the CelebA covered large pose variations and background clutter, including 10,177 of identities and 202,599 aligned face images. In the experiment, five state-of-the-art GANs were used to produce the training set of fake images based on the CelebA dataset, and they were as follows:

- DCGAN (Deep convolutional GAN) [15]
- WGAP (Wasserstein GAN) [16]
- WGAN-GP (WGAN with Gradient Penalty) [17]
- LSGAN (Least Squares GAN) [18]
- PGGAN [1]

By using the selected GANs, it was hard to synthesize realistic images with high-resolution, except for the PGGAN. In [15–18], the default size of the generated face images in the released source code was only 64 × 64 pixels. Specifically, if the size of the fake image was set to 128 × 128 pixels, many artifacts would be significant in the generated images, so the artifacts to recognize the fake image would be easily observed. In such a case, the fake image detector would not be needed. The most GANs could generate realistic fake images only of the smaller resolution, such as 64 × 64 pixels. To achieve a fair comparison of the performance of different image detectors in recognizing fake images generated by different GANs, the size of the input image was consistent, and it was set to 64 × 64 pixel. In the PGGAN, the best model released by the authors of the corresponding GAN was used. However, the PGGAN [1] can be used to generate the high-resolution fake face images, in which the size of the generated face image is different from the one used in our experiments. Therefore, in the experiments, we downsampled the fake face image generated by the PGGAN to 64 × 64. Note, the generated images are downloaded from the official website provided by the authors in PGGAN [1].

Each GAN randomly generated 40,000 fake images with the size of 64 × 64, which were recorded into the fake image pool. Since the fake image is generated by giving a random vector to GANs, the generated content will vary each time unless we set the random vector as a constant vector. To have a fair experiment result, we save the generated face images to a fake image pool to ensure the fake contents of the generated face images are consistent in each experiment. There were 200,000 fake images in total in the pool. We also randomly selected 200,000 real images from CelebA. Therefore, the total number of images, including real and fake images, was 400,000. To evaluate the performance of the proposed method, we split the image dataset into the training, validation, and test sets consisting

of 380,000, 10,000, and 10,000 images, respectively. In each set, the number of fake images was equal to the number of real images.

4.1.2. Experimental Settings

In the training of the proposed CFFN and fake face detector, we set the learning rate to $=1e-3$, and the total number of epochs to 15. The threshold value m of the contrastive loss was set to 0.5. Adam optimizer [25] was used for both the first- and second-step learning. The number of epochs of the first-step learning of the CFFN was set to 2, and the number of epochs of the classification learning of the classification sub-network was set to 13. The batch size was 88 for all the learning tasks. The parameters settings were as in [5,9,11].

In the experiments, we used the conventional image forgery method based on the sensor pattern noise [8] for performance comparison. The Baseline-I method was the jointly learning method based on the contrastive loss and binary cross-entropy loss without two-step learning. In the Baseline-II method, instead of the CFFN structure, the DenseNet [19] with two-step learning of the contrastive and binary cross-entropy loss functions was adopted.

We compared the performance regarding the precision P and recall R, which were defined as:

$$P = \frac{TP}{TP + FP} \tag{4}$$

$$R = \frac{TP}{TP + FN}, \tag{5}$$

where TP denoted the number of true positives, indicating that a real image was recognized as a real, FP denoted the number of false positives, denoting that a fake image was detected as a real image, and FN denoted the number of false negatives, showing that a real image was recognized as a fake one.

4.1.3. Objective Performance Comparison

To verify the effectiveness of the proposed method, we excluded one of the selected GANs from the training process and used it in the testing process instead to make the training and test sets be different. For instance, when the PGGAN was excluded from the training phase of the proposed DeepFD, the fake images generated by the PGGAN and the corresponding real images were used to evaluate the performance of the trained fake face detector. The objective performance comparison of the proposed fake face detector, two baseline methods, and methods proposed in [9–11] in terms of precision and recall, is presented in Table 3. As presented in Table 3, the proposed method significantly outperformed other state-of-the-art methods; thus, the CFFN can be used to capture the discriminative features of the fake images. The curves of the validation accuracy during the training phase are depicted in Figure 4. It demonstrated that the effectiveness of the proposed DeepFD. The proposed pairwise learning successfully captured the CFFs from the training images generated by different GANs. Thus, it was verified that the proposed method had higher generalization ability and effectiveness than the other methods.

Table 3. The objective performance comparison of the proposed and other fake face detectors.

Method/Target	WGAN-GP		DCGAN		WGAN		LSGAN		PGGAN	
	Precision	**Recall**	**Precision**	**Recall**	**Precision**	**Recall**	**Precision**	**Recall**	**Precision**	**Recall**
Method in [8]	0.322	0.373	0.334	0.349	0.371	0.391	0.350	0.396	0.345	0.378
Method in [9]	0.769	0.602	0.749	0.689	0.809	0.743	0.808	0.761	0.817	0.703
Method in [11]	0.792	0.684	0.820	0.811	0.864	0.881	0.848	0.869	0.868	0.853
Method in [10]	0.830	0.671	0.827	0.796	0.882	0.869	0.862	0.854	0.881	0.875
Method in [5]	0.832	0.690	0.871	0.847	0.885	0.920	0.866	0.898	0.922	0.909
Baseline-I	0.876	0.711	0.882	0.887	0.902	0.920	0.900	0.914	0.938	0.901
Baseline-II	0.901	0.728	0.822	0.838	0.864	0.881	0.920	0.919	0.917	0.887
The proposed	**0.986**	**0.751**	**0.929**	**0.916**	**0.988**	**0.927**	**0.947**	**0.986**	**0.988**	**0.948**

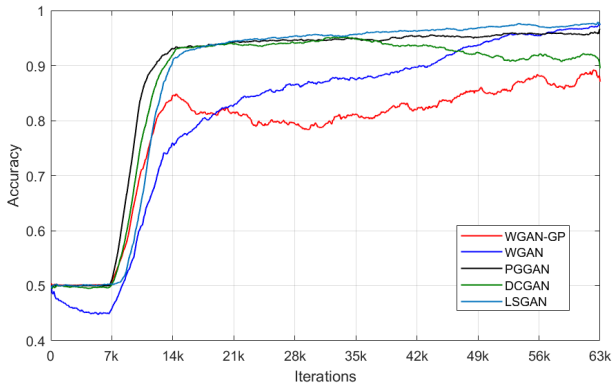


Figure 4. The curves of the validation accuracy during the training phase for the proposed DeepFD.

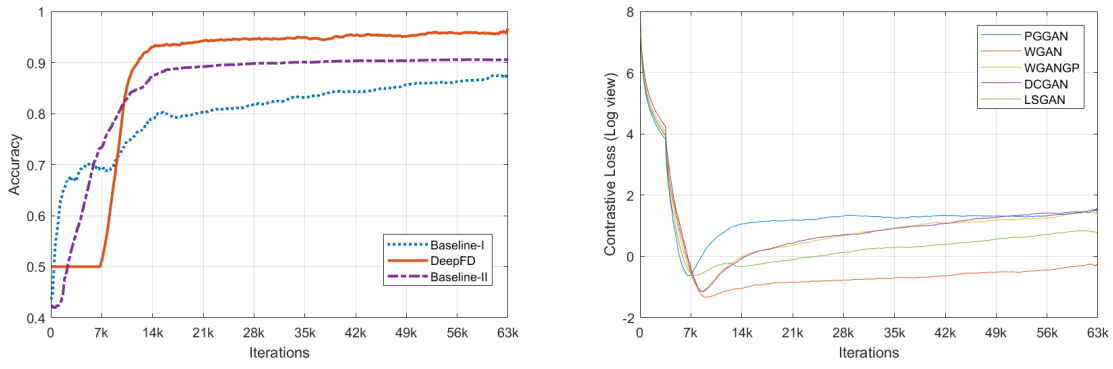
4.1.4. Visualized Result

As presented in [26], the object can be localized by designing the number of channels in the last convolution layers to be the same as the number of classes. Also, as suggested in [26], the channels of the last convolutional layer of the proposed CDNN enable the visualization ability. Therefore, the proposed model was used to visualize fake regions in the generated images by extracting the last convolution layer and mapping the responses to the image domain. Since the last convolution layer was designed to have two channels, the first channel was regarded as a feature response of the first class (i.e., real image), and the second channel corresponded to the second class (i.e., fake image). As a result, the proposed method could be used to visualize the fake regions, making a more intuitive interpretation of typical fake features generated by the GANs. Moreover, the heat map of the last convolutional layer is produced by the normalized response values in the feature map with the second channel (i.e., the feature map for fake feature). As a result in Figure 5, the higher feature response values can be observed in the corresponding regions with artifacts in the fake face images, while the real images have relatively lower feature response values. We map the response value to the original image to draw the artifact regions in red color.

Figure 5. The visualized fake feature map for localization of fake regions in face images generated by the (**a**–**e**) WGAN [16] and (**f**–**j**) PGGAN [1]. (**k**–**t**) are the real images and the corresponding feature responses.

4.1.5. Training Convergence

In the proposed method, it is necessary to guarantee the convergence of network training. Therefore, the convergence of the contrastive loss and CFFN training accuracy was analyzed, and the results are presented in Figure 6. In Figure 6a, the orange line depicts the accuracy curve during the training with supervised learning without contrastive loss, and the blue line indicates the validation accuracy curve of the training process with the proposed pairwise learning in two-step learning policy. It is clear that the proposed pairwise learning enhanced the method convergence significantly compared to the supervised learning strategy. On the other hand, it was necessary to ensure the contrastive loss converged as well. Figure 6b depicts the convergence of the proposed contrastive loss during the CFFN training. As the results show, the proposed contrastive loss was successfully decreased after only several iterations.



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

Figure 6. (**a**) The curves of the validation accuracy evaluated on the fake face images generated by PGGAN during the training phase for the proposed DeepFD, Baseline-I, and Baseline-II. (**b**) The curves of the contrastive loss values of the proposed DeepFD (in Log-view).

In general, over-fitting can be observed when the validation accuracy is dropping down as well as the training accuracy is still improving. In our experiments, the validation accuracy only slightly

improves after 21,000 iterations. It is clear that the validation accuracy in Figure 6a does not drop down after 21,000 iterations. Therefore, the total number of the training iterations can be higher than 21,000.

4.2. Fake General Image Detection

In the task of fake general image detection, we used three state-of-the-art GANs to generate high-quality fake images:

- BigGAN (Large Scale GAN Training for High Fidelity Natural Image Synthesis) [2]
- SA-GAN (Self-Attention GAN) [27]
- SN-GAN (Spectral Normalization GAN) [28]

The dataset was extracted from the ILSVRC12 [29]. We adopted the source code provided in [2,27,28] and its released model that was trained on the ILSVRC12 to generate the fake general images. Each GAN generated 100,000 fake images with a size of 128×128, which were recorded into the fake general image pool. Then, we randomly selected 300,000 real images from the ILSVRC12. Therefore, the total number of images was 600,000. To evaluate the performance of the proposed method, we split the image dataset into the training, validation, and test sets that consisted of 580,000, 10,000, and 10,000 images, respectively.

The objective performance comparison of the proposed method and other state-of-the-art image forgery detection methods is presented in Table 4. The results given in Table 4 show that the proposed CFFN with the pairwise learning strategy was significantly better than other state-of-the-art image forgery detectors. Compared to the supervised learning-based methods [9,11], the performance of the proposed method was significantly better. Accordingly, it was proven that the proposed method could learn the CFF of a fake general image.

Table 4. The subjective performance comparison of the proposed and other fake general image detectors.

Method/Target	BIGGAN		SA-GAN		SN-GAN	
	Precision	**Recall**	**Precision**	**Recall**	**Precision**	**Recall**
Method in [8]	0.358	0.409	0.430	0.509	0.354	0.424
Method in [9]	0.580	0.673	0.610	0.723	0.585	0.691
Method in [11]	0.650	0.737	0.682	0.762	0.653	0.691
Method in [5]	0.734	0.763	0.775	0.782	0.743	0.747
Baseline-I	0.769	0.789	0.787	0.811	0.798	0.791
Baseline-II	0.826	0.803	0.827	0.854	0.810	0.822
The proposed	**0.909**	**0.865**	**0.930**	**0.936**	**0.934**	**0.900**

4.3. Discussions and Limitations

In the proposed CFFN and DeepFD, the fake face and general image detection ability is provided using deep neural networks. Since the main contribution of this work is that the CFFs are learned from the pairwise training samples, the proposed CFFs may fail when the fake features of the results of a new generator are significantly different from most of those used in the training phase. In such a situation, the fake face and general image detector should be retrained. Another limitation of the proposed method is related to the collection of training samples. The technical details of some fake image generators maybe have not been revealed, so the training samples might be hard to collect in practice. In order to overcome this limitation, a few-shot learning policy should be employed in the learning of the CFF from a small-scale training set.

5. Conclusions

In this paper, a fake feature network-based pairwise learning is proposed to detect the fake face and general images generated by the state-of-the-art GANs. The proposed CFFN can be used to learn the middle- and high-level and discriminative fake features by aggregating the cross-layer feature representations. The proposed pairwise learning strategy enables the fake feature learning, which allows the trained fake image detector to have the ability to detect the fake image generated by a new GAN, even it was not included in the training phase. The experimental results demonstrated that the proposed method outperformed other state-of-the-art methods in terms of precision and recall rate. The fake video detection is also an important issue, so in our future work, we will extend the proposed method to fake video detection, incorporating the object detection and Siamese network structure.

Author Contributions: Conceptualization, C.-C.H.; Data curation, Y.-X.Z.; Formal analysis, C.-C.H., C.-Y.L.; Funding acquisition, C.-C.H.; Investigation, C.-C.H., C.-Y.L.; Methodology, C.-C.H.; Resources, Y.-X.Z.; Software, Y.-X.Z.; Supervision, C.-C.H., C.-Y.L.; Validation, Y.-X.Z.; Writing—original draft, C.-C.H. All authors have read and agreed to the published version of the manuscript.

Funding: This study were supported in part by the Ministry of Science and Technology, Taiwan, under Grants MOST 108-2634-F-007-009, 107-2218-E-020-002-MY3, and 107-2221-E-239-010-MY3.

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

Abbreviations

The following abbreviations are used in this manuscript:

CNN	Convolution neural network
CFF	Common fake feature
CFFN	Common fake feature network
DeepFD	Deep fake image detector
GAN	Generative adversarial nets

References

1. Karras, T.; Aila, T.; Laine, S.; Lehtinen, J. Progressive Growing of GANs for Improved Quality, Stability, and Variation. In Proceedings of the International Conference on Learning Representations, Vancouver, BC, Canada, 30 April–3 May 2018.
2. Brock, A.; Donahue, J.; Simonyan, K. Large scale GAN training for high fidelity natural image synthesis. In Proceedings of the International Conference on Learning Representations, New Orleans, LA, USA, 6–9 May 2019.
3. Zhu, J.; Park, T.; Isola, P.; Efros, A.A. Unpaired image-to-image translation using cycle-consistent adversarial networks. In Proceedings of the 2017 IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 22–29 October 2017; pp. 2242–2251. [CrossRef]
4. AI Can Now Create Fake Porn, Making Revenge Porn Even More Complicated. 2018. Available online: https://theconversation.com/ai-can-now-create-fake-porn-making-revenge-porn-even-more-complicated-92267 (accessed on 30 March 2019).
5. Hsu, C.; Lee, C.; Zhuang, Y. Learning to detect fake face images in the Wild. In Proceedings of the 2018 International Symposium on Computer, Consumer and Control (IS3C), Taichung, Taiwan, 6–8 December 2018; pp. 388–391. [CrossRef]
6. Chang, H.T.; Hsu, C.C.; Yeh, C.H.; Shen, D.F. Image authentication with tampering localization based on watermark embedding in wavelet domain. *Opt. Eng.* **2009**, *48*, 057002.
7. Hsu, C.C.; Hung, T.Y.; Lin, C.W.; Hsu, C.T. Video forgery detection using correlation of noise residue. In Proceedings of the IEEE Workshop on Multimedia Signal Processing, Cairns, Australia, 8–10 October 2008; pp. 170–174.
8. Farid, H. Image forgery detection. *IEEE Signal Process. Mag.* **2009**, *26*, 16–25. [CrossRef]

9. Huaxiao Mo, B.C.; Luo, W. Fake Faces Identification via Convolutional Neural Network. In Proceedings of the ACM Workshop on Information Hiding and Multimedia Security, Innsbruck, Austria, 20–22 June 2018; pp. 43–47.
10. Dang, L.; Hassan, S.; Im, S.; Lee, J.; Lee, S.; Moon, H. Deep learning based computer generated face identification using convolutional neural network. *Appl. Sci.* **2018**, *8*, 2610. [CrossRef]
11. Marra, F.; Gragnaniello, D.; Cozzolino, D.; Verdoliva, L. Detection of GAN-Generated Fake Images over Social Networks. In Proceedings of the IEEE Conference on Multimedia Information Processing and Retrieval, Miami, FL, USA, 10–12 April 2018, pp; 384–389. [CrossRef]
12. 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. 1610–02357.
13. Dang, L.M.; Hassan, S.I.; Im, S.; Moon, H. Face image manipulation detection based on a convolutional neural network. *Expert Syst. Appl.* **2019**, *129*, 156–168. [CrossRef]
14. 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.
15. Radford, A.; Metz, L.; Chintala, S. Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv* **2015**, arXiv:1511.06434.
16. Arjovsky, M.; Chintala, S.; Bottou, L. Wasserstein generative adversarial networks. In Proceedings of the International Conference on Machine Learning, Sydney, Australia, 6–11 August 2017; pp. 214–223.
17. Gulrajani, I.; Ahmed, F.; Arjovsky, M.; Dumoulin, V.; Courville, A.C. Improved training of wasserstein gans. In Proceedings of the Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, Long Beach, CA, USA, 4–9 December 2017; pp. 5767–5777.
18. Mao, X.; Li, Q.; Xie, H.; Lau, R.Y.; Wang, Z.; Smolley, S.P. Least squares generative adversarial networks. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 2813–2821.
19. Huang, G.; Liu, Z.; Van Der Maaten, L.; Weinberger, K. Densely connected convolutional networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 4700–4708.
20. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.
21. Liu, W.; Anguelov, D.; Erhan, D.; Szegedy, C.; Reed, S.; Fu, C.Y.; Berg, A.C. Ssd: Single shot multibox detector. In *European Conference on Computer Vision*; Springer: Cham, Switzerland, 2016; pp. 21–37.
22. Chopra, S.; Hadsell, R.; LeCun, Y. Learning a similarity metric discriminatively, with application to face verification. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, San Diego, CA, USA, 20–25 June 2005; Volume 1, pp. 539–546.
23. LeCun, Y.; Boser, B.E.; Denker, J.S.; Henderson, D.; Howard, R.E.; Hubbard, W.E.; Jackel, L.D. Handwritten digit recognition with a back-propagation network. In Proceedings of the Advances in Neural Information Processing Systems, Denver, CO, USA, 26–29 November 1990; pp. 396–404.
24. Liu, Z.; Luo, P.; Wang, X.; Tang, X. Deep Learning Face Attributes in the Wild. In Proceedings of the International Conference on Computer Vision, Santiago, Chile, 7–13 December 2015.
25. Sutskever, I.; Martens, J.; Dahl, G.; Hinton, G. On the importance of initialization and momentum in deep learning. In Proceedings of the International Conference on Machine Learning, Atlanta, GA, USA, 16–21 June 2013; pp. 1139–1147.
26. Oquab, M.; Bottou, L.; Laptev, I.; Sivic, J. Is object localization for free?-weakly-supervised learning with convolutional neural networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 685–694.
27. Zhang, H.; Goodfellow, I.; Metaxas, D.; Odena, A. Self-Attention generative adversarial networks. In *Proceedings of the 36th International Conference on Machine Learning*; Chaudhuri, K., Salakhutdinov, R., Eds.; PMLR: Long Beach, CA, USA, 2019; Volume 97, pp. 7354–7363.

28. Miyato, T.; Kataoka, T.; Koyama, M.; Yoshida, Y. Spectral normalization for generative adversarial networks. In Proceedings of the International Conference on Learning Representations, Vancouver, BC, Canada, 30 April–3 May 2018.
29. Russakovsky, O.; Deng, J.; Su, H.; Krause, J.; Satheesh, S.; Ma, S.; Huang, Z.; Karpathy, A.; Khosla, A.; Bernstein, M.; et al. ImageNet large scale visual recognition challenge. *Int. J. Comput. Vis. (IJCV)* **2015**, *115*, 211–252. [CrossRef]