Frequency-aware Discriminative Feature Learning Supervised by Single-Center Loss for Face Forgery Detection

Jiaming Li[1] Hongtao Xie[1*] Jiahong Li[2] Zhongyuan Wang[2] Yongdong Zhang[1]
[1]University of Science and Technology of China [2]Kuaishou Technology
ljmd@mail.ustc.edu.cn {htxie,zhyd73}@ustc.edu.cn {lijiahong,wangzhongyuan}@kuaishou.com

Abstract

Face forgery detection is raising ever-increasing interest in computer vision since facial manipulation technologies cause serious worries. Though recent works have reached sound achievements, there are still unignorable problems: a) learned features supervised by softmax loss are separable but not discriminative enough, since softmax loss does not explicitly encourage intra-class compactness and inter-class separability; and b) fixed filter banks and hand-crafted features are insufficient to capture forgery patterns of frequency from diverse inputs. To compensate for such limitations, a novel frequency-aware discriminative feature learning framework is proposed in this paper. Specifically, we design a novel single-center loss (SCL) that only compresses intra-class variations of natural faces while boosting inter-class differences in the embedding space. In such a case, the network can learn more discriminative features with less optimization difficulty. Besides, an adaptive frequency feature generation module is developed to mine frequency clues in a completely data-driven fashion. With the above two modules, the whole framework can learn more discriminative features in an end-to-end manner. Extensive experiments demonstrate the effectiveness and superiority of our framework on three versions of the FF++ dataset.

1. Introduction

Benefiting from the great progress made in deep learning, the Variational AutoEncoders [21, 35] and Generative Adversarial Networks based [15] face manipulation technology [38, 23, 44] enables ordinary people without professional skills and equipment to generate high-quality forged faces. Derived from that, certain free apps [2] and open-source projects [1, 3] quickly arise and gain popularity explosively. Unluckily, the technology may be abused for malicious purposes, causing severe trust issues in our society. Although digital forensics experts can analyze some influential videos for evidence of manipulation, they will be helpless in reviewing countless videos uploaded to the Internet every day. Thus, it is of high significance to develop efficient automatic detection algorithms.

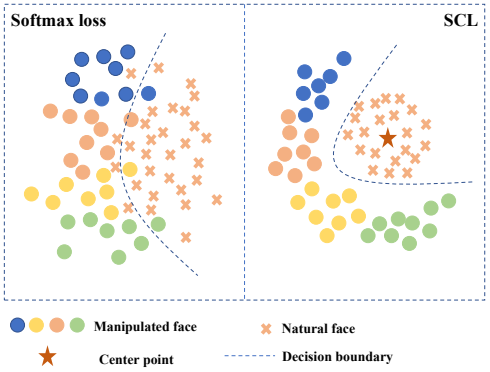


Figure 1. The feature distribution of samples in the embedding space. Left: learned features supervised by softmax loss are broadly separable but not discriminative enough, since the intra-class compactness and inter-class separability are not explicitly constrained. Right: our SCL only encourages the intra-class compactness of natural faces when constraining inter-class separability.

Towards such a concern, many methods have been proposed successively. Early research is keen on utilizing hand-crafted features or modifying the structure of existing neural networks[47, 5, 4, 34]. However, with remarkable progress made in facial synthesis technology [20, 46, 12], such methods have been unable to reliably detect face forgery. After that, the research mainstream is gradually turning to methods that introduce different information and prior knowledge into backbone networks [10, 33, 29]. For example, DeepRhythm [33] utilizes the minuscule periodic changes of skin color due to blood pumping through the face.

In essence, all current popular detection methods are using the powerful data fitting capability of neural networks to extract discriminative features for face forgery detection. And detection methods based on deep learning usually pose face forgery detection as a binary classification

*Corresponding author

problem and use softmax loss[1] to supervise the training of CNN networks. However, learned features supervised by softmax loss are not discriminative enough, since softmax loss does not explicitly encourage intra-class compactness and inter-class separability, as illustrated in the left of Figure 1. Recent work [22] has noticed this problem and attempted to utilize triplet loss [37] to extract discriminative features. However, regular metric learning methods usually indiscriminately encourage the intra-class compactness of natural and manipulated faces in the embedding space. Additionally, feature distributions of manipulated faces vary from one manipulation method to another considering various GAN fingerprints [48] and some unique operations, as shown in the left of Figure 1, making it nontrivial to aggregate all the manipulated faces. Therefore, constraining intra-class compactness of samples generated by varied manipulation methods usually leads to a sub-optimal solution because of optimization difficulty and even damages the performance owing to overfitting.

In addition, frequency-related cues are increasingly important for forgery detection since it's hard to find visual forgery clues. Although some studies [29, 9, 41, 13] have introduced frequency information and achieved remarkable results, their abilities to extract discriminative features are limited because of employing fixed filter banks and handcrafted features. These methods based on incomprehensive prior knowledge are insufficient to capture subtle forgery patterns from the frequency domain due to the diversity of background, gender, age, manipulation methods, *etc*.

With the above thoughts in mind, we propose a novel Frequency-aware Discriminative Feature Learning framework(FDFL). Explicitly, our framework mainly addresses two problems: a) how to adopt metric learning to learn more discriminative features for face forgery detection; and b) how to adaptively extract frequency-related features. Corresponding to the two problems, two sub-modules are developed: single-center loss (SCL) and adaptive frequency feature generation module (AFFGM), as shown in Figure 2. In specific, our single-center loss aims at only reducing intra-class variations of natural faces while increasing inter-class differences in the embedding space, as shown in the right of Figure 1. To this end, SCL minimizes the distance from representations of natural faces to the center point. Meanwhile, SCL encourages the distance from manipulated faces to the center point greater than from natural faces by at least a margin. Unlike regular metric learning methods, SCL does not restrict the intra-class compactness of manipulated faces, which agrees better with the characteristics of feature distribution of manipulated faces. Therefore, the network supervised by SCL can learn more discriminative features with less optimization difficulty. As for frequency-related features, we develop an AFFGM consisting of a special data preprocessing and adaptive frequency information mining block (AFIMB). The data preprocessing keeps the position relationship of image blocks in the spatial domain consistent with their position relationship in the frequency domain. In such a case, the preprocessed data is able to directly employ the existing convolution network. The AFIMB adaptively mines frequency clues in a data-driven fashion, which avoids utilizing too much incomprehensive prior knowledge. Compared to fixed filter banks and handcrafted features, AFFGM can capture forgery clues more flexibly in the frequency domain.

Extensive experiments demonstrate the effectiveness and superiority of our framework and we achieve state-of-the-art results on three versions of the FF++ dataset [36]. Our contributions can be summarized as follows:

- We propose a novel Frequency-aware discriminative feature learning framework which adopts metric learning and adaptive frequency features learning for face forgery detection.
- A single-center loss is designed to only compress intra-class variations of natural faces while boosting inter-class differences in the embedding space.
- An adaptive frequency feature generation module is developed to mine subtle artifacts from the frequency domain in a data-driven fashion.

2. Related work

With the development of neural networks and computer graphics, a new generation of face manipulation technology based on the Variational AutoEncoders [21, 35] and Generative Adversarial Networks [15] has been widely used. Correspondingly, face forgery detection has gradually become a research hotspot. In this section, we will briefly review previous works.

Face forgery detection Early works focus on utilizing hand-crafted features or modifying the structure of existing neural networks [47, 5, 19, 4, 34] to detect face forgery. Yang *et al*. [47] utilize the inconsistency of the head pose estimated from the central face and the whole face to identify manipulated faces. MesoNet [4] designs a shallow neural network that consists of two inception modules and two classic convolution layers. Though sound performances were achieved at that time, those methods are incapable of reliably detecting face forgery now due to the rapid development of face forgery technology. Especially when powerful general feature extractors like xception [7] are applied to forgery detection, the performance of early works is even more unsatisfactory. Therefore, the research mainstream is gradually turning to approaches which introduce different

[1]Following [27], we define the softmax loss as the combination of the last fully connected layer, softmax function, and cross-entropy loss.

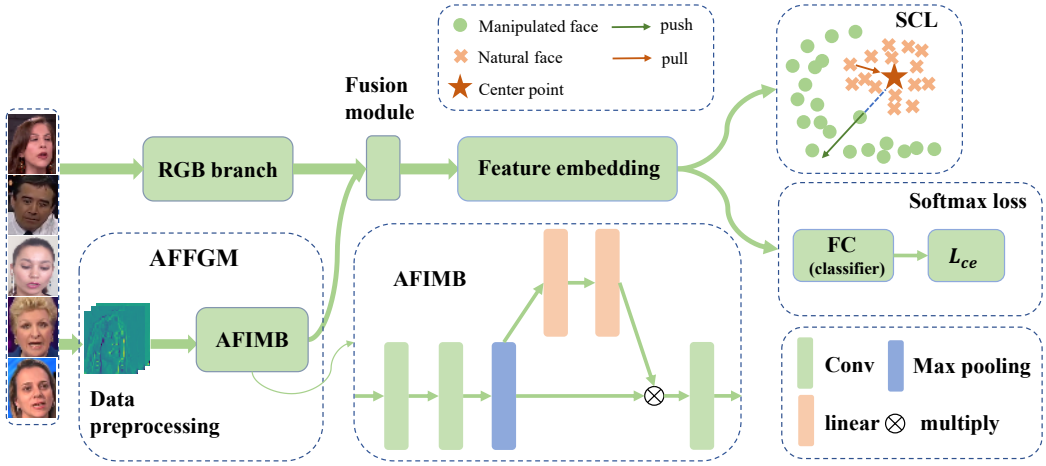


Figure 2. The Frequency-aware Discriminative Features Learning framework. AFFGM stands for the adaptive frequency feature generation module. AFIMB represents the adaptive frequency information mining block. FC represents the fully connected layer and L_{ce} represents the cross-entropy loss. The whole framework is trained end-to-end under the joint supervision of SCL and softmax loss.

information and prior knowledge into the backbone network to detect face forgery [10, 33, 29]. Dang *et al*. [10] introduce location information of manipulated regions to guide the network to focus on key regions. Qi *et al*. [33] exploit bioinformatics that skin color will present minuscule changes periodically due to blood pumping through the face. Face X-ray [24] innovatively uses self-generated data to train the network to locate blending boundaries, which greatly improves the generalization ability. Two-branch [29] utilizes fixed filter banks to extract frequency information, which limits the ability to extract discriminative features. In our work, we exploit a simple and effective module to adaptively mine frequency clues.

Metric learning Although metric learning has shown its advantages in face recognition [37] and person re-identification (re-ID) [17], learning discriminative features with deep metric learning for face forgery detection is more or less neglected. Center loss [43] and triplet loss [37] are the two most relevant metric learning methods to our work. Center loss [43] is designed to learn a center for features of each class and drive features of the same class closer to their corresponding center. Obviously, one disadvantage of center loss is that it ignores inter-class separability. Triplet loss [37] encourages features of data points with the same identity to get closer than those with different identities. However, triplet loss may suffer from the problem of time-consuming mining of hard triplets and dramatic data expansion. Kumar *et al*. [22] utilizes the network with the supervision of triplet loss to detect face forgery. But triplet loss performs poorly on the imagenet pre-trained backbone. Two-branch [29] proposes a novel loss which compresses the variability of natural faces and pushes away the manipulated faces. But its motivation comes from anomaly detection and the approach is very different from our SCL in many aspects. For example, our center point is updatable, while the center point of the two-branch is fixed. Additionally, two-branch constrains the absolute distance from all samples to the center point, whereas our SCL constrains the relative distance between natural and manipulated samples to the center point.

3. Proposed method

3.1. Overview

Aiming at solving the problems of previous methods in discriminative feature learning and frequency information mining, we propose a frequency-aware discriminative feature learning framework. As illustrated in Figure 2, our framework extracts features from the RGB domain and frequency domain at the same time and merges them in the early stage of the entire framework. After going through a feature embedding, high-level representations are obtained. At the end of the framework is a classifier that outputs the prediction results of input samples. The mining of frequency clues is achieved by our AFFGM (see Sec. 3.2). We fuse the frequency domain features and RGB domain features with a simple point-wise convolution block, which contributes to the reduction of parameters and computational expense. Finally, with the joint supervision of our single-center loss (see Sec. 3.3) and softmax loss, the network learns an embedding space where natural faces are clustered around the center point, while manipulated ones are far away from the center point.

3.2. Adaptive frequency features generation module

With the success in synthesizing realistic faces, it's harder to find visual forgery clues. But the discrepancy be-

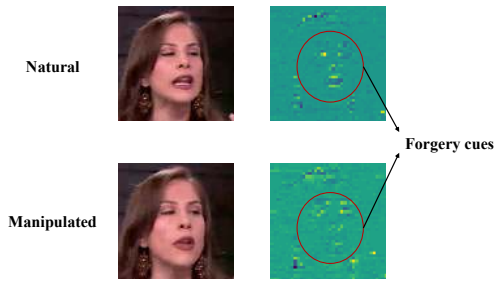


Figure 3. Inconsistency in the frequency domain could serve as an important forgery cue. The visualization of energy distribution in a certain frequency band is shown in the right column.

tween natural and manipulated faces in the frequency domain, especially in middle and high frequency bands, is pretty apparent as illustrated in Figure 3. Previous studies mostly use fixed filter banks or hand-crafted methods from other fields to extract frequency information [9, 13, 41, 29]. However, considering the diversity of background, gender, age, skin color, and especially manipulation methods, these methods based on incomprehensive prior knowledge are insufficient to capture forgery patterns of frequency. In order to tackle this problem, inspired by [16, 45, 31, 26, 42], we propose an adaptive frequency feature generation module (AFFGM) to efficiently mine subtle artifacts from the frequency domain. Our AFFGM consists of two parts: data preprocessing and adaptive frequency information mining block. Next, we will introduce them respectively.

Data preprocessing The pipeline of data preprocessing is shown in Figure 4. First, input RGB images are transformed into YCbCr color space. Next, the 2D DCT transformation is applied to each 8×8 block of images. It's worth noting that the two steps above are also widely used in current popular image compression standards, *e.g.*, JPEG. We think that will contribute to forgery detection from two aspects. On the one hand, the acceleration tools of existing compression algorithms can help improve the computational efficiency of our preprocessing. On the other hand, it will make our method more compatible with traces caused by compression. After that, the DCT-transformed coefficients from the same frequency band in all 8×8 blocks are grouped into a channel with their original position relationship retained. Therefore, the transformed images can directly exploit existing neural networks. Finally, all frequency channels are concatenated together to form one tensor. The shape of input images will change before and after preprocessing. Suppose the shape of the original input image is $H \times W \times 3$, then the shape of the input tensor becomes $H/8 \times W/8 \times 192$ after data preprocessing. Moreover, most energy of transformed images is concentrated on the low-frequency bands while the middle-frequency and high-frequency bands play

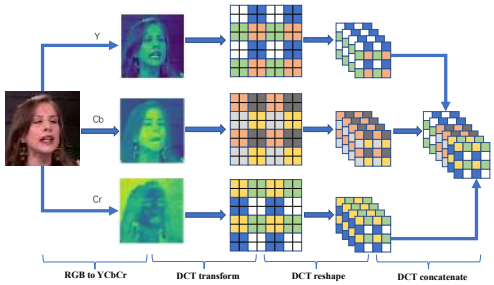


Figure 4. The pipeline of data preprocessing of AFFGM.

more significant roles in forgery detection. Therefore, every frequency channel is normalized by the mean and variance calculated from the training dataset.

Adaptive frequency information mining block Unlike previous methods, our AFFGM learns the frequency feature in a data-driven way, which avoids overly depending on incomprehensive prior knowledge. As illustrated in Figure 2, we empirically design a simple and effective network block to extract frequency features. In specific, the preprocessed data first passes through a layer of 3×3 convolution block with three groups[18]. That means the data from different channels of Y, Cb, Cr is processed separately. Then, it goes through an ordinary 3×3 convolution block and a max-pooling layer successively. In that process, information from different channels of Y, Cb, Cr interacts with each other. After that, we employ a channel attention block which consists of the aforementioned max-pooling layer and two linear layers for the sake of feature enhancement. Finally, an ordinary 1×1 convolution layer is used to further extract frequency-related features.

3.3. Single-center loss

Current face forgery detection methods based on deep learning usually use softmax loss to supervise network training. However, the learned features supervised by softmax loss are essentially not discriminative enough, since softmax loss only focuses on finding a decision boundary to separate different classes. The intra-class compactness and inter-class separability are not explicitly considered. Obviously, deep metric learning is a promising solution. However, most metric learning methods, such as triplet loss [37] and center loss [43], usually indiscriminately compress intra-class variations of natural and manipulated faces in embedding space. While feature distributions of manipulated faces vary from one manipulation method to another. That is because GAN fingerprints [48], manipulated region, and other unique operations, *e.g.*, post-processing techniques, lead to specific artifacts for each manipulation method. For example, Deepfakes [1] generates the whole face while NeuralTextures [39] only manipulates the mouth

region of the target person. Intuitively, their distribution in the embedding space should be evidently different. A side evidence would be that the generalization ability of features learned by supervised learning is significantly weakened on unseen manipulation methods. This implies that features learned by supervised learning are highly related to manipulation methods. The feature differences of samples generated by different manipulation methods make it difficult to aggregate all the manipulated faces. Therefore, indiscriminately constraining intra-class compactness in embedding space usually leads to a sub-optimal solution due to optimization difficulty and even damages the performance owing to overfitting. In order to solve this problem, we devise a novel single-center loss.

Definition As Figure 2 indicates that the goal of SCL is to minimize the distance from the representations of natural faces to the center point and to simultaneously push the representations of manipulated ones away from the center point. Let the given training dataset $(x^i, y^i)_{i=1}^N$ consist of N samples $x_i \in X$ with the associated labels $y_i \in \{0, 1\}$. And these samples are embedded into D-dimensional vectors with a neural network denoted by $f_\theta(\cdot)$. In our SCL, we just set the center point C of natural faces. For simplicity, we adopt f_i to represent $f(x^i)$ in the following paper. Similar to center loss, our method updates the parametric centers C at each iteration based on a mini-batch. Given a batch of training data, we define SCL as:

$$L_{sc} = M_{nat} + max(M_{nat} - M_{man} + m\sqrt{D}, 0) \quad (1)$$

where M_{nat} represents the mean Euclidean distance between representations of natural faces and the center point C in a batch. And M_{man} represents the mean Euclidean distance between representations of manipulated faces and center point C. Their functions are denoted as:

$$M_{nat} = \frac{1}{|\Omega_{nat}|} \sum_{i \in \Omega_{nat}} \parallel f_i - C \parallel_2 \quad (2)$$

$$M_{man} = \frac{1}{|\Omega_{man}|} \sum_{i \in \Omega_{man}} \parallel f_i - C \parallel_2 \quad (3)$$

where Ω_{nat} and Ω_{man} represent the representation sets of natural faces and manipulated faces respectively. As Eq. (1) shows, our SCL makes representations of natural faces aggregated around the center point. And it also pushes the distance from representations of manipulated faces to the center point greater than from natural faces by a margin. The Euclidean distance we employ is related to the arithmetic square root of feature dimension, and hence in order to set the hyperparameter easily, the margin is designed as m$\sqrt{D}$.

To compute the back-propagation gradients of the input feature embeddings and the center point, we assume there are s natural faces and t manipulated faces in a batch. And $y_i = 0$ and $y_i = 1$ represent i-th sample is a natural face and manipulated face respectively. The $\mathbb{1}[condition]$ is an indicator function which outputs 1 if the condition is satisfied and outputs 0 otherwise. For simplicity, we define

$$L = M_{nat} - M_{man} + m\sqrt{D}.$$

Then the derivatives of our SCL loss Eq. (1) with respect to the feature embedding of i-th sample $\frac{\partial L_{sc}}{\partial f_i}$ and center point $\frac{\partial L_{sc}}{\partial C}$ can be calculated as follows:

$$\frac{\partial L_{sc}}{\partial f_i} = \begin{cases} \frac{f_i - C}{s \cdot ||f_i - C||_2} \cdot (1 + \mathbb{1}[L > 0]), & y_i = 0; \\ -\frac{f_i - C}{t \cdot ||f_i - C||_2} \cdot \mathbb{1}[L > 0], & y_i = 1. \end{cases} \quad (4)$$

$$\frac{\partial L_{sc}}{\partial C} = -\frac{1}{s}\Big(\sum_{i \in \Omega_{nat}} \frac{f_i - C}{||f_i - C||_2}\Big) \cdot (1 + \mathbb{1}[L > 0]) + \frac{1}{t}\Big(\sum_{i \in \Omega_{man}} \frac{f_i - C}{||f_i - C||_2}\Big) \cdot \mathbb{1}[L > 0]. \quad (5)$$

The parametric center of SCL is randomly initialized and updated based on the mini-batches instead of the whole datasets, which will cause unstable training. Therefore, we introduce softmax loss with global information to guide the update of the center point. Moreover, softmax loss focuses on mapping the samples to discrete labels and our SCL aims to apply metric learning to the learned embeddings directly. Combining the two losses is beneficial to achieve more discriminative embeddings. The total loss can be written as:

$$L_{total} = L_{softmax} + \lambda L_{sc} \quad (6)$$

where λ is a hyper-parameter which controls the trade-off between the SCL and softmax loss.

4. Experiments

In this section, we first introduce the overall experiment setup and then present extensive experimental results to demonstrate the effectiveness and superiority of our approach.

4.1. Experimental setup

Dataset In order to facilitate comparison, our experiments are conducted on the FF++ [36] dataset. FF++ is a large-scale video dataset consisting of 1000 original videos that have been manipulated by four face manipulation methods: DeepFakes [1], Face2Face [40], FaceSwap[3], and NeuralTextures [39]. According to various compression factors, there are three versions of FF++ dataset: c0 (raw), c23 (light compression), and c40 (heavy compression). Our experiments are mainly conducted on the c40 version, the most

challenging case. As for dataset preprocessing, we sample 20 frames from each manipulated video and 80 frames from each original video. Compared to the setting of [36], the number of frames we use for training is pretty less. Besides, we utilize retinaface[11] to detect faces in each frame.

Evaluation metrics Following [29], we report video-level AUC score and pAUC [30] score by respectively averaging the AUC scores and pAUC scores of each frame in a video. pAUC is a global metric at a low false alarm rate. Given the significant class imbalance in the real world, pAUC can better reflect the performance of methods in the real world. To facilitate comparison with other methods, we also report the accuracy score. Besides, some visualizations (t-SNE [28]) are also reported to further evaluate the performance.

Implementation detail Our framework is implemented by PyTorch [32]. We use xception [7] pre-trained on imagenet, including the final fully connected layer, as our RGB branch and feature embedding of FDFL, which means D in Eq. (1) is equal to 1000. The fusion module is inserted between the entry flow and the middle flow of xception and the face forgery classifier is a simple FC layer with two nodes. The proposed modules, including the center point of SCL, are all initialized randomly. More details and hyper-parameters are provided in the supplementary material.

4.2. Ablation study

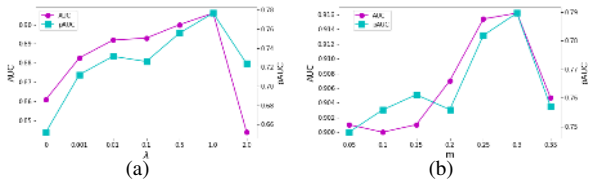
We perform the ablation study to analyze the effects of each component in FDFL, especially our SCL. All experiments are conducted on the challenging c40 version of the FF++ dataset.

4.2.1 SCL

In this section, we will show the relevant results of SCL experiments in detail to validate the effectiveness and superiority of SCL. We conduct all experiments including triplet loss and center loss only based on xception [7].

Parameter influence As indicated by the loss function in Eq. (6), the margin m and the weight λ may affect the final combination of the losses. Specifically, λ in Eq. (6) controls the trade-off between softmax loss and SCL loss. And m controls the relative distance between natural and manipulated faces to the center point in the embedding space. To study the impact of the two hyper-parameters, we present an empirical analysis on the c40 version of the FF++ dataset.

The influence of hyper-parameter λ is presented in Figure 5(a). The experimental results show that our SCL is quite robust to this parameter. For values from 0.001 to 1, the trained models consistently achieve promising results.



(a) (b)

Figure 5. The detection performances achieve by (a) varying λ when m is fixed as 0.1 and (b) varying m when λ is fixed as 0.5.

We assume that is because SCL and softmax loss are complementary losses. SCL focuses on feature representations directly, while softmax loss focuses on how to map feature representations into a discrete label space. What's more, the global information retained by softmax loss can guide the update of the center point of SCL. When λ is set to be 0, which means the model is trained by using only softmax loss, the performance is worst, only achieving an AUC of 0.861. But a $4\% \sim 6\%$ improvement of AUC could be reached by combining our SCL with softmax loss. In order to investigate the influence of m, we fix λ to be 0.5, and then take seven values from 0.05 to 0.35 at 0.05 intervals as m. It should be emphasized that m is a scale factor and the margin of the distance is proportional to the arithmetic square root of the dimension of the feature space, as shown in Eq. (1). As illustrated in Figure 5(b), our SCL can effectively improve the performance when m changes within a large range. When m is set to be 0.3 and λ to be 0.5, we get the best results, an AUC of 0.916 and a $\text{pAUC}_{0.1}$ of 0.790.

Comparison with other losses To validate the proposed SCL loss, we conduct additional experiments on various losses, including triplet loss with softmax loss and center loss with softmax loss. Similar to Eq. (6), both the weight of triplet loss and center loss are set as 0.01 and the margin of triplet loss is set as 0.3. As can be seen from Table 1, our SCL loss with softmax loss performs best among these losses, obtaining an AUC of 0.916 and a $\text{pAUC}_{0.1}$ of 0.790. In addition, both triplet loss with softmax loss and center loss with softmax loss improve subtly, compared to only using softmax loss.

Loss function	AUC	$\text{pAUC}_{0.1}$
softmax loss	0.860	0.652
center + softmax loss	0.868	0.666
triplet + softmax loss	0.863	0.655
SCL + softmax loss	**0.916**	**0.790**

Table 1. The performance of different losses on the c40 version of FF++.

Visualization of learned representations In order to explore the influence of different losses on feature distribution more thoroughly, we adopt t-SNE [28] to visualize features of the samples from the FF++ dataset. As is shown in

Figure 6, some properties can be observed: a) The learned features supervised by softmax loss appear as two clusters with neighboring boundaries. b) The triplet loss has little effect on feature distribution. We have tried to increase the weight of triplet loss, but in such a case the network cannot converge normally. c) The center loss significantly changes the distribution of features. However, constraining the intra-class compactness of manipulated faces leads to overfitting to some extent. Hence, the performance gain is very small. d) With the combination of SCL + softmax loss, the representations of natural faces are gathered compactly and separated from those of manipulated faces which are distributed less compactly.

Results analysis As shown in Table 1 and Figure 6, our SCL outperforms other losses, *i.e.*, softmax loss, center loss, and triplet loss. It is no wonder that softmax loss performs poorly since it only focuses on finding a decision boundary to separate different classes. As for triplet loss and center loss, though they explicitly consider intra-class compactness and inter-class separability, the results show that indiscriminately constraining intra-class compactness of natural and manipulated faces usually leads to a sub-optimal solution. This validates our analysis in Sec. 3.3 that different face manipulation methods will produce different forgery features due to GAN fingerprints [48] and some unique operations, making it nontrivial to aggregate all of the manipulated faces together. Compared to them, our SCL adopts an asymmetric optimization goal for natural and manipulated faces to learn discriminative features, which is more compatible with the feature distribution of samples.

4.2.2 Fusion module

We have studied the effects of different structures of the fusion module on performance and all experiments are conducted only with the supervision of softmax loss. As shown in Table 2, we explore concatenation, sum, and convolution block with different kernel sizes and group numbers. From the experimental results, We can see that: a) When a 1×1 convolution block is used as a fusion module and its group is set as 1, the performance reaches best in terms of AUC and $\text{pAUC}_{0.1}$; b) Even simple concatenation and sum operation can still achieve good performance. This fully reflects the effectiveness of our adaptive frequency feature generation module. In order to achieve the best results, we utilize the 1×1 convolution block, whose group is set to be 1, as the fusion module in all reference experiments.

4.2.3 The performance gain of each component

In order to evaluate the performance improvement from each component, we quantitatively evaluate our FDFL framework and its variants: 1) the baseline (xception); 2) FDFL w/o SCL; 3) FDFL w/o AFFGM. The quantitative results are listed in Table 3. It can be seen that both SCL and AFFGM can boost performance in terms of AUC and $\text{pAUC}_{0.1}$. Specifically, only with our SCL, AUC and $\text{pAUC}_{0.1}$ increased to 0.916 and 0.79 with an improvement of 5.5% and 13.8% individually. And AFFGM can also contribute an improvement of 4.5% in terms of AUC and 11.7% in terms of $\text{pAUC}_{0.1}$. These improvements prove the effectiveness of our SCL and AFFGM. In addition, when SCL and AFFGM are simultaneously integrated into the baseline to form the complete FDFL framework, an AUC of 0.924 and a $\text{pAUC}_{0.1}$ of 0.810 can be obtained. This fully validates the ability of our SCL to supervise the network to learn more discriminative features.

fusion module	AUC	$\text{pAUC}_{0.1}$
concatenation	0.892	0.727
sum	0.893	0.732
3×3, group=1	0.894	0.708
1×1, group=1	**0.906**	**0.769**
3×3, group=2	0.899	0.735
1×1, group=2	0.892	0.742

Table 2. The performance of different fusion modules on the c40 version of FF++.

SCL	AFFGM	AUC	$\text{pAUC}_{0.1}$
-	-	0.861	0.652
√	-	0.916	0.790
-	√	0.906	0.769
√	√	**0.924**	**0.810**

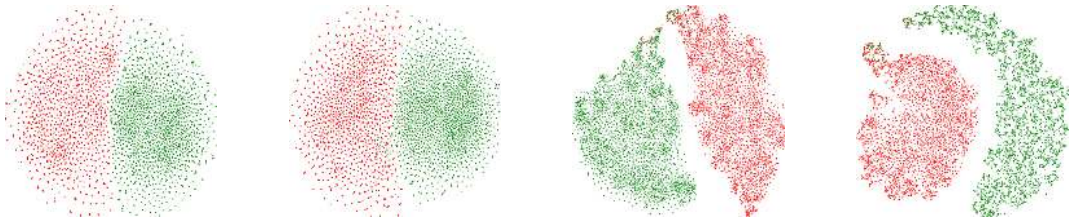
Table 3. The performance of different variants of FDFL on the c40 version of FF++.

4.3. Comparison with previous methods

We compare our approach with previous face forgery detection methods on the FF++ dataset. The results are listed in Table 4. Xception [36] and Face X-ray [24] are currently state-of-the-art image-based detection methods. Our method outperforms them on various versions of FF++ dataset in terms of AUC, $\text{pAUC}_{0.1}$, and accuracy. In the most challenging c40 version, we achieve a 6.4%, 15.8%, and 2.86% improvement in AUC, $\text{pAUC}_{0.1}$, and accuracy respectively. Although two-branch [29] is a video-based detection method, we still surpass it, especially in the c40 version. The results demonstrate the effectiveness and superiority of our framework.

5. Limitations

Although we achieve remarkable results on the FF++ dataset, there exist limitations of our framework. On the one hand, our framework lacks generalization ability on un-



(a) (b) (c) (d)

Figure 6. The visualization of features supervised by (a) softmax loss, (b) triplet + softmax loss, (c) center + softmax loss, (d) SCL + softmax loss. We randomly selected 5000 natural faces and manipulated ones respectively from the training dataset of the FF++ c40 version. Red dots represent natural faces and green dots represent manipulated faces (Best viewed in color).

Methods	c0			c23			c40		
	Acc	AUC	$pAUC_{0.1}$	Acc	AUC	$pAUC_{0.1}$	Acc	AUC	$pAUC_{0.1}$
Steg. Features + SVM[14]	97.63%	-	-	70.97%	-	-	55.98%	-	-
Cozzolino *et al.*[8]	98.57%	-	-	78.45%	-	-	58.69%	-	-
Bayar and Stamm[6]	98.75%	-	-	82.97%	-	-	66.84%	-	-
Rahmouni *et al.*[34]	97.03%	-	-	79.08%	-	-	61.18%	-	-
DSP-FWA[25]	-	-	-	-	0.575	0.516	-	0.623	0.519
MesoNet[4]	95.23%	-	-	83.10%	-	-	70.47%	-	-
Xception[36]	99.26%	-	-	95.73%	-	-	81.00%	-	-
Face X-ray[24]	-	0.988	-	-	0.874	-	-	0.616	-
Two-branch[29]	-	-	-	96.43%	0.991	0.984	86.34%	0.911	0.766
Xception[36] †	98.14%	0.997*	0.995	94.24%	0.972	0.903	86.14%	0.861	0.652
FDFL(our)	**99.43**%	**0.997***	**0.998**	**96.69**%	**0.993**	**0.985**	**89.00**%	**0.924**	**0.810**

Table 4. Quantitative results on the FF++ dataset with all three versions. c0 represents videos without compression, c23 represents videos with light compression, c40 represents videos with heavy compression and † represents the results of our baseline. Two-branch [29] is a video-based detection method and all others are image-based detection methods. The bold results are the best. The symbol * represents there is a difference at the fourth decimal place and more precise data are provided in the supplementary material.

seen manipulation methods (the results are provided in supplementary material). Our work and Face X-ray [24] imply that the discriminative features learned by supervised learning are highly related to manipulation methods. In our view, that is because forgery evidence is customized for specific manipulation methods due to GAN fingerprints [48] and some unique operations. If this explanation holds, not only our approach, but all approaches based on supervised learning will lack generalization ability on unseen manipulation methods. On the other hand, our framework ignores inter-frame information. Current face manipulation methods generally do not impose constraints on the temporal dimension. Therefore, the inconsistency between frames should be a valuable cue for video face forgery detection.

6. Conclusion

In this paper, we propose a novel frequency-aware discriminative feature learning framework that applies metric learning and adaptive frequency features learning to face forgery detection. Specifically, our single-center loss only compresses intra-class variations of natural faces when boosting inter-class separability in the embedding space. In such a case, the network can learn more discriminative features with less optimization difficulty. Besides, our adaptive frequency features generation module can effectively mine subtle artifacts from the frequency domain in a data-driven fashion, which avoids overly depending on incomprehensive prior knowledge. Extensive experiments demonstrate the effectiveness and superiority of our FDFL and we achieve state-of-the-art results on three versions of FF++ dataset.

In the future, we will explore how to effectively exploit inter-frame information and improve the generalization ability of detection methods by semi-supervised and unsupervised learning. In addition, it is worth studying the application of SCL in other fields, such as face anti-spoofing.

Acknowledgements This work is supported by the National Key Research and Development Program of China (2017YFC0820600), the National Nature Science Foundation of China (62022076, U1936210), the Youth Innovation Promotion Association Chinese Academy of Sciences (2017209), JD AI RESEARCH.

References

[1] Deepfakes. https://www.github.com/deepfakes/faceswap. Accessed: 2019-09-18. 1, 4, 5

[2] Faceapp. http://faceapp.com/app. Accessed: 2019-09-04. 1

[3] Faceswap. https://www.github.com/MarekKowalski/FaceSwap. Accessed: 2019-09-30. 1, 5

[4] Darius Afchar, Vincent Nozick, Junichi Yamagishi, and Isao Echizen. Mesonet: a compact facial video forgery detection network. In *2018 IEEE International Workshop on Information Forensics and Security (WIFS)*, pages 1–7. IEEE, 2018. 1, 2, 8

[5] Shruti Agarwal, Hany Farid, Yuming Gu, Mingming He, Koki Nagano, and Hao Li. Protecting world leaders against deep fakes. In *CVPR Workshops*, pages 38–45, 2019. 1, 2

[6] Belhassen Bayar and Matthew C Stamm. A deep learning approach to universal image manipulation detection using a new convolutional layer. In *Proceedings of the 4th ACM Workshop on Information Hiding and Multimedia Security*, pages 5–10, 2016. 8

[7] François Chollet. Xception: Deep learning with depthwise separable convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1251–1258, 2017. 2, 6

[8] Davide Cozzolino, Giovanni Poggi, and Luisa Verdoliva. Recasting residual-based local descriptors as convolutional neural networks: an application to image forgery detection. In *Proceedings of the 5th ACM Workshop on Information Hiding and Multimedia Security*, pages 159–164, 2017. 8

[9] Davide Cozzolino, Justus Thies, Andreas Rössler, Christian Riess, Matthias Nießner, and Luisa Verdoliva. Forensictransfer: Weakly-supervised domain adaptation for forgery detection. *arXiv preprint arXiv:1812.02510*, 2018. 2, 4

[10] Hao Dang, Feng Liu, Joel Stehouwer, Xiaoming Liu, and Anil K Jain. On the detection of digital face manipulation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5781–5790, 2020. 1, 3

[11] Jiankang Deng, Jia Guo, Evangelos Ververas, Irene Kotsia, and Stefanos Zafeiriou. Retinaface: Single-shot multi-level face localisation in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5203–5212, 2020. 6

[12] Yu Deng, Jiaolong Yang, Dong Chen, Fang Wen, and Xin Tong. Disentangled and controllable face image generation via 3d imitative-contrastive learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5154–5163, 2020. 1

[13] Ricard Durall, Margret Keuper, Franz-Josef Pfreundt, and Janis Keuper. Unmasking deepfakes with simple features. *arXiv preprint arXiv:1911.00686*, 2019. 2, 4

[14] Jessica Fridrich and Jan Kodovsky. Rich models for steganalysis of digital images. *IEEE Transactions on Information Forensics and Security*, 7(3):868–882, 2012. 8

[15] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680, 2014. 1, 2

[16] Lionel Gueguen, Alex Sergeev, Ben Kadlec, Rosanne Liu, and Jason Yosinski. Faster neural networks straight from jpeg. In *Advances in Neural Information Processing Systems*, pages 3933–3944, 2018. 4

[17] Alexander Hermans, Lucas Beyer, and Bastian Leibe. In defense of the triplet loss for person re-identification. *arXiv preprint arXiv:1703.07737*, 2017. 3

[18] Yani Ioannou, Duncan Robertson, Roberto Cipolla, and Antonio Criminisi. Deep roots: Improving cnn efficiency with hierarchical filter groups. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1231–1240, 2017. 4

[19] Tackhyun Jung, Sangwon Kim, and Keecheon Kim. Deepvision: Deepfakes detection using human eye blinking pattern. *IEEE Access*, 8:83144–83154, 2020. 2

[20] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4401–4410, 2019. 1

[21] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013. 1, 2

[22] Akash Kumar, Arnav Bhavsar, and Rajesh Verma. Detecting deepfakes with metric learning. In *2020 8th International Workshop on Biometrics and Forensics (IWBF)*, pages 1–6. IEEE, 2020. 2, 3

[23] Cheng-Han Lee, Ziwei Liu, Lingyun Wu, and Ping Luo. Maskgan: Towards diverse and interactive facial image manipulation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5549–5558, 2020. 1

[24] Lingzhi Li, Jianmin Bao, Ting Zhang, Hao Yang, Dong Chen, Fang Wen, and Baining Guo. Face x-ray for more general face forgery detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5001–5010, 2020. 3, 7, 8

[25] Yuezun Li and Siwei Lyu. Exposing deepfake videos by detecting face warping artifacts. *arXiv preprint arXiv:1811.00656*, 2018. 8

[26] Chuanbin Liu, Hongtao Xie, Zhengjun Zha, Lingyun Yu, Zhineng Chen, and Yongdong Zhang. Bidirectional attention-recognition model for fine-grained object classification. *IEEE Transactions on Multimedia*, 22(7):1785–1795, 2019. 4

[27] Weiyang Liu, Yandong Wen, Zhiding Yu, and Meng Yang. Large-margin softmax loss for convolutional neural networks. In *ICML*, volume 2, page 7, 2016. 2

[28] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(Nov):2579–2605, 2008. 6

[29] Iacopo Masi, Aditya Killekar, Royston Marian Mascarenhas, Shenoy Pratik Gurudatt, and Wael AbdAlmageed. Two-branch recurrent network for isolating deepfakes in videos. *arXiv preprint arXiv:2008.03412*, 2020. 1, 2, 3, 4, 6, 7, 8

[30] Donna Katzman McClish. Analyzing a portion of the roc curve. *Medical Decision Making*, 9(3):190–195, 1989. 6

[31] Shaobo Min, Hantao Yao, Hongtao Xie, Zheng-Jun Zha, and Yongdong Zhang. Multi-objective matrix normalization for fine-grained visual recognition. *IEEE Transactions on Image Processing*, 29:4996–5009, 2020. 4

[32] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. In *Advances in neural information processing systems*, pages 8026–8037, 2019. 6

[33] Hua Qi, Qing Guo, Felix Juefei-Xu, Xiaofei Xie, Lei Ma, Wei Feng, Yang Liu, and Jianjun Zhao. Deeprhythm: Exposing deepfakes with attentional visual heartbeat rhythms. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 4318–4327, 2020. 1, 3

[34] Nicolas Rahmouni, Vincent Nozick, Junichi Yamagishi, and Isao Echizen. Distinguishing computer graphics from natural images using convolution neural networks. In *2017 IEEE Workshop on Information Forensics and Security (WIFS)*, pages 1–6. IEEE, 2017. 1, 2, 8

[35] Danilo Jimenez Rezende, Shakir Mohamed, and Daan Wierstra. Stochastic backpropagation and approximate inference in deep generative models. *arXiv preprint arXiv:1401.4082*, 2014. 1, 2

[36] Andreas Rossler, Davide Cozzolino, Luisa Verdoliva, Christian Riess, Justus Thies, and Matthias Nießner. Faceforensics++: Learning to detect manipulated facial images. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1–11, 2019. 2, 5, 6, 7, 8

[37] Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 815–823, 2015. 2, 3, 4

[38] Yujun Shen, Jinjin Gu, Xiaoou Tang, and Bolei Zhou. Interpreting the latent space of gans for semantic face editing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9243–9252, 2020. 1

[39] Justus Thies, Michael Zollhöfer, and Matthias Nießner. Deferred neural rendering: Image synthesis using neural textures. *ACM Transactions on Graphics (TOG)*, 38(4):1–12, 2019. 4, 5

[40] Justus Thies, Michael Zollhofer, Marc Stamminger, Christian Theobalt, and Matthias Nießner. Face2face: Real-time face capture and reenactment of rgb videos. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2387–2395, 2016. 5

[41] Sheng-Yu Wang, Oliver Wang, Richard Zhang, Andrew Owens, and Alexei A Efros. Cnn-generated images are surprisingly easy to spot... for now. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, volume 7, 2020. 2, 4

[42] Yuxin Wang, Hongtao Xie, Zheng-Jun Zha, Mengting Xing, Zilong Fu, and Yongdong Zhang. Contournet: Taking a further step toward accurate arbitrary-shaped scene text detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11753–11762, 2020. 4

[43] Yandong Wen, Kaipeng Zhang, Zhifeng Li, and Yu Qiao. A discriminative feature learning approach for deep face recognition. In *European conference on computer vision*, pages 499–515. Springer, 2016. 3, 4

[44] Rongliang Wu, Gongjie Zhang, Shijian Lu, and Tao Chen. Cascade ef-gan: Progressive facial expression editing with local focuses. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5021–5030, 2020. 1

[45] Kai Xu, Minghai Qin, Fei Sun, Yuhao Wang, Yen-Kuang Chen, and Fengbo Ren. Learning in the frequency domain. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020. 4

[46] Chao Yang and Ser-Nam Lim. One-shot domain adaptation for face generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5921–5930, 2020. 1

[47] Xin Yang, Yuezun Li, and Siwei Lyu. Exposing deep fakes using inconsistent head poses. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 8261–8265. IEEE, 2019. 1, 2

[48] Ning Yu, Larry S Davis, and Mario Fritz. Attributing fake images to gans: Learning and analyzing gan fingerprints. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 7556–7566, 2019. 2, 4, 7, 8