CONCENTRIC RING LOSS FOR FACE FORGERY DETECTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Due to growing societal concerns about indistinguishable deepfake images, face forgery detection has received an increasing amount of interest in computer vision. Since the differences between actual and fake images are frequently small, improving the discriminative ability of learned features is one of the primary problems in deepfake detection. In this paper, we propose a novel Concentric Ring Loss (CRL) to encourage the model to learn intra-class compressed and inter-class separated features. Specifically, we independently add margin penalties in angular and Euclidean space to force a more significant margin between real and fake images, and hence encourage better discriminating performance. Compared to softmax loss, CRL explicitly encourages intra-class compactness and inter-class separability. Extensive experiments demonstrate the superiority of our methods over multiple datasets. We show that CRL consistently outperforms the state-of-the-art by a large margin.

1 INTRODUCTION

With remarkable progress made in face manipulation techniques Pumarola et al. (2018); Deepfakes (2020); Kowalski (2020), we are able to synthesize realistic deepfake images that reach an impressive quality level and are difficult to distinguish by a human. As the quality of forgery images reaches a higher level, the difference between real and fake images becomes more subtle. These forged images may be maliciously abused, leading to serious security and ethical issues. Therefore, it is of great significance to develop efficient and effective methods for automatic face forgery detection.

Many approaches have been proposed to tackle this issue. Using hand-crafted features or modifying the structure of existing networks were popular solutions in earlier studies Rahmouni et al. (2017); Afchar et al. (2018); Li & Lyu (2018). Later, more works attempted and succeeded in introducing other types of information (*i.e.*, frequency information Frank et al. (2020); Zhang et al. (2019); Durall et al. (2020); Dong et al. (2022), 3d geometry, and phase spectrum of frequency) and prior knowledge into the backbone network to boost performance Masi et al. (2020); Liu et al. (2021); Zhu et al. (2021). However, since the differences between real and fake images are usually too subtle to be identified, one of the urgent problems in deep forgery detection is to enhance the discriminative power of the learned features.

Most of the recent works on face forgery detection Qian et al. (2020); Li et al. (2020a); Liu et al. (2021); Zhao et al. (2021) use the traditional loss function (*i.e.*, softmax loss) adopted from classification. We argue that features learned by softmax loss are not discriminative enough to distinguish such subtle differences, which is also mentioned in Li et al. (2021). Some researchers Masi et al. (2020); Li et al. (2021) are aware of the problem and made efforts toward discriminative representation learning. They tried to combine softmax loss with pair-based loss function (*e.g.*, center loss or triplet loss) to increase intra-class compactness and inter-class separability. However, most current methods lack transferability and perform poorly on unseen manipulation methods.

To solve the problem, in this paper, we proposed a novel Concentric Ring Loss (CRL) to further improve the discriminative power and generalization ability of the forgery detection model. In order to maximize class separability of real and fake images, we fully explore the feature information in the angular and Euclidean space. To this end, we decouple the magnitude and direction information of embedding vectors, and independently add margin penalties in angular and Euclidean space to force a larger margin between real and fake images. As shown in Figure 1, the proposed CRL

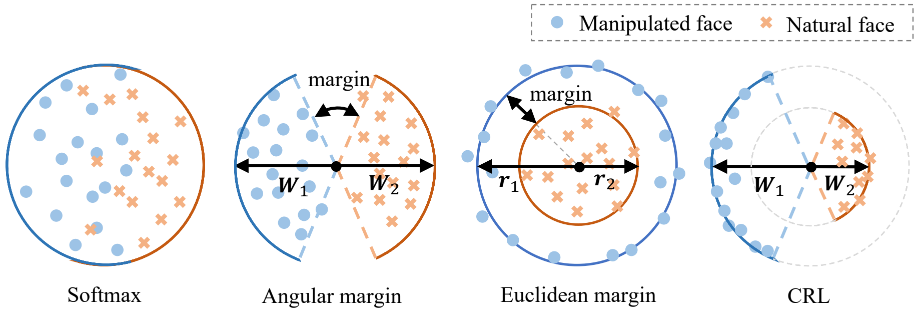

Figure 1: The feature distribution of samples in the embedding space. Features learned by softmax loss are separable but not discriminative enough. Angular margin loss maximizes the decision margin in the angular space. Euclidean margin loss reduces the distance between natural face embeddings and the center point while simultaneously pushing manipulated face embeddings away from the center point. The proposed CRL loss integrates both angular and Euclidean margin loss to further increase the discriminative power of the model. Note that our CRL only encourages the intra-class compactness of natural faces in order to enhance the model's generalization ability.

consists of an angular margin loss and a Euclidean margin loss. Both losses encourage the inter-class separability and intra-class compactness of the learned features. Joint supervision of angular and Euclidean margin loss further boosts the discriminative power and generalization ability of the model and stabilized the training process. With the supervision of CRL, the network learns an embedding space where natural faces are clustered around one side of the concentric ring space with a smaller radius, while manipulated ones are clustered around the other side of the concentric ring space with a larger radius (Figure 1). Furthermore, a frequency-aware feature learning framework is proposed to exploit high-frequency features and further improve the generalization ability of the model. We extract features from both RGB and frequency domains using a two-branch framework. Then these features are further fused to provide richer forgery clues for classification.

In summary, we make the following contributions:

1. A novel CRL is proposed to further improve the discriminative power and generalization ability of the forgery detection model. Margin penalties in angular and Euclidean space are applied independently to force a more significant margin between real and fake images.
2. We propose a two-branch framework to exploit both low- and high-frequency features and fused them to provide richer forgery clues for classification.
3. We claim state-of-the-art and provide extensive analysis to study individual contributions of angular and Euclidean margin loss, as well as the contributions of high-frequency features.

2 Related Work

Face forgery detection. Most existing methods treat face forgery detection as a universal binary classification problem. Some previous works introduce different types of information (*i.e.*, frequency information, 3d geometry, and phase spectrum of frequency) and prior knowledge into the network to boost classification performance Masi et al. (2020); Liu et al. (2021); Zhu et al. (2021). In particular, frequency information was found to be useful for identifying subtle forgery clues. F^3-Net Qian et al. (2020) integrate frequency statistics to the model using Discrete Cosine Transform. To increase transferability, SPSL Liu et al. (2021) combines spatial image and phase spectrum to capture the up-sampling artifacts in the manipulated faces. FDFL Li et al. (2021) transforms images into YCbCr color space and applies then 2D DCT transformation. Other works focus on modifying the network structure. Zhao et al. (2021) proposed a multi-attention framework to capture local discriminative features. Multi-task learning is also adopted for better generalization ability. Face X-ray Li et al. (2021) simultaneously predicts face forgery and localizes blending boundaries.

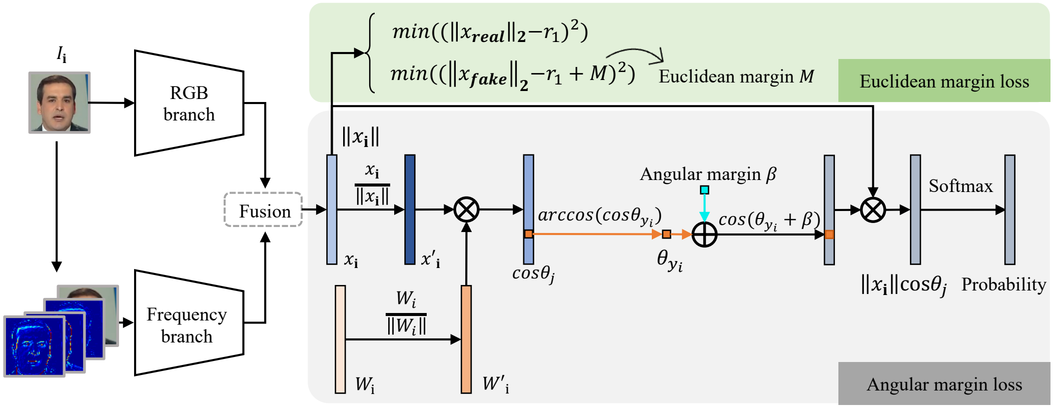

Figure 2: The framework of the proposed method. Features are extracted from both RGB and frequency domains using a two-branch framework. Then these features are further fused to provide richer forgery clues for classification. To further increase the inter-class separability and intra-class compactness of the learned features, a novel Concentric Ring Loss (CRL) consisting of *Euclidean* and *Angular* margin losses is proposed. With the supervision of CRL, the model learns an embedding space where natural faces are clustered around one side of the concentric ring space with a smaller radius, while manipulated ones are clustered around the other side of the concentric ring space with a larger radius.

Metric learning. Similar to face recognition or person re-identification, face forgery detection can largely benefit from more discriminative features. However, learning discriminative features via metric learning is highly overlooked in face forgery detection. Recently, some researchers are aware of the problem and made efforts to explore discriminative features via metric learning. FDFL Li et al. (2021) proposed a single-center loss to reduce the distance between natural face representations and the center point while increasing the distance between modified faces and the center point by at least a margin. Two-branch Masi et al. (2020) presents a new loss that compresses natural faces variability while pushing manipulated faces away. Different from existing work, we incorporate angular-based losses (*i.e.*, SphereFace Liu et al. (2017), CosFace Wang et al. (2018), and ArcFace Deng et al. (2019)) with Euclidean distance-based losses (*i.e.*, Center loss Wen et al. (2016) and triplet loss Schroff et al. (2015)) to further encourage intra-class compactness and inter-class discrepancy of the learned feature.

3 Methodology

3.1 Overview

Since the differences between real and fake images are usually too subtle to be identified, one of the main challenges in deepfake detection is to enhance the discriminative power of learned features. In this paper, we proposed a novel Concentric Ring Loss (CRL) to further increase the inter-class separability and intra-class compactness of the features learned from different categories. Specifically, we independently add margin penalties in angular and Euclidean space to force a larger margin between real and fake images, and hence encourage better discriminating performance. Furthermore, a frequency-aware feature learning module is proposed to exploit high-frequency features and further improve the generalization ability of the model. As shown in Figure 2, the proposed method extracts features from both RGB and frequency domains using a two-branch framework. Then these features are further fused to provide richer forgery clues for classification. Finally, with the supervision of CRL, the network learns an embedding space where natural faces are clustered around one side of the concentric ring space with a smaller radius, while manipulated ones are clustered around the other side of the concentric ring space with a larger radius (Figure 1).

3.2 Concentric Ring Loss

Most existing works on face forgery detection rely on the traditional softmax loss to classify real and fake images. However, feature embeddings learned by the traditional softmax loss are neither discriminative nor compact enough to divide various classes. Since the differences between real and fake images are often too subtle to be identified, it is necessary to enhance the discriminative power of the face forgery detection model. Recently, different metric learning methods have been proposed to encourage inter-class separability and intra-class compactness from either angular or euclidean distance perspective. Angular-based methods, such as SphereFace Liu et al. (2017), CosFace Wang et al. (2018), ArcFace Deng et al. (2019), achieve promising results by introducing the angular margin penalty to maximize the decision margin in the angular space. To achieve intra-class compactness and inter-class discrepancy, Euclidean distance-based approaches, such as Center loss Wen et al. (2016) and triplet loss Schroff et al. (2015), try to penalize the distance between the learned representations and their corresponding class centers while maximizing the distance between different class centers in the Euclidean space. However, most of these approaches suffer from poor generalization ability. Since the feature distribution of synthetic faces varies depending on the manipulation method, the learned features supervised by these approaches are separable for the known manipulation methods but not sufficient to distinguish fake images generated by unseen manipulation methods. To further improve the discriminative power of the trained model, we proposed a new loss function called CRL, which incorporates angular margins in the Softmax loss functions to maximize inter-class separability and penalize Euclidean distance between the deep features and their corresponding class centers to minimize intra-class compactness.

Definition. The proposed CRL consists of an angular margin penalty and Euclidean margin penalty to simultaneously enforce additional inter-class disparity and intra-class compactness. Let $\{I_i, y_i\}_{i=1}^{N}$ be N training samples, where I_i is the i-th image, belonging to the y_i-th class ($y_i \in \{0, 1\}$). We first embed the input image I_i into a d-dimensional vector $x_i \in \mathbb{R}^d$ using the proposed frequency-aware feature leaning network. Then our concentric ring loss is applied to encourage larger angular margin and Euclidean margin between real and fake images. Formally, we define the CRL as:

$$\mathcal{L}_{cr} = \mathcal{L}_{ang} + \lambda \mathcal{L}_{euc},$$

where $\mathcal{L}_{ang}$ and $\mathcal{L}_{euc}$ represent angular margin loss and Euclidean margin loss, respectively. And λ ($\lambda = 1$) is the trade-off weight between $\mathcal{L}_{ang}$ and $\mathcal{L}_{euc}$.

We start by revisiting the softmax loss, which is the most commonly used loss function for face forgery detection.

$$\mathcal{L}_{softmax} = -\frac{1}{N} \sum_{i=1}^{N} log(\frac{e^{W_{y_i}^T x_i + b_{y_i}}}{\sum_{j=1}^{C} e^{W_j^T x_i + b_j}}),$$

where C is the number of classes ($i.e., C = 2$), $W_j \in \mathbb{R}^d$ and $b_j \in \mathbb{R}^d$ is the weight and bias of the activation function in the softmax loss. However, traditional softmax loss does not explicitly optimize feature embedding to ensure the compactness of intra-class samples and the variety of inter-class samples, resulting in a performance gap for deepfake detection when intra-class variations are considerable (e.g. manipulated face images generated by various synthesis methods).

We then introduce the angular margin loss $\mathcal{L}_{ang}$ to improve the generalization ability and discriminative power of the model. For simplicity, the bias is first set to 0 as in Wang et al. (2018); Deng et al. (2019). Then the logit $W_j^T x_i + b_j$ can be simplified and written as $||W_j|| \cdot ||x_i|| \cos\theta$, where θ is the angle between the weight W_j and the feature x_i. Following ArcFace Deng et al. (2019), we normalize W_j and x_i using l_2 norm, then add an angular margin penalty β between W_{y_i} and x_i to improve the discriminative power of learned features. Different from ArcFace, we do not re-scale the normalized features to a fixed scale and project all features to a hypersphere with a fixed radius. Instead, we re-scale the feature x_i to its original magnitude before normalization. By doing so, we are able to leverage the information in Euclidean space and further boost the intra-class compactness with the joint supervision of Euclidean margin loss. The proposed angular margin loss can be expressed as:

$$\mathcal{L}_{ang} = -\frac{1}{N} \sum_{i=1}^{N} log \frac{e^{||x_i|| \cos(\theta_{y_i} + \beta)}}{e^{||x_i|| \cos(\theta_{y_i} + \beta)} + \sum_{j \neq y_i} e^{||x_i|| \cos(\theta_j)}}.$$

The angular margin loss encourages better discriminating and compacting performance than softmax loss. However, it constrains the intra-class compactness for *both* natural and manipulated faces, which reduces the model's generalization ability and leads to overfitting to some extent. To this end, we integrate Euclidean margin loss in our CRL to boost the model's discriminative power and generalization ability.

The purpose of Euclidean margin loss is to reduce the distance between natural face embeddings and the center point while simultaneously pushing manipulated face embeddings away from the center point. Natural faces will be constrained in the hypersphere with a radius of r_1, while manipulated faces are kept outside the hypersphere with a radius of $r_2 = r_1 + M$. r_1 is the target norm value for real face embeddings and is learned during training. M is a hyperparameter and denotes the margin between r_1 and r_2. The proposed Euclidean margin loss can be expressed as:

$$\mathcal{L}_{euc} = D_{real} + D_{fake},$$

where D_{real} and D_{fake} denote the difference between the feature norm and target norm value for real and fake face embeddings, respectively. They can be computed as:

$$D_{real} = \frac{1}{|\Omega_{real}|} \sum_{i \in \Omega_{real}} (\|x_i\|_2 - r_1)^2,$$

$$D_{fake} = \frac{1}{|\Omega_{fake}|} \sum_{i \in \Omega_{fake}} (\|x_i\|_2 - r_1 + M)^2.$$

3.3 Frequency-aware feature leaning

By introducing important forgery clues, wavelet transformation has achieved remarkable success in face forgery detection. Following this idea, a frequency-aware feature learning framework is proposed to utilize high-frequency features to further improve generalization and discriminative power of the model. As shown in Figure 2, a two-branch framework is used to extract features from both RGB and frequency domains. We adopt Xception-Net Chollet (2017) as the backbone for both branches. Given an input image I_i, the features (*i.e.*, x_{rgb} and x_{freq}) generated by the RGB branch $g(\cdot)$ and the frequency branch $g(\cdot)$ can be expressed as:

$$\begin{aligned} x_{rgb} &= g(I_i), \\ x_{freq} &= h(wavelet(I_i). \end{aligned} \tag{1}$$

We adopt a classic wavelet transformation method (*i.e.*, the Haar wavelet) to provide new clues in the frequency domain. The Haar wavelet contains four kernels (*i.e.*, LL, LH, HL, HH), where L and H are the low- and high- pass filters. They can be expressed as:

$$L = \frac{1}{\sqrt{2}} \begin{bmatrix} 1 \\ 1 \end{bmatrix}, \quad H = \frac{1}{\sqrt{2}} \begin{bmatrix} -1 \\ 1 \end{bmatrix}.$$

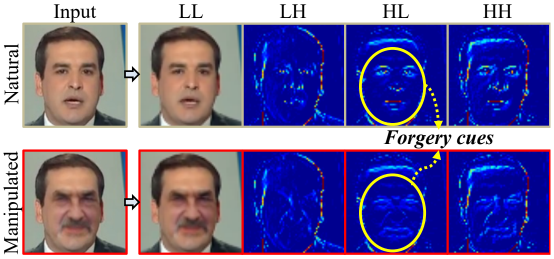

Figure 3: Inconsistency in the frequency domain can be considered a useful forgery indicator. We show the information of four frequency domains (*i.e.*, , LL, LH, HL, and HH) decomposed from a fake and a real image by implementing the Haar wavelet transformation. LL mainly consists of information in the low-frequency domain, depicting the overall appearance of an image, while LH, HL and HH contain information representing rich details.

As shown in Figure 3, the low-pass filter (*i.e.*, LL) focuses on smooth surfaces associated with low-frequency signals. In contrast, the high-pass filters (*i.e.*, LH, HL, HH) mainly capture the high-frequency details, such as vertical (*i.e.*, LH), horizontal (*i.e.*, HL), and diagonal (*i.e.*, HH) edges. We find that inconsistency in the frequency domain can be considered a useful forgery indicator. High-level features of the two modalities (*i.e.*, RGB and frequency signals) are then fused to provide richer forgery clues for classification. We simply concatenate RGB features x_{rgb} and frequency features x_{freq} and then adopt a point-wise convolution block Chollet (2017) to fuse them and make the prediction.

Table 1: Comparison on the FF++ dataset with different compression settings. Light compression videos are denoted as HQ (c23), and heavy compression videos are denoted as LQ (c40). Xception † is our baseline. All methods in the table are image/frame-based detection methods. However, some of them only report video-level results instead of frame-level results. To conduct a more comprehensive comparison, we show both frame- and video-level results. Following previous methods, we compute video-level scores by averaging the frame score in each video.

	HQ (c23)				LQ (c40)			
	Frame-level		Video-level		Frame-level		Video-level	
Methods	ACC	AUC	ACC	AUC	ACC	AUC	ACC	AUC
Bayar and Stamm (Bayar & Stamm, 2016)	82.97	–	–	–	66.84	–	–	–
MesoNet (Afchar et al., 2018)	83.10	–	–	–	70.47	–	–	–
Xception (Rossler et al., 2019)	92.39	94.86	95.73	96.30	80.32	81.76	86.86	89.30
Face X-ray (Li et al., 2020a)	–	87.35	–	–	–	61.60	–	–
F^3-Net (Qian et al., 2020)	–	–	97.52	98.10	–	–	90.43	93.30
SPSL (Liu et al., 2021)	91.50	95.32	–	–	81.57	82.82	–	–
Multi-attention (Zhao et al., 2021)	–	–	96.37	98.97	–	–	86.95	87.26
FDFL (Li et al., 2021)	–	–	96.69	99.30	–	–	89.00	92.40
Xception †	90.45	95.01	94.76	96.04	73.11	81.09	80.95	88.49
Ours	**94.04**	**98.23**	**97.57**	**99.53**	**82.31**	**84.10**	**90.55**	**94.36**

4 Experiments

4.1 Experimental settings

Dataset. We conduct our experiments on two large-scale benchmark deepfake datasets: FaceForensics++ (FF++) (Rossler et al., 2019) and Celeb-DF (Li et al., 2020b). FF++ consists of 1,000 original video sequences and 4,000 manipulated videos generated by four types of common face manipulation methods. Following (Luo et al., 2021; Zhu et al., 2021), we split the dataset into 720, 140, and 140 videos for training, validation, and testing, respectively. Celeb-DF (Li et al., 2020b) is one of the most challenging datasets for deepfake detection methods. It includes 408 real and 795 fake videos. For model generalization assessment, we mainly evaluate on the Celeb-DF dataset.

Evaluation metrics. Following previous work (Li et al., 2021; Liu et al., 2021; Zhao et al., 2021; Masi et al., 2020), we mainly report results on accuracy rate (ACC) and the area under the receiver operating characteristic curve (AUC). Note that we report both frame-level and video-level results. Following (Masi et al., 2020; Li et al., 2021), the ACC and AUC score at the video level are computed by averaging the ACC and AUC scores of each frame in a video, respectively.

Implementation details. To prepare the training and testing data, we first cropped facial images about the head region, which were then resized to 299×299. In the training process, we augment the original frames 4 times for real/fake label balance. We adopt Xception (Chollet, 2017) as the backbone network for both RGB and frequency branches. We initialize model parameters by Xception, which is pre-trained on ImageNet. We set hyper-parameters angular margin $\beta = 0.5$ and Euclidean margin $M = 10$. Optimization was done with SGD with a learning rate of 1.0×10^{-2} that dropped 0.5 at $40,000^{th}$ iterations. The model was trained with a batch size of 32 and for a total epoch of $60,000$ iterations. Implementation was done using PyTorch.[1]

4.2 Comparison with state-of-the-art methods

In this section, we compare the detection results of CRL with state-of-the-art methods on FF++ (Rossler et al., 2019). Moreover, we compare with face forgery detection methods that leverage metric learning-based strategies to show the superiority of the proposed method. Then, we also perform a cross-dataset evaluation on Celeb-DF (Li et al., 2020b) to demonstrate the generalizability and transferability of the proposed method.

[1] Source code and trained models available (removed for review).

Table 2: Cross-dataset evaluation on Celeb-DF dataset. The model is trained on FF++ and tested on Celeb-DF. We report AUC(%) scores on both datasets. Numbers are cited from [Zhao *et al.,* 2021]. It can be seen that our method achieves the best result on Celeb-DF and comparable results on FF++, which validates the discriminative power and generalization ability of the proposed method.

Method	FF++	Celeb-DF
Two-stream (Zhou et al., 2017)	70.10	53.80
Meso (Afchar et al., 2018)	84.70	54.80
DSP-FWA (Li & Lyu, 2018)	93.00	64.60
Multi-task (Nguyen et al., 2019a)	76.30	54.30
Xception (Rossler et al., 2019)	99.70	65.30
Capsule (Nguyen et al., 2019b)	96.60	57.50
Two Branch (Masi et al., 2020)	93.18	73.41
Face X-ray (Li et al., 2020a)	98.52	74.76
F^3-Net (Qian et al., 2020)	98.10	65.17
SRM-CBAM (Luo et al., 2021)	–	79.40
SPSL (Liu et al., 2021)	96.91	76.88
SLADD (Chen et al., 2022)	98.40	79.70
Ours	99.53	**83.57**

Table 3: Comparison with methods leveraging various metric learning-based strategies. The FF++ (LQ (c40)) dataset is used for training and testing.

Model	Loss function	AUC
Xception (Rossler et al., 2019)	softmax loss	86.00
Xception (Rossler et al., 2019)	softmax + triplet loss	86.30
Xception (Rossler et al., 2019)	softmax + center loss	86.80
Xception (Rossler et al., 2019)	softmax + SCL	91.60
FDFL (Li et al., 2021)	softmax + SCL	92.40
Ours	CRL	**94.36**

Evaluation on FF++. We report the performance of our method on different video compression settings (*i.e*., high-quality compression (HQ (c23)) and low-quality compression (LQ (c40))) of FF++ (Rossler et al., 2019). The comparison results are listed in Table 1. Since some methods only report frame-level or video-level results, we show both results for a more comprehensive comparison. Following (Qian et al., 2020; Zhao et al., 2021; Li et al., 2021), we compute the video-level scores by averaging the frame scores in a video.

On both HQ and LQ settings, the results in Table 1 show that our method delivers the state-of-the-art (SOTA) performance. Compared to SOTA works on the HQ setting, we achieve 1.79% and 3.05% improvements on ACC and AUC, respectively. Compared on the LQ setting, we also achieve a large improvement (*i.e*., 0.91% on ACC and 1.55% on AUC). We also notice that methods utilizing high-frequency features, such as F^3-Net (Qian et al., 2020), SPSL (Liu et al., 2021), and FDFL (Li et al., 2021), have relatively better performance than methods that only use RGB images, especially on low-quality videos. One possible reason is that low-quality videos are highly compressed and the information in the RGB channels is attenuated due to the large amount of noise introduced in the compression process. In this case, frequency information can provide additional forgery clues and thus greatly benefit forgery detection models.

Comparison with methods adopting metric learning. We next compare the proposed method with many methods that leverage metric learning-based strategies to improve intra-class compactness and inter-class separability, such as softmax loss, triplet loss (Schroff et al., 2015), center loss (Wen et al., 2016), and single-center loss (SCL) (Li et al., 2021). The comparison results are listed in Table 3. Numbers in the table are cited from (Li et al., 2021). We train and test our model on the challenging LQ setting of FF++. As shown in Table 3, we achieve significantly better performance than Xception (Rossler et al., 2019) supervised by various metric learning. We also show a comparison with FDFL (Li et al., 2021), which proposed a novel single-center loss for face forgery detection. Still, the proposed method outperforms its competitors by a large margin.

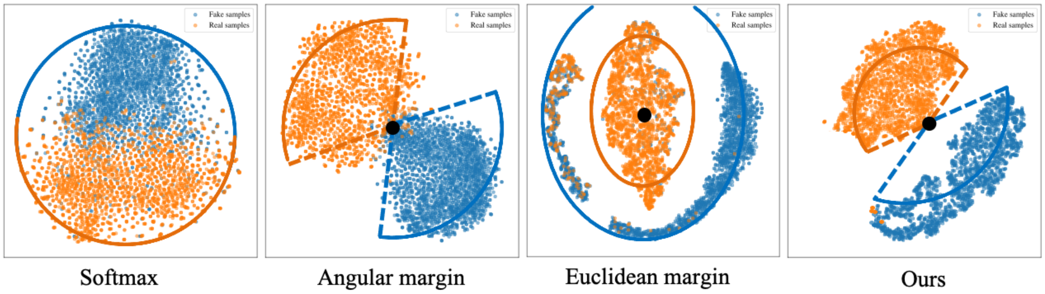

Figure 4: Feature visualization. Natural faces are denoted with orange dots and manipulated faces are blue dots. We can see that softmax and Euclidean loss are not discriminative enough to distinguish subtle differences between classes (*e.g.*, lots of orange and blue dots are mixed together). Angular loss improves feature compactness and discriminative power. Still, mixed dots can be observed in the central region. The proposed CRL loss achieves the best result by forcing natural faces gathered compactly and separated from those of manipulated faces which are distributed less compactly. Joint supervision of angular and Euclidean margin loss increases both the discriminative power and generalization ability of the model.

Cross-dataset evaluation. We perform a cross-dataset evaluation on Celeb-DF (Li et al., 2020b) to demonstrate the generalizability and transferability of the proposed method. The model is trained on FF++ (HQ) and tested on Celeb-DF. Following (Zhao et al., 2021), we sample 30 frames from each video and calculate AUC scores. As shown in Table 2, our method achieves the best result on Celeb-DF, while getting slightly worse results than SOTA on FF++. Although two methods (*i.e.*, Xception (Rossler et al., 2019) and Multi-attention (Zhao et al., 2021)) achieve higher in-dataset AUC than our method, their generalization and transferability are far less than ours. More results on generalizability comparisons are shown in the supplementary material.

4.3 Ablation Study

Effectiveness of CRL. To demonstrate the effectiveness of CRL, we conduct additional experiments on various losses, including traditional softmax loss, angular margin loss, and Euclidean margin loss. We first provide the feature visualization (t-SNE) of different losses in Figure 4. It shows that softmax and Euclidean loss are insufficient for distinguishing subtle differences between classes (*e.g.*, lots of orange and blue dots are mixed together). Angular loss increases the compactness and discriminative power of features. Even so, mixed dots can be seen in the center region. The proposed CRL achieves the best results by forcing natural faces to be gathered and separated from manipulated faces, which are distributed less compactly.

Table 4: Effectiveness of CRL. We report frame-level results on FF++ (HQ (c23)) dataset. Compared to softmax loss, both angular and Euclidean margin loss encourage better discriminating and compacting performance. Still, the proposed CRL achieves the best results.

Loss function	ACC	AUC
Softmax loss	90.45	95.01
Angular margin loss	92.32	97.85
Euclidean margin loss	91.06	95.63
CRL	**94.04**	**98.23**

The angular margin loss encourages better discriminating and compacting performance than softmax loss. However, it constrains the intra-class compactness for *both* natural and manipulated faces, which reduces the model's generalization ability and leads to overfitting to some extent. Compared to angular margin loss, CRL uses joint supervision of angular and Euclidean loss to boost the model's discriminative power and generalization ability while also stabilizing the training process.

Since the Euclidean margin loss does not have a fully-connected layer to output final predictions, the model jointly is supervised by softmax loss and Euclidean margin loss. We set angular margin $\beta = 0.5$ and Euclidean margin $M = 10$ for all losses. As can be seen from Table 4, our CRL loss significantly outperforms the other losses, with an ACC of 94.04 and an AUC of 98.23. Additionally, we noticed that angular margin loss is more efficient than Euclidean margin loss. Compared to using softmax loss only, the improvement of adding Euclidean margin loss is subtle.

Table 5: Performance gain of each component. Video-level results of FF++ (HQ (c23)) dataset are reported. We use Xception (Chollet, 2017) as the classification model for single branches (*i.e.*, RGB branch and frequency branch). Compared to baseline models without CRL, both single-branch and two-branch models (*i.e.*, RGB, Freq., RGB + Freq.) can benefit from the proposed CRL loss and achieve better performance. Moreover, the results of frequency branch (*i.e.*, row 3 and 6) validates the efficacy of frequency information in detecting manipulated faces.

RGB	Freq.	CRL	ACC	AUC
✓			94.76	96.04
	✓		94.90	96.86
✓	✓		95.84	97.51
✓		✓	96.42	99.27
	✓	✓	96.90	99.01
✓	✓	✓	**97.57**	99.53

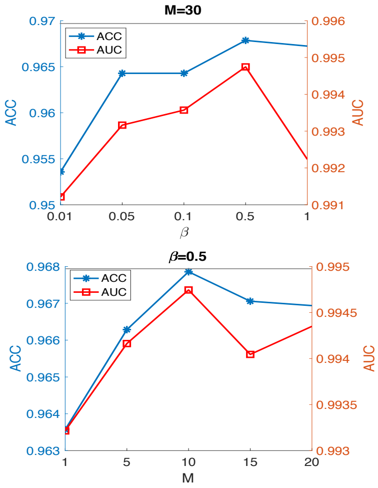

Figure 5: Ablation study for two hyper-parameters: angular margin β and Euclidean margin M. We first set $M = 10$, and show the detection performances of various β (top). Then we fix $\beta = 0.5$, and show the detection performances of various M (bottom).

Performance gain of framework components. We next measure the contribution of each component in our framework. To this end, we quantitatively evaluate the proposed framework and its variants: 1) RGB branch + softmax loss (baseline); 2) Frequency branch + softmax loss; 3) Both branch + softmax loss; 4) RGB branch + CRL loss; 5) Frequency branch + CRL loss; 6) Both branch + CRL loss. Table 5 lists the evaluation results (*i.e.*, ACC and AUC scores) of six variants used. As can be observed from the table, compared to the baseline model (*i.e.*, RGB branch + softmax loss), both frequency branch and CRL loss can boost the performance of the model. The CRL loss, in particular, contributes an improvement of 1.80% in ACC and 2.07% in AUC. This verifies the discriminative ability of the CRL loss to supervise the network to learn intra-class compressed and inter-class separated features. Moreover, we also validates the efficacy of frequency information in detecting manipulated faces. As shown in Table 5, models with frequency branch perform better than models without one by a large margin. More results using different frequency clues (*e.g.*, Cosine or Fourier Transform) are shown in the supplementary material.

Parameter influence. As shown in the CRL loss function, there are two hyper-parameters (*i.e.*, the angular margin β and the Euclidean margin M) that may affect the effectiveness of the loss. To investigate the influence of these two hyper-parameters, we conduct an empirical analysis on the FF++ HQ (c23) dataset. Figure 5 shows video-level ACC and AUC results on various pairs of (β, M). It can be seen that the proposed approach may effectively enhance performance when β and M varies within a broad range. The best results are obtained when $beta$ is set to 0.5 and M is set to 10, with an ACC of 0.976 and an AUC of 0.995.

5 Conclusion

Due to the growing social concerns over indistinguishable deepfake pictures, deepfake detection has gotten a lot of attention in computer vision. Since the distinction between the natural and manipulated image is often subtle, enhancing the discriminative power of learned features is one of the primary issues in deepfake detection. In this paper, we propose a novel Concentric Ring Loss (CRL) to explicitly encourage intra-class compactness and inter-class separability of learned features by adding angular and Euclidean margin penalties. Moreover, a frequency-aware feature learning module is proposed to exploit high-frequency features and further improve generalization ability of the model. Extensive experiments demonstrate the superiority of our methods over different datasets. We show that CRL consistently outperforms the state-of-the-art by a large margin.

REFERENCES

Darius Afchar, Vincent Nozick, Junichi Yamagishi, and Isao Echizen. Mesonet. In *WIFS*, pp. 1–7. IEEE, 2018.

Belhassen Bayar and Matthew C Stamm. A deep learning approach to universal image manipulation detection using a new convolutional layer. In *ACM Workshop on IHMS*, pp. 5–10, 2016.

Liang Chen, Yong Zhang, Yibing Song, Lingqiao Liu, and Jue Wang. Self-supervised learning of adversarial example: Towards good generalizations for deepfake detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18710–18719, 2022.

François Chollet. Xception. In *CVPR*, pp. 1251–1258, 2017.

Deepfakes. Deepfakes github. `http-s://github.com/deepfakes/faceswap`, 2020.

Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. Arcface. In *CVPR*, pp. 4690–4699, 2019.

Chengdong Dong, Ajay Kumar, and Eryun Liu. Think twice before detecting gan-generated fake images from their spectral domain imprints. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7865–7874, 2022.

Ricard Durall, Margret Keuper, and Janis Keuper. Watch your up-convolution: Cnn based generative deep neural networks are failing to reproduce spectral distributions. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 7890–7899, 2020.

Joel Frank, Thorsten Eisenhofer, Lea Schönherr, Asja Fischer, Dorothea Kolossa, and Thorsten Holz. Leveraging frequency analysis for deep fake image recognition. In *International conference on machine learning*, pp. 3247–3258. PMLR, 2020.

Marek Kowalski. Faceswap github. `https://github.com/MarekKowalski/FaceSwap`, 2020.

Jiaming Li, Hongtao Xie, Jiahong Li, Zhongyuan Wang, and Yongdong Zhang. Frequency-aware discriminative feature learning supervised by single-center loss for face forgery detection. In *CVPR*, pp. 6458–6467, 2021.

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 *CVPR*, pp. 5001–5010, 2020a.

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

Yuezun Li, Xin Yang, Pu Sun, Honggang Qi, and Siwei Lyu. Celeb-df. In *CVPR*, pp. 3207–3216, 2020b.

Honggu Liu, Xiaodan Li, Wenbo Zhou, Yuefeng Chen, Yuan He, Hui Xue, Weiming Zhang, and Nenghai Yu. Spatial-phase shallow learning. In *CVPR*, pp. 772–781, 2021.

Weiyang Liu, Yandong Wen, Zhiding Yu, Ming Li, Bhiksha Raj, and Le Song. Sphereface: Deep hypersphere embedding for face recognition. In *CVPR*, pp. 212–220, 2017.

Yuchen Luo, Yong Zhang, Junchi Yan, and Wei Liu. Generalizing face forgery detection with high-frequency features. In *CVPR*, pp. 16317–16326, 2021.

Iacopo Masi, Aditya Killekar, Royston Mascarenhas, Shenoy Gurudatt, and Wael AbdAlmageed. Two-branch recurrent network for isolating deepfakes in videos. In *ECCV*, pp. 667–684. Springer, 2020.

Huy H Nguyen, Fuming Fang, Junichi Yamagishi, and Isao Echizen. Multi-task learning for detecting and segmenting manipulated facial images and videos. *arXiv preprint arXiv:1906.06876*, 2019a.

Huy H Nguyen, Junichi Yamagishi, and Isao Echizen. Capsule-forensics: Using capsule networks to detect forged images and videos. In *ICASSP*, pp. 2307–2311. IEEE, 2019b.

Albert Pumarola, Antonio Agudo, Aleix Martinez, Alberto Sanfeliu, and Francesc Moreno-Noguer. Ganimation. In *ECCV*, pp. 818–833, 2018.

Yuyang Qian, Guojun Yin, Lu Sheng, Zixuan Chen, and Jing Shao. Thinking in frequency: Face forgery detection by mining frequency-aware clues. In *ECCV*, pp. 86–103. Springer, 2020.

Nicolas Rahmouni, Vincent Nozick, Junichi Yamagishi, and Isao Echizen. Distinguishing computer graphics from natural images using convolution neural networks. In *WIFS Workshop*, pp. 1–6. IEEE, 2017.

Andreas Rossler, Davide Cozzolino, Luisa Verdoliva, Christian Riess, Justus Thies, and Matthias Nießner. Faceforensics++: Learning to detect manipulated facial images. In *ICCV*, pp. 1–11, 2019.

Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet. In *CVPR*, pp. 815–823, 2015.

Hao Wang, Yitong Wang, Zheng Zhou, Xing Ji, Dihong Gong, Jingchao Zhou, Zhifeng Li, and Wei Liu. Cosface. In *CVPR*, pp. 5265–5274, 2018.

Yandong Wen, Kaipeng Zhang, Zhifeng Li, and Yu Qiao. A discriminative feature learning approach for deep face recognition. In *ECCV*, pp. 499–515. Springer, 2016.

Xu Zhang, Svebor Karaman, and Shih-Fu Chang. Detecting and simulating artifacts in gan fake images. In *2019 IEEE international workshop on information forensics and security (WIFS)*, pp. 1–6. IEEE, 2019.

Hanqing Zhao, Wenbo Zhou, Dongdong Chen, Tianyi Wei, Weiming Zhang, and Nenghai Yu. Multi-attentional deepfake detection. In *CVPR*, pp. 2185–2194, 2021.

Peng Zhou, Xintong Han, Vlad I Morariu, and Larry S Davis. Two-stream neural networks for tampered face detection. In *CVPRW*, pp. 1831–1839. IEEE, 2017.

Xiangyu Zhu, Hao Wang, Hongyan Fei, Zhen Lei, and Stan Z Li. Face forgery detection by 3d decomposition. In *CVPR*, pp. 2929–2939, 2021.