CDNET: CLUSTER DECISION FOR DEEPFAKE DETECTION GENERALIZATION

Zeming Hou[⋆] *Zhongyun Hua*[⋆] *Kuiyuan Zhang*[⋆] *Yushu Zhang*[†]

[⋆] Harbin Institute of Technology, Shenzhen, Guangdong 518055, China
[†]Nanjing University of Aeronautics and Astronautics, Nanjing, Jiangsu 210016, China

2023 IEEE International Conference on Image Processing (ICIP) | 978-1-7281-9835-4/23/$31.00 ©2023 IEEE | DOI: 10.1109/ICIP49359.2023.10223180

ABSTRACT

The fast development of deepfake generation technology has caused serious security threats to human society. Many deepfake detection methods have been proposed recently, but most of them can only show high detection performance for the deepfakes generated by the similar techniques with the training dataset. To improve the ability of detecting unseen types of deepfakes, some deepfake detection methods have constructed self-generated datasets to train their models. However, the artifacts on these self-generated datasets are usually caused by some specific face-blending algorithms and lack of generality. In this paper, we propose cluster decision network (CDNet) to improve the deepfake detection generalizability. We design a selective attention module that decides the attention areas by manually cropping the facial areas (e.g., eyes, nose, and lips), which greatly reduce the model size and ensure a small model size. Inspired by the contrastive learning, we also propose a cluster classifier to equally utilize the feature representation. Extensive experiments show that our method outperforms existing state-of-the-art methods in deepfake detection generalizability and has the minimum model size.

Index Terms— Deepfake detection, Generalizability, Contrastive learning.

1. INTRODUCTION

With the fast development of deepfake generation technologies, the generated deepfakes are harder and harder to be recognized by human eyes [1, 2]. This may cause a huge security threat to human society in different aspects. Researchers have developed many deepfake detection methods to deal with this threat [3, 4, 5]. These methods can achieve a high detection performance for the deepfakes generated by the same or similar techniques with the training dataset but show poor performance on new types of deepfakes.

Recently, some methods have been proposed to improve the generalizability of deepfake detection, and they can be roughly divided into three kinds. The first kind introduces self-generated deepfake datasets to guide the model to learn the artifact features that widely exist in most deepfakes [6, 7]. These detection methods show good generalizability in previous detection tasks. However, their self-generated data are composed of face-blending artifacts that can be easily fixed by new deepfake generation methods, making these detection methods perform poorly on detecting the deepfakes generated by some advanced generation methods. The second kind uses domain adaptation to transfer the model to a new deepfake dataset to achieve higher generalizability [8]. These methods can achieve high performance on the target domain, however, may lose the knowledge of the source domain. The third kind designs additional architecture or strategy to guide the model focus on the artifact-relevant features [4]. These methods ensure that the models can pay attention to the areas of the artifacts but have complex architectures and training strategies.

In this paper, we design and evaluate CDNet, a new deepfake detection model for improving the detection generalization. CDNet consists of the selective attention module (SAM) and the cluster decision module (CDM). We design SAM to ensure that our model focus on the artifacts of some selective facial features, which exist in most deepfakes and are hard to be eliminated [9]. Specifically, SAM uses a special attention mechanism to directly take the cropped facial sub-images as input. Inspired by the contrastive learning, we design CDM to enlarge the difference between the real and fake samples. The designed cluster classifier uses a novel decision-making approach to decide the sample class according to its distance to the centers of the two kinds of samples (real or fake). Since all the features are treated equally in the distance calculation, our module can avoid over-reliance on certain features. Extensive experiments are conducted to validate the superiority of our model in deepfake detection generalizability. We summarize the contributions of this paper as follows: 1) We propose a new deepfake detection generalizing method called CDNet and it contains SAM and CDM. Compared with existing methods, our model can pay attention to the artifacts that exist in general deepfakes, and has a simple model; 2) We design a new cluster classifier based on the contrastive learning. It can improve the deepfake detection generalizability by enlarging the sample discrimination and treating all the features equally; 3) We conduct experiments to test out CDNet and the results show that it can obtain state-of-the-art generalization performance and has a small model size.

Corresponding author: Zhongyun Hua. This paper is supported by the National Natural Science Foundation of China under Grant 62071142.

2. RELATED WORK

2.1. Generalizable Deepfake Detection Methods

Recently, many deepfake detection methods have been developed to improve the detection generalizability for unseen types of deepfakes, and these methods can be roughly divided into three kinds.

Self-generated data-based methods. The methods in [6, 7] train their models using the samples with face-blending artifacts generated by different strategies. These methods can reduce the model dependence on some specific datasets while ensuring the model to pay attention to the deepfake artifacts that appeared in the training sets. However, the face-blending operation is not indispensable in deepfake generation. These detection methods inevitably show poor performance in detecting some kinds of deepfakes.

Domain adaptation-based methods. In [8], Kim $et\ al.$ use knowledge distillation on the representation of the previous model as a training restriction, which prevents the performance degradation on the previous task. In [10], Tariq $et\ al.$ train the model using the target domain data and use data from the source domain to repair the damaged knowledge. These domain adaptation-based methods can improve the detection generalizability effectively. However, they have to retrain their model with data from previous task.

General artifact attention-based methods In [11], Zhao $et\ al.$ generate multiple attention maps that autonomously concentrate the model on the areas of general anomalies, rather than the discriminative areas only. In [12], Wang $et\ al.$ first adopt a feature map to cover parts of the face that the classifier relies on, and drive the model to learn more general features from the rest of face areas. These methods can help the model concentrate on the areas where the general deepfake artifacts exist. However, they usually have very complicated model architectures or training strategies.

2.2. Contrastive Learning

The contrastive learning is an effective tool in differentiating the samples from different kinds in the feature map [13]. It has been verified that the contrastive learning is highly effective in assisting classification. These methods with contrastive learning [14, 13] adopt fully connected layers as their final classifier. As a result, the detection result tends to rely on some discriminative features and has a limited performance for generalizability.

3. CDNET

3.1. Overview

Fig. 1 shows the structure of our CDNet, which consists of a selective attention module (SAM) and a cluster decision module (CDM). As can be seen, SAM uses the local information (i.e., left eye, right eye, nose, and mouth) as the main discrimination basis and the whole image as a complement. Both the spatial and frequency features are extracted and then fused for classification. In CDM, we utilize two different classifiers and switch them in different training stages. Specifically, we design a distinctive cluster classifier to classify an image according to the feature distance between this image with the centers of the real and fake images.

3.2. Selective Attention Module

According to previous studies [9], the artifacts existing in the facial features are hard to be eliminated. Based on this property, we design SAM that uses the face features as the main basis for detecting deepfakes. This manual selection can ensure that our model pays more attention to the features related to the artifacts. At the same time, the model can keep lightweight, and the face-blending artifacts can be omitted.

Given an input face image, we first use BlazeFace [15] to crop four face feature images. Considering the different significance between the whole face image and its partial face feature images, we design two kinds of modified Xception [3] blocks to extract the spatial features. The first one is for the face feature images. In contrast, the second one for the face images holds fewer channels and parameters, and it serves only for global information extraction. The face feature blocks transform each face feature image into a 1×256 vector, while the face image blocks also transform the face image into a 1×256 vector. To enhance the feature representation ability, we also extract the frequency features using the discrete cosine transform (DCT) which was adopted by recent works[16, 17]. For each image, the average amplitudes of each frequency are obtained to form a 1×128 frequency vector. Finally, all frequency and spatial feature vectors are concatenated and then projected by a fully connected layer into a feature vector F with size 1×200 for classification.

3.3. Cluster Decision Module

Most previous deepfake detection methods [18] use a fully connected layer as the binary classifier to detect deepfakes. To achieve the least loss, this classifier tends to greatly rely on the discriminative features (e.g., the inconsistency of light contrast, specific face blending artifacts, and identity of faces) However, these features may be irrelevant to the deepfake artifacts, making the model show a poor performance in generalizability. To address this disadvantage, we propose a new cluster classifier in CDM. As can be seen in the right part of Fig. 1, we set a classifier switch to switch two different classifiers for the final classification. The traditional binary classifier is to learn how to capture the features, while the proposed cluster classifier is to improve the detection generalizability. Specifically, we first use the traditional binary classifier to train our model until it performs the best in the

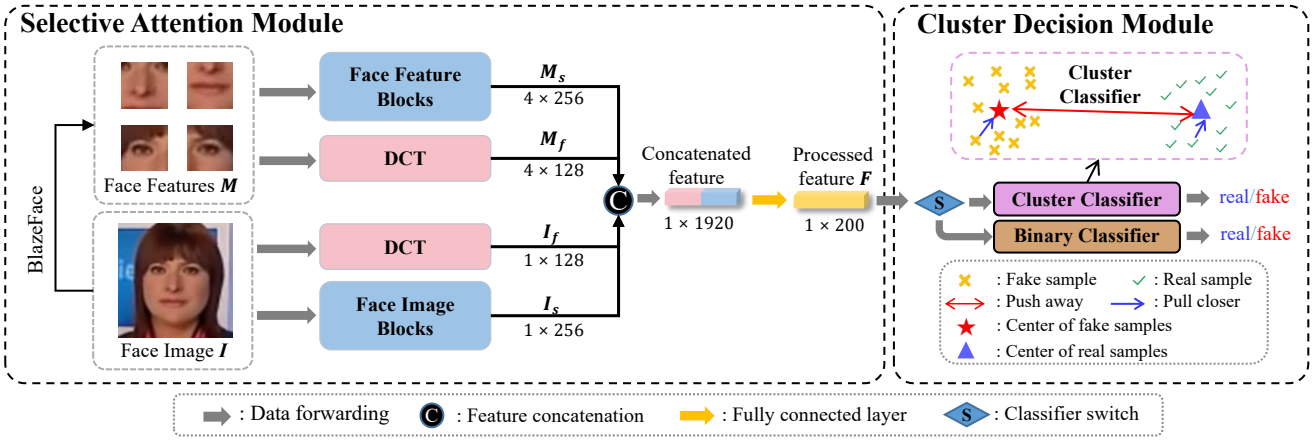

Fig. 1: Overview framework of our CDNet. Four face feature images are cropped from each face image.

validation dataset. The intermediate model generates two features centers as follows:

$$CF = \text{Avg}(\sum_{F \in Fake} F), \quad CR = \text{Avg}(\sum_{F \in Real} F), \tag{1}$$

where $CF \in \mathbb{R}^{1\times 200}$ and $CR \in \mathbb{R}^{1\times 200}$ are the feature centers of the fake and real images, *Fake* and *Real* means that the image with feature F is a fake image and real image, respectively.

After obtaining the sample centers of the whole training dataset, CDNet switches to the cluster classifier to continue training. Inspired by the contrastive learning [13], our model aims to enlarge the distance of the feature centers of the real image and fake images as much as possible. This operation can improve the detection generalizability by separating the features of the unseen real and fake samples. We design the loss function as follows:

$$\begin{aligned} L_F &= \sum_{F \in Fake} \max(||CF - F||_2^2 - \beta||CF - CR||_2^2, 0), \\ L_R &= \sum_{F \in Real} \max(||CR - F||_2^2 - \beta||CF - CR||_2^2, 0), \end{aligned} \tag{2}$$

$$L = \frac{L_F + L_R}{N \times ||CF - CR||_2^2}, \tag{3}$$

where β is the sample scaling factor to avoid over-concentration in feature distribution, and N is the training batch size. By minimizing the loss L, our CDNet can enlarge the distance of the feature centers of the real images and the fake images, and gather the samples with the same labels. During this training process, we update the feature centers using the features of the current batch and a center moving rate α. The updating process is calculated as follows:

$$\begin{aligned} CF^t &= (1-\alpha)CF^{t-1} + \alpha \text{Avg}(\sum_{F \in Fake} \text{F}) \\ CR^t &= (1-\alpha)CR^{t-1} + \alpha \text{Avg}(\sum_{F \in Real} \text{F}) \end{aligned} \tag{4}$$

The cluster classifier classifies the input image as:

$$CC(F) = \begin{cases} Fake, & \text{if } ||CF - F||_2^2 < ||CR - F||_2^2; \\ Real, & else. \end{cases} \tag{5}$$

In this decision strategy, the cluster classifier treats all features equally, which avoids over-reliance on certain discriminative features.

4. EXPERIMENTS

4.1. Datasets

Training and testing datasets. We evaluate the performance of our CDNet on the most widely used Faceforencis++ (FF++) dataset [1] and Celeb-DF dataset [2]. The FF++ dataset contains 1000 original videos and 4000 deepfake videos generated by four deepfake generation methods. Following the settings of the recent studies [19, 7], we choose the C23 compression level, and use 720 and 140 original videos as the training set and testing set, respectively. The Celeb-DF dataset contains 408 original videos and 795 deepfake videos, and it is used for testing the detection generalizability of our CDNet.

Preprocessing. We construct our training set and testing set using the following steps. (a) Uniformly extract 32 frames from each video. (b) Use BlazeFace [15] to capture the face from each frame and then crop four face feature images (i.e., left eye, right eye, nose, and mouth). (c) Resize each face image as 299×299 and each face feature image as 64×64.

Training settings. We use the Adam optimizer with a learning rate of 0.0001. For the parameters of the cluster classifier, we experimentally set $\alpha = 0.0001$ and $\beta = 0.4$. We set the training batch size as 32. We train our model using the traditional binary classifier for 20 epochs in the first stage of training. We choose the model with the best validation accuracy in FF++ dataset, replace its classifier with the cluster classifier, and train it for 20 epochs in the second stage of training. We implement our model using PyTorch and run all

Table 1: Performance comparison with some state-of-the-art deepfake detection methods. All the models are trained on the FF++ dataset and tested on the FF++ dataset and the Celeb-DF dataset, respectively.

Methods	FF++	Celeb-DF	Avg	Params(M)
Xception [3]	0.997	0.653	0.825	22.9
EfficientNet B4 [20]	0.997	0.643	0.82	19.3
Two-branch [14]	0.932	0.734	0.833	-
SPSL [21]	0.969	0.724	0.847	22.9
MADD [11]	**0.998**	0.674	0.836	417.6
F^3-Net[5]	**0.998**	0.697	0.848	42.5
Ours	0.979	**0.770**	**0.875**	**5.6**

the experiments on a computer with an Intel i9-10900X CPU and an RTX3090 GPU.

4.2. Performance Comparison

We compare our CDNet with six state-of-the-art methods, including Xception [3] , EfficientNet B4 [20], Two-branch [14], SPSL [21], MADD [11], and F^3-Net [5]. These methods are trained without introducing self-generated data. The area under the ROC curve (AUC) was chosen as the metric. We implement these models following their original papers. The AUC results of all the competing models are referred to [7] except F^3-Net, and their model sizes are tested in our implementation. Table 1 lists the comparison results.

Intra-dataset performance. We first compare the AUC of different deepfake detection methods within the FF++ dataset, which means that the training set and testing set are all from the FF++ dataset. As seen from the first column of Table 1, most methods have AUC scores larger than 95% except for the Two-branch. This indicates that these methods can effectively capture the discriminative features of the deepfakes generated using the same or similar technologies.

Cross-dataset generalizability. We then compare the AUC of these deepfake detection methods across different datasets to test their generalizability. This means that the training set is from the FF++ dataset, while the testing set is from the Celeb-DF dataset. As can be seen from the second column of Table 1, our CDNet can obtain the best AUC score on the Celeb-DF dataset. This demonstrates that our CDNet has the best deepfake detection generalizability.

Model size. Previous models usually construct complex structures or strategies to improve generalizability. Compared with these models, our model is much smaller (only 5.6 M), which can be seen from the fifth column of Table 1. This advantage makes our model friendly to be deployed on some terminal devices with limited computation and storage resources.

Table 2: Ablation study of the selective attention module (SAM) and cluster decision module (CDM).

SAM	CDM	FF++	Celeb-DF	Avg
		0.997	0.653	0.825
✓		0.972	0.674	0.823
	✓	0.992	0.757	0.874
✓	✓	0.979	**0.770**	**0.875**

4.3. Ablation Study

4.3.1. Effectiveness of SAM and CDM

We conduct experiments to study the effect of our selective attention module (SAM) module and cluster decision module (CDM). We choose Xception as the backbone when disabling SAM, and use a fully connected layer as classifier when disabling CDM. Table 2 shows the results of our ablation study.

SAM. By comparing the first and second rows in Table 2, one can observe that the performance on the FF++ dataset decreases and that on the Celeb-DF dataset increases when applying our SAM. This is because SAM discards some dataset-relevant information to improve the cross-dataset generalizability. However, these discarded information contains some discriminative features that are helpful for intra-data classification. As a result, SAM slightly reduces the intra-dataset performance but improves the cross-dataset generalizability.

CDM. By comparing the first and third rows in Table 2, one can observe that performance on the FF++ dataset drops slightly while that on the Celeb-DF dataset increases more than 0.1. This result demonstrates the superior performance of our cluster classifier compared to the traditional classifier using fully connected layers. The improvement is achieved from the following two aspects. (a) The difference of different samples has been enlarged. (b) The features used for decision are treated equally.

The fourth row shows the performance of our CDNet with both SAM and CDM. By comparing it with the other settings, one can observe that CDNet greatly improve the generalizability with a small cost on intra-dataset performance.

5. CONCLUSION

In this paper, we propose CDNet as a new deepfake detection model to improve the deepfake detection generalizability without using self-generated data. CDNet manually selects face features as the main feature source and designs a new cluster classifier based on contrastive learning. The selective attention design can ensure that our model can pay attention to the areas of the artifacts precisely and thus learn the artifact-relevant features. This design also leads to a very small model size (only 5.6M), which makes it suitable to be deployed in devices with limited resources. Our designed cluster classifier can enlarge the difference of the features between the real and fake samples in the unseen datasets. It also offers a novel decision making mechanism to address the disadvantage of the traditional binary classifier. Extensive experiments proved that our CDNet owns better generalizability than some state-of-the-art deepfake detection methods.

6. REFERENCES

[1] 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/CVF International Conference on Computer Vision (ICCV)*, 2019, pp. 1–11.

[2] Yuezun Li, Xin Yang, Pu Sun, Honggang Qi, and Siwei Lyu, "Celeb-df: A large-scale challenging dataset for deepfake forensics," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 3207–3216.

[3] Francois Chollet, "Xception: Deep Learning with Depthwise Separable Convolutions," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 1800–1807.

[4] Yuchen Luo, Yong Zhang, Junchi Yan, and Wei Liu, "Generalizing Face Forgery Detection with High-frequency Features," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021, pp. 16312–16321.

[5] Yuyang Qian, Guojun Yin, Lu Sheng, Zixuan Chen, and Jing Shao, "Thinking in Frequency: Face Forgery Detection by Mining Frequency-Aware Clues," in *European Conference on Computer Vision (ECCV)*.

[6] Lingzhi Li, Jianmin Bao, Ting Zhang, Hao Yang, Dong Chen, Fang Wen, and Baining Guo, "Face x-ray for more general face forgery detection," *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pp. 5001–5010, 2020.

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

[8] Minha Kim, Shahroz Tariq, and Simon S Woo, "Fretal: Generalizing deepfake detection using knowledge distillation and representation learning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition(CVPR)*, 2021, pp. 1001–1012.

[9] Shu Hu, Yuezun Li, and Siwei Lyu, "Exposing gan-generated faces using inconsistent corneal specular highlights," *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 2500–2504, 2021.

[10] Shahroz Tariq, Sangyup Lee, and Simon Woo, "One Detector to Rule Them All: Towards a General Deepfake Attack Detection Framework," in *Proceedings of the Web Conference 2021*, 2021, pp. 3625–3637.

[11] Hanqing Zhao, Wenbo Zhou, Dongdong Chen, Tianyi Wei, Weiming Zhang, and Nenghai Yu, "Multi-attentional deepfake detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021, pp. 2185–2194.

[12] Chengrui Wang and Weihong Deng, "Representative forgery mining for fake face detection," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition(CVPR)*, 2021, pp. 14923–14932.

[13] 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," *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6454–6463, 2021.

[14] Iacopo Masi, Aditya Killekar, Royston Marian Mascarenhas, Shenoy Pratik Gurudatt, and Wael AbdAlmageed, "Two-branch recurrent network for isolating deepfakes in videos," in *European Conference on Computer Vision (ECCV)*, 2020, pp. 667–684.

[15] Valentin Bazarevsky, Yury Kartynnik, Andrey Vakunov, Karthik Raveendran, and Matthias Grundmann, "Blazeface: Sub-millisecond neural face detection on mobile gpus," *arXiv preprint arXiv:1907.05047*, 2019.

[16] Sara Concas, Gianpaolo Perelli, Gian Luca Marcialis, and Giovanni Puglisi, "Tensor-based deepfake detection in scaled and compressed images," in *2022 IEEE International Conference on Image Processing (ICIP)*, 2022, pp. 3121–3125.

[17] Frank Joel, Eisenhofer Thorsten, Schönherr Lea, Fischer Asja, Kolossa Dorothea, and Holz Thorsten, "Leveraging frequency analysis for deep fake image recognition," *International Conference on Machine Learning (ICML)*, pp. 3247–3258, 2020.

[18] Peng Zhou, Xintong Han, Vlad I. Morariu, and Larry S. Davis, "Two-Stream Neural Networks for Tampered Face Detection," in *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2017, pp. 1831–1839.

[19] Tianchen Zhao, Xiang Xu, Mingze Xu, Hui Ding, Yuanjun Xiong, and Wei Xia, "Learning Self-Consistency for Deepfake Detection," in *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, Montreal, QC, Canada, Oct. 2021, pp. 15003–15013, IEEE.

[20] Mingxing Tan and Quoc V Le, "EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks," *Proceedings of the 36th International Conference on Machine Learning (ICML)*, pp. 6105–6114, 2019.

[21] Honggu Liu, Xiaodan Li, Wenbo Zhou, Yuefeng Chen, Yuan He, Hui Xue, Weiming Zhang, and Nenghai Yu, "Spatial-phase shallow learning: rethinking face forgery detection in frequency domain," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021, pp. 772–781.