Fully Unsupervised Deepfake Video Detection Via Enhanced Contrastive Learning

Tong Qiao, Shichuang Xie, Yanli Chen, Florent Retraint, and Xiangyang Luo

***Abstract*—Nowadays, Deepfake videos are widely spread over the Internet, which severely impairs the public trustworthiness and social security. Although more and more reliable detectors have recently sprung up for resisting against that new-emerging tampering technique, some challengeable issues still need to be addressed, such that most of Deepfake video detectors under the framework of the supervised mechanism require a large scale of samples with accurate labels for training. When the amount of the training samples with the true labels are not enough or the training data are maliciously poisoned by adversaries, the supervised classifier is probably not reliable for detection. To tackle that tough issue, it is proposed to design a fully unsupervised Deepfake detector. In particular, in the whole procedure of training or testing, we have no idea of any information about the true labels of samples. First, we novelly design a pseudo-label generator for labeling the training samples, where the traditional hand-crafted features are used to characterize both types of samples. Second, the training samples with the pseudo-labels are fed into the proposed enhanced contrastive learner, in which the discriminative features are further extracted and continually refined by iteration on the guidance of the contrastive loss. Last, relying on the inter-frame correlation, we complete the final binary classification between real and fake videos. A large scale of experimental results empirically verify the effectiveness of our proposed unsupervised Deepfake detector on the benchmark datasets including FF++, Celeb-DF, DFD, DFDC, and UADFV. Furthermore, our proposed well-performed detector is superior to the current unsupervised method, and comparable to the baseline supervised methods. More importantly, when facing the problem of the labeled data poisoned by malicious adversaries or insufficient data for training, our proposed unsupervised Deepfake detector performs its powerful superiority.**

***Index Terms*—Contrastive learning, data augmentation, Deepfake detection, pseudo-label.**

I. Introduction

With the advancement of deep learning technique, many challengeable studies have achieved the unprecedented success in the community of artificial intelligence and computer vision. On the downside, more and more new multimedia forgery paradigms have emerged endlessly, leading to the severe political threats and social problems as the abuse of them proliferates over the social network platform. Undoubtedly, Deepfake, which is generally defined as fake multimedia synthesized by a deep learning technique, has recently become one of the most fashionable video forgery manners [1]. One can replace the face of a person in the true video with the face of another person. For clarity, by collecting some source facial images and video clips of the target person, the deep learning technique can easily complete the task of fake facial synthesization.

In the past, the traditional hand-crated manipulations such as resampling [2], splicing [3], copy-moving [4], and etc., usually require the professional tampering tools, and meanwhile the malicious attacker who might have some anti-forensic background often carries out the post-processing operation, in order to perfectly hide the tampering traces. However, currently, various manipulation algorithms, mainly targeting face forgery, have been designed such as "Deepfake" (DF) [5], "FaceSwap" (FS) [5], "Face2Face" (F2F) [6], and "NeuralTextures" (NT) [7] (see Fig. 1 for illustration). Moreover, some new modern forgery tools installed in the smart phone, such as FaceApp [8] and ZAO [9], largely simplify the procedure of face tampering. Basically, different from the traditional forgery manners, new automatic end-to-end forgery method indeed gradually supersedes its forerunner in the aspect of realism and efficiency.

To address that challenge, many reliable Deepfake detectors have been devised. Specifically, many detectors focus on the synthesized facial image traces, where the establishments of the hand-crafted feature extraction are inspired by the traditional image forensics, such Principal Component Analysis (PCA) [10], Local Binary Pattern (LBP) [11], image quality [12], or textural traces [13]. Meanwhile, the biological features are also helpful for distinguishing between real and Deepfake video, such as blinking frequency [14], head pose [15], heart ratio [16]. Recently, most researchers pay more attention on the Deep Neural Networks (DNN). The design of diverse feature extraction

Manuscript received 16 August 2023; revised 11 January 2024; accepted 17 January 2024. Date of publication 22 January 2024; date of current version 5 June 2024. This work was supported in part by the Zhejiang Provincial Natural Science Foundation of China under Grants LZ23F020006 and LTGG24F020008, in part by the National Natural Science Foundation of China under Grants U23A20305 and 62172435, and in part by the National Key Research and Development Program of China under Grant 2022YFB3102900.e Recommended for acceptance by C.G.M. Snoek. *(Corresponding author: Xiangyang Luo.)*

Tong Qiao is with the School of Cyberspace, Hangzhou Dianzi University, Hangzhou 310005, China, and also with the State Key Laboratory of Mathematical Engineering and Advanced Computing, Zhengzhou Science and Technology Institute, Zhengzhou 450064, China (e-mail: tong.qiao@hdu.edu.cn).

Shichuang Xie and Yanli Chen are with the School of Cyberspace, Hangzhou Dianzi University, Hangzhou 310005, China (e-mail: shichuang_xie@163.com; yanli.chen@hdu.edu.cn).

Florent Retraint is with the Laboratory of Computer Science and Digital Society, University of Technology of Troyes, 10300 Troyes, France (e-mail: florent.retraint@utt.fr).

Xiangyang Luo is with the State Key Laboratory of Mathematical Engineering and Advanced Computing, Zhengzhou Science and Technology Institute, Zhengzhou 450064, China (e-mail: xiangyangluo@126.com).

Our source codes have been released at https://github.com/bestalllen/Unsupervised_DF_Detection/.

This article has supplementary downloadable material available at https://doi.org/10.1109/TPAMI.2024.3356814, provided by the authors.

Digital Object Identifier 10.1109/TPAMI.2024.3356814

Fig. 1. Illustration of real and fake samples. Left: the identity is falsified from fake video generated by "Deepfake" (DF) or "Faceswap" (FS); right: the expression is tampered from fake videos generated by "Face2Face" (F2F) or "NeuralTextures" (NT).

networks and representation learning methods indeed bring the remarkable improvement of detection accuracy [17], [18], [19], [20], [21], [22], [23]. However, most of current detectors are established under the framework of the supervised mechanism, which require a large scale of samples with accurate labels for training. When the amount of the training samples with accurate labels are not enough or the training samples are incorrectly labeled by malicious adversaries, the supervised classifier probably becomes unreliable for Deepfake detection. Besides, most well-trained efficient detectors possibly suffer from a lack of generalization capability, resulting in the performance degradation when tackling the unknown Deepfake videos [24].

To address the aforementioned issues, to our knowledge, the authors of [25] first proposed to establish the unsupervised framework of Deepfake detection, which indeed opens a new way in this research field. In particular, inspired by the study of source camera identification [26], the core idea of [25] is that facial region from the real video is captured by digital imaging device such as digital camera or smartphone while the synthesized facial region from the Deepfake video is from computer software; the different origins of two kinds of videos are perfectly characterized by the different noise patterns. Furthermore, dependent of two very effective noise patterns, referring to Photo-Response Non-Uniformity (PRNU) [27] and noiseprint [28], two rounds of clustering, involving the first multi-classification and the second binary classification, are effectively conducted in sequence. However, the detection accuracy and generalization performance need to be further improved. Besides, another literature [29] also tried to investigate the possibility of unsupervised detection for Deepfake video. Relying on the framework of contrastive learning, the trained feature extractor on the upstream is actually unsupervised while the classifier on the downstream still needs to be trained with the known true-labeled data. Thus, without knowing any prior information about training samples, it hardly holds true that the proposed method of [29] can realize the fully unsupervised detection.

Nevertheless, both pioneer studies [25], [29] indeed inspire us in this context to continue addressing one of the most challengeable scenarios, referring to as *fully unsupervised detection*. It should be noted that we design a fully unsupervised Deepfake detector; in the whole procedure of feature extraction or classification, one has no idea of any information about the true correct labels of samples. For clarity and simplicity, in this paper, the main contributions of our proposed unsupervised detection mechanism include:

- A novel unsupervised Deepfake detection method is proposed; regardless of the training or testing procedure, both feature extractor and binary classifier cannot acquire the true labels of the data.
- A pseudo-label generator is designed, in which the traditional hand-crafted features are adopted for characterizing both types of samples, in order to assign the pseudo-labels (0 or 1) to two primitive clusters. The samples from the same cluster should be assigned the same pseudo-labels, regardless of real or fake.[1]
- It is proposed to design the enhanced contrastive learner, where the discriminative features of the data with pseudo-label are continually refined by iteration on the guidance of the contrastive loss. Moreover, the well-designed scheme of data augmentation is proposed, which further improves the performance of discriminative feature extraction.
- The extensive experimental results empirically verify the effectiveness and superiority of our proposed unsupervised method on the benchmark datasets. On the one hand, compared with the prior unsupervised detector, our method performs its higher detection accuracy; On the other hand, compared with the baseline supervised detectors, our unsupervised detector rivals them. More importantly, when facing the problem of the labeled data poisoned by malicious adversaries or insufficient data for training, our unsupervised detector performs its powerful superiority.

The rest of this paper is organized as follows. First of all, Section II reviews the prior-art methods and Section III mainly describes the general framework of our proposed unsupervised detection towards Deepfake. Next, Section IV mainly elaborates the specific procedure of establishing the pseudo-label generator. Next, it is proposed to establish the enhanced contrastive learner in Section V, together with the well-designed scheme of data augmentation. Then, the binary classifier is devised in Section VI. Section VII demonstrates and analyzes the extensive experimental results. Finally, let us draw the conclusion in Section VIII.

II. Related Work

Generally, it is proposed to divide the current Deepfake detectors into three categories: based on traditional hand-crafted features, based on biometric features, and based on deep learning.

In the first type of methods, the researchers mainly focus on the idea of image forensics and analyzes the video frames,

[1]In this context, the real sample denotes the real video clip, and the fake sample denotes the fake video clip.

with regards to PCA [10], LBP [11], image quality [12], or textural traces [13], eyes or teeth artifacts [30], in which Support Vector Machine (SVM) is usually trained with the extracted hand-crafted feature vectors.

In the second type of methods, the biological characteristics mainly serve as the unique biometric information for Deepfake detection. [14] proposes to carry out the task of detection using the physiological signal of blinking. By estimating the movement of 3D head poses, [15] successfully reveals the abnormal traces of facial region caused by Deepfake video. Besides, [16] makes full use of heart rate bio-signals to expose the artifacts of Deepfake video. [31] proposes an efficient method through temporal modeling on precise geometric features. Aiming at high-level semantic irregularities in mouth movements, [32] proposes a spatiotemporal network to learn internal representations associated with natural mouth movements for identifying Deepfake videos. In addition, targeting the specific individuals, [33] designs the Deepfake detector based on the biometric information. However, the weaknesses of current generation models are possibly overcome in the next generation of Deepfake.

In the last type of methods, the detectors are established relying on the feature extracted from Deep Neural Networks (DNN). For instance, the Recurrent Neural Network (RNN) incorporates temporal feature to detect Deepfake videos [17]. [18] designs a lightweight DNN with a small number of layers. By combining Convolutional Neural Network (CNN) and RNN, [34] proposes to distinguish between real and fake facial region. Next, [35] designs the representative capsule network by digging out the position discrimination features existing between two types of videos. Currently, the design of two-stream network for Deepfake detection becomes more and more popular. [36] utilizes GoogLeNet and triplet network to construct a two-stream network architecture. Besides, [37] exploits a two-stream network by fusing the RGB color space and the spatial rich model feature [38]. [39] aims to localize the blending boundary in a self-supervised mechanism for Deepfake videos detection. Next, [40] builds the direct mapping from instance embeddings to bag prediction for addressing the potential risks of training. [41] proposes a series of detection systems, where the generalizable and explainable performance of detection is improved. In the framework of multi-attentional mechanism, the detection performance is further improved [19]. By combining self-consistency with contrastive learning, [42] effectively solves the overfitting problem. Additionally, [43] proposes a weakly supervised Second Order Local Anomaly (SOLA) learning module to improve the generalization ability of detection.

Generally, it is worth noting that nearly all the detectors except [25] are devised under the the supervised mechanism. Once the training data are poisoned by the malicious attacker, referring to as assigning the incorrect labels, the high detection accuracy cannot be guaranteed. Thus, in this context, we propose the fully unsupervised method to address that challenging issue.

III. Preliminary

In this section, we intend to outline the overall scheme of the proposed detector. In general, we mainly propose the enhanced contrastive learning to address the problem of unsupervised Deepfake video detection. For simplicity and clarity, as Fig. 2 illustrates, the overall scheme generally involves three main stages.

1) On the stage 1, referring to as the upstream task, we intend to establish a pseudo-label generator, which is in charge of labeling the inquiry original samples, regardless of the types. In particular, by adopting the simple yet effective feature extractor, the traditional hand-crafted feature is directly used for roughly clustering. Last, relying on the predicted classification of primitive clustering, the pseudo-label generator assigns the binary labels to the inquiry original samples, where each cluster has the same labels. It should be addressed that in the whole procedure, we have no idea of the true labels of the inquiry original samples. Thus, we define the label from the pseudo-label generator as "pseudo-label".
2) On the stage 2, referring to as the downstream task, we utilize the data with pseudo-labels to train the enhanced contrastive learner, which can help us to convert the primitive feature on the stage 1 to the discriminative feature on this stage. It should be noted that during the enhanced contrastive learning, let us maximize the similarity between samples with the the same pseudo-labels while minimizing the similarity between samples with the different pseudo-labels. Meanwhile, in order to further refine the pseudo-label, we are prone to retain the confidence samples closer to the centroid, which are more possibly used for each round iteration. Last, through various data augmentation modes, the proposed enhanced contrastive learner is effectively trained on the guidance of the contrastive loss with pseudo-label information.
3) On the stage 3, referring to as binary classification and authentication, we intend to accomplish the testing task of distinguishing between real and fake videos; the well-trained encoder network of the enhanced contrastive learner on the stage 2 serves as a feature extractor, in order to extract the discriminative features of all the inquiry data. Next, without loss of generality, the extracted representation are used for binary clustering by Kmeans algorithm, where the predicted labels are accurately assigned. Within each cluster, it is proposed to calculate the inter-frame correlation of each video sample, in which the cluster with the smaller average correlation as authentic video while the larger one as Deepfake video.

It should be noted that neither real nor fake samples are manually labeled prior to training from the stage 1 to the stage 3. Meanwhile, on the whole procedure of Deepfake video detection, we cannot acquire any information about the inquiry data. Fortunately, under the framework of the proposed detection architecture, mainly relying on the proposed enhanced contrastive learning algorithm coupled with the well-designed pseudo-label generator, the task of unsupervised Deepfake detection can be successfully completed. In the following sections, we will specifically elaborate the proposed unsupervised Deepfake detector.

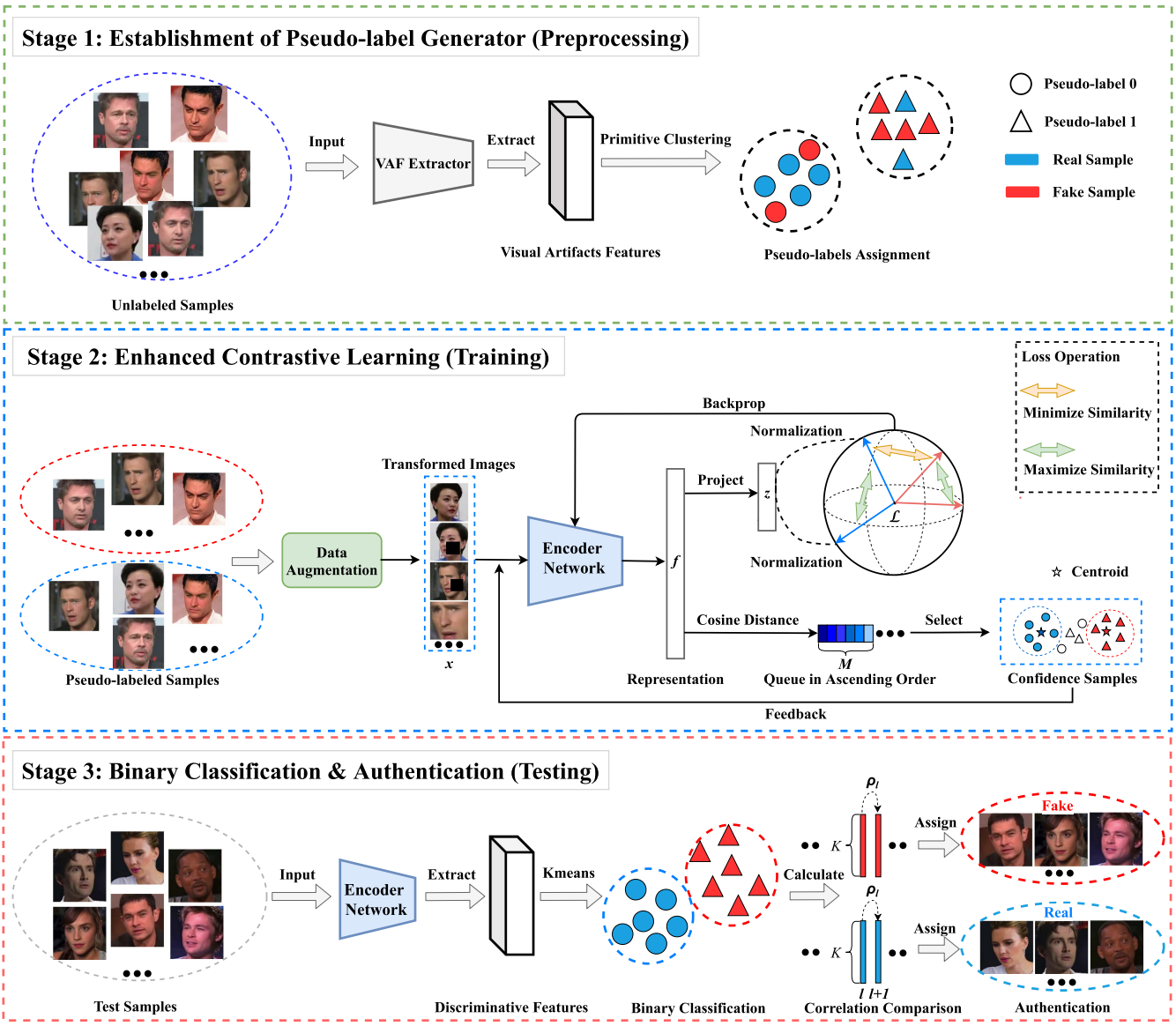


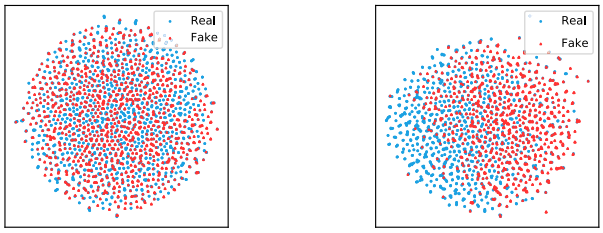
Fig. 2. Illustration of the proposed fully unsupervised framework: on the stage 1, the proposed pseudo-label generator is established, where the primitive clustering is completed; on the stage 2, the proposed enhanced contrastive learning is conducted, where the discriminative features are extracted; on the stage 3, based on the inter-frame correlation, binary classification and authentication are realized in order to distinguish between real and fake video.

IV. Establishment of Pseudo-Label Generator

In the downstream task, our proposed enhanced contrastive learning indeed requires both real and fake samples, even though the true labels are totally unknown in the training stage. Thus, it is proposed to assign the pseudo-labels to the training data. However, we cannot randomly assign labels, which is not helpful to the learning procedure. In our assumption, the learning procedure is successfully conducted based on the enough samples with basically correct labels. In other words, it is required to assign the pseudo-labels to the samples with to some degree clustering purity. To this end, it is proposed to primitively cluster the training samples, and then two primitive clusters are assigned with pseudo-labels. Moreover, the assigned pseudo-labels are only used to distinguish two types of clusters, where we cannot identify which cluster is real or fake since that any prior information about data is unknown in our proposed unsupervised framework.

A. Primitive Clustering

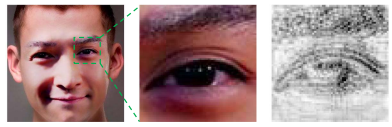
In this section, our task is to primitively cluster two types of data with to some degree purity. In fact, in the unsupervised clustering, the metric *purity* is used to evaluate the performance of clustering. Due to the requirement of the following enhanced contrastive learning, we have to guarantee the purity of the primitive clustering. To this end, it is proposed to extract the representative features which can effectively characterize the discriminations between two types of inquiry videos. Straightforward, we initially adopt the strategy of deep clustering [44],



(a) Purity 0.59 (b) Purity 0.74

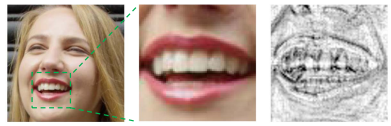
Fig. 3. T-SNE visualization comparison between DNN (a) and hand-crafted (b) features.

(a) Facial region from real sample (left), zoom-in eye region (middle), and VAF visualization (right)

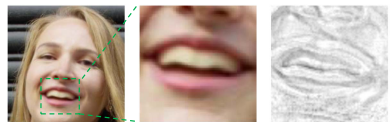


(b) Facial region from fake sample (left), zoom-in eye region (middle), and VAF visualization (right)

Fig. 4. Visual artifact comparison of eye region between real and fake sample.



(a) Facial region from real sample (left), zoom-in mouth region (middle), and VAF visualization (right)



(b) Facial region from fake sample (left) and zoom-in mouth region (middle), and VAF visualization (right)

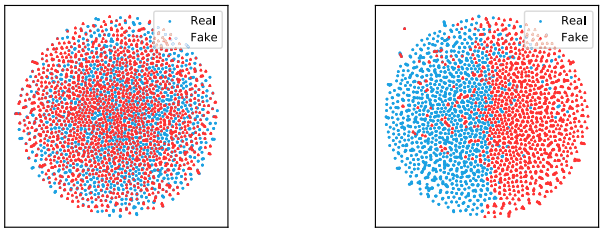
Fig. 5. Visual artifact comparison of mouth region between real and fake sample.

which is originally designed for multiply clustering various types of images mainly through a pre-trained DNN model. The core idea of deep clustering is to map original high-dimensional data into the low-dimensional feature space. The representations by pre-training are then fine-tuned via the compositive loss functions, to make them more discriminative in the procedure of iteration.

However, in our task of binary clustering for real and fake video, that strategy cannot be feasible. On the one hand, the pre-trained model is not applied to our specific unsupervised task; on the other hand, two types of videos behave very similarly which is hardly to be distinguished visually. Otherwise, the following iteration and optimization become invalid. In this context, we need to extract the discriminative features used for clustering, which are probably hidden in the shallow layers while not in the deep layers. Basically, regardless of real or fake video, the face attribute from two types of videos is nearly consistent, which belongs to the same category, leading to the similar high-level feature. To address the problem of Deepfake detection, let us extract features in the semantic level, not in the understanding level as the classical task of the image classification. To this end, we resort to the traditional hand-crafted feature extraction. Compared with the advanced deep clustering [44], the hand-crafted features are likely appropriate for our proposed primitive clustering. As Fig. 3 illustrates, the hand-crafted features with higher purity, imply the better pseudo-label assignment; on the contrary, the DNN features with lower purity perform worse.

Without loss of generality, many hand-crafted features have been proposed to deal with the problem of distinguishing between real and Deepfake video, especially in the early era of Deepfake detection. In fact, in the current study, the hand-crafted features are gradually replaced by the end-to-end extracted features from the DNN model, which can bring the remarkable improvement of detection rate in the supervised mechanism. However, in our proposed unsupervised framework, the hand-crafted features, such as Visual Artifact Feature (VAF) [45], are discriminative enough to complete the task of primitive clustering. Here, it should be noted that in the primitive clustering, the simple but effective feature extraction, which can meet the requirement of pseudo-label assignment, is our choice. Next, it is proposed to adopt the classical Kmeans algorithm for binary clustering.

Specifically, on this stage, we mainly exploit the eye and mouth regions for characterizing the discrimination caused by Deepfake, as shown in Figs. 4 and 5. It can be observed that the VAF of fake samples is visually different from that of real samples. It is proposed to adopt the algorithm of [45] to extract the visual artifact feature from eye and mouth region separately. Firstly, relying on the facial landmarks, the facial region is effectively cropped. It should be noted that all the facial regions are re-scaled to the same size, and both eye and mouth regions are successfully segmented according to the corresponding landmarks. It is worth noting that the segmented regions, referring to as eye and mouth, are used for feature extraction by the texture energy approach. Specifically, the visual artifact features are extracted by 16 fixed 5×5 convolution kernels. Last, by aggregating all the feature vectors from both regions, we can obtain the fused hand-crafted features for primitive clustering. In particular, it is proposed to adopt the simple but effective Kmeans algorithm to classify all the samples into two clusters, in which each cluster is assigned a pseudo-label, referring to as "0" or "1". Here, we cannot clearly distinguish which cluster contains real or fake samples.



(a) Purity 0.52 (b) Purity 0.85

Fig. 6. T-SNE visualization comparison of the features directly from contrastive learning, where the pseudo-label generator is not adopted (a) and the pseudo-label generator is adopted (b).

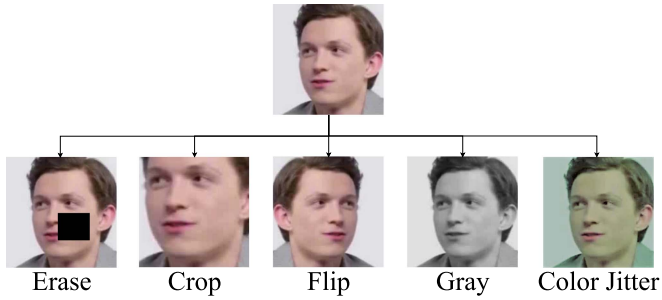


Fig. 7. Illustration of augmentation manners.

B. Pseudo-Label Generation

Next, let us assign the pseudo-labels to two clusters. In fact, we cannot guarantee that each cluster is perfectly assigned the correct labels. That is the reason why we define the label on this stage as pseudo-label while not true label. Nevertheless, it should be noted that even though the true labels are totally unknown, the following training on the stage 2 cannot be completely impacted. That is because the only target on this stage is to cluster two types; it is not required to clearly determine which cluster is real or fake. On the stage 3, referring to as "binary classification and authentication" (see Section VI for details), that tough issue can be effectively addressed relying on the correlation among inter-frames of inquiry videos. Thereafter, the samples with pseudo-labels are fed into the model of the proposed enhanced contrastive learning, in order to further refine the discriminative features.

Moreover, we would like to address the importance of the pseudo-label generator. The establishment of the pseudo-label generator serves for the following enhanced contrastive learning on the stage 2. In fact, in the traditional task of classification, the vanilla contrastive learning scheme [46] tries to search for the representative features among different objects, such as *airplane*, *cat*, *dog*, and etc., where the discriminative boundary is remarkable. However, in our challenging unsupervised Deepfake detection, the discriminative boundary between real and fake face behaves very obscurely. Therefore, only relying on the contrastive learning scheme, the features between real and fake samples are hardly distinguished (see Fig. 6(a) for illustration), implying the failure clustering; by adopting the proposed pseudo-label generator, the very promising results can efficiently help us further refine the features in the following steps (see Fig. 6(b) for illustration). In other words, the vanilla contrastive learning scheme cannot directly characterize the discrimination caused by Deepfake. Thus, we novelly propose to adopt the scheme of pseudo-label generation on the stage 1, leading to that the pre-processed data with pseudo-labels can further enhance the performance of contrastive learning. Even though the inquiry data with inaccurate labels occupy at a certain percentage on the stage 1, the powerful capability of contrastive learning on the stage 2, can effectively refine the more discriminative feature, meeting the requirement of binary classification and authentication on the stage 3.

V. Enhanced Contrastive Learning

In this section, our target is to further refine the rough features by the proposed enhanced contrastive learning. To our knowledge, the scheme of [46] is one of the most effective contrastive learning strategies. However, in our Deepfake detection, only dependent on the vanilla contrastive learning framework [46], we cannot directly refine the rough features from the stage 1 (See the specific analysis in Appendix A, available online). To tackle with that problem, it is proposed to feed the data with pseudo-labels on the stage 1, into our designed enhanced contrastive learner, where the well-designed data augmentation scheme, and the high-efficient contrastive approach customized for Deepfake detection, together with the strategy of confidence sample selection, are introduced.

A. Data Augmentation

In the vanilla contrastive learning, the augmentation manners are generally designed for image classification or object detection, which cannot be directly applied to deal with the problem of Deepfake detection. Without loss of generality, the main task of binary classification between real and Deepfake video is to tackle the problem of the more close discrimination boundary in the feature space. The selection of augmentation manner is prone to more diverse, leading to more discriminative features.

In addition, the operation of data augmentation indeed plays an important role dealing with the problem of feature description. In fact, the contrastive learning requires much stronger data augmentation than supervised learning [46]. In other words, the unsupervised learning benefits more than supervised learning from data augmentation. Based on the aforementioned discussion, we need carefully and empirically address the importance of data augmentation in this section.

Specifically, on this stage, let us adopt five data augmentation manners: *erase*, *crop*, *flip*, *gray*, and *color jitter*, as shown in Fig. 7. The specific procedure is as follows:

- *Erase:* The local region of the facial image is randomly erased, where the pixel intensity equals to zero.
- *Crop:* The portion of the facial image is randomly cropped, and then the cropped region is re-scaled to the original size.
- *Flip:* The original facial image is flipped horizontally.
- *Gray:* The facial color image is converted to grayscale image.

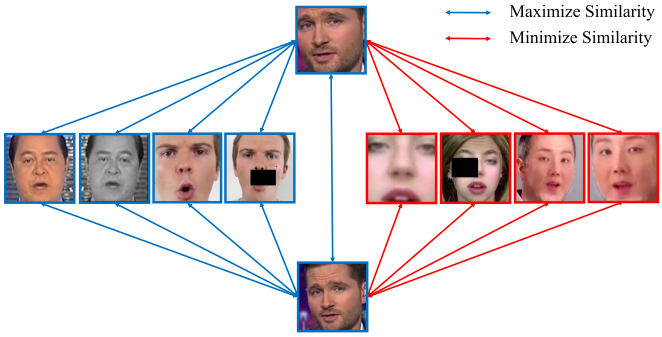


Fig. 8. Illustration of similarity operation between samples, where the distance of the samples with the same pseudo-labels is minimized, and the distance of the samples with different pseudo-labels is maximized.

- *Color jitter:* The properties of the facial image, referring to as brightness, contrast, saturation and hue, are randomly changed.

For instance, for the input images with the batch size N, we randomly select two different augmentation manners to obtain $2\,N$ images. Then the augmented images are fed into the encoder network for obtaining the high-dimensional $2\,N$ representation vectors. Last, all the representation vectors are fed to the projection layer for outputting low-dimensional features.

B. Enhanced Contrastive Learner

The core idea of contrastive learning is to learn the representation of samples by comparing real samples with fake samples in the embedding space, in which the distance between samples from the same attribute is minimized, and the distance between samples from the different attributes is maximized. The main difficulty lies in how to construct the samples with attribute information. In particular, we use pseudo-labels on the stage 1 to provide attribute information for our proposed enhanced contrastive learning. For clarity, as Fig. 8 illustrates, we demonstrate the main operation in the contrastive learning framework of constructing samples with pseudo-labels. In a batch, we minimize the distance between samples with the same pseudo-labels while maximizing the distance between samples with different pseudo-labels. More details are elaborated in the design of loss function. Next, let us extend the specific procedure of enhanced contrastive learning proposed in this context.

1) Encoder Network: Specifically, let us denote the encoder network as $f_{enc}(\cdot)$, in order to extract representation from data augmented image x. Then the high-dimensional representation vector $f = f_{enc}(x)$ can be obtained by the convolution and pooling operation of the encoder network. Next, the representation features are fed into the projection head, where the vector dimension is reduced. The network is a linear stack of depthwise separable convolutional modules with residual connections, which enables us completely decouple the mapping of cross-channel correlation and spatial correlation in the feature map of the encoder.

2) Projection Head: The high-dimensional features bring high computation cost, especially during the procedure of loss computation on the training phase. Thus, it is proposed to introduce the module of projection head. The projection head, namely a multi-layer perceptron denoted as $g(\cdot)$, consists of a linear layer with 2048 neurons, a ReLu layer, and a linear layer with 128 neurons, in which all the layers are stacked. In particular, the projection head maps high-dimensional representations into low-dimensional features, denoted as $z = g(f)$. The features z are used to calculate the contrastive loss after normalization (see (3)). It should be noted that on the stage 3, referring to testing phase, we adopt the high-dimensional representations directly from the trained encoder network, in order to retain the rich discrimination. Thus, this projection head will be discarded on the stage 3.

3) Confidence Sample Selection: In order to further improve the purity of each cluster, we novelly propose to select the confidence samples for training the encoder network, which is evaluated by cosine distance. In details, the representation vector of each cluster is assigned with pseudo-labels. For instance, f_i^{pl0}, $i \in \{1, \ldots, I\}$ represents f with "pseudo-label 0", extracted from the encoder network; f_j^{pl1}, $j \in \{1, \ldots, J\}$ represents f with "pseudo-label 1", extracted from the encoder network. It is worth noting that all the representations are divided into two clusters, where the summation of I and J is $2\,N$. Next, for all the vectors $\{f_i^{pl0}\}$ in the cluster with pseudo-label 0, the cosine distance $D_i^{\cos}$ between f_i^{pl0} and the centroid f_c^{pl0}, $c \neq i$, in the cluster is straightforward formulated as:

$$D_i^{\cos} = 1 - \frac{f_i^{pl0} \cdot f_c^{pl0}}{\| f_i^{pl0} \| \| f_c^{pl0} \|} \tag{1}$$

where $||\cdot||$ denotes L2 norm calculation, and I cosine distances are obtained in total. Then let us sort $\{D_i^{\cos}\}$ in descending order to form a queue in each cluster, where the confidence samples are selected based on the top M^{pl0} cosine distances in the queue. For the cluster with pseudo-label 1, the same operation is conducted for obtaining M^{pl1} selected confidence samples. In the next round training, the $M \leq 2\,N$ samples in total are fed into the encoder network, formulated as:

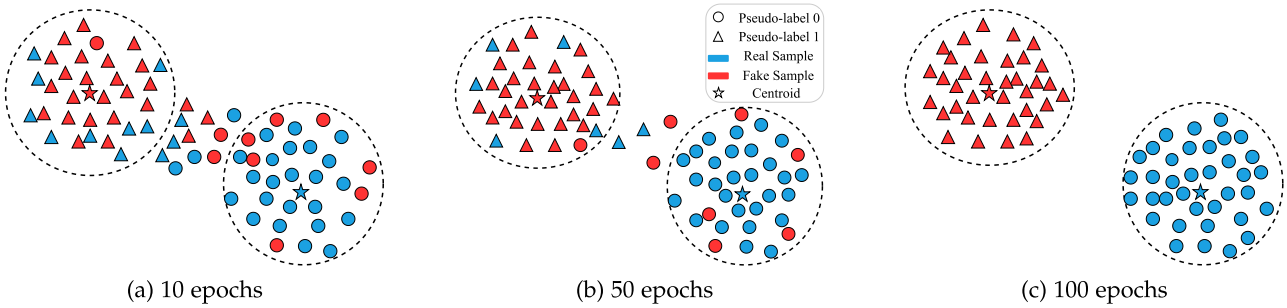
$$M = M^{pl0} + M^{pl1} \tag{2}$$

4) Loss Function: It is proposed to establish the loss function customized for our proposed unsupervised detection framework. In particular, the total loss consists of two sub-losses, respectively from the cluster $\{z_i^{pl0}\}$, $i \in \{1, \ldots, M^{pl0}\}$ with pseudo-label 0 and the cluster $\{z_j^{pl1}\}$, $j \in \{1, \ldots, M^{pl1}\}$ with pseudo-label 1. Straightforward, the total loss function $\mathcal{L}$ for the enhanced contrastive learning, is formulated as:

$$\mathcal{L} = \mathcal{L}^{pl0} + \mathcal{L}^{pl1} \tag{3}$$

where $\mathcal{L}^{pl0}$ denotes the sub-loss function from the cluster with pseudo-label 0, which is calculated by:

$$\mathcal{L}^{pl0} = \sum_{i=1}^{M^{pl0}} \frac{-1}{M^{pl0} - 1} \mathcal{L}_i^{pl0},$$



(a) 10 epochs (b) 50 epochs (c) 100 epochs

Fig. 9. Performance visualization of enhanced contrastive learner at different epochs on the stage 2.

$$\mathcal{L}_i^{pl0} = \log \frac{\sum_{i \neq k} \exp(\frac{\text{sim}(z_i^{pl0}, z_k^{pl0})}{T})}{\sum_{i \neq k} \exp(\frac{\text{sim}(z_i^{pl0}, z_k^{pl0})}{T}) + \sum_{j=1}^{M^{pl1}} \exp(\frac{\text{sim}(z_i^{pl0}, z_j^{pl1})}{T})} \tag{4}$$

where $\text{sim}(\cdot)$ is defined as the operation of dot product, denoting the similarity illustrated in Fig. 8. T denotes the temperature value. Similarly, the another sub-loss function from the cluster with pseudo-label 1 is formulated as:

$$\mathcal{L}^{pl1} = \sum_{j=1}^{M^{pl1}} \frac{-1}{M^{pl1} - 1} \mathcal{L}_j^{pl1},$$

$$\mathcal{L}_j^{pl1} = \log \frac{\sum_{j \neq k} \exp(\frac{\text{sim}(z_j^{pl1}, z_k^{pl1})}{T})}{\sum_{j \neq k} \exp(\frac{\text{sim}(z_j^{pl1}, z_k^{pl1})}{T}) + \sum_{i=1}^{M^{pl0}} \exp(\frac{\text{sim}(z_j^{pl1}, z_i^{pl0})}{T})}. \tag{5}$$

In fact, when the purity of each cluster is not very satisfying on the stage 1, that probably misleads the binary classification and authentication on the stage 3. Thus, on the stage 2, in order to refine the discriminative features, we exploit the advantage of enhanced contrastive learning, which can make the representation of samples from each cluster distribute more densely in the embedding space. That effectively mitigates the impact from noisy pseudo-labels. Furthermore, we select confidence samples to strengthen the robustness of our proposed contrastive learning. For clarity, it is proposed to illustrate the performance visualization of enhanced contrastive learner in Fig. 9. As the increment of iteration times, the purity of each cluster is further optimized, implying that the extracted features are more discriminative serving for the task on the stage 3.

VI. Binary Classification & Authentication

In this section, let us establish a simple but very effective binary classifier and authentication for completing the task of distinguishing between real and fake videos. In fact, when the discriminative features are successfully extracted, we cannot acquire the classification results, since that the true labels of the features cannot be provided on the whole procedure of classification. Then, it is proposed to distinguish two clusters of videos based on the inter-frames correlation. As Fig. 10 illustrates, in the real video clip, the inter-frames behave very

(a) Consecutive inter-frames extracted from a real video clip

(b) Consecutive inter-frames extracted from a fake video clip

Fig. 10. Comparison of inter-frames from both real and fake videos. (a) The natural face movement between inter-frames from real video clip; (b) the noticeable artifact between inter-frames from fake video clip.

natural and no manipulated traces are visualized; by contrast, in the fake video clip, the noticeable artifact can be easily captured. Then the inter-frame correlation can be adopted to authenticate the inter-frames from different types of videos.

The encoder network trained on the stage 2, is used to extract discriminative features on the stage 3, and the Kmeans algorithm is also adopted to cluster two types of videos. It should be noted that we adopt the high-dimensional representations directly from the trained encoder network while discarding the module of projection head, in order to retain the rich discrimination. After binary classification, we conduct the procedure of authentication. Specifically, for a video clip with L frames, $\mathbf{f}_l = \{f_{l,1}, f_{l,2}, \ldots, f_{l,K}\}$ represent the extracted feature vector from the frame l, where K denotes the dimension of the feature vector. Here, it is proposed to calculate the inter-frame correlation by the spearman correlation, which is formulated as:

$$\rho_l = 1 - \frac{6 \sum_{k=1}^{K} (f_{l,k} - f_{l+1,k})^2}{K(K^2 - 1)} \tag{6}$$

where the correlation value of frame l is mainly evaluated by the differences between the adjacent inter-frame vector, referring to

$\mathbf{f}_l$ and $\mathbf{f}_{l+1}$. Next, we need calculate the average value of all the frames correlation:

$$\bar{\rho} = \frac{\sum_{1}^{L-1} \rho_l}{L-1} \tag{7}$$

In the procedure of the final authentication, we select the correlation $\bar{\rho}$ from more than half of the inquiry videos in each cluster, where its average result is calculated, denoted as *representative correlation* of each cluster. Then by comparing the representative correlation between two clusters, we assign the cluster with the smaller inter-frame correlation as the real class, and the cluster with the larger inter-frame correlation as the fake class.

VII. Experimental Results

To verify the effectiveness of the proposed unsupervised Deepfake detector, let us conduct the numerical experiments in this section. First, it is proposed to describe the benchmark datasets used in our evaluation experiments, and the implementation details are also elaborated. Next, it is proposed to conduct the ablation studies for addressing the importance of data augmentation and enhanced contrastive learner on the stage 2, and meanwhile we also evaluate the module of projection head, the selection of confidence sample and backbone network. Next, we mainly compare the proposed unsupervised Deepfake detector with the baseline methods including the supervised and unsupervised detectors. Moreover, it should be noted that we first propose to consider two practical scenarios, referring to as "training data with correct labels" and "training data with incorrect labels", in order to comprehensively verify the superiority of our proposed unsupervised method. Additionally, the generalization capability and robustness performance are also evaluated.

A. Datasets

We conduct extensive experiments on the current benchmark datasets: FaceForensics++ (FF++) [5], UADFV [15], Celeb-DF [47], DeepFake Detection Challenge (DFDC) [48], and DeepFake Detection (DFD) [49]. Specifically, FF++ is a representative large-scale forgery face dataset. It contains 1,000 original realistic videos downloaded from the YouTube-8 M dataset [50], and 4,000 fake videos generated by four different face synthesis approaches including two deep leaning-based methods Deepfake (DF) and NeuralTextures (NT) and two graphics-based methods Face2Face (F2F) and FaceSwap (FS). It should be noted that for fair comparison, the providers of FF++ dataset strictly regulate the ratio of training, validation, and testing as 720:140:140. Besides, this dataset contains uncompressed (C_{00}), H.264 compressed format with high quality (C_{23}) and low quality (C_{40}), in order to simulate the practical scenarios over the social network platform.

UADFV contains relevant small data, consisting of 49 real videos and 49 fake videos. Celeb-DF contains 590 real videos and 5,639 more realistic synthetic videos. DFD is released as a complement to the FF++ dataset, containing 363 real videos and 3,068 fake videos. It should be noted the forgery manner in the aforementioned three datasets only adopts DF.

The very large scale of DFDC dataset is released for the DeepFake Detection Challenge containing 128,154 videos in total, and 5 facial forgery manners. It is worth noting that in DFDC, it mainly consists of low quality videos and the face region of most videos is very small, making the detection exceptionally challenging. In our evaluation experiments, by considering the computation cost, we randomly select 1,000 manipulated videos and 1,000 real videos.

B. Implementation Details

1) Data Preprocessing: For each inquiry video, we randomly select 32 frames per video, and crop the face region by the commonly-used method MTCNN [51], and then resize the cropped face region as the size 299×299, in order to match the input layer of the backbone network.

2) Data Augmentation: We use five standard data augmentation manners as Fig. 7 illustrates, including erase, crop, flip, gray, and color jitter. For erase augmentation, the scale factor (0.02, 0.2) and the aspect ratio (0.5, 2) determine the size of the erasing region. For crop augmentation, the scale factor (0.5, 1) and the aspect ratio (0.9, 1.1) control the cropped region of the image. For flip augmentation, the horizontal flip is carried out. For gray augmentation, the gray level image is transformed from the color image. For color jitter augmentation, the parameters of brightness, contrast, saturation and hue follow (0.4, 0.4, 0.4, 0.1), respectively.

3) Enhanced Contrastive Learning: In our enhanced contrastive learning, the adopted encoder network is trained for 1,000 epochs using Adam optimizer. The learning rate is set as 0.0001, and the batch size is set as 64. The temperature parameter T of (4) is set as 0.07.

4) Evaluation Metric: We report the widely-used Area Under Curve (AUC) as the main metric, which is calculated by the Receiver Operating Characteristic (ROC) curve.

C. Ablation Study of Our Proposed Unsupervised Detector

In this context, the main contribution of our proposed method is to leverage the enhanced contrastive learning. To address the importance of it, in this subsection, we mainly discuss the effectiveness of the proposed unsupervised framework on the stage 2, where the scheme of data augmentation is adopted and the enhanced contrastive learner is designed. Then the ablation studies are conducted to demonstrate the detection results. Additionally, we propose to evaluate the effectiveness of the proposed scheme of discarding the module of projection head on the stage 3. Besides, it is worth noting that the only DF forgery is discussed in this subsection.

1) Ablation Study of Data Augmentation: To comprehensively evaluate the effectiveness of the various data augmentation manners, we conduct the ablation experiments on the FF++ dataset. As Table I illustrates, with increasing the types of augmentation manners, the detection performance is gradually improved. It can be observed that the manner of erase shows its remarkable gain to detection results, which helps the AUC

TABLE I
ABLATION STUDY OF DATA AUGMENTATION ON THE FF++ DATASET

Data Augmentation					AUC
Crop	Flip	Color Jitter	Erase	Gray	
✓					0.81
✓	✓				0.82
✓	✓	✓			0.83
✓	✓	✓	✓		0.99
✓	✓	✓	✓	✓	**1.00**

TABLE II
ABLATION STUDY OF ENHANCED CONTRASTIVE LEARNER ON THE STAGE 2 ON THE FF++ DATASET

Pseudo-label Generator	Enhanced Contrastive Learner			AUC
	Data Augmentation	Contrastive Learning	Confidence Sample Selection	
	✓	✓		0.55
✓				0.71
✓	✓	✓		0.91
✓	✓	✓	✓	**1.00**

from 0.83 to 0.99. Moreover, with adopting all the proposed augmentation manners, our proposed unsupervised detector can achieve 100% detection accuracy.

2) Ablation Study of Enhanced Contrastive Learner: To further explore the impact of different modules of our proposed method, it is proposed to split each step separately for comprehensive validation. In this ablation study, we still adopt the FF++ dataset for completing the task of Deepfake detection. The specific comparison results are illustrated in Table II. By observation, if the vanilla contrastive learning with data augmentation is deployed, the very low detection rate 0.55 indicates that the unsupervised detector nearly becomes invalid, which meanwhile empirically verifies the results of Fig. 6 (See the specific analysis in Appendix A, available online). Similarly, if the only module of pseudo-label generator is deployed, the detection accuracy only arrives at 0.71. Noticeably, as the aforementioned modules are integrated together, the detection results are remarkably improved, reaching to as high as 0.91. Moreover, the module of confidence sample selection further helps the proposed unsupervised detector improve the detection accuracy.

3) Ablation Study of Projection Head: On the stage 2, it is proposed to adopt the module of projection head to improve the efficiency of loss computation. On the stage 3, we discard it to retain the rich discriminative feature extracted from the trained encoder network. Then let us empirically verify the effectiveness of our design. Straightforward, we conduct the comparison experiments on the various datasets, referring to FF++, DFD, Celeb-DF, UADFV. As Fig. 11 illustrates, when testing the inquiry samples, as the features extracted from the trained encoder network while not from the projection head, the detection results are better.

D. Selection of Confidence Sample

In the training phase, we propose to adopt the module of confidence sample selection, in order to further refine the pseudo-label. To validate the effectiveness of the proposed strategy of confidence sample selection, let us carry out the visualization experiments with varying confidence levels on the DFD dataset. In our proposed strategy, we arbitrarily quantify the confidence into five levels, ranging from 1 (the highest confidence) to 5 (the lowest confidence). Within each cluster, the assigned level for each sample is determined by its distance from the centroid of the cluster. For instance, we arrange the samples in descending order. The top 20% of samples in the queue are assigned a confidence level 1; the remaining samples are assigned the confidence level at the interval 20%, corresponding to 2∼5.

In Fig. 12, it is proposed to illustrate the visualization results of the extracted features from the training samples with various confidence levels. By observation, the feature extractor can be better trained based on the samples with high confidence, leading to that the discriminative features from the facial regions gain much more attention measured by Grad-CAM [52]. In such case, the T-SNE visualization of the features between real and fake samples is more separated, which is illustrated in Fig. 13.

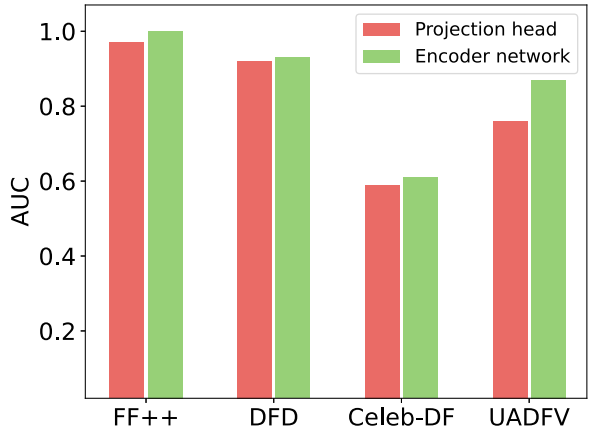


Fig. 11. Ablation study of projection head.

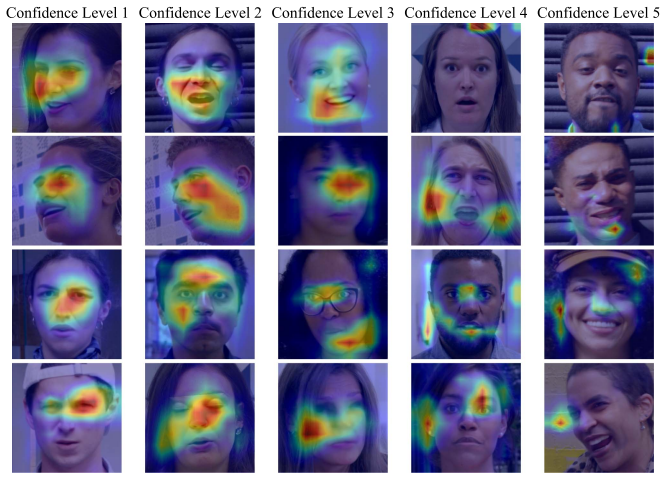


Fig. 12. Visualization analysis of facial region by Grad-CAM.

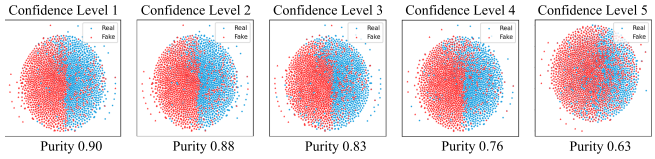


Fig. 13. Visualization analysis of features by T-SNE.

TABLE III
SELECTION OF BACKBONE NETWORK

Backbone Network	Dataset				Number of Training Samples			
	FF++	DFD	UADFV	Avg.	720	180	45	11
Xception [53]	**1.00**	0.93	0.89	0.94	**1.00**	0.99	**0.97**	**0.85**
Inception [54]	0.99	0.94	0.74	0.89	0.99	0.98	0.73	0.61
EfficientNet [55]	0.97	0.92	0.80	0.90	0.97	0.95	0.72	0.68
ViT [56]	0.95	0.96	0.96	0.96	0.95	0.99	0.94	0.64
MaxViT [57]	**1.00**	0.96	0.97	0.98	**1.00**	0.99	0.96	0.67
PiT [58]	**1.00**	**0.98**	**0.98**	**0.99**	**1.00**	**1.00**	0.96	0.64

E. Selection of Backbone Network

In our establishment of the unsupervised framework, the encoder of the backbone network indeed plays an important role for feature extraction. Thus, it is proposed to discuss the baseline backbone networks [53], [54], [55], [56], [57], [58] to verify the validity of our selection for Deepfake video detection. Specifically, three benchmark datasets, referring to FF++, DFD, UADFV, are used for the comparison experiments. As Table III illustrates, the more advanced PiT brings higher detection accuracy superior to the others; meanwhile, the Xception also highlights its effectiveness and efficiency when the insufficient data are used for training. Nevertheless, we need to address that in the proposed unsupervised framework, the backbone network can be continuously upgraded as the deep learning technique advances in the future study, which manifests the flexibility of the proposed method.

F. Comparison to SOTAs

In the practical detection, we cannot perfectly guarantee that the training samples are labeled correctly. On the one hand, as the advancement of deep learning technique, the falsified facial videos become more and more realism, which largely increases the difficulty of crowdsourcing annotation for labeling the training data, leading to innocently incorrect labels; On the other hand, the malicious attacker possibly poisons the training data by purposefully assigning the incorrect labels to the training data, in order to mislead the classifier. In this context, not only the training samples with correct labels for detection are analyzed, but also the training samples with incorrect labels for detection are discussed. To our knowledge, for Deepfake video detection, we first propose to evaluate the detector performance when the labels of training samples are incorrect.

1) Performance of Training Samples With Correct Labels: Generally, we first report the performance of our method when the training samples are labeled correctly. To comprehensively evaluate the effectiveness of the proposed unsupervised detector, it is proposed to use the FF++ dataset, containing four different tampering types, referring to DF, F2F, FS, and NT. Besides, for compared methods, we select the recently-published baselines, including the supervised methods [5], [12], [15], [18], [29], [30], [31], [35], [39], [40], [41], [42], [43], [55], and the unsupervised method [25]. The results of performance comparison are shown in Table IV. By observation, when dealing with both DF and F2F forgery, our proposed unsupervised detector with the perfect 100% detection accuracy rivals the new benchmark supervised methods CORE [42] and SOLA [43], and even outperforms most of the advanced supervised methods. More importantly, compared to the unsupervised detector ULF [25], our proposed unsupervised method performs its powerful strength, especially detecting DF and F2F forgery. However, when dealing with both FS and NT forgery, our proposed unsupervised detector performs worse than all the supervised baselines. Nevertheless, the proposed unsupervised indeed achieves the very promising results in this practical detection scenario. Besides, we also evaluate the detection performance for identifying the DF forgery when the training samples are insufficient. As Fig. 14 illustrates, with the number of the training samples decreases, all the compared supervised methods perform worse and worse. Fortunately, our unsupervised detector nearly cannot be impacted, competing with the training-free ULF [25].

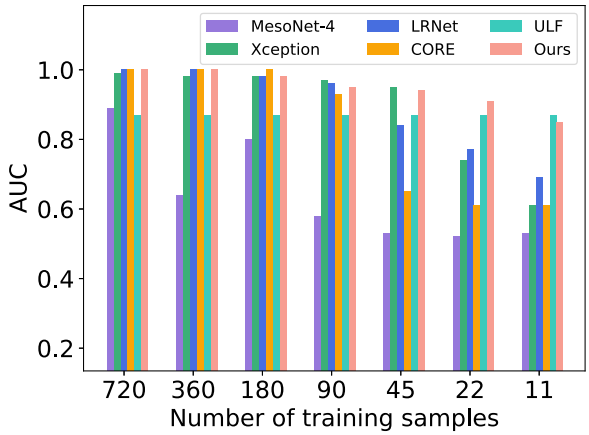


Fig. 14. Illustration of insufficient samples during training.

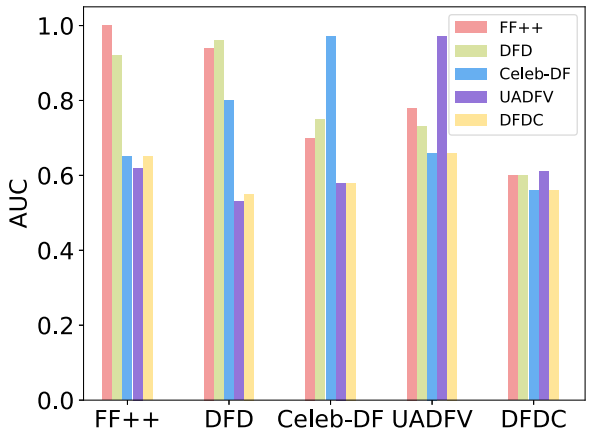


Fig. 15. Generalization performance on the different benchmark datasets.

TABLE IV
PERFORMANCE COMPARISON OF TRAINING SAMPLES WITH CORRECT LABELS

Learning manner	Methods	Reference	DF	F2F	FS	NT
Supervised Learning-based	MesoNet-4 [18]	WIFS 2018	0.89	0.95	–	–
	MesoInception-4 [18]	WIFS 2018	0.92	0.97	–	–
	VA-MLP [30]	WACVW 2019	0.85	0.87	–	–
	VA-LogReg [30]	WACVW 2019	0.78	0.82	–	–
	Headpose [15]	ICASSP 2019	0.54	0.44	0.53	0.51
	FWA [12]	CVPRW 2019	0.86	0.64	0.73	0.39
	Capsule [35]	ICASSP 2019	0.92	0.90	0.93	–
	Xception [5]	ICCV 2019	0.99	**1.00**	0.99	**1.00**
	EfficientNet [55]	PMLR 2019	0.99	**1.00**	0.99	0.96
	Face X-ray [39]	CVPR 2020	0.99	0.99	0.99	0.99
	SMIL [40]	MM 2020	**1.00**	0.99	**1.00**	0.99
	PBD [41]	CVPR 2021	0.97	0.95	–	–
	LRNet [31]	CVPR 2021	**1.00**	0.92	0.88	0.92
	DeepfakeUCL [29]	IJCNN 2021	0.93	0.92	0.92	0.95
	SOLA [43]	CVPR 2022	**1.00**	**1.00**	**1.00**	**1.00**
	CORE [42]	CVPR 2022	**1.00**	**1.00**	**1.00**	**1.00**
Unsupervised Learning-based	ULF [25]	IEEE TMM 2022	0.87	0.48	0.56	0.90
	Ours		**1.00**	**1.00**	0.71	0.77

TABLE V
PERFORMANCE COMPARISON OF TRAINING SAMPLES WITH INCORRECT LABELS, EVALUATED BY MEAN, AND STANDARD VARIATION OF AUC (%)

Methods	Ratio of training samples with incorrect labels										
	10%	20%	30%	40%	50%	60%	70%	80%	90%	100%	Avg.
MesoNet-4 [18]	94.28±2.89	88.03±2.66	79.37±4.76	61.13±2.58	51.81±2.12	38.78±4.47	17.12±1.77	10.09±4.09	3.71±3.22	4.81±0.98	44.91±2.95
MesoInception-4 [18]	94.57±4.69	97.22±1.58	88.29±3.42	67.75±4.38	49.07±1.73	35.73±7.26	8.14±4.49	3.07±2.47	2.19±2.38	3.90±3.85	44.99±3.65
Xception [5]	99.35±0.04	95.36±2.74	89.64±2.66	72.29±3.46	54.41±3.70	25.55±3.14	9.61±1.30	2.67±1.33	0.83±0.12	0.25±0.01	45.00±1.85
EfficientNet [55]	99.29±0.08	99.08±0.19	95.50±0.02	58.85±2.41	50.98±1.51	16.55±1.46	0.34±0.06	0.35±0.18	0.41±0.09	0.32±0.39	42.52±0.65
DeepfakeUCL [29]	85.73±1.46	76.23±0.25	63.57±0.68	53.14±0.19	51.37±3.21	39.03±2.01	19.35±4.73	15.73±0.41	0.93±0.50	0.46±0.15	40.55±1.36
LRNet [31]	99.47±0.11	98.96±1.12	96.66±1.06	79.58±2.06	49.66±3.70	20.35±4.53	3.04±0.39	0.86±0.41	0.22±0.14	0.03±0.01	44.88±1.35
CORE [42]	99.55±0.01	97.29±0.40	86.35±1.40	71.22±2.10	51.06±1.54	30.48±2.04	13.08±4.26	2.53±2.07	0.54±0.02	0.31±0.02	45.24±1.39
ULF [25]	87.00±0.00	87.00±0.00	87.00±0.00	87.00±0.00	87.00±0.00	87.00±0.00	87.00±0.00	87.00±0.00	87.00±0.00	87.00±0.00	87.00±0.00
Ours	**99.59±0.06**	**99.58±0.02**	**99.55±0.01**	**99.50±0.06**	**99.43±0.03**	**99.49±0.06**	**99.52±0.06**	**99.38±0.09**	**99.42±0.20**	**99.53±0.05**	**99.49±0.06**

TABLE VI
GENERALIZATION PERFORMANCE COMPARISON ON THE CROSS-DATASETS

Training Set	Methods	Testing Set				Avg.
		FF++	DFD	Celeb-DF	UADFV	
FF++	Xception [5]	–	0.91	0.62	0.57	0.70
	CORE [42]	–	0.91	**0.72**	0.65	0.76
	M2TR [23]	–	0.81	**0.72**	**0.93**	**0.82**
	Ours	–	**0.94**	0.70	0.78	0.81
DFD	Xception [5]	0.77	–	0.55	0.47	0.60
	CORE [42]	**0.95**	–	0.73	**0.88**	**0.85**
	M2TR [23]	0.85	–	**0.78**	0.84	0.82
	Ours	0.92	–	0.75	0.73	0.80

2) Performance of Training Samples With Incorrect Labels: Next, let us consider the very challenging scenario, in which the training samples are labeled incorrectly at the different ratios. Here, we adopt the DF forgery from FF++ dataset to evaluate the effectiveness of our proposed method. Besides, we define the ratio of training samples with incorrect labels as $p/(p+q)$, in which p represents the number of samples with incorrect labels and q represents the number of samples with correct labels. For fairness, we carry out the large-scale experiments with five training runs, in which the performance of each compared detector is comprehensively evaluated by average detection accuracy and standard variance.

The comparison results are illustrated in Table V . As the ratio increases, all the supervised detectors nearly become invalid while our proposed unsupervised always remains its perfect detection accuracy with low standard variance. Noticeably, due to that the ULF [25] is training-free, it is also immune to the training samples with incorrect labels. Although the satisfying results are obtained by ULF, it is still not good enough to compete with our proposed unsupervised detector. Besides, it should be noted that LRNet [31] and EfficientNet [55] also perform to some extent robustness as the ratio is not larger than 30%. As we expect, when resisting against the attack from the samples with incorrect labels, our proposed unsupervised detector indeed demonstrates its unparallel superiority. In the practical detection, the malicious attackers probably inject the incorrect labels into the training set, in order to set the poison attack; otherwise, the samples from crowdsourcing annotation are carelessly incorrectly labeled, especially in the case that the Deepfake video become more and more indistinguishable as the technique of forgery facial synthesization continuously advances. Relying on our proposed unsupervised Deepfake detection framework, the detection accuracy can always be guaranteed regardless of whether the samples are labeled correctly or not. Besides, it is proposed to address that the assigned labels in the experiments are the original ones, which are not utilized in our unsupervised framework since the inquiry samples can be assigned the new labels from the proposed pseudo-label generator.

G. Generalization Performance of Our Proposed Unsupervised Detector

To demonstrate the generalization of our proposed method, we evaluate the performance of the proposed unsupervised detector

TABLE VII
GENERALIZATION PERFORMANCE ENHANCEMENT OF SOTAS BY ADOPTING OUR PROPOSED ANNOTATION MODULE WHEN TRAINING SAMPLES WITH INCORRECT LABELS

Testing Set	Methods	Ratio of training samples with incorrect labels										
		10%	20%	30%	40%	50%	60%	70%	80%	90%	100%	Avg.
DFD	Xception [5]	0.87	0.83	0.73	0.56	0.51	0.37	0.17	0.18	0.12	0.04	0.44
	Xception [5]+Ours	0.91	0.91	0.91	0.91	0.91	0.91	0.91	0.91	0.91	0.91	0.91
	Enhancement	0.04 ↑	0.08 ↑	0.18 ↑	0.35 ↑	0.40 ↑	0.54 ↑	0.74 ↑	0.73 ↑	0.79 ↑	0.87 ↑	0.47 ↑
	CORE [42]	0.84	0.87	0.75	0.68	0.50	0.31	0.22	0.18	0.17	0.07	0.46
	CORE [42]+Ours	0.91	0.91	0.91	0.91	0.91	0.91	0.91	0.91	0.91	0.91	0.91
	Enhancement	0.07 ↑	0.04 ↑	0.16 ↑	0.23 ↑	0.41 ↑	0.60 ↑	0.69 ↑	0.73 ↑	0.74 ↑	0.84 ↑	0.45 ↑
	M2TR [23]	0.78	0.80	0.63	0.57	0.48	0.42	0.36	0.31	0.24	0.19	0.48
	M2TR [23]+Ours	0.81	0.81	0.81	0.81	0.81	0.81	0.81	0.81	0.81	0.81	0.81
	Enhancement	0.03 ↑	0.01 ↑	0.18 ↑	0.24 ↑	0.33 ↑	0.39 ↑	0.45 ↑	0.50 ↑	0.57 ↑	0.62 ↑	0.33 ↑
Celeb-DF	Xception [5]	0.61	0.61	0.56	0.52	0.47	0.50	0.52	0.49	0.50	0.51	0.53
	Xception [5]+Ours	0.62	0.62	0.62	0.62	0.62	0.62	0.62	0.62	0.62	0.62	0.62
	Enhancement	0.01 ↑	0.01 ↑	0.06 ↑	0.10 ↑	0.15 ↑	0.12 ↑	0.10 ↑	0.13 ↑	0.12 ↑	0.11 ↑	0.09 ↑
	CORE [42]	0.69	0.63	0.60	0.53	0.48	0.43	0.40	0.39	0.02	0.01	0.42
	CORE [42]+Ours	0.72	0.72	0.72	0.72	0.72	0.72	0.72	0.72	0.72	0.72	0.72
	Enhancement	0.03 ↑	0.09 ↑	0.12 ↑	0.19 ↑	0.24 ↑	0.29 ↑	0.32 ↑	0.33 ↑	0.70 ↑	0.71 ↑	0.30 ↑
	M2TR [23]	0.69	0.70	0.69	0.64	0.50	0.45	0.34	0.29	0.24	0.27	0.48
	M2TR [23]+Ours	0.72	0.72	0.72	0.72	0.72	0.72	0.72	0.72	0.72	0.72	0.72
	Enhancement	0.03 ↑	0.02 ↑	0.03 ↑	0.08 ↑	0.22 ↑	0.27 ↑	0.38 ↑	0.43 ↑	0.48 ↑	0.45 ↑	0.24 ↑

on the cross-datasets, consisting of FF++, DFD, Celeb-DF, UADFV, and DFDC. In this section, we adopt the more advanced backbone network PiT [58] in our proposed unsupervised framework. As shown in Fig. 15, the training set is illustrated in the horizontal axis, and the testing results are distributed in order as bar graph. Basically, when the training set and testing set remain matched, the proposed unsupervised detector can achieve better results than the others. Meanwhile, when detecting the Deepfake videos from unseen datasets, the detector performs to some degree generalization capability. Besides, we also calculate the average AUC from each bar, our proposed unsupervised detector trained on FF++ dataset can bring the better result than the others. Moreover, compared with the prior arts (see Table VI), our unsupervised detector rivals the supervised detectors on the cross-datasets.

One of the main contributions is that our proposed unsupervised method can deal with the problem of the training data with incorrect labels. More importantly, our proposed method can serve as a plug-and-play annotation module, which re-annotates the incorrectly-labeled data and enables the SOTAs learn the more generalized representation. That can help the SOTAs resist against the malicious attack, and restore their generalization performance. In Table VII, when the detectors are trained on the FF++ dataset with increasing the ratio of the incorrect labels, the generalization capability of the original SOTAs becomes gradually invalid, since that the samples with incorrect labels disturb the model training. Fortunately, when adopting our proposed module assisting the SOTAs, the generalization capability of them is remarkably enhanced, and restores to the original level. The enhanced detection accuracy indeed be3nefits much from our proposed unsupervised annotation module.

H. Robustness Performance of Our Proposed Unsupervised Detector

As Table VIII illustrates, it is proposed to further evaluate the robustness performance of our unsupervised detector. Then on

TABLE VIII
ROBUSTNESS PERFORMANCE OF OUR PROPOSED UNSUPERVISED DETECTOR ON THE FF++ DATASET

Training \ Testing	C_{00}	C_{23}	C_{40}
C_{00}	1.00	0.95	0.77
C_{23}	1.00	1.00	0.96
C_{40}	0.99	0.99	0.98

the baseline FF++ dataset with DF forgery, three types of compression ratio are provided to conduct the experiments. In details, C_{00} denotes uncompressed format; C_{23} denotes compression with low ratio; C_{40} denotes compression with high ratio. As the compression ratio increases, the video quality is reduced, which to some degree impacts the detection performance. By observation, when training in the C_{00} version, the detection performance of detecting the compressed videos, such as C_{23} or C_{40} version, is unavoidably degraded. On the contrary, when training in the C_{23} or C_{40} version, the proposed unsupervised detector basically remains its high detection accuracy when resisting against compression attacks. Nevertheless, our proposed unsupervised method can withstand to some extent video compression.

VIII. CONCLUSION

In the current Deepfake detection, most of detectors arbitrarily assume that all the training samples are labeled correctly. However, when the training data are maliciously poisoned by the unknown adversaries, referring to as the samples with incorrect labels, the current supervised detectors probably become invalid. To address that issue, in this context, it is proposed to establish the Deepfake video detector in the fully unsupervised mechanism relying on the enhanced contrastive learning. Specifically, the overall framework contains three main stages. First, the pseudo-label generator is designed, in which the primitive clustering is adopted. Next, dependent of the backbone network, the enhanced contrastive learner is established, together with

the scheme of confidence sample selection, in order to refine discriminative feature extraction. Last, we complete the task of binary classification based on the features extracted from the encoder network, and authenticate the Deepfake video by calculating the inter-frame correlation. The numerical experimental results verify the effectiveness of the proposed unsupervised detector, especially in the case of samples with incorrect labels.

To the best of our knowledge, our study is the first work to address the problem of Deepfake video detection when the training samples are poisoned, referring to as data with incorrect labels. In fact, in the practical detection, the data is possibly attacked. Our proposed unsupervised detector is completely independent of the true labels, leading to its immunization to the interference from the attacker. Moreover, it should be noted that in the case of the training samples with correct labels, our proposed unsupervised detector rivals to the SOTAs on DF and F2F forgery; in the case of the training samples with incorrect labels, our unsupervised detector performs remarkably better than the SOTAs. In addition, our unsupervised detector is comprehensively superior to another unsupervised detector [25], and further enriches the Deepfake detection solution in the unsupervised mechanism.

However, we have to admit that the proposed detector cannot perform very well when detecting FS and NT forgery. In fact, in the primitive clustering of the stage 1, when assigning the pseudo-labels, we have to guarantee to some extent purity of clustering. As dealing with FS and NT forgery manners, all the extracted rough features behave kind of similar, leading to the degraded performance of the enhanced contrastive learner on the stage 2. Thus, in the future study, we need to design more robust feature extractor on the stage 1.

REFERENCES

[1] T. T. Nguyen, C. M. Nguyen, D. T. Nguyen, D. T. Nguyen, and S. Nahavandi, "Deep learning for Deepfakes creation and detection," 2019, *arXiv: 1909.11573*.

[2] T. Qiao, R. Shi, X. Luo, M. Xu, N. Zheng, and Y. Wu, "Statistical model-based detector via texture weight map: Application in re-sampling authentication," *IEEE Trans. Multimedia*, vol. 21, no. 5, pp. 1077–1092, May 2019.

[3] X. Bi, Z. Zhang, and B. Xiao, "Reality transform adversarial generators for image splicing forgery detection and localization," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 14 294–14 303.

[4] Q. Lyu, J. Luo, K. Liu, X. Yin, J. Liu, and W. Lu, "Copy move forgery detection based on double matching," *J. Vis. Commun. Image Representation*, vol. 76, 2021, Art. no. 103057.

[5] A. Rossler, D. Cozzolino, L. Verdoliva, C. Riess, J. Thies, and M. Nießner, "FaceForensics : Learning to detect manipulated facial images," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2019, pp. 1–11.

[6] J. Thies, M. Zollhofer, M. Stamminger, C. Theobalt, and M. Nießner, "Face2face: Real-time face capture and reenactment of RGB videos," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 2387–2395.

[7] J. Thies, M. Zollhöfer, and M. Nießner, "Deferred neural rendering: Image synthesis using neural textures," *ACM Trans. Graph.*, vol. 38, no. 4, pp. 1–12, 2019.

[8] A. Neyaz, A. Kumar, S. Krishnan, J. Placker, and Q. Liu, "Security, privacy and steganographic analysis of faceapp and TikTok," *Int. J. Comput. Sci. Secur.*, vol. 14, no. 2, 2020, Art. no. 38.

[9] "Zao," 2018. [Online]. Available: https://apps.apple.com/cn/app/zao/id1465199127

[10] P. Korshunov and S. Marcel, "Deepfakes: A new threat to face recognition? Assessment and detection," 2018, *arXiv:1812.08685*.

[11] A. Khodabakhsh, R. Ramachandra, K. Raja, P. Wasnik, and C. Busch, "Fake face detection methods: Can they be generalized?," in *Proc. Int. Conf. Biometrics Special Int. Group*, 2018, pp. 1–6.

[12] Y. Li and S. Lyu, "Exposing Deepfake videos by detecting face warping artifacts," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops*, 2019, pp. 46–52.

[13] Z. Xia, T. Qiao, M. Xu, N. Zheng, and S. Xie, "Towards Deepfake video forensics based on facial textural disparities in multi-color channels," *Inf. Sci.*, vol. 607, pp. 654–669, 2022.

[14] Y. Li, M.-C. Chang, and S. Lyu, "In ICTU oculi: Exposing AI created fake videos by detecting eye blinking," in *Proc. IEEE Int. Workshop Inf. Forensics Secur.*, 2018, pp. 1–7.

[15] X. Yang, Y. Li, and S. Lyu, "Exposing deep fakes using inconsistent head poses," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, 2019, pp. 8261–8265.

[16] S. Fernandes et al., "Predicting heart rate variations of Deepfake videos using neural ode," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. Workshops*, 2019, pp. 1721–1729.

[17] D. Güera and E. J. Delp, "Deepfake video detection using recurrent neural networks," in *Proc. 15th IEEE Int. Conf. Adv. Video Signal Based Surveill.*, 2018, pp. 1–6.

[18] D. Afchar, V. Nozick, J. Yamagishi, and I. Echizen, "MesoNet: A compact facial video forgery detection network," in *Proc. IEEE Int. Workshop Inf. Forensics Secur.*, 2018, pp. 1–7.

[19] H. Zhao, W. Zhou, D. Chen, T. Wei, W. Zhang, and N. Yu, "Multi-attentional Deepfake detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 2185–2194.

[20] J. Hu, X. Liao, W. Wang, and Z. Qin, "Detecting compressed Deepfake videos in social networks using frame-temporality two-stream convolutional network," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 32, no. 3, pp. 1089–1102, Mar. 2022.

[21] Z. Gu, T. Yao, Y. Chen, S. Ding, and L. Ma, "Hierarchical contrastive inconsistency learning for Deepfake video detection," in *Proc. Eur. Conf. Comput. Vis.*, 2022, pp. 596–613.

[22] D. Zhang, F. Lin, Y. Hua, P. Wang, D. Zeng, and S. Ge, "Deepfake video detection with spatiotemporal dropout transformer," in *Proc. 30th ACM Int. Conf. Multimedia*, 2022, pp. 5833–5841.

[23] J. Wang et al., "M2TR: Multi-modal multi-scale transformers for Deepfake detection," in *Proc. Int. Conf. Multimedia Retrieval*, 2022, pp. 615–623.

[24] Z. Wang, Y. Guo, and W. Zuo, "Deepfake forensics via an adversarial game," *IEEE Trans. Image Process.*, vol. 31, pp. 3541–3552, 2022.

[25] L. Zhang, T. Qiao, M. Xu, N. Zheng, and S. Xie, "Unsupervised learning-based framework for Deepfake video detection," *IEEE Trans. Multimedia*, vol. 25, pp. 4785–4799, 2022.

[26] Y. Chen, T. Qiao, F. Retraint, and G. Hu, "Efficient privacy-preserving forensic method for camera model identification," *IEEE Trans. Inf. Forensics Security*, vol. 17, pp. 2378–2393, 2022.

[27] J. Lukas, J. Fridrich, and M. Goljan, "Digital camera identification from sensor pattern noise," *IEEE Trans. Inf. Forensics Security*, vol. 1, no. 2, pp. 205–214, Jun. 2006.

[28] D. Cozzolino and L. Verdoliva, "Noiseprint: A CNN-based camera model fingerprint," *IEEE Trans. Inf. Forensics Security*, vol. 15, pp. 144–159, 2019.

[29] S. Fung, X. Lu, C. Zhang, and C.-T. Li, "Deepfakeucl: Deepfake detection via unsupervised contrastive learning," in *Proc. Int. Joint Conf. Neural Netw.*, 2021, pp. 1–8.

[30] F. Matern, C. Riess, and M. Stamminger, "Exploiting Visual Artifacts to Expose Deepfakes and Face Manipulations," in *Proc. IEEE Winter Appl. Comput. Vis. Workshops*, 2019, pp. 83–92.

[31] Z. Sun, Y. Han, Z. Hua, N. Ruan, and W. Jia, "Improving the efficiency and robustness of Deepfakes detection through precise geometric features," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 3609–3618.

[32] A. Haliassos, K. Vougioukas, S. Petridis, and M. Pantic, "Lips don't lie: A generalisable and robust approach to face forgery detection," in *Proc. IEEE/CVF Conf. Comput. Vis. pattern Recognit.*, 2021, pp. 5039–5049.

[33] S. Agarwal, H. Farid, Y. Gu, M. He, K. Nagano, and H. Li, "Protecting world leaders against Deepfakes," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops*, 2019, pp. 38–45.

[34] E. Sabir, J. Cheng, A. Jaiswal, W. AbdAlmageed, I. Masi, and P. Natarajan, "Recurrent convolutional strategies for face manipulation detection in videos," *Interfaces*, vol. 3, no. 1, pp. 80–87, 2019.

[35] H. H. Nguyen, J. Yamagishi, and I. Echizen, "Capsule-forensics: Using capsule networks to detect forged images and videos," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, 2019, pp. 2307–2311.

[36] P. Zhou, X. Han, V. I. Morariu, and L. S. Davis, "Two-stream neural networks for tampered face detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops*, 2017, pp. 1831–1839.
[37] B. Han, X. Han, H. Zhang, J. Li, and X. Cao, "Fighting fake news: Two stream network for Deepfake detection via learnable SRM," *IEEE Trans. Biom., Behav., Ident. Sci.*, vol. 3, no. 3, pp. 320–331, 2021.
[38] J. Kodovskỳ and J. Fridrich, "Steganalysis of JPEG images using rich models," in *Media Watermarking, Security, and Forensics* Bellingham, WA, USA: International Society for Optics and Photonics, 2012.
[39] L. Li et al., "Face X-ray for more general face forgery detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 5001–5010.
[40] X. Li et al., "Sharp multiple instance learning for Deepfake video detection," in *Proc. 28th ACM Int. Conf. Multimedia*, 2020, pp. 1864–1872.
[41] S. Schwarcz and R. Chellappa, "Finding facial forgery artifacts with parts-based detectors," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 933–942.
[42] Y. Ni, D. Meng, C. Yu, C. Quan, D. Ren, and Y. Zhao, "CORE: Consistent representation learning for face forgery detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 12–21.
[43] J. Fei, Y. Dai, P. Yu, T. Shen, Z. Xia, and J. Weng, "Learning second order local anomaly for general face forgery detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 20270–20280.
[44] M. Caron, P. Bojanowski, A. Joulin, and M. Douze, "Deep clustering for unsupervised learning of visual features," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 132–149.
[45] K. I. Laws, "Textured image segmentation," University of Southern California Los Angeles Image Processing INST, Tech. Rep. 940, 1980.
[46] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," *Int. Conf. Mach. Learn.*, 2020, pp. 1597–1607.
[47] Y. Li, X. Yang, P. Sun, H. Qi, and S. Lyu, "Celeb-DF: A large-scale challenging dataset for Deepfake forensics," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 3207–3216.
[48] B. Dolhansky et al., "The Deepfake detection challenge (DFDC) dataset," 2020, *arXiv: 2006.07397*.
[49] N. Dufour and A. Gully, "Contributing data to Deepfake detection research," *Google AI Blog*, vol. 1, no. 2, 2019, Art. no. 3.
[50] S. Abu-El-Haija et al., "Youtube-8 m: A large-scale video classification benchmark," 2016, *arXiv:1609.08675*.
[51] K. Zhang, Z. Zhang, Z. Li, and Y. Qiao, "Joint face detection and alignment using multitask cascaded convolutional networks," *IEEE signal Process. Lett.*, vol. 23, no. 10, pp. 1499–1503, Oct. 2016.
[52] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: Visual explanations from deep networks via gradient-based localization," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 618–626.
[53] F. Chollet, "Xception: Deep learning with depthwise separable convolutions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 1251–1258.
[54] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 2818–2826.
[55] M. Tan and Q. Le, "Efficientnet: Rethinking model scaling for convolutional neural networks," in *Proc. Int. Conf. Mach. Learn.*, 2019, pp. 6105–6114.
[56] A. Dosovitskiy et al., "An image is worth 16x16 words: Transformers for image recognition at scale," 2020, *arXiv: 2010.11929*.
[57] Z. Tu et al., "MaxViT: Multi-axis vision transformer," in *Proc. Eur. Conf. Comput. Vis.*, 2022, pp. 459–479.
[58] B. Heo, S. Yun, D. Han, S. Chun, J. Choe, and S. J. Oh, "Rethinking spatial dimensions of vision transformers," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 11 936–11 945.
[59] O. Russakovsky et al., "ImageNet large scale visual recognition challenge," *Int. J. Comput. Vis.*, vol. 115, no. 3, pp. 211–252, 2015.

Tong Qiao received the BS degree in electronic and information engineering from Information Engineering University, Zhengzhou, China, in 2009, and the MS degree in communication and information system from Shanghai University, Shanghai, China, in 2012, and the PhD degree from the University of Technology of Troyes, Laboratory of Systems Modeling and Dependability, Troyes, France, in 2016. He currently works as an associate professor with the School of Cyberspace, Hangzhou Dianzi University. His current research interests include focus on media forensics, AI security and data hiding. He has published more than 60 peer-reviewed papers on journals and conferences.

Shichuang Xie received the BS degree in computer science and technology from Henan University, Kaifeng, China, in 2020, the MS degree from the School of Cybersapce, Hangzhou Dianzi University, Hangzhou, Zhejiang, China, in 2023. His current research interests include focus on multimedia forensics and AI security.

Yanli Chen received the BS degree in electronic information science and technology from Qingdao University, Qingdao, China, in 2015, the MS degree in signal and information processing from Xidian University, Xi'an, China, in 2018, and the PhD degree from the Laboratory of Computer Science and Digital Society, University of Technology of Troyes, France, in 2022. Her research interest includes focus on media forensics.

Florent Retraint received the engineering Diploma degree in computer science from the Compiegne University of Technology, in 1993, the MS degree in applied mathematics from ENSIMAG, in 1994, and the PhD degree in applied mathematics from the National Institute of Applied Sciences of Lyon, France, in 1998. He held a postdoctoral position with CEA Grenoble for one year. He was a research engineer with Thomsom CSF for two years. Since 2002, he has been with Laboratory of Computer Science and Digital Society, University of Technology of Troyes, France. He is currently a full professor. His research interests include image modeling, statistical image processing, hypothesis testing theory, anomaly detection, and localization.

Xiangyang Luo received the BS, MS, and PhD degrees from the State Key Laboratory of Mathematical Engineering and Advanced Computing, Zhengzhou, China, in 2001, 2004, and 2010, respectively. He is the author or co-author of more than 200 refereed international journal and conference papers. He is currently a full professor with the State Key Laboratory of Mathematical Engineering and Advanced Computing. His research interests include network security and multimedia security.