Detecting Deepfakes with Multi-Metric Loss

Ziwei Zhang[1], Xin Li[1], Rongrong Ni[1] and Yao Zhao[1]

[1]*Beijing Jiaotong University*

Abstract

In recent years, DeepFake techniques have advanced to generate so realistic forged content that it could jeopardize personal privacy and national security. We observe the distribution discrepancy between genuine faces and tampered faces manipulated by DeepFake techniques. It can be described that embedding vectors of genuine faces are tightly distributed in the embedding space, while tampered faces are comparatively scattered. We, therefore, propose a novel DeepFake detection method based on Multi-metric Loss. Specifically, real and fake faces are mapped onto the embedding space, which is of intra-class compactness and inter-class separation. Then by adding Weight-Center Loss to project genuine faces onto a more compact region in the embedding space, the distance between the two types of sample clusters is further expanded, thereby improving the separability of genuine and tampered samples. Moreover, the Adaptive Hardness-aware Expander is designed to further improve feature description ability of the model because the metric is always challenged with proper difficulty. Extensive experiments show that our approach can achieve state-of-the-art performance on present datasets.

Keywords

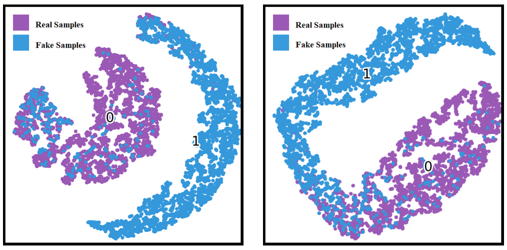
Deepfakes, Multi-metric Loss, Adaptive Hardness-aware Expander

1. Introduction

Of various digital media, videos containing digital human faces, especially the ones involving personal identification information, are most vulnerable to be attacked. These assaults are collectively referred to as DeepFake manipulations. Therefore, to develop effective methods capable of detecting DeepFake videos carries substant weight. Since the existing manipulations tamper with specific areas frame by frame, the artifacts and noises appear in the spurious videos. So previous researchers have proposed many handcrafted methods [1, 2, 3, 4] and data-driven methods [5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16] to find manipulation traces.

Due to uncertain counterfeit methods and manipulation quality in DeepFake videos, the spurious data is scattered in the whole feature space. Relatively, genuine human faces concentrate close to a non-linear low-dimensional manifold [17] in the feature space. As shown in Figure 1, the vectors of real faces are tightly distributed, while the fakes are comparatively scattered. Therefore, we consider that this distribution discrepancy also exists in the embedding space obtained by the feature space mapping. The existing detection schemes, however, do not consider the distribution discrepancy between the two types of samples.

To this end, we propose the DeepFake detection framework with Multi-metric Loss, as shown in Figure 2. Triplet Loss, Cross-Entropy Loss and Weight-Center Loss together constitute Multi-metric Loss acting on different levels and face sample cluster with diverse labels (real/fake). Under the restriction of Triplet Loss and Cross-Entropy Loss, the real faces and fake faces are mapped onto the embedding space, which is intra-class compactness and inter-class separation. Then through adding Weight-Center Loss, the real faces are projected to a more compact region. The method of excavating fundamental distinction between the two types of samples is, therefore, to extend the distance between the two types of sample clusters in the embedding space, thereby improving the separability of genuine and spurious videos. In the end-stage of training, in order to further improve the feature description ability of the model, we designed the Adaptive Hardness-aware Expander (AHE). The rigorous experiments on FaceForensics++ [6], DFDC [18] and Celeb-DF [19] datasets show that the proposed method based on Multi-metric Loss is highly effective and achieves state-of-the-art performance.



(a) DFDC (b) Celeb-DF

Figure 1: DFDC and Celeb-DF dataset distribution visualization by t-SNE. The projections of real face features are tightly distributed, while the fakes are comparatively scattered.

International Workshop on Safety & Security of Deep Learning, August 19th -26th, 2021, Montreal-themed Virtual Reality

rrni@bjtu.edu.cn (R. Ni); yzhao@bjtu.edu.cn (Y. Zhao)

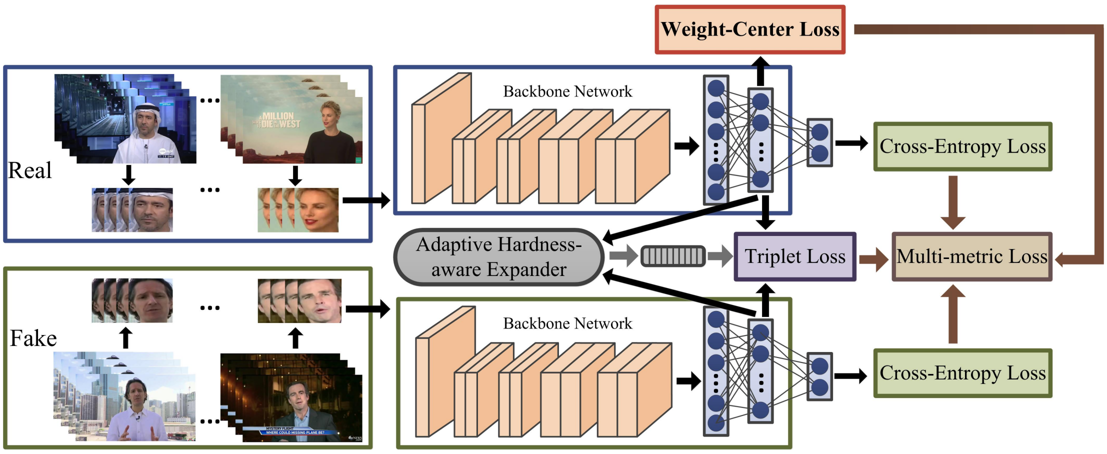


Figure 2: The framework of our method. Use MTCNN to crop the video frames into facial area maps and send them to the backbone network to get the embedding vectors. The Cross-Entropy Loss and Triple Loss of all original embedding vectors and the Weight-Center Loss of genuine embedding vectors, are calculated. In the end-stage of training, Adaptive Hardness-aware Expander continuously synthesizes samples with adaptive hardness.

2. Related works

With huge risks posed by face forgery technology, there is currently an urge to investigate DeepFake detection methods. Existing detection techniques mainly fall into two categories: handcrafted and data-driven methods.

Handcrafted Methods. For the limited face manipulation techniques at that time, early works achieved the DeepFake detection through handcraft features. This methods mainly include eye blinking [1], incomplete details in the eye and teeth [2], face warping [3] and head poses [4]. With the development of generative adversarial network (GAN) [20],a variety of tampering technologies have emerged and forgery faces have become more realistic. Therefore, the effectiveness of formerly handcrafted methods has gradually been weakened.

Data-driven Methods. Given the powerful feature representation capabilities of deep neural network, the data-driven methods have received widespread attention. Firstly, some classification networks were applied to detect fake faces like MesoNet [5], XceptionNet [6], Capsule network [7], R3D and C3D [8] etc. Then Zhou et al. [9] proposed to use the Two-stream neural network to capture tampering artifacts and local noise residuals. The adaptive face weighting layer [10] was designed with it focus forgery details. The model [11] was trained to mark the blending boundary for forged images. Considering inconsistent warping left by manipulation in the inter-frame, the methods [12, 13, 14] were proposed. The methods [15, 16] introduced the Deep Metric Learning to DeepFake detection for the first time.

Kumar et al. [15] mainly explored the method's effectiveness for detecting videos with high compression factor. Feng et al. [16] used the difference of the full face image in videos as the feature for DeepFake detection. Although they also mapped data onto the embedding space based on Deep Metric Learning, they just followed the traditional metric strategy and imposed the same constraint on two types of samples. In our work, considering the distribution discrepancy of real and fake data, different levels of classification constraints are imposed on these two kinds of sample clusters. Specifically, we design the Multi-metric Loss to further widen the distance between the real cluster and the fakes by capturing fundamental distinction between spurious videos and genuine videos, and the Adaptive Hardness-aware Expander to further improve the feature description ability of the model.

3. Proposed Approach

In this section, we give an overview of our framework. As aforementioned, the embedding vectors of real faces are aggregative in the embedding space, while the fakes are relatively scattered. Motivated by this observation, two key components are integrated into the framework: 1) Multi-metric Loss is designed to mine fundamental distinction between real and fake faces so as to improve separability; 2) Adaptive Hardness-aware Expander can be used to further improve the feature description ability of the model. The framework is depicted in Figure 2.

3.1. Multi-metric Loss

Let $\boldsymbol{\mathcal{X}}$ denote the data space where we sample a set of facial area maps $\mathbf{X} = [x_1, x_2, \cdots, x_N]$. Each data x_i has a label $l_i \in \{0, 1\}$ representing real or fake. Let $h : \boldsymbol{\mathcal{X}} \xrightarrow{h} \boldsymbol{\mathcal{Y}}$ be the mapping from the data space to the feature space, where the extracted feature y_i preserves semantic characteristics of its corresponding data point x_i. Then the feature is projected onto the embedding space $\boldsymbol{\mathcal{Z}}$ with the mapping $g : \boldsymbol{\mathcal{Y}} \xrightarrow{g} \boldsymbol{\mathcal{Z}}$. Since the

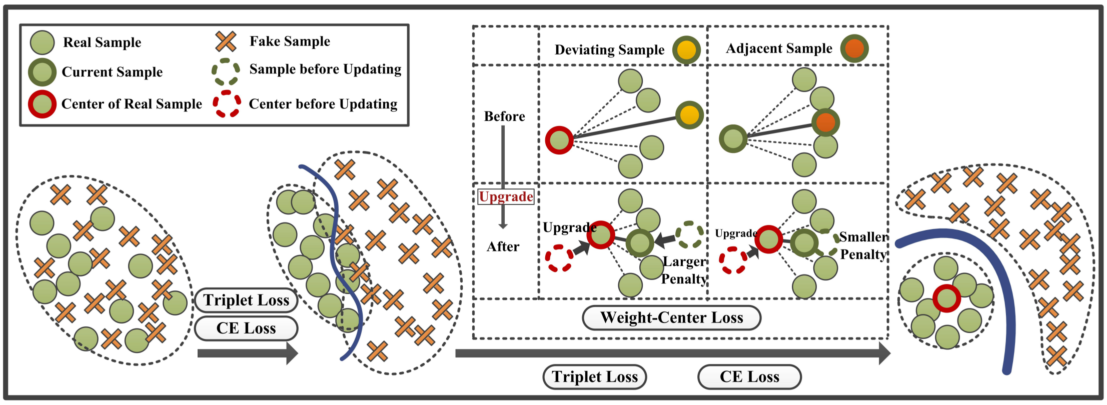


Figure 3: The proposed Multi-metric Loss with Triplet Loss, Cross-Entropy Loss and Weight-Center Loss. Weight-Center Loss, which only acts on the cluster of real samples, imposes larger penalty on samples that deviate from the center, and imposes smaller penalty on adjacent samples, while continuously updating the center of the real sample cluster.

projection can be incorporated into the deep network, we can directly learn the mapping $f(\cdot;\theta) = h \circ g : \boldsymbol{\mathcal{X}} \xrightarrow{f} \boldsymbol{\mathcal{Z}}$ from the data space to the embedding space, where θ is network parameters.

Based on the data distribution discrepancy, namely, embedding vectors of real faces are tightly distributed, while the fakes are comparatively scattered. We deem that various levels of classification constraints should be imposed, so as to mine fundamental distinction between spurious videos and genuine videos, as shown in the Figure 3. Multi-metric Loss is formulated as follows:

$$Loss = \mathcal{L}_{Weight-Center} + \beta\mathcal{L}_{Triplet} + \chi\mathcal{L}_{CE} \tag{1}$$

3.1.1. Triplet Loss

Under the constraint of Triplet Loss, the mapping from high-dimensional sparse features into low-dimensional dense vectors is learned. Reflected in the embedding space, the distribution of data is characterized by intra-class compactness and inter-class separation.

Let $f(x_a;\theta)$ be the anchor embedding vector. The embedding vector with the same and different label relative to $f(x_a;\theta)$, are defined as $f(x_p;\theta)$ and $f(x_n;\theta)$, respectively. Triplet Loss is formulated as follows:

$$\mathcal{L}_{Triplet} := [S_{an} - S_{ap} + \kappa]_+ \tag{2}$$

where $S_{ap} = \langle f(x_a;\theta), f(x_p;\theta)\rangle$ indicates the similarity of positive pair, $S_{an} = \langle f(x_a;\theta), f(x_n;\theta)\rangle$ is the similarity of negative pair, $\langle\cdot,\cdot\rangle$ denotes dot product, and κ is metric margin.

3.1.2. Cross-Entropy Loss

In our approach, Cross-Entropy (CE) Loss and the Triplet Loss act jointly. Specifically, CE Loss encourages the separation of real embedding vectors from the fakes. Simultaneously, the Triplet Loss is used to achieve intra-class compactness and inter-class separation, so as to initially separate the two types of sample clusters.

3.1.3. Weight-Center Loss

Considering the distribution discrepancy of genuine and tampered data, we hope to further widen the distance between two categories of sample clusters by capturing the fundamental distinction between real videos and fake videos. Under the action of Triplet Loss and CE Loss, the network has acquired preliminary classification capability. On this basis, we design Weight-Center Loss for real sample cluster to capture the fundamental distinction between two types of samples.

Some embedding vectors are far from the center of the real sample cluster, it may be due to certain interference, which has nothing to do with judging real and fake videos. Therefore, Weight-Center Loss is proposed which only acts on the cluster of real samples. We define the sample that is far from the center of the real sample cluster compared to the surrounding samples as the deviating sample. It adaptively imposes larger penalty on deviating samples, and imposes smaller penalty on adjacent samples. Simultaneously, the center of the real sample cluster is continuously updated. Based on the above operations, real faces are projected to a more compact region in the embedding space, so as to broaden the distance between the real sample cluster and the fake sample cluster. Weight-Center Loss is formulated as follows:

$$\mathcal{L}_{Weight-Center} = \frac{1}{\alpha}\log\left[1 + \sum_{k\in P} e^{-\alpha(S_{kc}-\lambda)}\right] \tag{3}$$

where P is the collection of real embedding vectors, S_{kc} is the similarity of the center sample pair $\{f(x_k;\theta), f(x_c;\theta)\}$, $f(x_k;\theta)$ and $f(x_c;\theta)$ are real embedding vectors and the iterative center and α, λ are fixed hyperparameters. It is worth noting that the center is iterated continuously.

Based on [21], we can get the generic definition about the penalty weight of sample pair. Then the penalty weight of the center sample $\{f(x_k;\theta), f(x_c;\theta)\}$ in

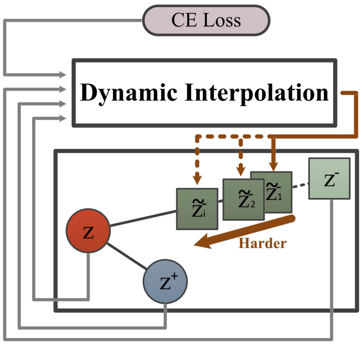


Figure 4: Adaptive Hardness-aware Expander module. The synthetic samples $\left\{\tilde{z}_1^-, \tilde{z}_2^-, \ldots, \tilde{z}_i^-\right\}$, which are generated by the distance of tuple $\left\{z, z^+, z^-\right\}$ and CE Loss.

$\mathcal{L}_{Weight-Center}$ is calculated as follows:

$$w_{kc} = \frac{1}{e^{-\alpha(\lambda - S_{kc})} + \sum_{i \in P, i \neq k} e^{-\alpha(S_{ic} - S_{kc})}} \tag{4}$$

where $f(x_k; \theta)$ is one embedding vector of set, $f(x_i; \theta)$ is the other in the set except sample $f(x_k; \theta)$. S_{kc} and S_{ic} indicate the similarity of sample pair $\{f(x_k; \theta), f(x_c; \theta)\}$ and $\{f(x_i; \theta), f(x_c; \theta)\}$.

Eq.4 shows that the penalty weight for sample pair is determined by its relative similarity, measured by comparing it with the distance from surrounding samples with the center, which is fundamentally different from Center Loss [22]. According to the relative position relationship in the set, there are two different situations, as shown in Figure 3. Firstly, the embedding vector $f(x_k; \theta)$ is far from the center of set relative to other samples $f(x_i; \theta)$, as described by the deviating sample in Figure 3, and the formula is expressed as $S_{ic} > S_{kc}$. We consider that the current embedding vector extracted contains certain interference, which has nothing to do with judging real and fake videos, so the larger penalty weight is imposed. When the embedding vector is closer to the center of set relative to other samples, as described by the adjacent sample in Figure 3, the formula is expressed as $S_{ic} \leq S_{kc}$. Smaller penalty weight is imposed and the network parameters are fine-tuned to find the features that could best represent the fundamental distinction between spurious videos and authentic videos.

3.2. Adaptive Hardness-aware Expander

In the end-stage of training, considering that original samples are already well separable under the action of Multi-metric Loss. Continuing to train original samples cannot further improve the model's feature description ability. To address this limitations, we propose the Adaptive Hardness-aware Expander, as shown in Figure 4.

We construct the hardness-aware triplet $\left\{z, z^+, \tilde{z}_i^-\right\}$ in the embedding space, where manipulation of the distances among samples will directly alter the hard level of the triple. The distances of negative pairs $\left\{z, \tilde{z}_i^-\right\}$ is manipulated, and for other samples $\left\{z, z^+\right\}$, we perform no transformation. Then the reduction in the distance between negative pairs will create rise of the hard level, so that the measurement process is always at an appropriate level of difficulty during the training cycle. As shown in the Figure 4, in order to simplify the representation, we use z, z^+, z^- to represent the anchor embedding vector $f(x_a; \theta)$, the positive embedding vector $f(x_p; \theta)$, and the negative embedding vector $f(x_n; \theta)$, respectively.

Firstly, a toy example that constructs an augmented harder negative sample $\tilde{z}^-$ by linear interpolation, is presented:

$$\tilde{z}^- = z + \omega\left(z^- - z\right), \omega \in [0, 1] \tag{5}$$

However, samples too close to the anchor are likely to cause confusion in the label. Therefore, we exploit the CE Loss in the previous section to control the hardness of the generated negative samples, since it is a good indicator of training process. If the CE Loss is small, the generated negative sample will be closer to the anchor point, but will not cross the positive sample. Adaptive Hardness-aware Expander can be represented as:

$$\tilde{z}^- = \begin{cases} z + \left[\eta d^- + (1-\eta) d^+\right] \frac{\left(z^- - z\right)}{d^-} & \text{if } d^- > d^+ \\ z^- & \text{if } d^- \leq d^+ \end{cases} \tag{6}$$

where $\eta = e^{-\frac{\gamma}{\mathcal{L}_{CE}}}$ is a balance factor to control the hardness of the generated negative samples, γ is the pulling factor used to balance the scale of $\mathcal{L}_{CE}$, $d^+ = \left\|z^- - z\right\|_2$ and $d^- = \left\|z^+ - z\right\|_2$ are the distance between positive pair and negative pair, respectively.

In the early stage of training, the generated hard samples can not represent related face information, considering that the embedding space has no accurate semantic structure. It may even cause the model to be trained in the wrong direction from the beginning. As the training progresses, however, the model is growing more tolerant of hard samples, that is, the metric is always challenged with proper difficulty. Thereby Adaptive Hardness-aware Expander can improve the feature description ability of the model.

4. Experiments

In this section, we first explore the optimal settings for our approach and then present extensive experimental results to demonstrate the effectiveness of our method.

4.1. Implement Details

For all real/fake video frames, we use face extractor MTCNN to detect faces and save the aligned facial images as inputs with the size of 256 × 256. β, χ in Eq.1 and α in Eq.3 is set to 2.0, 1.0, 2.0 to impose different levels of classification constraints. The margin of Triplet Loss in Eq.2 is set to 1.0. Optimization is performed using SGD optimizer with weight decay $5e^{-4}$. The initial learning rate is kept at 0.01 and divided by 10 after every 3000 iterations. We adopt ResNet-34, which is pre-trained on

Table 1
Testing ACC(%) and AUC(%) score of our method and other methods on FaceForensics++ dataset.

Methods	FF++/df		FF++/ff		FF++/fs		FF++/nt	
	ACC	AUC	ACC	AUC	ACC	AUC	ACC	AUC
MesoNet [5]	0.827	0.853	0.562	0.634	0.611	0.679	0.502	0.596
XceptionNet [6]	0.948	0.986	0.928	0.972	0.903	0.933	0.807	0.835
Li et al. [3]	0.969	0.995	0.972	0.987	0.963	0.990	0.890	0.913
Capsule[7]	0.941	0.960	0.963	0.958	0.972	0.974	0.887	0.948
Feng et al. [16]	0.953	0.991	0.938	0.957	0.921	0.940	0.841	0.902
Kumar et al. [15]	0.960	0.990	0.932	0.962	0.944	0.978	0.832	0.872
Bonettini et al.[10]	0.981	0.992	0.955	0.970	0.973	0.980	0.845	0.863
Ours	**0.985**	**0.998**	**0.974**	**0.991**	**0.995**	**1.000**	**0.938**	**0.968**

Table 2
Testing ACC(%) and AUC(%) score of our method and other methods on DFDC and Celeb-DF dataset.

	DFDC		Celeb-DF	
	ACC	AUC	ACC	AUC
MesoNet [5]	0.746	0.818	0.482	0.536
XceptionNet [6]	0.845	0.909	0.788	0.832
Li et al. [3]	0.793	0.861	0.571	0.628
Capsule [7]	0.861	0.933	0.791	0.879
Feng et al. [16]	0.883	0.963	0.814	0.867
Kumar et al. [15]	0.825	0.899	0.792	0.943
Bonettini et al.[10]	0.944	0.967	0.903	0.959
Ours	**0.962**	**0.979**	**0.927**	**0.968**

Table 3
The ablation study about Multi-metric Loss.

	df	ff	fs	nt
Triplet Loss	0.946	0.925	0.939	0.810
+ Cross-Entropy Loss	0.962	0.933	0.942	0.870
+ Weight-center Loss	0.985	0.974	0.995	0.938

the ImageNet dataset, as the backbone network. Our model is trained on 4 RTX 2080Ti GPUs with batch size 16 and the total number of iterations is set to 10, 000.

4.2. Comparsion with Previous Methods

In this section, we compare our method with previous DeepFake detection methods. The performance of various methods on FaceForensics++ [6], DFDC [18] and Celeb-DF [19] dataset is shown. We adopt ACC (accuracy) and AUC (area under Receiver Operating Characteristic Curve) as the evaluation metrics for experiments.

The evaluation results of the individual datasets are shown in Table 1 and Table 2. The results indicate that our model trained with Multi-metric Loss and AHE have significant improvement over previous methods with metric learning [15, 16], especially in DFDC and Celeb-DF dataset. The reason is that different levels of classification constraints based on the phenomena of distribution discrepancy is imposed to mine the fundamental distinction between spurious videos and genuine videos, so that it can still work on tampered videos without obvious artifacts. At the same time, the generation of adaptive hardness-aware samples forces the network to pay more attention to some key features that characterize the truth and counterfeit under the constraint of Multi-metric Loss, thereby improving the feature description ability of the model and achieving better classification performance. Therefore, our method can achieve state-of-the-art performance on FaceForensics++, DFDC and CelebDF datasets.

4.3. Ablation Study

To verify the effectiveness of Multi-metric Loss and Adaptive Hardness-aware Expander, we conduct ablation studies and results are shown in Table 3, Figure 5, Figure 6.

4.3.1. Effectiveness of Multi-metric Loss

To confirm the effectiveness of Multi-metric Loss, we evaluate how different levels classification constraints affect the detection accuracy. We train the model on FF++ (c23), other hyperparameters are kept the same as settings in Table 1.

The t-SNE plots of four different manipulation methods in FF++ datasets are reported in Figure 5. It can be found that the separability of the sample is poor when Triple Loss acts independently, as shown in the first row of Figure 5. The reason is that the data selection in the batch results in uneven data distribution, which makes it difficult to divide the interface. When Cross-entropy Loss is introduced, the data distribution of different manipulation in FF++ datasets is shown in the second row of Figure 5. Among them, Cross-Entropy Loss encourages the separation of real embedding vectors from fake embedding vectors, and the Triple Loss helps constrain the intra-class compactness and inter-class separation, thereby improving the separability of samples. In the third row of Figure 5, Weight-Center Loss is added and it only acts on the real cluster. By mining the features representing authenticity, the real sample clusters are tightly clustered, thereby further extending distance between two types of sample clusters in the embedding space. The ACC ablation studys about Multi-metric Loss on FF++ are reported in Table 3, which further confirm the effectiveness of Multi-metric Loss.

Note that Triplet Loss and Cross-Entropy Loss work during the entire training stage, while Weight-Center

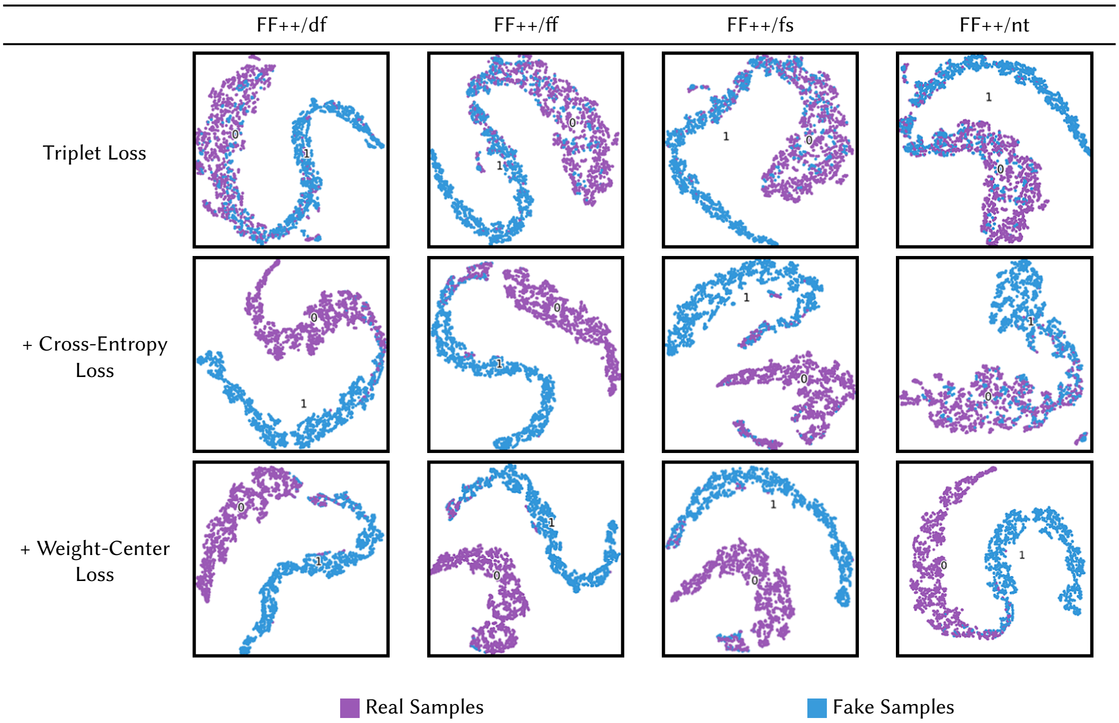


Figure 5: t-SNE plots of ablation study about Triplet Loss, Cross-Entropy Loss and Weight-Center Loss in FF++ dataset.

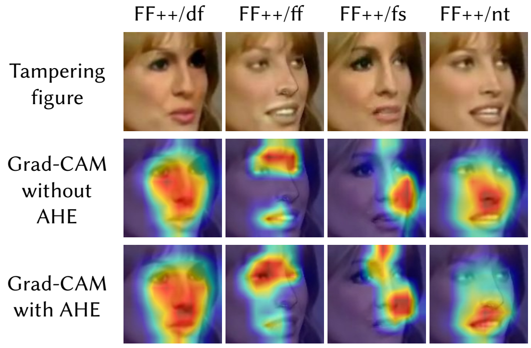


Figure 6: Heatmaps generated by Grad-CAM about with or without AHE in four manipulation methods.

Loss only works in the middle and end stages of training. The main reason is that the center point of the real samples is unstable at the beginning of the training, which will cause the network to optimize in the wrong direction.

4.3.2. Effectiveness of AHE

To confirm the effectiveness of Adaptive Hardness-aware Expander, we analyze the class activation maps for four different manipulation methods, as shown in Figure 6.

Class activation maps corresponding to the operation of Expander indicate that synthetic samples with adaptive hardness force the network paying more attention to some key features that characterize the authenticity and the counterfeit under the constraint of Multi-metric Loss, thereby improving the feature description ability of the model. For example, in Figure 6, NeuralTextures (nt), a tampering scheme only modifies the mouth area. Before the Adaptive Hardness-aware Expander is used, class activation map shows that the nose and mouth regions together provide evidence that the video is tampered. After the Adaptive Hardness-aware Expander is used, class activation map shows that the network will pay more attention to the mouth area tampered, which demonstrates the interpretability of our proposed method.

5. Conclusion

In this work, we propose the DeepFake detection method based on Multi-metric Loss, considering the distribution discrepancy that the embedding vectors of genuine faces are tightly distributed in the embedding space, while tampered faces are comparatively scattered. Multi-metric Loss improves the separability of genuine and tampered samples through further widening distance between the two types of sample clusters. Besides, adaptive hardness-aware samples is generated to make the metric be always in the proper difficulty, so as to improve the feature description ability of the model. Our method achieves good improvements in extensive metrics.

Acknowledgments

This work was supported in part by the National Key Research and Development of China (2018YFC0807306), National NSF of China (U1936212), Beijing Fund-Municipal Education Commission Joint Project (KZ202010015023).

References

[1] Y. Li, M. Chang, S. Lyu, In Ictu Oculi: Exposing AI Created Fake Videos by Detecting Eye Blinking, in: 2018 IEEE International Workshop on Information Forensics and Security (WIFS), 2018, pp. 1–7.

[2] F. Matern, C. Riess, M. Stamminger, Exploiting Visual Artifacts to Expose Deepfakes and Face Manipulations, in: 2019 IEEE Winter Applications of Computer Vision Workshops (WACVW), 2019, pp. 83–92.

[3] Y. Li, S. Lyu, Exposing DeepFake Videos By Detecting Face Warping Artifacts, in: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops, 2019, pp. 46–52.

[4] X. Yang, Y. Li, S. Lyu, Exposing Deep Fakes Using Inconsistent Head Poses, in: IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP, 2019, pp. 8261–8265.

[5] D. Afchar, V. Nozick, J. Yamagishi, I. Echizen, MesoNet: a Compact Facial Video Forgery Detection Network, in: 2018 IEEE International Workshop on Information Forensics and Security (WIFS), 2018, pp. 1–7.

[6] A. Rössler, D. Cozzolino, L. Verdoliva, C. Riess, J. Thies, M. Niessner, FaceForensics++: Learning to Detect Manipulated Facial Images, in: 2019 IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 1–11.

[7] H. Nguyen, J. Yamagishi, I. Echizen, Use of a Capsule Network to Detect Fake Images and Videos, 2019. arXiv:1910.12467v2.

[8] I. Ganiyusufoglu, L. M. Ngô, N. Savov, S. Karaoglu, T. Gevers, Spatio-temporal Features for Generalized Detection of Deepfake Videos, 2020. arXiv:2010.11844.

[9] P. Zhou, X. Han, V. I. Morariu, L. S. Davis, Two-Stream Neural Networks for Tampered Face Detection, in: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops, 2017.

[10] N. Bonettini, E. D. Cannas, S. Mandelli, L. Bondi, P. Bestagini, S. Tubaro, Video Face Manipulation Detection Through Ensemble of CNNs, in: 2020 25th International Conference on Pattern Recognition (ICPR), 2020, pp. 5012–5019.

[11] L. Li, J. Bao, T. Zhang, H. Yang, D. Chen, F. Wen, B. Guo, Face X-Ray for More General Face Forgery Detection, in: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 5000–5009.

[12] D. Güera, E. J. Delp, Deepfake Video Detection Using Recurrent Neural Networks, in: 2018 15th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS), 2018, pp. 1–6.

[13] E. Sabir, J. Cheng, A. Jaiswal, W. AbdAlmageed, I. Masi, P. Natarajan, Recurrent Convolutional Strategies for Face Manipulation Detection in Videos, in: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), 2019, pp. 80–87.

[14] K. Chugh, P. Gupta, A. Dhall, R. Subramanian, Not Made for Each Other- Audio-Visual Dissonance-Based Deepfake Detection and Localization, in: Proceedings of the 28th ACM International Conference on Multimedia, 2020, p. 439–447.

[15] A. Kumar, A. Bhavsar, R. Verma, Detecting Deepfakes with Metric Learning, in: 2020 8th International Workshop on Biometrics and Forensics (IWBF), 2020, pp. 1–6.

[16] K. Feng, J. Wu, M. Tian, A Detect method for deepfake video based on full face recognition, in: 2020 IEEE International Conference on Information Technology,Big Data and Artificial Intelligence (ICIBA), 2020, pp. 1121–1125.

[17] N. Lei, Z. Luo, S. Yau, X. D. Gu, Geometric Understanding of Deep Learning, 2018. arXiv:1805.10451.

[18] B. Dolhansky, J. Bitton, B. Pflaum, J. Lu, R. Howes, M. Wang, C. Canton, The DeepFake Detection Challenge Dataset, 2020. arXiv:2006.07397.

[19] Y. Li, X. Yang, P. Sun, H. Qi, S. Lyu, Celeb-DF: A Large-Scale Challenging Dataset for DeepFake Forensics, in: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR, 2020, pp. 3204–3213.

[20] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, Y. Bengio, Generative Adversarial Nets, in: 2014 Annual Conference on Neural Information Processing Systems (NIPS), 2014, pp. 2672—-2680.

[21] X. Wang, X. Han, W. Huang, D. Dong, M. R. Scott, Multi-Similarity Loss With General Pair Weighting for Deep Metric Learning, in: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 5017–5025.

[22] Y. Wen, K. Zhang, Z. Li, Y. Qiao, A Discriminative Feature Learning Approach for Deep Face Recognition, in: Computer Vision - ECCV 2016 - 14th European Conference, Amsterdam, 2016, pp. 499–515.