Multi-Perspective Features Learning for Face Anti-Spoofing

Zhuming Wang[1], Yaowen Xu[1], Lifang Wu*[1,2], Hu Han[3], Yukun Ma[4], and Guozhang Ma[1]

[1]Beijing University of Technology, Beijing 100124, China
[2]Beijing Key Laboratory of Computational Intelligence and Intelligent System
[3]Key Laboratory of Intelligent Information Processing of Chinese Academy of Sciences (CAS),
Institute of Computing Technology, CAS, Beijing 100190, China
[4]Henan Institute of Science and Technology, Henan, China

Abstract

Face anti-spoofing (FAS) is important to securing face recognition. Most of the existing methods regard FAS as a binary classification problem between bona fide (real) and spoof images, training their models from only the perspective of Real vs. Spoof. It is not beneficial for a comprehensive description of real samples and leads to degraded performance after extending attack types. In fact, the spoofing clues in various attacks can be significantly different. Furthermore, some attacks have characteristics similar to the real faces but different from other attacks. For example, both real faces and video attacks have dynamic features, and both mask attacks and real faces have depth features. In this paper, a Multi-Perspective Feature Learning Network (MPFLN) is proposed to extract representative features from the perspectives of Real + Mask vs. Photo + Video and Real + Video vs. Photo + Mask. And using these features, a binary classification network is designed to perform FAS. Experimental results show that the proposed method can effectively alleviate the above issue of the decline in the discrimination of extracted features and achieve comparable performance with state-of-the-art methods.

1. Introduction

In recent years, with the wide application of face recognition technology [10] in identity authentication systems, such as financial payment and access control unlock, the attempts of presentation attacks (PAs) against it are also increasing. Using PAs like photo-print, video-replay, and 3D masks [11] , attackers can deceive the face recognition system easily, thus posing a severe threat to users' private property and even the public security of society. Therefore, It is needed to develop an effective face anti-spoofing (FAS) method.

Presentation attack detection (PAD) has been studied for over a decade. In the early days, researchers attempted to use human liveness cues [13, 21, 17] or handcrafted texture features [3, 6, 5, 15, 22, 27] to perform binary classification on bona fide (real) and spoof faces. Recently, with the rapid development of deep learning technology, convolutional neural networks (CNN) based methods [7, 16, 34, 18, 32] have shown significant performance improvement compared with traditional methods. Feng *et al*. [7], Li *et al*. [16] trained a deep neural network to learn real/spoof faces classification. Face depth map and rPPG signal were utilized in [18] as the auxiliary supervision to train the proposed model. A well-designed Central Difference Convolutional Network (CDCN) was used in [34] to extract face depth information for PAD which achieved competitive performance.

However, most existing methods considered FAS as a binary classification problem between real faces and spoof faces, ignoring the differences between different attack types and their commonness with real faces. Although some methods utilize auxiliary information as supervision, their perspective of group-level classification is still Real vs. Spoof in the training stage. It is feasible when there are only photo-print and video replay attacks due to their similar spoofing clues. However, when the 3D mask attack is considered, the limitation of binary classification is distinctly presented because 3D masks have quite different spoofing clues from photo and video attacks.

In fact, different attack types present different spoofing features, and spoofing features that can detect one particular attack type well (features this type of attack has while real faces do not) may not exist in another attack type, and trying to find some universal features that can detect multiple

*Corresponding author: Lifang Wu E-mail: lfwu@bjut.edu.cn

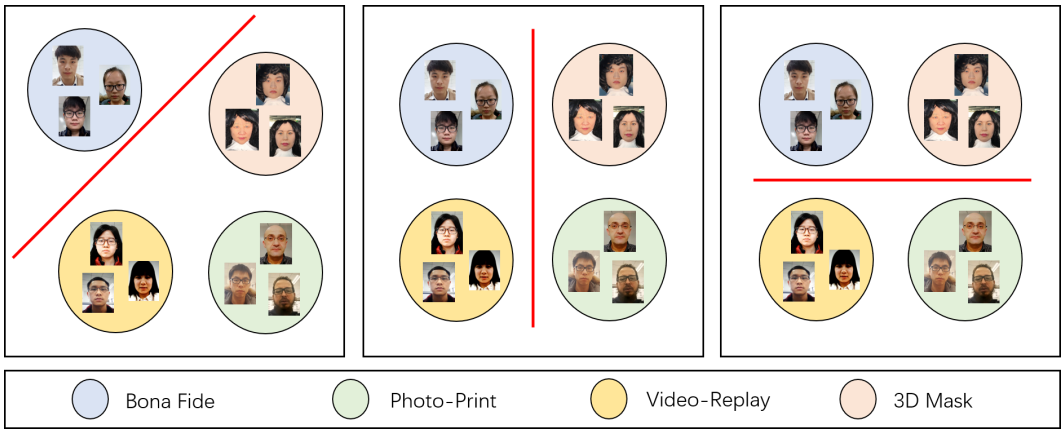

Figure 1. Sample division from different perspectives. Taking photos, videos, and masks as examples, both real faces and 3D masks have depth information while photos and videos do not, both real faces and videos have dynamic information while photos and 3D masks do not.

attack types possibly is a tradeoff of different attacks, which may not be optimal to each attack type. On the other hand, considered from different perspectives, there are common features between the attack types and real faces. For example, as shown in Fig.1, both real faces and 3D masks have depth information while photos and videos do not; both real faces and videos have dynamic information while photos and 3D masks do not. Therefore, by selecting appropriate perspectives, it is possible to obtain a more comprehensive description of real faces and avoid the network neglecting some features that have a strong discriminating ability between real and spoof faces yet are not universal among a variety of attack types, thus improve the performance of FAS.

Motivated by the discussions above, we propose a Multi-Perspective Features Learning strategy to extract more comprehensive description to real faces using a Real + 1 vs. Rest group-level classification training strategy, i.e., images of real faces and a given attack type are used as positive samples, and images of the rest of the attack types are used as negative samples for group-level classification. Furthermore, considering the wide application of depth and temporal information in PAD, two perspectives of Real + Mask vs. Rest and Real + Video vs. Rest are selected, and a novel Multi-Perspective Features Learning Network (MPFLN) is designed to extract the common features between real faces and video or mask attacks for FAS. The proposed network consists of two components: the Common Features Extraction Unit (CFEU) to extract the representative features and the Binary Classification Unit (BCU) that utilizes common features to classify real and spoof faces.

The main contributions of this work can be summarized as follows:

- We find the limitation of binary classification to FAS and propose a novel idea of Multi-Perspective Feature Learning, which can extract the features of real faces more comprehensively. These features have a strong discriminating ability between real and spoof faces yet are not universal among a variety of attack types.
- A novel neural network called Multi-Perspective Features Learning Network (MPFLN) is designed, which can simultaneously detect multiple types of attacks, including photos, videos, and masks, and effectively carry out the FAS task.
- The proposed MPFLN achieves competitive performance in the test database. Well-designed comparative experiments demonstrate the negative impact on the model's classification performance of retraining with the introduction of new attack types and the effectiveness of the method of Multi-Perspective Feature Learning to alleviate this problem.

2. Related work

Traditional approaches. Traditional methods usually focus on the texture difference between real and spoof faces brought by spoof media, using handcrafted features extracted from images, such as LBP [3, 5], HOG [15], SIFT [22], and DoG [27], to carry out binary classification of real and spoof faces. However, these features are usually based on 2D features, such as the noise brought by printing paper or video replay equipment, which is difficult to be applied on the 3D mask attacks. Similarly, the texture features extracted for masks [23, 1] are also difficult to be applied to the detection of photos and videos. There are also several methods considering temporal information and using dynamic texture [14, 28] or spontaneous human movement [13, 21] like blinking and head motion to perform PAD. These methods have achieved good performance in detecting photos and

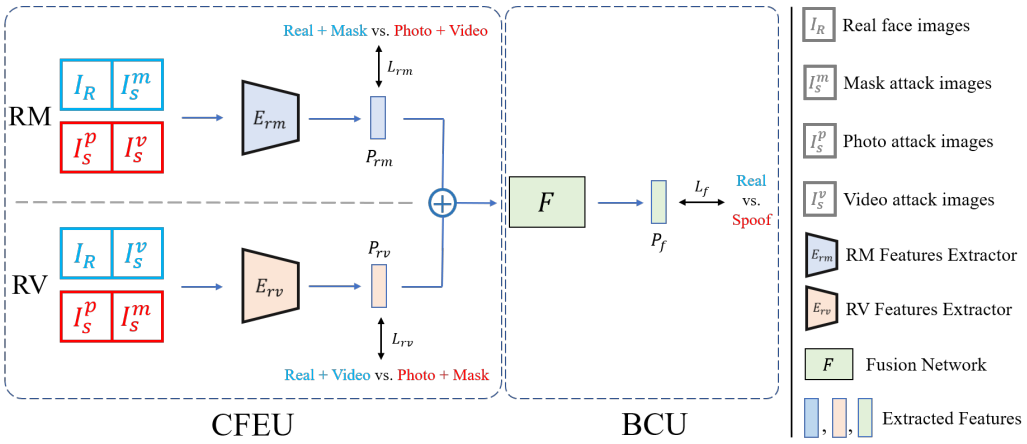

Figure 2. The overall framework of the proposed MPFLN (Taking photos, videos, and masks as examples).

static videos, and mask attack detection methods using dynamic information [24, 25] also have certain effects. Still, these methods are difficult to detect attacks that naturally contain dynamic information such as video replay.

Deep learning approaches. With the rapid development of deep learning, some deep learning based approaches were proposed. Feng *et al*. [7], Li *et al*. [16] using pre-trained CNN as a feature extractor to distinguish real and spoofing faces. Xu *et al*. [31] using LSTM architecture to combine spatial information with temporal information for attack detection. In addition to methods considering FAS as a simple binary classification problem supervised by the binary cross-entropy loss, some works using pseudo depth label [2, 35], rPPG signal [20], binary mask label [9, 26, 19], Fourier map [12], LBP texture Map [35], etc., as auxiliary supervision, have also achieved good performance. Liu *et al*. [18] introduced face depth map and rPPG signal as auxiliary supervision guiding feature extraction. Yu *et al*. [34] using a well-designed Central Difference Convolutional Network (CDCN) to extract face depth information for PAD and achieved good performance.

However, most existing work only focuses on dealing with photo and video attacks and has difficulty covering mask attacks. Also, approaches aimed to mask attacks can not deal well with the former attacks either. The reason is that the differences between them are so significant that it's hard to find some universal clues which can identify multiple attacks at the same time. Trying to find such clues can lead to a decline in the discrimination on spoofing features of each attack type. Therefore, we use a Multi-Perspective Feature Learning method to retain clues that have a strong discriminating ability for each attack type, thus preventing a decrease in the discrimination on features.

3. Methodology

The main idea of the proposed method is to extract the representative features from multiple perspectives to describe real faces more discriminatively. As shown in Fig.2, the proposed network consists of two modules: Common Features Extraction Unit (CFEU) and Binary Classification Unit (BCU). For the convenience of description, we only take common attacks, including photo-print, video-replay, and 3D mask as examples, to introduce the proposed network in this paper.

3.1. Common Features Extraction Unit

The Common Features Extraction Unit is used to extract the common features with strong discrimination between real faces and various attack types from multiple perspectives. Therefore, unlike the traditional binary classification, CFEU uses the perspective of Real + 1 vs. Rest instead of Real vs. Spoof for group-level classification training. In addition, considering the wide application of depth information and timing information in PAD, we only build two parts in CFEU: Real_Mask part (RM) and Real_Video part (RV), which are beneficial to the extraction of depth and temporal information, respectively, but not the Real_Photo part whose features are not prominently.

Real_Mask part. The function of RM is to extract the common features with strong discrimination between real faces and mask attacks. For this purpose, we adopt the Real + Mask vs. Rest group-level classification training strategy. Specifically, let I_R, I_S^p, I_S^v, I_S^m denote the set of the real face images, photo attack images, video attack images, and mask attack images in the whole sample space respectively. For any input image $X \in I$, we use PRNet [8] to generate its pseudo depth map $D \in \mathbb{R}^{d \times d}$, and its corresponding group-

level classification label Y_{rm} can be formulated as

$$Y_{rm} = \begin{cases} D, & X \in (I_R \cup I_S^m) \\ Z, & X \in (I_S^p \cup I_S^v) \end{cases}, \quad (1)$$

where $Z \in \mathbb{R}^{d \times d}$ is a 'zero map' with the same shape as D.

Then, we extract the feature map $P_{rm} = E_{rm}(X)$ as the predicted depth map of X by RM Features Extractor (E_{rm}), and Y_{rm} is used as the supervision to train the extraction capacity of group-level classification features (the common features between Real and Mask) of E_{rm}. Here, E_{rm} uses the famous CDCN [34] as the backbone network. Therefore, the loss function of RM can be formulated as

$$L_{rm} = \sum_{i=1}^{N} L_{MSE}(P_{rm}^i, Y_{rm}^i) + L_{CDL}(P_{rm}^i, Y_{rm}^i), \quad (2)$$

where N is the total samples number of the training set, L_{MSE} and L_{CDL} denote mean square error loss and contrastive depth loss [29], respectively. Notice that, although the network structure and the loss function of RM are basically the same as the backbone network, the meaning of RM's group-level classification label (Real + Mask vs. Rest) has a completely different meaning than the traditional one (Real vs. Spoof).

Real_Video part. Similar to RM, RV is used to extract the common features with strong discrimination between real faces and video attacks by adopting a Real + Video vs. Rest group-level classification training strategy. Since CDCN is mainly used to extract depth information but is not good at extracting temporal information, we use 3D-CDCN [30], a variant of CDCN, as the backbone network of E_{rm}. Compared with CDCN, 3D-CDCN can extract the temporal information of input frame sequences, which is more suitable for the group-level classification criterion of RV.

Therefore, the process of RV is modified as: for any input frame sequence $X \in I$, we extract the feature map $P_{rv} = E_{rv}(X)$ of X by RV Features Extractor (E_{rv}) and use Y_{rv} as the supervision to train the extraction capacity of group-level classification features (the common features between Real and Video) of E_{rv}. Consistent with 3D-CDCN, the corresponding group-level classification label Y_{rv} of X can be formulated as

$$Y_{rv} = \begin{cases} O, & X \in (I_R \cup I_S^v) \\ Z, & X \in (I_S^p \cup I_S^m) \end{cases}, \quad (3)$$

where $O \in \mathbb{R}^{d \times d}$ is a 'one map' with the same shape as Z.

The loss function of RV can be formulated as

$$L_{rv} = \sum_{i=1}^{N} L_{MSE}(P_{rv}^i, Y_{rv}^i), \quad (4)$$

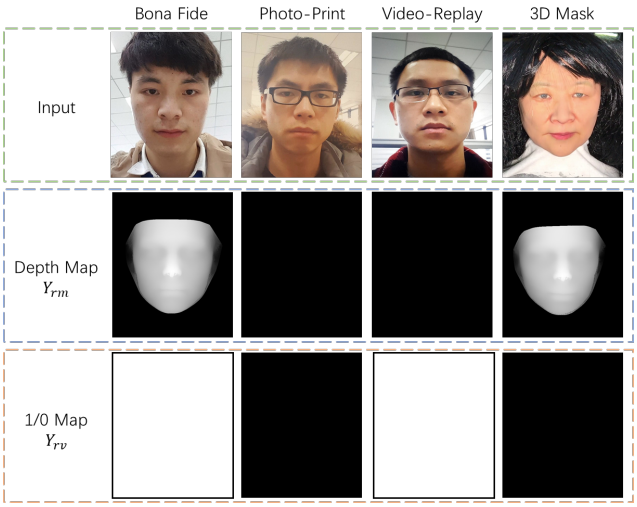

Figure 3. Input images and their corresponding pseudo depth maps and 1/0 maps. Here, samples of bona fide faces, photo-prints, and video-replays are from OULU-NPU, and samples of 3D masks are from CASIA-SURF 3DMask.

3.2. Binary Classification Unit

BCU is used to classify the real faces and the spoof attacks through the E_{rm} and E_{rv} obtained from CFEU. Specifically, for any input X, we concatenate its corresponding P_{rm}, P_{rv} obtained from CFEU, then fed them into the fusion network F consisting of three convolution layers and get the output $P_f = F(P_{rm}, P_{rv})$. The corresponding real/spoof classification label Y_f of X can be formulated as

$$Y_f = \begin{cases} O, & X \in I_R \\ Z, & X \in (I_S^p \cup I_S^v \cup I_S^m) \end{cases}, \quad (5)$$

The loss function of BCU module formulated as

$$L_f = \sum_{i=1}^{N} L_{MSE}(P_f^i, Y_f^i), \quad (6)$$

In the testing stage, we calculate the mean value of P_f as the final score of the real/spoof binary classification task.

Notice that, we suggest that CFEU and BCU should be trained in two-stage rather than end-to-end format, i.e., the CFEU part should be trained first, and its parameters should be fixed during training the BCU part. The reason is that the classification labels of RM, RV conflict with F, thus optimizing F leads to an offset towards the corresponding negative samples of the mask class in RM and the video class in RV, which is contrary to the motivation of our proposed method.

4. Experiments

4.1. Datasets and Protocols

Databases. Two datasets, including OULU-NPU [4] and CASIA-SURF 3DMask [33], are used in our experiments. Oulu-NPU contains high-resolution attack samples of photo-prints and video-replays, but does not contain 3D mask attack samples. To this end, we introduce the mask attack samples from CASIA-SURF 3DMask as a supplementary set to the former dataset.

Protocols. Based on the four test protocols of OUlU-NPU, we extend the attack type of the 3D mask and design two sets of experiments. In the first set, we only expanded the training set by mask samples with the same number of photos and videos, while the validation set and testing set remained unchanged. This set of experiments is named Experiment-O, which is used to prove that our method can achieve comparable performance with state-of-the-art methods. In the second set, we expanded mask samples in not only the training set but the validation set and the testing set also. This set of experiments is named Experiment-M, which is used to prove that, by extracting representative features from multiple perspectives, the decline in the discrimination after extending attack types suffered by existing methods can be alleviated effectively.

Performance Metrics. We adopt Attack Presentation Classification Error Rate (APCER), Bona Fide Presentation Classification Error Rate (BPCER), and Average Classification Error Rate (ACER) as the performance metrics.

4.2. Implementation Details

Ground Truth Generation. Although MPFLN adopts the fully convolutional structure and has no requirements on the size of input images, for the convenience of analysis, we detect the face regions of input images then crop and resize their scales to 256×256. Consistent with CDCN, we adopt PRNet [8] to generate pseudo depth maps of images, resize their scales to 32×32, and normalized them to [0, 1]. Accordingly, the scales of all the 1/0 maps used in the model are also 32×32. Input images and their corresponding pseudo depth maps and 1/0 maps are shown in Fig.3.

Experimental Setting. Our proposed method is implemented with Pytorch. In the training stage, RM, RV, and BCU are trained with Adam optimizer, and the initial learning rate (lr) is 1e-4, 1e-3, and 5e-4, respectively, and weight decay (wd) is 1e-5. We train RM, RV with a maximum of 600 epochs while lr halves every 300 epochs and set 500/200 for BCU. The batch size is 16.

4.3. Experimental Comparison

Experiment-O. Since most of the comparison methods only used photo and video attacks in training, although MPFLN has the ability to detect mask attacks, we still only considered the photo and video attacks in this set of experiments. In other words, mask samples were only introduced in the training stage of MPFLN, while the testing stage was consistent with the comparison experiments, using OULU-NPU samples only and following its four protocols. Notice that, although mask samples are introduced into the training set, they are not conducive to be used as expanded data to enhance the discrimination of the other two attack types due to the significant difference between them. On the contrary, this large difference leads to a deviation of the feature extraction direction of the model after introducing mask attack samples in training, which eventually raises the risk of the decrease of the model's overall performance.

As shown in Table 1, our method can achieve comparable performance with the state-of-the-art methods on most protocols, only has a decline in the performance on protocol 4. It means that, although our method can alleviate the decline of feature discrimination caused by introducing mask attack samples in most conditions, it is still impacted severely on protocol 4. The reason for this is, while

Prot.	Method	APCER(%)	BPCER(%)	ACER(%)
1	STASN [32]	1.2	2.5	1.9
	Auxiliary [18]	1.6	1.6	1.6
	FaceDs [12]	1.2	1.7	1.5
	Disentangled [35]	1.7	0.8	1.3
	CDCN [34]	0.4	1.7	**1.0**
	MPFLN	1.0	1.7	1.3
2	FaceDs [12]	4.2	4.4	4.3
	Auxiliary [18]	2.7	2.7	2.7
	Disentangled [35]	1.1	3.6	2.4
	STASN [32]	4.2	0.3	2.2
	CDCN [35]	1.5	1.4	1.5
	MPFLN	2.2	0.6	**1.4**
3	FaceDs [12]	4.0±1.8	3.8±1.2	3.6±1.6
	Auxiliary [18]	2.7±1.3	3.1±1.7	2.9±1.5
	STASN [32]	4.7±3.9	0.9±1.2	2.8±1.6
	CDCN [34]	2.4±1.3	2.2±2.0	2.3±1.4
	Disentangled [35]	2.8±2.2	1.7±2.6	2.2±2.2
	MPFLN	2.2±4.0	1.1±3.9	**1.7±1.5**
4	Auxiliary [18]	9.3±5.6	10.4±6.0	9.5±6.0
	STASN [32]	6.7±10.6	8.3±8.4	7.5±4.7
	CDCN [34]	4.6±4.6	9.2±8.0	6.9±2.9
	FaceDs [12]	1.2±6.3	6.1±5.1	5.6±5.7
	Disentangled [35]	5.4±2.9	3.3±6.0	**4.4±3.0**
	MPFLN	10.0±15.0	5.0±5.0	7.5±5.0
	MPFLN+	7.1±7.1	5.0±10.0	6.0±3.5

Table 1. Results of Experiment-O on four protocols of Oulu-NPU.

Prot.	Method	APCER(%)	BPCER(%)	ACER(%)
1	CDCN-rs	1.7	4.2	2.9
	MPFLN-rs	0.2	5.0	2.6
	MPFLN	1.5	0.8	**1.2**
2	CDCN-rs	6.6	0.3	3.5
	MPFLN-rs	5.5	0.8	3.2
	MPFLN	1.9	1.1	**1.5**
3	CDCN-rs	2.5±2.2	5.3±14.8	3.9±6.3
	MPFLN-rs	2.0±2.7	2.2±6.1	2.1±2.4
	MPFLN	1.3±2.0	1.1±3.9	**1.2±1.5**
4	CDCN-rs	8.6±8.1	5.8±9.2	7.2±4.5
	MPFLN-rs	4.5±3.9	5.0±10.0	4.7±3.9
	MPFLN	7.2±12.8	4.2±5.8	5.7±4.3
	MPFLN+	4.2±4.2	4.2±5.8	**4.2±2.5**

Table 2. Results of Experiment-M on four protocols of Oulu-NPU.

MPFLN can extract highly discriminating common features between real faces and attack types, some features unique to real faces themselves are ignored. In the first three protocols with a large number of training samples, MPFLN can learn enough real face sample features for classification. Still, in the fourth protocol, which has a small number of training samples, MPFLN could not learn enough features for classification.

Therefore, we propose an extended network version MPFLN+, which adds a Real vs. Spoof perspective to the original two perspectives to supplement some features unique to real faces. Specifically, we added the R branch paralleling RM and RV in CFEU, whose structure was consistent with CDCN, taking real faces as positive samples, various attack types as negative samples. As can be seen from both Table 1 and Table 2, MPFLN+ has a significant performance improvement over the MPFLN.

Experiment-M. In structure, MPFLN can easily be confused with the Model Stacking strategy. However, the two are fundamentally different in method. The purpose of MPFLN is to extract the representative features from multiple perspectives for classification. Hence the label of CFEU is in Real + 1 vs. Rest form, rather than the traditional Real vs. Spoof form. To demonstrate that the performance improvement of MPFLN is due to the group level classification strategy rather than the Model Stacking strategy, we design this set of experiments. We compare the performance of three models: CDCN-rs, whose structure is identical to CDCN, but is retrained after introducing mask attacks in the negative sample; MPFLN-rs, whose structure is identical to MPFLN, but the CFEU module classify the samples by Real vs. Spoof perspective; MPFLN, the proposed method, in which CFEU module classify the samples by Real + 1 vs. Rest perspective.

As shown in Table 2, CDCN-rs has a significant decline in performance compared with CDCN. This decline comes from the fact that CDCN is good at extracting depth information as clues to classify the positive and negative samples, but such clues could not distinguish real faces well from the newly introduced mask samples. On the contrary, setting masks as negative samples will lead to confusion of the depth estimation, resulting in the weakening of extracted features' discrimination. This experimental result supports our assumption of model performance's degradation because of the introduction of new attack types. Furthermore, the performance of MPFLN-rs improves over CDCN-rs due to the added 3D-CDCN structure and Model Stacking strategy. However, although the performance of MPFLN-rs is improved compared to CDCN-rs, it is still much lower than MPFLN, suggesting that the alleviation on feature discrimination's decline is indeed due to the strategy of extracting features by group-level classification on multiple perspectives, rather than Model Stacking strategy.

5. Conclusions

In this paper, we design a Multi-Perspective Features Learning method to extract the representative features from multiple perspectives by Real+1 vs. Rest group level classifying. Base on this method, we design a neural network named Multi-Perspective Features Learning Network (MPFLN), which learned features from Real+Mask and Real+Video perspectives to perform PAD. Experiments prove the effectiveness of the proposed strategy and the model.

The possible future directions include: 1) We only performed experiments on OULU-NPU and proved the effectiveness of our method, but the studies on other datasets are still necessary; 2) Although we designed MPFLN+ to improve the performance on protocol 4, experiments on other protocols should still be performed. 3) The method of Multi-Perspective Features Learning could theoretically be extended to any backbone network or any number of attack types, which has not yet been proven.

Acknowledgements

This work was supported by the Beijing Municipal Education Committee Science Foundation under Grant KM201910005024, and the Beijing Postdoctoral Research Foundation under Grant Q6042001202101.

References

[1] A. Agarwal, R. Singh, and M. Vatsa. Face anti-spoofing using haralick features. In *2016 IEEE 8th International Conference on Biometrics Theory, Applications and Systems (BTAS)*, pages 1–6, 2016. 2

[2] Y. Atoum, Y. Liu, A. Jourabloo, and X. Liu. Face anti-spoofing using patch and depth-based cnns. In *2017 IEEE International Joint Conference on Biometrics (IJCB)*, pages 319–328, 2017. 3

[3] Z. Boulkenafet, J. Komulainen, and A. Hadid. Face anti-spoofing based on color texture analysis. In *2015 IEEE International Conference on Image Processing (ICIP)*, pages 2636–2640, 2015. 1, 2

[4] Z. Boulkenafet, J. Komulainen, L. Li, X. Feng, and A. Hadid. Oulu-npu: A mobile face presentation attack database with real-world variations. In *IEEE International Conference on Automatic Face & Gesture Recognition*, pages 612–618, 2017. 5

[5] T. de Freitas Pereira, A. Anjos, J. M. De Martino, and S. Marcel. Lbp top based countermeasure against face spoofing attacks. In *Asian Conference on Computer Vision*, pages 121–132, 2012. 1, 2

[6] T. de Freitas Pereira, A. Anjos, J. M. De Martino, and S. Marcel. Can face anti-spoofing countermeasures work in a real world scenario? In *2013 International Conference on Biometrics (ICB)*, pages 1–8, 2013. 1

[7] L. Feng, L. Po, and Y. Li. Integration of image quality and motion cues for face anti-spoofing: A neural network approach. *Journal of Visual Communication & Image Representation*, 38:451–460, 2016. 1, 3

[8] Y. Feng, F. Wu, X. Shao, Y. Wang, and X. Zhou. Joint 3d face reconstruction and dense alignment with position map regression network. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 557–574, 2017. 3, 5

[9] A. George and S. Marcel. Deep pixel-wise binary supervision for face presentation attack detection. In *2019 International Conference on Biometrics (ICB)*, pages 1–8, 2019. 3

[10] J. Guo, X. Zhu, C. Zhao, D. Cao, Z. Lei, and S. Z. Li. Learning meta face recognition in unseen domains. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 6162–6171, 2020. 1

[11] S. Jia, G Guo, and Z. Xu. A survey on 3d mask presentation attack detection and countermeasures. *Pattern recognition*, 98:107032–107032, 2020. 1

[12] A. Jourabloo, Y. Liu, and X. Liu. Face de-spoofing: Anti-spoofing via noise modeling. In *European Conference on Computer Vision*, pages 290–306, 2018. 3, 5

[13] K. Kollreider, H. Fronthaler, M. I. Faraj, and J. Bigun. Real-time face detection and motion analysis with application in liveness assessment. *IEEE Transactions on Information Forensics and Security*, 2(3):548–558, 2007. 1, 2

[14] J. Komulainen, A. Hadid, and M. Pietikinen. Face spoofing detection using dynamic texture. In *Asian Conference on Computer Vision*, pages 146–157. Springer, 2012. 2

[15] J. Komulainen, A. Hadid, and M. Pietikinen. Context based face anti-spoofing. In *IEEE Biometrics Theory, Applications, and Systems*, pages 1–8, 2013. 1, 2

[16] L. Li, X. Feng, Z. Boulkenafet, Z. Xia, M. Li, and A. Hadid. An original face anti-spoofing approach using partial convolutional neural network. In *2016 Sixth International Conference on Image Processing Theory, Tools and Applications (IPTA)*, pages 1–6, 2016. 1, 3

[17] X. Li, J. Komulainen, G. Zhao, P. C. Yuen, and M. Pietikinen. Generalized face anti-spoofing by detecting pulse from face videos. In *2016 23rd International Conference on Pattern Recognition (ICPR)*, pages 4244–4249, 2016. 1

[18] Y. Liu, A. Jourabloo, and X. Liu. Learning deep models for face anti-spoofing: Binary or auxiliary supervision. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 389–398, 2018. 1, 3, 5

[19] Y. Liu, J. Stehouwer, A. Jourabloo, and X. Liu. Deep tree learning for zero-shot face anti-spoofing. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4675–4684, 2019. 3

[20] X. Niu, H. Han, S. Shan, and X. Chen. Continuous heart rate measurement from face: A robust rppg approach with distribution learning. In *2017 IEEE International Joint Conference on Biometrics (IJCB)*, pages 642–650, 2017. 3

[21] G. Pan, L. Sun, Z. Wu, and S. Lao. Eyeblink-based anti-spoofing in face recognition from a generic webcamera. In *2007 IEEE 11th International Conference on Computer Vision*, pages 1–8, 2007. 1, 2

[22] K. Patel, H. Han, and A. K. Jain. Secure face unlock: Spoof detection on smartphones. *IEEE Transactions on Information Forensics and Security*, 11(10):2268–2283, 2016. 1, 2

[23] R. Shao, X. Lan, and P. C. Yuen. Deep convolutional dynamic texture learning with adaptive channel-discriminability for 3d mask face anti-spoofing. In *2017 IEEE International Joint Conference on Biometrics (IJCB)*, pages 748–755, 2017. 2

[24] R. Shao, X. Lan, and P. C. Yuen. Joint discriminative learning of deep dynamic textures for 3d mask face anti-spoofing. *IEEE Transactions on Information Forensics and Security*, 14(4):923–938, 2019. 3

[25] T. A. Siddiqui, S. Bharadwaj, T. I. Dhamecha, A. Agarwal, M. Vatsa, R. Singh, and N. Ratha. Face anti-spoofing with multifeature videolet aggregation. In *2016 23rd International Conference on Pattern Recognition (ICPR)*, pages 1035–1040, 2016. 3

[26] W. Sun, Y. Song, C. Chen, J. Huang, and A. C. Kot. Face spoofing detection based on local ternary label supervision in fully convolutional networks. *IEEE Transactions on Information Forensics and Security*, 15:3181–3196, 2020. 3

[27] X. Tan, L. Yi, J. Liu, and J. Lin. Face liveness detection from a single image with sparse low rank bilinear discriminative model. In *European Conference on Computer Vision*, pages 504–517. Springer, 2010. 1, 2

[28] Z. Wang, Z. Yu, C. Zhao, X. Zhu, Y. Qin, Q. Zhou, F. Zhou, and Z. Lei. Deep spatial gradient and temporal depth learning for face anti-spoofing. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5041–5050, 2020. 2

[29] Z. Wang, C. Zhao, Y. Qin, Q. Zhou, and Z. Lei. Exploiting temporal and depth information for multi-frame face anti-spoofing. arXiv preprint arXiv:1811.05118. 4

[30] Y. Xu, Z. Wang, H. Han, L. Wu, and Y. Liu. Exploiting non-uniform inherent cues to improve presentation attack detection. In *2021 IEEE International Joint Conference on Biometrics (IJCB)*, pages 1–8, 2021. 4

[31] Z. Xu, S. Li, and W. Deng. Learning temporal features using lstm-cnn architecture for face anti-spoofing. In *2015 3rd IAPR Asian Conference on Pattern Recognition (ACPR)*, pages 141–145, 2015. 3

[32] X. Yang, W. Luo, L. Bao, Y. Gao, D. Gong, S. Zheng, Z. Li, and W. Liu. Face anti-spoofing: Model matters, so does data. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 3502–3511, 2019. 1, 5

[33] Z. Yu, J. Wan, Y. Qin, X. Li, S. Z. Li, and G. Zhao. Nas-fas: Static-dynamic central difference network search for face anti-spoofing. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 1–1, 2020. 5

[34] Z. Yu, C. Zhao, Z. Wang, Y. Qin, Z. Su, X. Li, F. Zhou, and G. Zhao. Searching central difference convolutional networks for face anti-spoofing. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 5294–5304, 2020. 1, 3, 4, 5

[35] K. Y. Zhang, T. Yao, J. Zhang, Y. Tai, S. Ding, J. Li, F. Huang, H. Song, and L. Ma. Face anti-spoofing via disentangled representation learning. In *European Conference on Computer Vision*, pages 641–657. Springer, 2020. 3, 5