A Dual-stream Framework for 3D Mask Face Presentation Attack Detection

Shen Chen[1]*, Taiping Yao[1]*, Keyue Zhang[1], Yang Chen[1], Ke Sun[1,2]
Shouhong Ding[1]†, Jilin Li[1], Feiyue Huang[1], Rongrong Ji[2]
[1] YouTu Lab, Tencent [2] Media Analytics and Computing Lab, Xiamen University
{kobeschen, taipingyao, zkyezhang, wizyangchen, ericshding, jerolinli, garyhuang}@tencent.com
skjack@stu.xmu.edu.cn, rrji@xmu.edu.cn

Abstract

*Face presentation attack detection (PAD) plays a vital role in face recognition systems. Many previous face anti-spoofing methods mainly focus on the 2D face representation attacks, which however, suffer from great performance degradation when facing high-fidelity 3D mask attacks. To address this issue, we propose a novel dual-stream framework consisting of the vanilla convolution stream and the central difference convolution stream. These two streams complement each other and learn more comprehensive features for 3D mask attacks detection. Moreover, we extend 3D PAD to a multi-classification task that contains real face, plaster attack and transparent attack, and utilize various data augmentations and label smoothing techniques to improve the generalizability on unseen attacks. The proposed method achieved the second place in the Chalearn 3D High-Fidelity Mask Face Presentation Attack Detection Challenge@ICCV2021 with a score of **3.15** (ACER).*

1. Introduction

Face recognition technologies [41, 11, 17] have been widely used in personal verification and identification due to their convenience and remarkable accuracy. Despite the recent noticeable advances, the security of face recognition systems (FRS) is still vulnerable to presentation attacks (PA) [63, 25, 29]. An impostor can fool the FRS simply by presenting a face artifact, which is also known as a presentation attack instrument [45].

Based on the way of generating face artifacts, face presentation attacks can be divided into 2D attacks (*e.g.*, print attacks [63] or video replay [9]) and 3D attacks (*e.g.*, by wearing a mask [35]). Existing research on FRS has paid more attention to 2D attacks due to its simplicity, efficiency, and low cost. However, as material science and 3D printing technology advance, creating face-like 3D structures or materials has become easier and more affordable. Compared with traditional 2D attacks, 3D face masks are more realistic in terms of color, texture, and geometry structure, making them more challenging to be detected. In Figure 1, we show some samples from the recently released large-scale High-Fidelity Mask dataset [26], including resin, transparent, and plaster 3D attacks.

Figure 1. **The samples from the HiFiMask dataset** [26].

The vulnerability of current FRS to realistic face presentation attacks has facilitated a series of studies [54, 55, 57] on 3D face presentation attack detection (PAD). Early methods tried to explore the difference between real face skin and 3D fake face materials based on the reflectance properties [46], texture analysis [29] or shape descriptors [52]. They have achieved considerable performance on several coarse 3D face masks datasets [28, 33, 16], but are not robust to the high-fidelity mask attacks [26]. Some recent studies [3, 43, 27] utilized deep features for 3D PAD and achieved promising detection performance. However, these methods suffer from performance degradation in the face of unseen 3D attacks. In this paper, inspired by the lightweight network ResNet9 [10], we develop an efficient dual-stream framework that includes a vanilla convolution branch and

*Equal contribution.
†Corresponding authors.

a central difference convolution (CDC) [59] branch. CDC has been proven in previous work [59] to effectively capture the intrinsic detailed patterns. Combined with the vanilla convolution, it can further promote the distinction between real skin and high-fidelity 3D masks. Unlike most methods that treat PAD as a binary classification task, we extend it to a multi-classification task, *i.e.*, distinguishing between real face and each material type of attack. Through multi-class learning, our method learns the nature of attacks and has a stronger discriminative ability. We also utilize extensive data augmentations and label smoothing technique [47] during training to further improve generalizability on unseen attacks. As a result, our method achieves a score of **3.15** (ACER) in the Chalearn 3D High-Fidelity Mask Face Presentation Attack Detection Challenge@ICCV2021[1], which is ranked second place over all teams.

In summary, our main contribution is three-fold:

• We propose a novel dual-stream framework that combines vanilla convolutional stream and central difference convolutional stream, which can complement each other to detect high-fidelity 3D mask attacks.

• We address the 3D face presentation attacks detection task via multi-class learning and use extensive data augmentations and label smoothing techniques to improve the generalizability on unseen attacks.

• Our method achieved the second place in the Chalearn 3D High-Fidelity Mask Face Presentation Attack Detection Challenge@ICCV2021 with a score of **3.15** (ACER).

2. Related work

In recent decades, PAD technologies [56, 58, 57, 61, 31] developed rapidly, which could be categorized into two stages. Initially, some researchers utilized traditional handcrafted features, such as LBP [5], SIFT [37], HOG [22], to extract the related information from the facial images, then trained a classifier to distinguish the fake and real faces. However, such methods didn't work very well because of the limited representation abilities. To solve this limitation, deep learning was introduced into PAD. The work in [14] trained CNNs to learn a binary classifier, which was easily overfitting on specfic attacks. Some auxiliary information [60, 51, 62, 7] was introduced to regularize the feature spaces, such as depth map, reflection map, rPPG signals. Besides, some works [40, 39, 8, 30] adopted domain generalization or meta-learning further to improve the generalization.

However, the performance of the above methods may degrade encountering the 3D attacks, since they mainly focus on 2D attacks and detection of 3D fake faces is more challenging than detecting fake faces with 2D planar surfaces. To specifically defend such 3D attacks, earlier studies [19, 53, 46] extracted the reflectance difference between real face skin and mask material. Texture-based methods explore the texture pattern difference of real faces and masks with the help of texture feature descriptors, such as the widely used LBP [35, 13] and Haralick features [2]. Shape-based 3D mask PAD methods use shape descriptors [23, 49, 18] or 3D reconstruction [52] to extract discriminative features from faces and 3D masks. Different from reflectance-based or texture-based detection methods, these schemes only require standard color images without the need for special sensors. However, their detection performances rely on the quality of 3D mask attacks, and may not be robust to super realistic 3D face presentation attacks. Instead of extracting hand-crafted features, deep-feature based methods automatically extract features from face images. Two deep representation approaches were investigated in [34] for spoofing detection in different biometric modalities. Image quality cues (Shearlet) and motion cues (dense optical flow) were fused in [15] using a hierarchical neural network for mask spoofing detection. A network based on transfer learning using a pre-trained VGG-16 model architecture is presented in [32] to recognize the photo, video, and 3D mask attacks. Based on the observation of the importance of dynamic facial texture information, a deep convolutional neural network-based approach was developed in [43, 44].

Despite these advances in 3D face anti-spoofing, there are limitations in each category of detection methods. For example, the main limitation of reflectance-based methods is the requirement of special and expensive devices to acquire multispectral images at varying wavelengths. Although the texture and shape-based methods are easy-to-implement, their robustness to different mask spoofing attacks needs further investigation. Deep-feature based approaches are generally sensitive to dataset sizes and lack transparency. In addition, most of them still suffer from performance degradation when applied to databases with more realistic face spoofing attacks. Differently, in this paper, we combine vanilla convolution and central difference convolution to construct a more robust representation and use techniques such as multi-class learning the improve the generalizability on unseen attacks.

3. Method

In this section, we first present the data preprocessing for the HIFIMask dataset used in Chalearn 3D High-Fidelity Mask Face Presentation Attack Detection Challenge@ICCV2021 [26], including four strategies: face detection, noise removal, black edge removal and data augmentations. Then we introduce a dual-stream network that combines a vanilla convolution branch and a central differential convolution branch. Next, we describe the loss function used in our method, namely multi-class loss, multi-

[1] https://competitions.codalab.org/competitions/30910

Figure 2. **Three typical bad images in the HIFIMask dataset.**

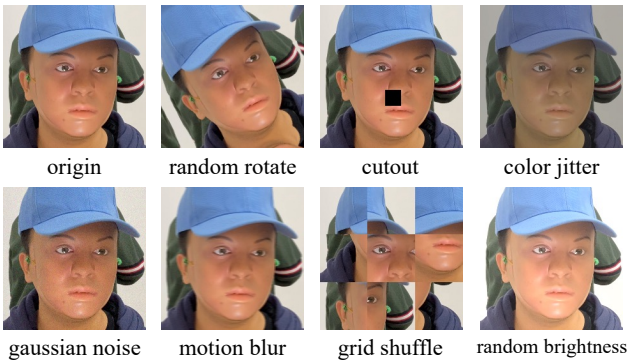

Figure 3. **The examples of augmented images.**

head loss and label smoothing. Finally, we introduce post-processing to minimize the distribution gap between validation and test sets, called logits temperature scaling. The entire framework of our method is illustrated in Figure 4.

3.1. Data Preprocessing

Although the HIFIMask dataset has provided face crops detected by Dlib [20], there are still some problems with these images. Figure 2 summarizes three typical *bad images* in the HIFIMask dataset, *i.e.*, small face, noise images and black edge. In this work, we propose several strategies to solve the above problems by using a high-precision face detector, noise removal and black edge removal, respectively.

Face Detection. Dual Shot Face Detector (DSFD) [24] is a one-stage efficient face detector that achieves state-of-the-art on several benchmarks. To this end, we use DSFD to detect faces in the HIFIMask dataset and enlarge the face box by 1.5 times to include more face information.

Noise Removal. We calculate the face confidences for the HIFIMask dataset via DSFD, where we observe that some images have small face confidence, even close to 0. Therefore, we remove all the images with face confidence less than 0.9 in the train set to avoid the interference caused by noises.

Black Edge Removal. We design a simple but effective black edge removal algorithm to delete the black edge region of the image. In particular, we first convert the input image x to a gray map x_{gray}, then find all pixels with value 0 to obtain a binary map x_{binary}. Subsequently, we scan x_{binary} from four directions to find the black edge. During the scanning process, if the average pixel value of a row (or column) is larger than a predefined threshold t_1 (set as 0.9 in our work), the row (or column) is considered as a black border, otherwise, the scanning will stop; if the number of black borders is less than a predefined threshold t_2 (set to 5 in our work), we treat it as a normal image. When the scanning process is finished, we obtain a box $[a_1, a_2, b_1, b_2]$ to remove the black edge of the image.

Data Augmentations. During training, we use extensive data augmentations, as follows: *random rotate*, *cutout* [12], *color jitter*, *gaussian noise*, *motion blur*, *grid shuffle*, *random brightness contrast*, *etc*. All the above augmentations are implemented through the albumentations [6] library. Figure 3 presents some examples of augmented faces, and most of them are visually difficult to recognize and closer to real-world scenarios.

3.2. Network Architecture

In this paper, we use the lightweight network ResNet9 [10] as the backbone and combine vanilla convolution and central difference convolution (CDC) [59] to develop a dual-stream network that can better discover the intrinsic patterns in high-fidelity 3D mask attacks.

ResNet9. As stated in [10], ResNet9 is a lightweight network which has only nine layers, and uses a smooth CELU [4] instead of ReLU [1] as the activation function for better optimization. In addition, the pooling layer is placed behind the convolution operation, which can effectively reduce the inference time.

Vanilla Convolution. As 2D spatial convolution is the basic operation in CNN for vision tasks, here we denote it as vanilla convolution and review it shortly first. There are two main steps in the 2D convolution: 1) sampling local receptive field region R over the input feature map x; 2) aggregation of sampled values via weighted summation. Hence, the output feature map f can be formulated as:

$$f(p_0) = \sum_{p_n \in \mathcal{R}} w(p_n) \cdot x(p_0 + p_n) \tag{1}$$

where p_0 denotes current location on both input and output feature maps while p_n enumerates the locations in $\mathcal{R}$. For instance, the local receptive field region for convolution operation with 3×3 kernel and dilation 1 is $\mathcal{R} = \{(-1,-1),(-1,0),\cdots,(0,1),(1,1)\}$.

Central Difference Convolution. As stated in previous work [59], the intensity-level semantic information and gradient-level detailed message are both crucial for distinguishing the living and spoofing faces. The central differ-

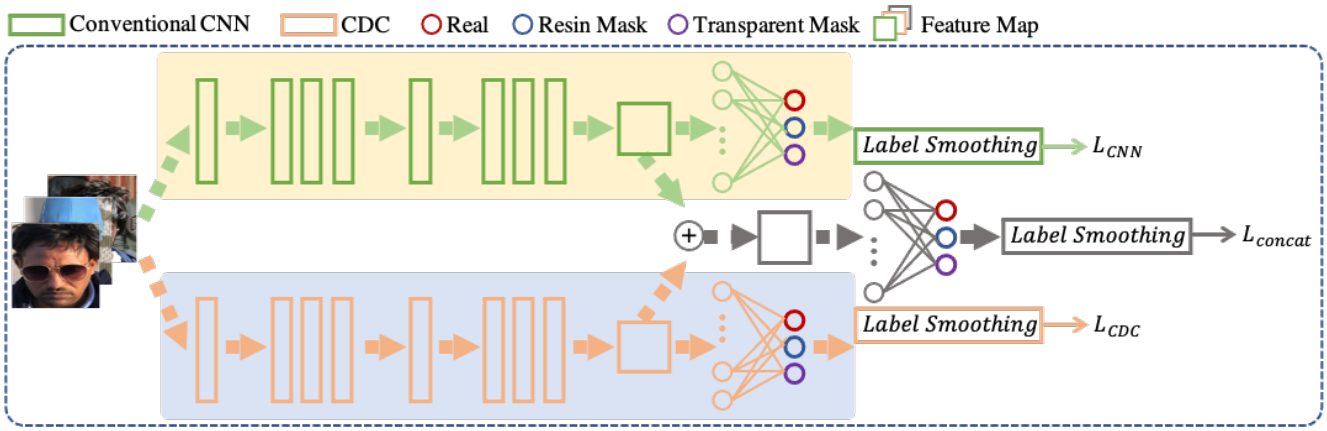

Figure 4. **The framework of our proposed method.** We use the lightweight network ResNet9 [10] as the backbone, and combine vanilla convolution and central difference convolution (CDC) [59] to develop a dual-stream network that can better discover the intrinsic patterns in high-fidelity 3D mask attacks.

ence convolution (CDC) [59] enhances the representation and generalization capacity of the network. Similarly, central difference convolution also consists of two steps, *i.e.*, sampling and aggregation. The sampling step is similar to that in vanilla convolution while the aggregation step is different, central difference convolution prefers to aggregate the center-oriented gradient of sampled values. Eq. 1 becomes:

$$f\left(p_0\right)=\sum_{p_n \in \mathcal{R}} w\left(p_n\right) \cdot\left(x\left(p_0+p_n\right)-x\left(p_0\right)\right). \quad (2)$$

When $p_n=(0,0)$, the gradient value always equals to zero with respect to the central location p_0 itself.

Dual-stream Network. Based on the above-mentioned vanilla convolution and CDC, we develop a novel dual-stream network, in which both branches use ResNet9 as the backbone, except that one branch uses vanilla convolution and the other branch uses CDC. we concatenate the extracted features extracted from both branches for classification. Furthermore, we inserted an auxiliary classifier behind each branch to enhance the learning of intensity-level information and gradient-level information.

3.3. Loss Function

Real-world 3D face presentation attacks vary in terms of materials and generation. Instead of binary classification, we utilize multi-class loss with multi-head loss, which is adjusted via label smoothing for better generalizability.

Label Smoothing. To address the overfitting problem on known attacks, we used the widely used label smoothing [47] regularization in the loss function with a smoothing parameter of 0.1. This technique significantly improves the generalizability in the phase-2 of the competition.

Multi-class Loss. Different from previous work that treats 3D PAD as a binary classification task, we extend it to a multi-classification task, *i.e.*, distinguishing between the real face and each type of attack. Since the train set of HIFI-Mask dataset contains only two types of attacks, *i.e.*, plaster and transparent attacks, we decide to adopt three-class learning, which is formulated as follows:

$$\mathcal{L}=-\sum_{i=1}^{C} p_i \log\left(y_i\right)+\left(1-p_i\right) \log\left(1-y_i\right), \quad (3)$$

where C is the number of classes, and $\mathbf{y}_i \in \{0,1\}^C$ denotes the one-hot encoding of ground-truth label. We set the label of real face, resin attack, and transparent attack to 0, 1, 2, respectively. During testing, we use the predicted probability p_0 of the real face as the final prediction result.

Multi-head Loss. Since our framework predicts three different values based on features from CNN branch, CDC branch, and concatenated branch separately, so three losses $\mathcal{L}_{CNN}$, $\mathcal{L}_{CDC}$, and $\mathcal{L}_{concat}$ are calculated following Eq. 3. Then the overall loss function for the whole training process is formulated as:

$$\mathcal{L}_{all}=\mathcal{L}_{concat}+\lambda_1 \mathcal{L}_{CNN}+\lambda_2 \mathcal{L}_{CDC}, \quad (4)$$

where λ_1 and λ_2 are the weights for balanceing the loss.

3.4. Logits Temperature Scaling

To minimize the distribution gap between validation and test sets, we use the logits temperature scaling technique as the post-processing step to calibrate the output distribution. In specific, we add all logits from each attack type to form one uniformed attack logits value. Then we divide the real logits by a factor of 3.6 and the attack logits by a factor of 5.0 before the softmax operation. As shown in Figure 5, temperature scaling softens the distribution on the test set and makes it more even-distributed, which results in better generalization.

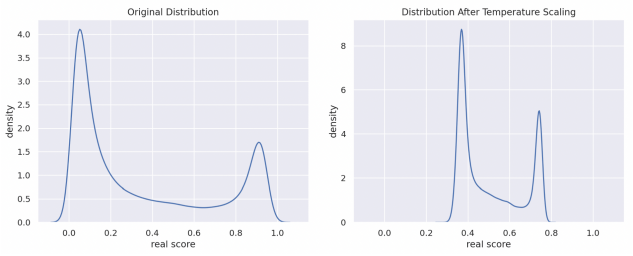

Figure 5. **The test set real score distribution curve.** Left: Original. Right: After logits temperature scaling.

4. Experiments

In this section, we describe the dataset setup, evaluation metrics, and implementation details. And we analyze in detail the performance of our method on the validation set and test set of the HIFIMask dataset [59].

4.1. HIFIMask Dataset

The High-Fidelity Mask dataset, namely CASIA-SURF HiFiMask (briefly HiFiMask), is currently the largest 3D face mask PAD dataset, which contains 54600 videos captured from 75 subjects of three skin tones. This dataset provides 3 high-fidelity masks with the same identity, which are made of transparent, plaster and resin materials, respectively. Besides, six complex scenes and different lighting directions are considered to simulate the real-world scenarios. Based on the HIFIMask dataset, the Chalearn 3D High-Fidelity Mask Face Presentation Attack Detection Challenge@ICCV2021 was launched. In this challenge, the HIFIMask dataset is separated into train set, validation set, and test set, which have 33767, 4645, and 173620 images, respectively. These images are sampled at an equal interval from the corresponding video, and the complex backgrounds are removed from the original images except face areas through Dlib. It is worth noting that the train set and validation set contain the same type of 3D attacks, while the test set includes the challenging unseen attacks. Therefore, generalizability is crucial in practical applications.

In the experiments, we use the HIFIMask dataset to train and evaluate models. For the train set, we remove the noisy images with low face confidence and the black edge of image. For the validation and test sets, we only use the high-precision DSFD for face cropping but do not remove the noise and black edge. The above strategy ensures the diversity of train set while improving the accuracy of test set.

4.2. Evaluation Metrics

Following the HIFIMask dataset, we selected the Attack Presentation Classification Error Rate (APCER), Bona Fide Presentation Classification Error Rate (BPCER), and Average Classification Error Rate (ACER) as the evaluation metric. APCER and BPCER are used to measure the error rate of fake or live samples, respectively. The ACER on the test set is determined by the Equal Error Rate (EER) thresholds on validation sets and calculated via $ACER = (APCER + BPCER)/2$. Additionally, Area Under Curve (AUC) is adopted as an evaluation criterion because the ACER is sensitive to the threshold which does not clearly indicate which classifier performs better. Specifically, APCER and BPCER are formulated as below:

$$APCER = \frac{FN}{TP + FN}, \quad BPCER = \frac{FP}{FP + TN}, \tag{5}$$

where TP and TN refer to the number of correctly classified attacks or real samples respectively; On the contrary, FP and FN refers to the number of incorrectly classified real or attack samples respectively.

Table 1. Performance on the phase-2 of Chalearn 3D High-Fidelity Mask Face Presentation Attack Detection Challenge.

Team	APCER	BPCER	ACER	Rank
VisionLabs	3.777	**2.330**	**3.053**	1
Ours	**1.858**	4.452	3.155	2
CLFM	3.708	2.722	3.215	3
oldiron666	4.944	2.653	3.798	4
Reconova Lab	2.126	6.367	4.247	5

Table 2. Ablation Study on the validation set of HIFIMask dataset.

Model	APCER	BPCER	ACER
CNN Branch	0.604	1.482	1.043
CDC Branch	0.885	1.019	0.952
Concatenate Branch	0.833	0.885	0.859

4.3. Implementation Details

We implement our method via open-source framework PyTorch [36], and train the network on 4 NVIDIA V100 GPUs. The whole training procedure takes around 5 hours and the inference takes 4 seconds per 1000 images. We use wandb [38] to automatically search the hyperparameters and set the batch size to 36, dropout rate to 0.2, and learning rate to 0.00067. We resize the input image to 224×224 and train the network using Adam optimizer [21] with total epochs 80. The learning rate was reduced to 0.2 times of original when the validation metric did not improve for 10 consecutive epochs. The λ_1 and λ_1 are set to 0.5 in Eq. 4.

4.4. Results

Results on Test Sets. We compare the performance of our method and the solutions of other teams on the validation set and test set in the Chalearn 3D High-Fidelity Mask Face Presentation Attack Detection Challenge. In phase-1, from the leaderboard on the competition website[2], we can see that most of the methods achieve high performance on the

[2]https://competitions.codalab.org/competitions/30910#results

Figure 6. Visualization of attention maps for different face types, including real face, transparent, plaster and resin attacks.

validation set due to the fact that the attacks on the validation set and the train set are homologous. In phase-2, as shown in Table 1, since the test set contains unseen attack, *e.g.*, plaster attack, the performance of all methods is significantly degraded. For example, the ACER of the Top-3 methods deteriorates by 2.676 on average; And in the challenging test set, our method achieves APCER, BPCER and ACER by 1.858, 4.452 and 3.155, respectively, ranking second place in this competition. Moreover, our APCER is the best. Although VisionLabs team achieves the best performance with ACER 3.05, five EfficientNet-b0s [48] are required in their method to fuse features from different local regions. In contrast, our method includes only two ResNet9 networks, which requires less inference time and facilitates practical applications.

Results on Validation Sets. We evaluate the effectiveness of each branch in our proposed approach through experiments on the HIFIMask validation set, as shown in Table 2. From the table we can see that: (1) three branches all perform well in the validation set because of the high similarity between the training set and validation set; (2) the concatenate branch outperforms the other two branches, illustrating the effectiveness of the dual-stream framework.

4.5. Visualization

Attention Map. As illustrate in Figure 6, we extract the final feature maps of our model and use GradCAM [42] to visualize the attention maps of different attack types, including transparent, plaster and resin attacks. From the figure, we can observe that our model focuses on different areas for each attack type. As we can see, in the first and second column of the transparent attacks, the nose regions in the attention map have the highest values since these attacks always present strong reflections in the nose area, which is different from live faces. For the plaster and resin attacks, although the texture of these attacks is visually close to real skin, it is still difficult to vividly imitate real human eyes. Thus, the regions of the eyes are an effective classification feature, and our model mainly focuses on these areas in the attention map. Besides, a colored hook exists in the ear region for most attacks. It can be seen from the attention map that our method can also make full use of this tiny region for discrimination. It is worth noting that the train set we used does not contain the plaster attack. Nevertheless, our method is still able to capture generalized features such as colored hooks and edges.

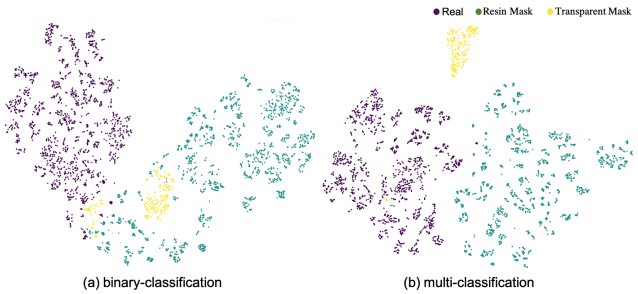

Figure 7. Visualization of feature distributions from binary-classification method and multi-classification method by t-SNE [50].

Feature Distribution. The distribution of features for the traditional binary-classification method and our proposed multi-classification method is shown in Figure 7 via t-SNE [50]. It is clear that the features from the multi-classification approach (Figure 7(b)) present more well-clustered behavior than that from the binary-classification approach (Figure 7(a)), which demonstrates the discrimination ability and generalization of multi-class learning for distinguishing the living faces from high-fidelity 3D mask presentation attacks.

5. Conclusion

In this paper, we propose a lightweight dual-stream network that combines the vanilla convolutional branch and the central difference convolutional branch for high-fidelity 3D face presentation attack detection. Corresponding multi-class loss and multi-head loss are introduced to facilitate the learning of the dual-stream network. Furthermore, we introduce label smoothing and logits temperature scaling techniques to improve the performance on unseen attacks. Experimental results show that the proposed method achieves promising performance and achieves the second place in the Chalearn 3D High-Fidelity Mask Face Presentation Attack Detection Challenge.

References

[1] Abien Fred Agarap. Deep learning using rectified linear units (relu). *arXiv preprint arXiv:1803.08375*, 2018.

[2] Akshay Agarwal, Richa Singh, and Mayank Vatsa. Face anti-spoofing using haralick features. In *IEEE ICB Theory, Applications and Systems*, 2016.

[3] Samet Akçay, Mikolaj E Kundegorski, Michael Devereux, and Toby P Breckon. Transfer learning using convolutional neural networks for object classification within x-ray baggage security imagery. In *ICIP*, 2016.

[4] Jonathan T Barron. Continuously differentiable exponential linear units. *arXiv preprint arXiv:1704.07483*, 2017.

[5] Zinelabidine Boulkenafet, Jukka Komulainen, and Abdenour Hadid. Face anti-spoofing based on color texture analysis. In *ICIP*, 2015.

[6] Alexander Buslaev, Vladimir I. Iglovikov, Eugene Khvedchenya, Alex Parinov, Mikhail Druzhinin, and Alexandr A. Kalinin. Albumentations: Fast and flexible image augmentations. *Information*, 2020.

[7] Shen Chen, Taiping Yao, Yang Chen, Shouhong Ding, Jilin Li, and R. Ji. Local relation learning for face forgery detection. In *AAAI*, 2021.

[8] Zhihong Chen, Taiping Yao, Kekai Sheng, Shouhong Ding, Ying Tai, Jilin Li, Feiyue Huang, and Xinyu Jin. Generalizable representation learning for mixture domain face anti-spoofing. *arXiv preprint arXiv:2105.02453*, 2021.

[9] Ivana Chingovska, André Anjos, and Sébastien Marcel. On the effectiveness of local binary patterns in face anti-spoofing. In *ICBSIP*, 2012.

[10] Davidcpage. How to train your resnet, 2018.

[11] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *CVPR*, 2019.

[12] Terrance Devries and Graham W. Taylor. Improved regularization of convolutional neural networks with cutout. *arXiv, preprint arXiv:2007.13723*, 2017.

[13] Nesli Erdogmus and Sebastien Marcel. Spoofing face recognition with 3d masks. *TIFS*, 2014.

[14] Litong Feng, Lai-Man Po, Yuming Li, Xuyuan Xu, Fang Yuan, Terence Chun-Ho Cheung, and Kwok-Wai Cheung. Integration of image quality and motion cues for face anti-spoofing: A neural network approach. *JVCIR*, 2016.

[15] Litong Feng, Lai-Man Po, Yuming Li, Xuyuan Xu, Fang Yuan, Terence Chun-Ho Cheung, and Kwok-Wai Cheung. Integration of image quality and motion cues for face anti-spoofing: A neural network approach. *JVCIR*, 2016.

[16] Anjith George, Zohreh Mostaani, David Geissenbuhler, Olegs Nikisins, André Anjos, and Sébastien Marcel. Biometric face presentation attack detection with multi-channel convolutional neural network. *TIFS*, 2019.

[17] Jianzhu Guo, Xiangyu Zhu, Chenxu Zhao, Dong Cao, Zhen Lei, and Stan Z Li. Learning meta face recognition in unseen domains. In *CVPR*, 2020.

[18] Bensenane Hamdan and Keche Mokhtar. The detection of spoofing by 3d mask in a 2d identity recognition system. *Egyptian Informatics Journal*, 2018.

[19] Youngshin Kim, Jaekeun Na, Seongbeak Yoon, and Juneho Yi. Masked fake face detection using radiance measurements. *JOSA A*, 2009.

[20] Davis E King. Dlib-ml: A machine learning toolkit. *The Journal of Machine Learning Research*, 2009.

[21] Diederik P. Kingma and Jimmy Ba. Adam: a method for stochastic optimization. *Computing Research Repository*, 2015.

[22] Jukka Komulainen, Abdenour Hadid, and Matti Pietikäinen. Context based face anti-spoofing. In *ICB*, 2013.

[23] Neslihan Kose and Jean-Luc Dugelay. On the vulnerability of face recognition systems to spoofing mask attacks. In *ICASSP*, 2013.

[24] Jian Li, Yabiao Wang, Changan Wang, Ying Tai, Jianjun Qian, Jian Yang, Chengjie Wang, Jilin Li, and Feiyue Huang. Dsfd: dual shot face detector. In *CVPR*, 2019.

[25] Lei Li, Zhaoqiang Xia, Abdenour Hadid, Xiaoyue Jiang, Haixi Zhang, and Xiaoyi Feng. Replayed video attack detection based on motion blur analysis. *TIFS*, 2019.

[26] Ajian Liu, Chenxu Zhao, Zitong Yu, Jun Wan, Anyang Su, Xing Liu, Zichang Tan, Sergio Escalera, Junliang Xing, Yanyan Liang, et al. Contrastive context-aware learning for 3d high-fidelity mask face presentation attack detection. *arXiv preprint arXiv:2104.06148*, 2021.

[27] Jun Liu and Ajay Kumar. Detecting presentation attacks from 3d face masks under multispectral imaging. In *CVPRW*, 2018.

[28] Siqi Liu, Baoyao Yang, Pong C Yuen, and Guoying Zhao. A 3d mask face anti-spoofing database with real world variations. In *CVPRW*, 2016.

[29] Siqi Liu, Pong C Yuen, Shengping Zhang, and Guoying Zhao. 3d mask face anti-spoofing with remote photoplethysmography. In *ECCV*, pages 85–100. Springer, 2016.

[30] Shubao Liu, Ke-Yue Zhang, Taiping Yao, Kekai Sheng, Shouhong Ding, Ying Tai, Jilin Li, Yuan Xie, and Lizhuang Ma. Dual reweighting domain generalization for face presentation attack detection. *arXiv preprint arXiv:2106.16128*, 2021.

[31] Yaojie Liu, Joel Stehouwer, and Xiaoming Liu. On disentangling spoof trace for generic face anti-spoofing. *arXiv preprint arXiv:2007.09273*, 2020.

[32] Oeslle Lucena, Amadeu Junior, Vitor Moia, Roberto Souza, Eduardo Valle, and Roberto Lotufo. Transfer learning using convolutional neural networks for face anti-spoofing. In *ICIAR*, 2017.

[33] Ishan Manjani, Snigdha Tariyal, Mayank Vatsa, Richa Singh, and Angshul Majumdar. Detecting silicone mask-based presentation attack via deep dictionary learning. *TIFS*, 2017.

[34] David Menotti, Giovani Chiachia, Allan Pinto, William Robson Schwartz, Helio Pedrini, Alexandre Xavier Falcao, and Anderson Rocha. Deep representations for iris, face, and fingerprint spoofing detection. *TIFS*, 2015.

[35] Erdogmus Nesli and Sébastien Marcel. Spoofing in 2d face recognition with 3d masks and anti-spoofing with kinect. In *IJCB*, 2013.

[36] Adam Paszke, S. Gross, Soumith Chintala, G. Chanan, E. Yang, Zachary Devito, Zeming Lin, Alban Desmaison, L.

Antiga, and A. Lerer. Automatic differentiation in pytorch. In *NeurIPS Autodiff Workshop*, 2017.

[37] Keyurkumar Patel, Hu Han, and Anil K Jain. Secure face unlock: Spoof detection on smartphones. *TIFS*, 2016.

[38] Carey Phelps. Wandb: Weights and biases, 2021.

[39] Yunxiao Qin, Zitong Yu, Longbin Yan, Zezheng Wang, Chenxu Zhao, and Zhen Lei. Meta-teacher for face anti-spoofing. *IEEE TPAMI*, 2021.

[40] Yunxiao Qin, Chenxu Zhao, Xiangyu Zhu, Zezheng Wang, Zitong Yu, Tianyu Fu, Feng Zhou, Jingping Shi, and Zhen Lei. Learning meta model for zero-and few-shot face anti-spoofing. In *AAAI*, pages 11916–11923, 2020.

[41] Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. In *CVPR*, 2015.

[42] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *ICCV*, pages 618–626, 2017.

[43] Rui Shao, Xiangyuan Lan, and Pong C Yuen. Deep convolutional dynamic texture learning with adaptive channel-discriminability for 3d mask face anti-spoofing. In *IJCB*, 2017.

[44] Rui Shao, Xiangyuan Lan, and Pong C Yuen. Joint discriminative learning of deep dynamic textures for 3d mask face anti-spoofing. *TIFS*, 2018.

[45] I Standard. Information technology–biometric presentation attack detection–part 3: testing and reporting. *International Organization for Standardization*, 2017.

[46] Holger Steiner, Andreas Kolb, and Norbert Jung. Reliable face anti-spoofing using multispectral swir imaging. In *ICB*, 2016.

[47] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *CVPR*, 2016.

[48] Mingxing Tan and Quoc Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In *ICML*, 2019.

[49] Yinhang Tang and Liming Chen. 3d facial geometric attributes based anti-spoofing approach against mask attacks. In *FG*, 2017.

[50] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *JMLR*, 2008.

[51] Xinyao Wang, Taiping Yao, Shouhong Ding, and Lizhuang Ma. Face manipulation detection via auxiliary supervision. In *ICONIP*, 2020.

[52] Yan Wang, Song Chen, Weixin Li, Di Huang, and Yuhong Wang. Face anti-spoofing to 3d masks by combining texture and geometry features. In *CCBR*, 2018.

[53] Yueyang Wang, Xiaoli Hao, Yali Hou, and Changqing Guo. A new multispectral method for face liveness detection. In *ACPR*, 2013.

[54] Zezheng Wang, Zitong Yu, Chenxu Zhao, Xiangyu Zhu, Yunxiao Qin, Qiusheng Zhou, Feng Zhou, and Zhen Lei. Deep spatial gradient and temporal depth learning for face anti-spoofing. In *CVPR*, pages 5042–5051, 2020.

[55] Zitong Yu, Xiaobai Li, Xuesong Niu, Jingang Shi, and Guoying Zhao. Face anti-spoofing with human material perception. In *ECCV*, pages 557–575. Springer, 2020.

[56] Zitong Yu, Xiaobai Li, and Guoying Zhao. Remote photoplethysmograph signal measurement from facial videos using spatio-temporal networks. In *BMVC*, page 277, 2019.

[57] Zitong Yu, Yunxiao Qin, Xiaobai Li, Chenxu Zhao, Zhen Lei, and Guoying Zhao. Deep learning for face anti-spoofing: A survey. *arXiv preprint arXiv:2106.14948*, 2021.

[58] 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 TPAMI*, pages 1–1, 2020.

[59] Zitong Yu, Chenxu Zhao, Zezheng Wang, Yunxiao Qin, Zhuo Su, Xiaobai Li, Feng Zhou, and Guoying Zhao. Searching central difference convolutional networks for face anti-spoofing. In *CVPR*, pages 5295–5305, 2020.

[60] Jian Zhang, Ying Tai, Taiping Yao, Jia Meng, Shouhong Ding, Chengjie Wang, Jilin Li, Feiyue Huang, and Rongrong Ji. Aurora guard: Reliable face anti-spoofing via mobile lighting system. *arXiv preprint arXiv:2102.00713*, 2021.

[61] Ke-Yue Zhang, Taiping Yao, Jian Zhang, Shice Liu, Bangjie Yin, Shouhong Ding, and Jilin Li. Structure destruction and content combination for face anti-spoofing. *arXiv preprint arXiv:2107.10628*, 2021.

[62] Ke-Yue Zhang, Taiping Yao, Jian Zhang, Ying Tai, Shouhong Ding, Jilin Li, Feiyue Huang, Haichuan Song, and Lizhuang Ma. Face anti-spoofing via disentangled representation learning. *arXiv preprint arXiv:2008.08250*, 2020.

[63] Zhiwei Zhang, Junjie Yan, Sifei Liu, Zhen Lei, Dong Yi, and Stan Z Li. A face antispoofing database with diverse attacks. In *ICB*, 2012.