Bandpass Filter Based Dual-stream Network for Face Anti-spoofing

Dingheng Zeng[1], Liang Gao[1], Hao Fang[2], Guohui Xiang[1], Yue Feng[1*], Quan Lu[1]
[1]Mashang Consumer Finance Co., Ltd., Chongqing, China
[2]School of Artificial Intelligence, University of Chinese Academy of Sciences, Beijing, China
{dingheng.zeng,liang.gao01,guohui.xiang,yue.feng,quan.lu}@msxf.com, fanghao21@mails.ucas.ac.cn

Abstract

Face Attack Detection (PAD) technology is crucial for protecting facial recognition systems. At present, methods for Face Anti-spoofing (FAS) mainly focus on short-distance applications, and algorithm performance can sharply decline when facing challenges such as low resolution, pedestrian obstruction, and blurriness in long-distance scenarios. To address these issues, we propose a dual-stream architecture that combines information from the images and its bandpass filtered image to distinguish attacks. Specifically, one branch extracts detailed facial structure and texture information from the original spatial domain of images. The other branch take the Gaussian bandpass filtered image as input to learn the complementary discriminative features. The filtering process was done in frequency domain by FFT/IFFT. We proposed a cross-attention fusion module to fuse the features extracted by the two network branches. Additionally, to further improve the model's generalization ability to data quality, we use automatic correction and lion optimizer. Finally, our method achieved a result of 6.22% on the ACER metric and ranked third in the 4th Face Anti-Spoofing Challenge @CVPR2023.

With the rapid development of computer vision technology, face recognition technology [1,11,26] has become very mature and widely used in personal identification scenarios. However, there are also increasing threats to face recognition technology, such as 3D masks, high-definition printed photos, high-definition video replays, and so on. These face representation attacks attempt to deceive face recognition systems using physical media in order to gain illegal benefits through these systems. Therefore, research on face representation attack detection technology is extremely critical for protecting user privacy and system security. This technology can detect in advance whether the input face is a live person or an attack, so that appropriate measures can be taken to protect the privacy and security of the system and citizens.

1. Introduction

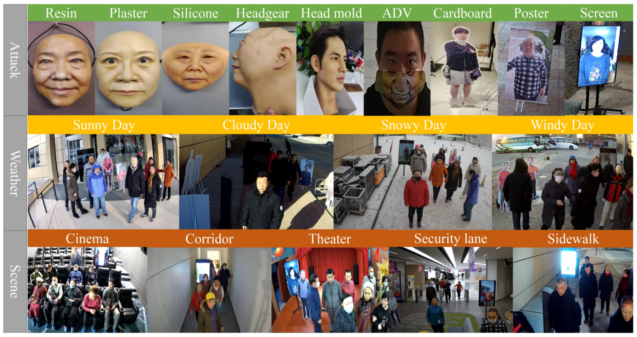

Figure 1. Examples from the SuHiFiMask dataset [7].

Based on the differences in structural features and texture features of human faces, the types of physical attacks can be roughly divided into two types: 2D and 3D. Common 2D attacks include electronic photos [4], video playback [5], printed photos [39], etc. The 3D attack includes HD masks, headgear and head models [22], etc. However, regardless of 2D or 3D attacks, existing FAS work is limited to detecting attack types in close-range restricted environments. With the popularization of remote sensors and the large-scale deployment of monitoring networks. The FAS community urgently needs to extend the face anti-counterfeiting algorithm to long-distance monitoring scenarios.

The task of facial anti-fraud in monitoring fields is relatively difficult because the targets in most scenes are unconstrained, making it difficult to ensure the quality of the face images. Compared with previous close-distance face anti-spoofing tasks, the face in surveillance scenes are small, occlusion, motion blur and pose variations. In Figure 1, we show some examples from large-scale Surveillance High-Fidelity Mask (SuHiFiMask) dataset [7], which has 101 subjects from different age groups with 232 3D masks,200 2D attacks,and 2 adversarial attacks.

With the widespread application of face recognition technology in surveillance scenarios, research on face anti-

[*]Corresponding author

spoofing in this field has been promoted. Previous CNN methods pay more attention to fine-grained feature, which includes image reflection, based on color-texture [10, 24, 28], based on depth map [27, 36, 37], or based on remote photoplethysmography [14, 35], which achieved very good results. However, they did not generalize well to low-quality, long-distance face images. In this paper, inspired by the lightweight network EfficientFormerV2 [13], we propose a dual-stream network framework that combines information from the images and its bandpass filtered image. One stream uses band-pass filters to enhance edges of image and denoise the image, which has been proven to be a very effective denoising method in previous studies [29, 33, 40, 41]. To avoid information loss caused by image denoising, the other stream extracts features from the original image. By fusing the features of the dual streams, we can improve the overall algorithm's generalization performance on low-quality images. In addition, we use autoaugmentation [6]strategies to reduce model overfitting and use the Lion [2] optimizer to ensure algorithm training stability. Finally, we achieved 6.22%(ACER) in the 4th Face Anti-spoofing Challenge@CVPR2023, ranking third among all participating teams.

Our main contributions can be summarized as follows:

- We proposed a dual-stream network architecture that extract complementary discriminative features from the images and its bandpass filtered image based on a visual transformer. The bandpass filter branch obtains features after image filtering, while the other branch preserves the original image features. We proposed a cross-attention fusion module to fuse the output features of dual-stream network, can enhance the model's generalization ability to low-quality facial images.

- We have demonstrated through experiments that using the strategy of automatic data augmentation and using the LION optimizer can significantly improve the generalization ability of the algorithm and ensure the stability of model iteration.

- Our method achieves 6.22%(ACER) and ranks third in the Face Anti-spoofing Challenge@CVPR2023. In addition, We have summarized the effectiveness of methods for low-quality data and identified future research directions.

2. Related work

The existing face anti-spoofing methods [8, 9, 30] can be broadly classified into two categories: hand-crafted feature and deep learning methods. The hand-crafted feature method uses manually designed features as inputs for classifiers such as LBP [34], HOG [12], SURF [25], etc., and then train a classifier similar to SVM [31]. These methods have advantages such as low computational complexity and fast speed, and have achieved good results in traditional simple scenarios. However, the manual feature extraction process is cumbersome, and when the input data volume doubles, the cost of manual feature extraction is high and often cannot be located to features that are beneficial for downstream tasks. Therefore, these methods will fail in complex or cross-domain tasks.

On the other hand, the deep learning-based methods use CNN to treat the live detection as a binary classification task. These methods mainly analyze fine-grained features of the face, such as color and texture [15, 17, 19, 21–23, 38], and are therefore suitable for high-quality face images captured at close-distance. However, when these methods are extended to long-distance monitoring scenarios, the fine-grained features they rely on in high-quality images contain only partial discriminative information, while also containing noise interference such as dynamic blur, occlusion, and pseudo-images, which makes the network unable to distinguish the correct optimization direction and thus leads to a decrease in performance.

To address the problem of low-quality face images, recent studies have proposed several solutions. For example, Chen et al. [3] consists of the depthwise separable attention module and the multi-modal based feature augment module to enhance the low-quality images. Fang et al. [7]proposes a Contrastive Quality-Invariance Learning network to eliminate the effect of image quality. This framework includes three modules: the Image Quality Variable module, which upscales the low-quality images to restore fine-grained information, the Contrastive Learning Branch, which uses generated samples to simulate quality distribution and learn live features that are not affected by quality interference, and the Separate Quality Network, which learns the quality factor from the input images. To alleviate performance degradation caused by large face pose, Liu et al. [20] designed a Pose-Independent Face Anti-Spoofing (PIFAS) framework to disentangle face into an appearance information and a pose code to capture liveness and liveness-irrelated features, respectively. Face Presentation Attack Detection (PAD) approaches based on multi-modal data have been studied extensively by its good performance. Liu et al. [18] proposed a Cross-modal Auxiliary (CMA) framework, via a generative model that maps inputs from one modality (i.e., RGB) to another (i.e. , NIR), to assist the performance improvement of VIS-based PAD. Liu et al. [16] present a single branch based Transformer framework, namely Modality-Agnostic Vision Transformer (MA-ViT), which aims to improve the performance of arbitrary modal attacks with the help of multi-modal data.

To effectively alleviate the negative impact of low-quality images, this paper proposes a dual-stream network framework take image and its bandpass filtered image as inputs, which has the following advantages: (1) One branch converts the image to the frequency domain and obtains an edge enhanced image through the Gaussian bandpass filter, which can extract features that are less related to image quality. (2) Another branch extracts features of the original spatial domain image, including detailed facial structural information and color texture. (3) The features of the two branches are fused by a cross attention fusion module, better utilizing the complementary advantages of the two features for classification. In addition, we use automatic augmentation and lion optimizer to improve the algorithm's generalization performance and training stability.

3. Methodology

In this section, we will first introduce the framework of our method, including the network architecture and proposed features fusion module, followed by the data preprocessing process, data augmentation techniques employed during training, the loss function, and the strategy used for the final outputs.The entire framework of our method is illustrated in Fig 2.

3.1. Network Architecture

Many studies have already demonstrated the effectiveness of vision transformers (ViT) for fine-grained classification tasks. In this work, we use the EfficientFormerV2 [13] as the backbone and designed a dual-stream architecture.

EfficientFormerV2. As stated in [13], EfficientFormerV2 is a neural network architecture that is designed and optimized for mobile devices, taking reference from MobileNet and making a series of mobile-specific optimizations to the Vision Transformer (VIT). The model's parameter count and latency are crucial for hardware with limited resources, so EfficientFormerV2 employs a fine-grained joint search strategy to create an efficient network with low latency and small size. Compared to MobileNetV2, this network achieves 4% higher performance on the validation set of the ImageNet dataset while maintaining the same parameter count and latency. EfficientFormerV2 thoroughly investigates mixed visual backbones and validates network structure designs that are more friendly to mobile devices. It follows the conventional ViT architecture, replacing the token mixer's average pooling layer with depthwise separable convolution layers with the same kernel size, which improves performance without increasing latency. The network design consists of four stages that capture features at resolutions of 1/4, 1/8, 1/16, 1/32 of the input resolution. Similar to its predecessor EfficientFormer, EfficientFormerV2 embeds the input image from a small kernel convolution stem, rather than using inefficient non-overlapping patches. The first two stages capture high-resolution local information, so the paper only uses a unified feedforward network (FFN), while local FFN and global MHSA blocks are used in the last two stages. Additionally, EfficientFormerV2 introduces fine-grained joint search for size and speed on top of its previous version, resulting in extremely fast inference and small model size, surpassing previous technologies and serving as a powerful backbone for various downstream tasks.

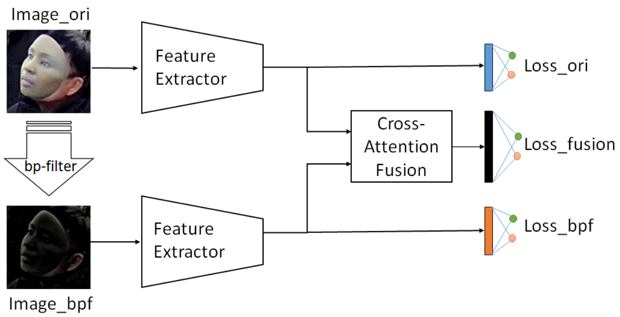

Figure 2. Framework of our proposed method.

Dual-stream Network. Dual-stream networks are commonly used in applications that involve processing of multimodal data. By processing each modality of data separately, the network can learn features that are specific to each modality, which can then be fused to obtain a more comprehensive representation of the data. In face anti-spoofing tasks, a dual-stream network can be used to process rgb and depth inputs separately and then fuse the extracted features to improve the overall accuracy of the classification. The dual-stream structure we used has two identical backbone with different inputs. One input is the original image, while the other input is the image that has been filtered using a Gaussian band-pass filter. A band-pass filter can remove signals below or above a certain frequency range, thus filtering out noise and interference. And it increase signals in certain frequency ranges, thereby changing the image's color and contrast, among other visual features. So we can regard it as data of another modality.

Cross Attention Fusion. In order to fuse the output features from the two different branches, a cross-attention fusion module is adopted which illustrated in Fig 3 This module composed by two regular multi-head attention blocks. The original outputs from the two branches and the fused features are then fed into three separate classification heads to generate three output results.

3.2. Data Preprocessing

The two branches of our network take as input the original RGB image and the image filtered using a Gaussian band-pass filter, respectively. The band-pass filtering process can be illustrated as shown in Fig 4

As shown in the diagram, the band-pass filtering process

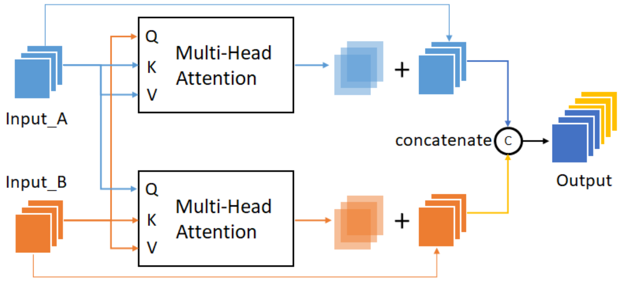

Figure 3. Cross attention fusion module.

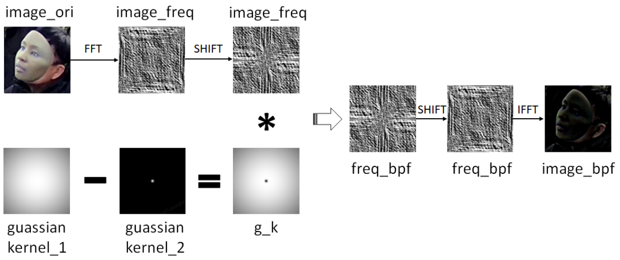

Figure 4. Band-pass filtering process.

involves the following steps: The Gaussian band-pass filtering process we used involves subtracting two Gaussian window with different standard deviations to obtain a band-pass filter kernel, denoted as g_k. The original image is then transformed into the frequency domain using the Fast Fourier Transform (FFT), and applied with shift operation to obtain centralized frequency domain image. To apply the band-pass filter to the frequency domain image, we multiply the bandpass filter kernel g_k with the frequency domain image, element-wise. This operation attenuates the low and high-frequency components of the image, while preserving the mid-frequency components within the range of the band-pass filter. The resulting filtered frequency domain image is then shifted back to its original position, and transformed back into the spatial domain using the Inverse Fast Fourier Transform (IFFT). This yields the final band-pass filtered image, which emphasizes the edges and high-frequency details within the range of the band-pass filter. The process of frequency domain filtering can be expressed by the following formula:

$$I(x, y) = Real\{IFFT[H(u, v)F(u, v)]\} \tag{1}$$

where I(x,y) is the filtered spatial domain image, Real refers to the operation of taking the real part of a complex number, IFFT is the inverse fast Fourier transform, H(u,v) is the frequency domain filtering function, and F(u,v) is the data of the input image after performing the Fourier transform to the frequency domain. H(u,v) here we used is Gaussian window which is shown in Eq2.

$$w(n) = e^{-\frac{1}{2}(\frac{n}{\sigma})^2} \tag{2}$$

We use a Gaussian window instead of a rectangular window as the filter because using a rectangular window to truncate the spectrum can result in striped patterns(Refer to Fig 5) due to spectral leakage. These extra patterns, when applied to both positive and negative samples, can cause training instability.

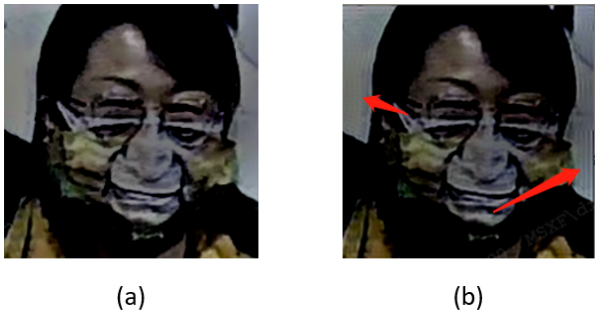

Figure 5. Filter comparison. (a) Gaussian window as filters for frequency domain filtering. (b) Rectangular window as filters, the area pointed by the red arrow is the stripe artifact.

3.3. Data Augmentation

We use the data augmentation module provided by the timm framework, which includes various techniques such as auto-augmentation, color jitter, random erase, random scaling and mixup etc,. In addition, we also added some operations such as random noise, image quality degradation to further diversify the training data.

To regularize the model and prevent overfitting, we employed mixup and label smoothing techniques during training. To ensure the consistency of the labels for the dual-input network after mixup, we replayed the random seed during each mixup operation. This ensured that the same pairs of images were mixed up in both branches of the network with the same mixing coefficients, resulting in consistent labels across the two branches.

Fig 6 shows the resulting data after mixup for the dual-branch inputs.

3.4. Loss Function

We employed three separate classification heads, each with its own cross-entropy loss function. These losses were combined using a weighted sum to form the final loss function, which is given by:

$$Loss = Loss\_fusion + \lambda 1 * Loss\_ori + \lambda 2 * Loss\_bpf \tag{3}$$

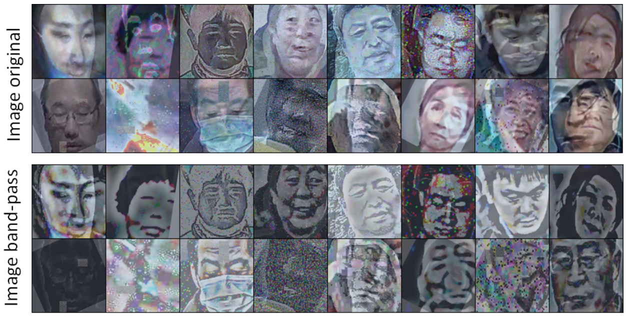

Figure 6. Network inputs after data augmentations and mixup.

where Loss_fusion is the cross-entropy losses of the classification head corresponding to the output of the cross-attention fusion module, Loss_ori refers to the output cross-entropy loss of the network branch corresponding to the original image, while Loss_bpf refers to the output cross-entropy loss of the network branch corresponding to the band-pass filtered image. $\lambda 1$, and $\lambda 2$ are the corresponding weights. The weights of each loss can be determined through cross-validation or grid search.

4. Experiments

In this section, we describe the dataset setup, evaluation metrics, and implementation details, results and visualization.

4.1. Dataset & Metrics

SuHiFiMask [7]. The large-scale Surveillance High-Fidelity Mask (SuHiFiMask) dataset captured under 40 surveillance scenes, which has 101 subjects from different age groups with 232 3D attacks (highfidelity masks), 200 2D attacks (posters, portraits, and screens), and 2 adversarial attacks. There are three protocols to evaluate the performance in surveillance environments: Protocol 1-ID, Protocol 2-Mask, and Protocol 3-quality. This challenge is based on Protocol 3 which evaluates the robustness of the algorithm to image quality degradation. Variable quality and disturbances are factors that affect the stability of the algorithm.

Performance Metrics. For the performance evaluation, selected the standardized ISO/IEC 30107-3 metrics: Attack Presentation Classification Error Rate (APCER), Normal/Bona Fide Presentation Classification Error Rate (NPCER/BPCER) and Average Classification Error Rate (ACER) as the evaluation metric, in which APCER and BPCER/NPCER are used to measure the error rate of fake or live samples, respectively. They can be formulated as

$$APCER = \frac{FP}{TN + FP} \tag{4}$$

$$BPCER = \frac{FN}{FN + TP} \tag{5}$$

$$ACER = \frac{APCER + BPCER}{2} \tag{6}$$

where FP, FN, TN and TP denote the false positive, false negative, true negative and true positive sample numbers, respectively. ACER is used to determine the final ranking in the 4th Face Anti-spoofing Challenge@CVPR2023. Our experiments based on Protocol 3 also, we use train and validation set to train and evaluation, get the threshold corresponding to the Equal Error Rate(EER) of validation set and then calculation the ACER for the test set.

4.2. Implementation Details

Our proposed method is implemented with Pytorch and timm library [32]. In the training stage, models are trained with Lion optimizer [2] , the initial learning rate is 1e-4 and minimal learning rate is 1e-6, respectively. We used cosine LR schedule, first 5 epochs used for warmup, total 100 epochs was trained. We use 4 NVIDIA V100 GPUs to train our models. The batch size is 64. Image resolution set as 224 × 224 for both training and testing. In the ablation study phase, we used EfficientFormerV2-S0 as the backbone to quickly verify the effectiveness of our approach. For the final submission of our results, we used EfficientFormerV2-S2 as the backbone to achieve higher performance. The $\lambda 1$, and $\lambda 2$ in the loss function are set to 0.5 and 0.5, respectively. The score from the classification heads obtained by fusing features from the two branches is used as the final output of the model.

4.3. Results

Ablation Study. During the ablation study phase, we evaluated the effectiveness of the proposed dual-stream network structure based on band-pass filtering and the Cross-attention fusion module. The experimental results are shown in Tab 1. We conducted two sets of experiments, all of which used a dual-stream network with the same backbone(EfficientFormerV2-S0) and hyperparameters. In Experiment No.1, the two inputs of the network were the original images, which underwent online data augmentation. The outputs of the two branch networks were concatenated for feature fusion, and then a classification head was added to obtain the network output ”cat”. In Experiment No.2, one of the network inputs was the original image, and the other input was the image filtered by a Gaussian bandpass filter(bpf). The output features of the two branch networks were send to the cross-attention fusion module, followed by a classification head to abtain the output ”caf”. From the results, we can find that introducing the filtered image as input and using the cross-attention fusion module

No.	Model	Branch	APCER	BPCER	ACER
1	effiv2-s0	img_ori	14.81	8.96	11.89
	wo/bpf	img_ori	16.69	7.54	12.11
	wo/caf	**cat**	15.1	7.81	11.45
2	effiv2-s0	img_ori	14.42	8.71	11.57
	w/bpf	img_bpf	15.44	11.99	13.71
	w/caf	**caf**	13.6	9.01	**11.30**

Table 1. Ablation Study on the SuHiFiMask dataset

Rank	Team	APCER	BPCER	ACER
1	**Baidu Inc**	5.07	4.38	4.73
2	**China Telecom**	9.20	1.90	5.56
3	**Ours**	8.17	4.26	6.22

Table 2. Final Submission Results

for feature fusion can effectively improve the final classification accuracy. In addition, all the experimental results have demonstrated that using a dual-stream network to extract features separately from inputs, and then fusing the features of the two branches, performs better than a single branch, especially when the two branches can extract complementary information.

Final Submission. In the final submission stage, we used EfficientFormerV2-S2 as the backbone of the dual-stream network and used the model pre-trained on the imagenet-1k dataset by the official. We trained the network on the train and validation datasets with more data augmentation strategies to obtained the optimal result. The final score on the leaderboard is shown in Tab 2. Our ACER is 6.22%, ranked third place in this competition. It is worth mentioning that we used EfficientFormerV2-S2 as the backbone of the dual-stream network, and the total number of parameters in the model was only 25M, with a computational cost of 5G FLOPs.

4.4. Visualization

We use GradCam, a visual explanation technique based on gradient-based localization, to analyze different types of attacks on a face recognition system. The types of attacks studied include 3D masks and head models, adversarial hats and masks, and 2D attacks. The results show that the model focuses on different areas depending on the type of attack. For example, when faced with a 3D mask, the model pays attention to the eyes and the area with clear boundaries. However, for a side profile mask, the model mainly relies on the mask boundary to make decisions due to the lack of facial information. When dealing with adversarial attacks, the model primarily relies on the adversarial pattern to make decisions, while in the case of 2D attacks, the model focuses on the eye area to make decisions. Please refer to Fig. 7 for details.

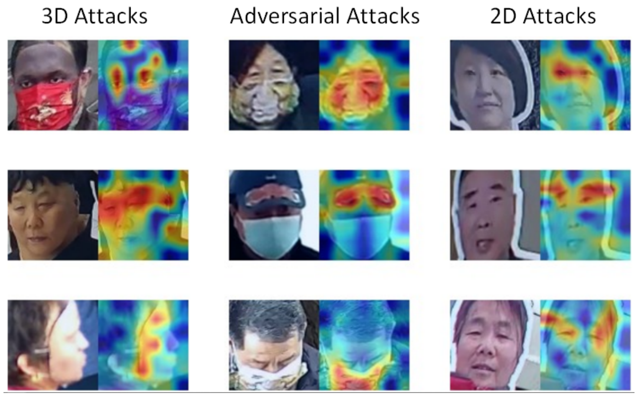

Figure 7. Visualization of attention maps for different attacks.

5. Conclusion

In this paper, we proposed a dual-stream network based on Gaussian bandpass filtering and a cross-attention fusion module, and demonstrated its effectiveness through experiments.The proposed approach wins the third place of the 4th Face Anti-spoofing Challenge@CVPR2023.

6. Acknowledgments

This work was supported by the National Key Research and Development Plan under Grant 2021YFF0602105, the External cooperation key project of Chinese Academy Sciences 173211KYSB20200002, the Chinese National Natural Science Foundation Projects 61876179 and 61961160704, the Science and Technology Development Fund of Macau Project (No.~0010/2019/AFJ, 0008/2019/A1, 0025/2019/AKP, 0019/2018/ASC).

References

[1] Q. Cao, L. Shen, W. Xie, O. M. Parkhi, and A. Zisserman. Vggface2: A dataset for recognising faces across pose and age. *IEEE Computer Society*, 2017. 1

[2] Xiangning Chen, Chen Liang, Da Huang, Esteban Real, Kaiyuan Wang, Yao Liu, Hieu Pham, Xuanyi Dong, Thang Luong, Cho-Jui Hsieh, Yifeng Lu, and Quoc V. Le. Symbolic discovery of optimization algorithms. *CoRR*, abs/2302.06675, 2023. 2, 5

[3] Xudong Chen, Shugong Xu, Qiaobin Ji, and Shan Cao. A dataset and benchmark towards multi-modal face anti-spoofing under surveillance scenarios. *IEEE Access*, 9:28140–28155, 2021. 2

[4] I. Chingovska, A. Anjos, and S. Marcel. On the effectiveness of local binary patterns in face anti-spoofing. *IEEE*, 2012. 1

[5] A. Costa-Pazo, S. Bhattacharjee, E. Vazquez-Fernandez, and S. Marcel. The replay-mobile face presentation-attack database. In *2016 International Conference of the Biometrics Special Interest Group (BIOSIG)*, 2016. 1

[6] Ekin D. Cubuk, Barret Zoph, Dandelion Mané, Vijay Vasudevan, and Quoc V. Le. Autoaugment: Learning augmentation strategies from data. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pages 113–123. Computer Vision Foundation / IEEE, 2019. 2

[7] Hao Fang, Ajian Liu, Jun Wan, Sergio Escalera, Chenxu Zhao, Xu Zhang, Stan Z. Li, and Zhen Lei. Surveillance face anti-spoofing. *CoRR*, abs/2301.00975, 2023. 1, 2, 5

[8] Haocheng Feng, Zhibin Hong, Haixiao Yue, Yang Chen, Keyao Wang, Junyu Han, Jingtuo Liu, and Errui Ding. Learning generalized spoof cues for face anti-spoofing. *CoRR*, abs/2005.03922, 2020. 2

[9] Anjith George and Sébastien Marcel. Deep pixel-wise binary supervision for face presentation attack detection. In *2019 International Conference on Biometrics, ICB 2019, Crete, Greece, June 4-7, 2019*, pages 1–8. IEEE, 2019. 2

[10] Yunpei Jia, Jie Zhang, Shiguang Shan, and Xilin Chen. Single-side domain generalization for face anti-spoofing. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, pages 8481–8490. Computer Vision Foundation / IEEE, 2020. 2

[11] M. Kim, A. K. Jain, and X. Liu. Adaface: Quality adaptive margin for face recognition. 2022. 1

[12] Jukka Komulainen, Abdenour Hadid, and Matti Pietikäinen. Context based face anti-spoofing. In *IEEE Sixth International Conference on Biometrics: Theory, Applications and Systems, BTAS 2013, Arlington, VA, USA, September 29 - October 2, 2013*, pages 1–8. IEEE, 2013. 2

[13] Yanyu Li, Ju Hu, Yang Wen, Georgios Evangelidis, Kamyar Salahi, Yanzhi Wang, Sergey Tulyakov, and Jian Ren. Rethinking vision transformers for mobilenet size and speed. *CoRR*, abs/2212.08059, 2022. 2, 3

[14] Bofan Lin, Xiaobai Li, Zitong Yu, and Guoying Zhao. Face liveness detection by rppg features and contextual patch-based CNN. In *Proceedings of the 3rd International Conference on Biometric Engineering and Applications, ICBEA 2019, Stockholm, Sweden, May 29-31, 2019*, pages 61–68. ACM, 2019. 2

[15] Ajian Liu, Xuan Li, Jun Wan, Yanyan Liang, Sergio Escalera, Hugo Jair Escalante, Meysam Madadi, Yi Jin, Zhuoyuan Wu, Xiaogang Yu, et al. Cross-ethnicity face anti-spoofing recognition challenge: A review. *IET Biometrics*, 10(1):24–43, 2021. 2

[16] Ajian Liu and Yanyan Liang. Ma-vit: Modality-agnostic vision transformers for face anti-spoofing. In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI-22*, pages 1180–1186, 2022. 2

[17] Ajian Liu, Zichang Tan, Jun Wan, Sergio Escalera, Guodong Guo, and Stan Z Li. Casia-surf cefa: A benchmark for multi-modal cross-ethnicity face anti-spoofing. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1179–1187, 2021. 2

[18] Ajian Liu, Zichang Tan, Jun Wan, Yanyan Liang, Zhen Lei, Guodong Guo, and Stan Z Li. Face anti-spoofing via adversarial cross-modality translation. *IEEE Transactions on Information Forensics and Security*, 16:2759–2772, 2021. 2

[19] Ajian Liu, Jun Wan, Sergio Escalera, Hugo Jair Escalante, Zichang Tan, Qi Yuan, Kai Wang, Chi Lin, Guodong Guo, Isabelle Guyon, et al. Multi-modal face anti-spoofing attack detection challenge at cvpr2019. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 0–0, 2019. 2

[20] Ajian Liu, Jun Wan, Ning Jiang, Hongbin Wang, and Yanyan Liang. Disentangling facial pose and appearance information for face anti-spoofing. In *2022 26th International Conference on Pattern Recognition (ICPR)*, pages 4537–4543. IEEE, 2022. 2

[21] Ajian Liu, Chenxu Zhao, Zitong Yu, Anyang Su, Xing Liu, Zijian Kong, Jun Wan, Sergio Escalera, Hugo Jair Escalante, Zhen Lei, et al. 3d high-fidelity mask face presentation attack detection challenge. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, pages 814–823, 2021. 2

[22] 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. *IEEE Transactions on Information Forensics and Security*, 17:2497–2507, 2022. 1, 2

[23] Shice Liu, Shitao Lu, Hongyi Xu, Jing Yang, Shouhong Ding, and Lizhuang Ma. Feature generation and hypothesis verification for reliable face anti-spoofing. In *Thirty-Sixth AAAI Conference on Artificial Intelligence, AAAI 2022, Thirty-Fourth Conference on Innovative Applications of Artificial Intelligence, IAAI 2022, The Twelveth Symposium on Educational Advances in Artificial Intelligence, EAAI 2022 Virtual Event, February 22 - March 1, 2022*, pages 1782–1791. AAAI Press, 2022. 2

[24] Youngjun Moon, Intae Ryoo, and Seokhoon Kim. Face anti-spoofing method using color texture segmentation on FPGA. *Secur. Commun. Networks*, 2021:9939232:1–9939232:11, 2021. 2

[25] Keyurkumar Patel, Hu Han, and Anil K. Jain. Secure face unlock: Spoof detection on smartphones. *IEEE Trans. Inf. Forensics Secur.*, 11(10):2268–2283, 2016. 2

[26] F. Schroff, D. Kalenichenko, and J. Philbin. Facenet: A unified embedding for face recognition and clustering. *IEEE*, 2015. 1

[27] Rui Shao, Xiangyuan Lan, Jiawei Li, and Pong C. Yuen. Multi-adversarial discriminative deep domain generalization for face presentation attack detection. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pages 10023–10031. Computer Vision Foundation / IEEE, 2019. 2

[28] Chien-Yi Wang, Yu-Ding Lu, Shang-Ta Yang, and Shang-Hong Lai. Patchnet: A simple face anti-spoofing framework via fine-grained patch recognition. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pages 20249–20258. IEEE, 2022. 2

[29] Haohan Wang, Xindi Wu, Zeyi Huang, and Eric P. Xing. High-frequency component helps explain the generalization of convolutional neural networks. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR*

2020, Seattle, WA, USA, June 13-19, 2020, pages 8681–8691. Computer Vision Foundation / IEEE, 2020. 2

[30] Zhuo Wang, Zezheng Wang, Zitong Yu, Weihong Deng, Jiahong Li, Tingting Gao, and Zhongyuan Wang. Domain generalization via shuffled style assembly for face anti-spoofing. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pages 4113–4123. IEEE, 2022. 2

[31] Di Wen, Hu Han, and Anil K. Jain. Face spoof detection with image distortion analysis. *IEEE Trans. Inf. Forensics Secur.*, 10(4):746–761, 2015. 2

[32] Ross Wightman. Pytorch image models. https://github.com/rwightman/pytorch-image-models, 2019. 5

[33] Zhi-Qin John Xu, Yaoyu Zhang, Tao Luo, Yanyang Xiao, and Zheng Ma. Frequency principle: Fourier analysis sheds light on deep neural networks. *CoRR*, abs/1901.06523, 2019. 2

[34] Jianwei Yang, Zhen Lei, Shengcai Liao, and Stan Z. Li. Face liveness detection with component dependent descriptor. In Julian Fiérrez, Ajay Kumar, Mayank Vatsa, Raymond N. J. Veldhuis, and Javier Ortega-Garcia, editors, *International Conference on Biometrics, ICB 2013, 4-7 June, 2013, Madrid, Spain*, pages 1–6. IEEE, 2013. 2

[35] Chenglin Yao, Shihe Wang, Jialu Zhang, Wentao He, Heshan Du, Jianfeng Ren, Ruibin Bai, and Jiang Liu. rppg-based spoofing detection for face mask attack using efficientnet on weighted spatial-temporal representation. In *2021 IEEE International Conference on Image Processing, ICIP 2021, Anchorage, AK, USA, September 19-22, 2021*, pages 3872–3876. IEEE, 2021. 2

[36] 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 *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, pages 5294–5304. Computer Vision Foundation / IEEE, 2020. 2

[37] 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. In *International IEEE Joint Conference on Biometrics, IJCB 2021, Shenzhen, China, August 4-7, 2021*, pages 1–6. IEEE, 2021. 2

[38] Shifeng Zhang, Ajian Liu, Jun Wan, Yanyan Liang, Guodong Guo, Sergio Escalera, Hugo Jair Escalante, and Stan Z Li. Casia-surf: A large-scale multi-modal benchmark for face anti-spoofing. *IEEE Transactions on Biometrics, Behavior, and Identity Science*, 2(2):182–193, 2020. 2

[39] Yuanhan Zhang, Zhenfei Yin, Yidong Li, Guojun Yin, and Ziwei Liu. Celeba-spoof: Large-scale face anti-spoofing dataset with rich annotations. 2020. 1

[40] Bolun Zheng, Shanxin Yuan, Gregory G. Slabaugh, and Ales Leonardis. Image demoireing with learnable bandpass filters. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, pages 3633–3642. Computer Vision Foundation / IEEE, 2020. 2

[41] Peng Zhou, Xintong Han, Vlad I. Morariu, and Larry S. Davis. Learning rich features for image manipulation detection. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, pages 1053–1061. Computer Vision Foundation / IEEE Computer Society, 2018. 2