3D mask presentation attack detection via high resolution face parts

Oleg Grinchuk, Aleksandr Parkin, Evgenija Glazistova
VisionLabs
https://github.com/AlexanderParkin/chalearn_3d_hifi

Abstract

3D mask presentation attack detection (PAD) is a long standing challenge in face anti-spoofing due to the high fidelity of attack artifacts and a limited number of samples available for training and evaluation. With the recent release of the large-scale and diverse CASIA-SURF HiFi-Mask dataset [19], it now becomes possible to address 3D mask PAD with deep neural networks. This paper introduces a new one-shot method for 3D mask PAD that extracts fine-grained information from appropriate parts of the human face and uses it to identify subtle differences between real and fake samples. The proposed method achieves state-of-the-art results of 3% ACER on the CASIA-SURF HiFi-Mask test set.

1. Introduction

Face anti-spoofing is a must-have component for the majority of face recognition applications. Presentation attack detection (PAD) can be efficiently realized with special cameras [18, 6, 33] given their ability to deliver useful features for face anti-spoofing (e.g. depth cameras extract 3D structure and infrared cameras filter out the light emitted by phone screens and monitors). The major drawback of such systems is the requirement of extra sensor equipment, which is rarely available in common scenarios of mobile and desktop authentication. Therefore, the demand for RGB-only based solutions is high. Moreover, many scenarios of face anti-spoofing require fast and simple authentication pipelines, while video-based PAD methods are slow and often involve human interaction [24].

PAD methods for printed and replay attacks [36, 4] have recently shown a great progress due to the availability of large and diverse datasets [22, 7, 20, 19]. At the same time, 3D printers became more accessible and accurate, decreasing the cost of producing realistic 3D face masks. These factors increase the importance of 3D mask PAD algorithms since cheap high fidelity plastic or silicone 3D masks with reproduced identity of a victim are becoming a real threat to face biometric systems.

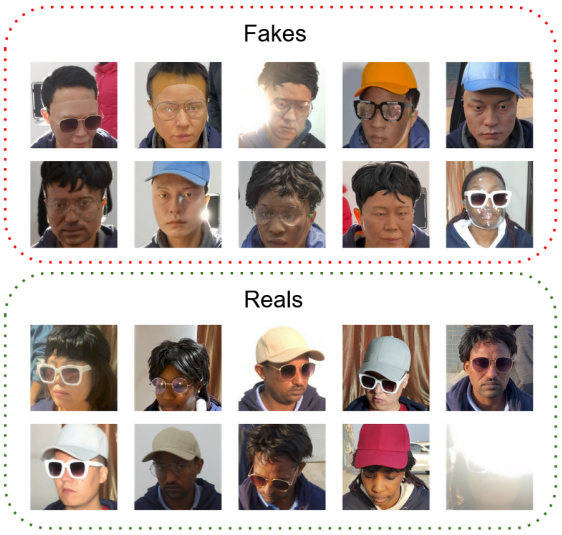

Figure 1. Fake and real samples from the CASIA-SURF HiFi Mask dataset.

In this paper we propose a new method for 3D high-fidelity face mask presentation attack detection that is capable of predicting liveness score from a single RGB image. Our method achieved the first place in 3D High-Fidelity Mask Face Presentation Attack Detection Challenge, based on the CASIA-SURF HiFiMask dataset [19].

A known issue for PAD is the poor generalization to new types of attacks and scenarios [26, 27, 32, 15]. The majority of available 3D PAD datasets lack the diversity in either number of subjects, skin tone, artifact materials or lighting conditions, causing low performance of algorithms in unseen domains. The recently introduced CASIA-SURF HiFiMask dataset provides the most diverse 3D mask attack images, reducing the problem of domain shift and allowing to move the focus on algorithmic improvements. To make this focus even more concrete, HiFiMask authors launched a challenge with strict conditions: no external data, no pre-trained models and no ensembling could be used. They also

fixed a test set protocol, where both seen and unseen 3D mask types are present. The protocol closely follows real life situations, where an algorithm that has been trained on one set of attacks must cope with new types of unseen presentation attacks.

The differences between high-fidelity 3D face masks and bona fide examples are highly subtle. A close view on a specific face region is often required to notice attack artifacts such as abnormal skin texture or eye glitter. Standard neural networks trained on whole face images could miss such fine-grained features due to the limited image resolution. To tackle this problem, we propose a part-based architecture that attends to multiple face regions at higher resolution. Besides the whole face image, we also consider ears, nose, eye and mouth regions. Each face part is processed by a different branch of a network, that learns region-specific features. For example, strong features for ears may encode the transition between the texture of a mask and the skin texture of an intruder. Since face parts provide limited information about the subject and do not always contain fake features, we use a modified binary cross entropy for each face part with an increased weight on bona fide class. Finally, to reduce the error of inaccurate face part crops, we extract features both from original and flipped images.

As a result, our method achieves state-of-the-art on one of the largest and most diverse liveness dataset – CASIA-SURF HiFiMask with 3% average classification error rate.

2. Related work

2.1. 3D Mask PAD methods

There are several publicly available 3D mask datasets with varying number of subjects, skin tones, capture conditions and mask types - 3DMAD [22], 3DFS-DB [7], HKBU-MARs [20], BRSU [29], WMCA [9], SiW-M [21], CASIA-SURF 3DMask [35]. The recently released CASIA-SURF HiFi Mask dataset surpasses previous datasets in an overall diversity and contains 54000 videos of 75 bona fide and 75 fake subjects with same identity.

3D mask PAD approaches, trained on these datasets, focus on loss function optimizations [8, 10, 17, 31], data sampling strategies [11], or try to utilize the difference in context features between paired bona fide and fake images [19]. However, due to the high fidelity of 3D masks and limited resolutions of common backbone architectures (e.g.ResNet [13], MobileNet[14], EfficientNet [30]), those approaches could miss the presence of strong local features in small face regions.

2.2. Face parts

The idea of splitting face images into multiple face parts is frequently used for face recognition [25] and is adopted for presentation attack detection in recent works [5, 28, 34].

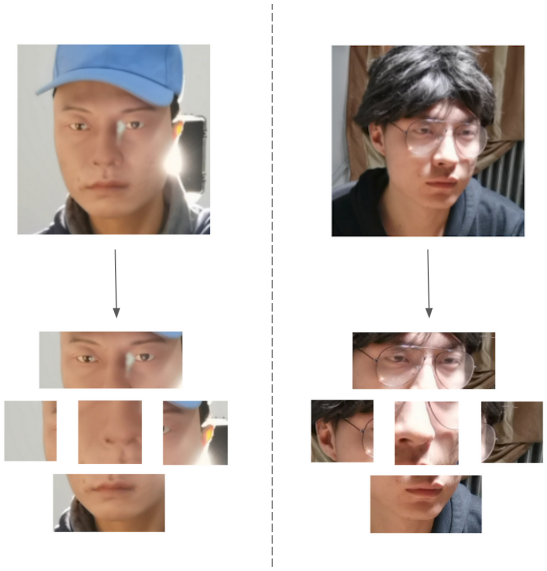

Figure 2. Proposed face parts: left ear, right ear, eyes, nose, mouth.

In [28] a bag of random face patches is fed to a single backbone. In [5] the authors split the image into 9 smaller squares and train different backbones for each branch. The drawback of such methods is that selected face parts are either too small and do not capture local semantic information or are focused only on local face parts, ignoring global semantics.

To tackle these issues, our proposed framework aggregates information from the whole image and appropriate local face parts, making use of both coarse and fine-gained features.

2.3. CASIA-SURF HiFi Mask dataset

HiFiMask is currently the largest 3D RGB face dataset: it contains 54,600 short videos captured from 75 subjects. 3 types of spoof attacks are present: transparent plastic, highly realistic plaster and resin masks. The examples of dataset images are shown in Fig. 1.

The structure of the HiFiMask dataset aims to reduce common problems for deep learning methods:

- All 3D mask identities match the bona fide identities thus preventing the models to overfit on person id.
- Each bona fide and fake sample is recorded under different lighting conditions (white light, green light, periodic three-color Light, outdoor sunshine, outdoor shadow, and motion blur) and light intensity (normal light, dim light, bright light, back light, side light and top light). This reduces model overfitting to scene-specific features.
- Videos are recorded on a wide range of high-resolution devices: iPhone11, iPhoneX, MI10, P40, S20, Vivo,

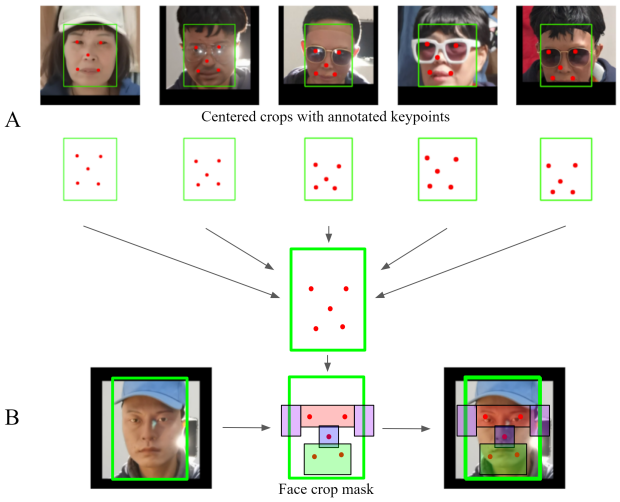

Figure 3. **A**. Centered face crops (top) are manually annotated with face keypoints and keypoint locations are then averaged to obtain average keypoint mask. **B**. Average keypoint mask is used to define face part regions for arbitrary face images.

HJIM, etc., which helps to develop robustness to camera type. Additional details like glasses, hats and wigs are used to make spoof attacks more plausible while adding these attributes to bona fide examples eliminates overfitting probability to these attributes.

The common problem of PAD methods is their generalization ability to unseen attacks. To address this issue, authors created a mixed protocol, where test set contains both seen and unseen attacks. This protocol is used in CASIA-SURF HiFiMask PAD challenge.

3. Proposed method

This section describes the proposed face part based method for 3D mask presentation attack detection. We start with the baseline model and then introduce in details the process of face part cropping, loss modification and post processing.

3.1. Face detector

Face centering is a common pre-processing step aiming to reduce the spatial variability of face images. We use a DSFD [16] face detector and center crop all training and test images such that the face bounding box is placed at the center of a square crop. Square crops are 1.3 times larger than detected face bounding boxes. If the obtained crop exceeds boundaries of the original image, we fill the missing parts with black pixels, see Fig. 3.

3.2. Baseline

We use a standard EfficientNet-B0 [30] architecture trained with binary cross-entropy loss as a baseline method. In our experiments we tried different architecture families as well as deeper versions inside same family and found that EfficientNet-B0 is better or on par with other models.

3.3. Face parts

High fidelity 3D masks are often very hard to distinguish from genuine person images. Image features that can help identifying fake faces are often local and require high resolution image information. Moreover, some features are strictly region specific - for example, unnatural eye glitter can only be discovered in eye region. Rubber band holding the mask can be typically found in ear regions. Though standard neural networks could process the whole face image and find these features, in some cases the features are so small that they are being blurred out when a pre-prosessing resizes image before passing it to the network.

To deal with this issue, we introduce face parts as illustrated in Fig. 2. Original image X is processed with DSFD detector to transform it into centered face crop X_c. We crop out semantically meaningful face parts: X_{eyes}, X_{nose}, X_{mouth}, $X_{leftear}$, $X_{rightear}$ and process them along with a whole face X_{face} at a higher resolution rate. Each face part is resized to 224×224 pixels. In order to obtain the face parts from the image, we need a prior information about face part locations.

We manually annotate 100 images from a training set with face keypoints and then average obtained keypoints, which resulted in 5 point coordinates $K = \{x_i, y_i\}_{i=1,\dots,5}$ of an averaged keypoint mask (see Fig. 3). We annotate eyes centers, nose and mouth corners and then use this information to define face part bounding boxes:

$$X_{facepart} = f(X_c, K, \theta),$$

where θ are manually defined shift and scale parameters for each face part. For exact values of θ please refer to our implementation on GitHub [1].

3.4. Network architecture

To use high-resolution image features, we propose an extension to the baseline method by adding extra branches that process different face parts. The method diagram is shown in Fig. 4. Original image X is processed with DSFD detector to transform it into a centered face crop X_c as described in Sec. 3.1. Note that we do not re-scale face images at this step. Face parts are then extracted from X_c using prior information about average keypoint locations K and then processed by a *shared* conv-bn-relu module. Each face part is then further processed by a copy of EfficientNetB0 body blocks with part-specific parameters. Left and right ears

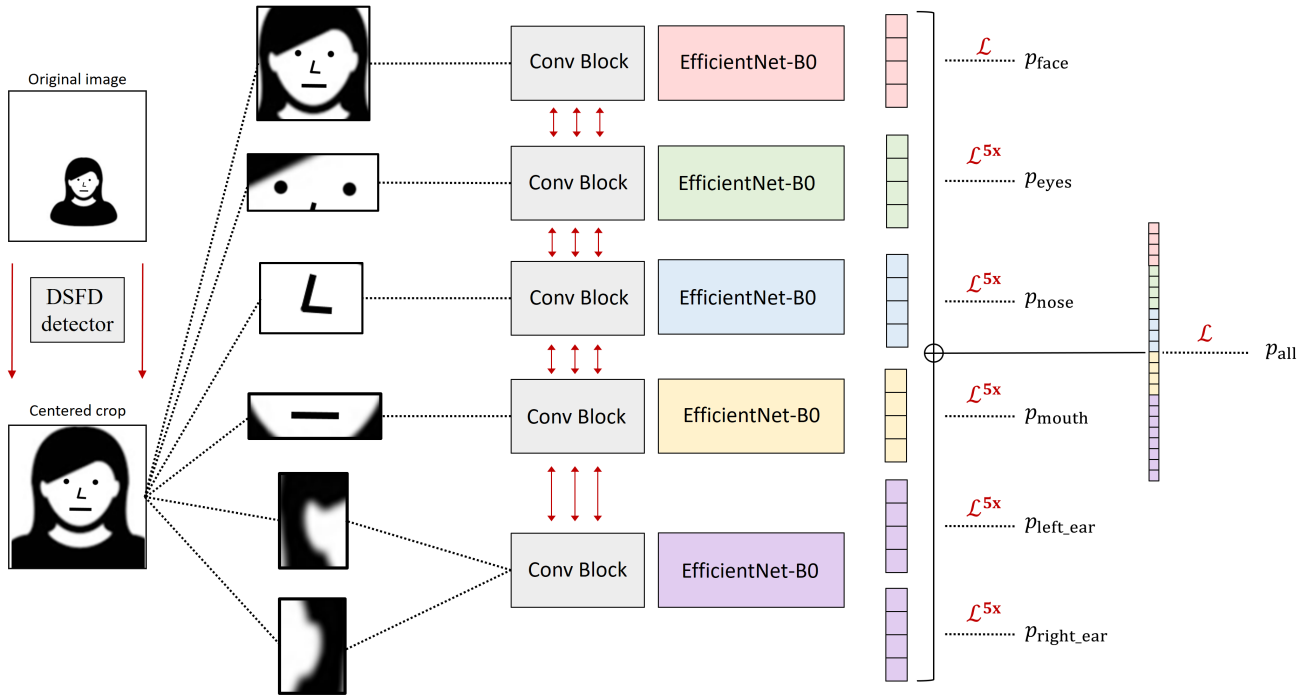

Figure 4. Method diagram. Centered crop is split into face parts which are further processed with different EfficientNet-B0 backbones. All backbones share the same conv-relu-bn block in the beginning. Each face part is trained separately with BCE or BCE weighted loss. At the same time the concatenated feature vector is trained with the BCE loss.

are fed into a common backbone with shared parameters. We also modify the standard B0 architecture so that the last fully connected layer returns a feature vector of length 320.

3.5. Loss function

Face parts may not always contain information about liveness, so training with regular BCE loss could result in a neural net confusion, reducing the accuracy of the whole pipeline. To tackle this, we propose to use weighted loss for face parts with a focus on bona fide class. This will reduce a penalty for misclassifying fake class for cases when face part does not contain any liveness related features.

Our face part architecture returns 6 descriptors of length 320: d_{face}, d_{eyes}, d_{nose}, d_{mouth}, $d_{\text{left ear}}$, $d_{\text{right ear}}$. Each descriptor is processed with sigmoid and fed into corresponding classification layers, resulting in liveness scores p_{face}, p_{eyes}, p_{nose}, p_{mouth}, $p_{\text{left ear}}$, $p_{\text{right ear}}$. In addition to that, a descriptor d_{all}, obtained by concatenating face part descriptors, is also fed into a classification layer, that returns p_{all} score.

d_{all} and d_{face} parts are trained with the standard BCE loss $\mathcal{L}$. All other face parts are trained using BCE weighted loss

$$\mathcal{L}^{\beta\times} = -\beta \cdot y \log(p) - (1-y) log(1-p), \qquad (1)$$

where we use $\beta = 5$ to increase the weight for bona fide class. The final loss function is a combination of losses for different face parts and for the whole descriptor:

$$\begin{aligned} \mathcal{L}_{total}(y, \mathbf{p}) = 5 \cdot \mathcal{L}(y, p_{\text{all}}) + 5 \cdot \mathcal{L}(y, p_{\text{face}}) \\ + \mathcal{L}^{5\times}(y, p_{\text{eyes}}) + \mathcal{L}^{5\times}(y, p_{\text{nose}}) + \mathcal{L}^{5\times}(y, p_{\text{mouth}}) \\ + \frac{1}{2} \cdot \mathcal{L}^{5\times}(y, p_{\text{rightear}}) + \frac{1}{2} \cdot \mathcal{L}^{5\times}(y, p_{\text{leftear}}) \end{aligned} \qquad (2)$$

3.6. Postprocessing

Global loss for concatenated descriptor aggregates information from different face part descriptors while local part specific losses push corresponding model branches. We further aggregate global and local predictions, allowing direct local part contribution to the final score, so that a strong signal from any of face parts will not be missed. To calculate the aggregated score, we use a formula similar to (2):

$$\begin{aligned} p_{\text{agg}} = \frac{1}{8}(2 \cdot p_{\text{all}} + 2 \cdot p_{\text{face}} + p_{\text{eyes}} + p_{\text{nose}} + p_{\text{mouth}} + \\ \frac{1}{2} \cdot p_{\text{rightear}} + \frac{1}{2} \cdot p_{\text{leftear}}) \end{aligned} \qquad (3)$$

Using face part locations cropped from averaged prior information could lead to inaccurate cropping. To mitigate this effect, we additionally compute (3) for horizontally flipped images and average predictions for the original and flipped images.

	Method	Val			Test		
		APCER, %	BPCER, %	ACER, %	APCER, %	BPCER, %	ACER, %
1	Baseline	0.40	1.02	0.71	5.57	5.12	5.34
2	Naive face parts	1.37	1.06	1.22	1.99	8.95	5.47
3	+Face part aggregation	1.09	0.88	0.98	2.70	7.24	4.97
4	+Shared conv block	0.69	1.43	1.06	7.12	2.68	4.90
5	+Weighted loss	1.05	1.53	1.29	2.44	4.94	3.69
6	+Postprocessing	0.85	1.25	1.05	3.78	2.33	3.05

Table 1. Results on CASIA-SURF HiFiMask validation and test subsets.

4. Experiments

4.1. Experimental settings

All experiments are conducted on the CASIA-SURF HiFiMask dataset following the protocol 3 [19]. For performance measurement we use Attack Presentation Classification Error Rate (APCER), Bona fide Presentation Classification Error Rate (BPCER), and ACER metrics. We report results on the validation and test subsets of the dataset. We fix the learning rate schedule, the number of training epochs and image augmentations for all our experiments.

Implementation details. All our code is written in python with the pytorch framework and is available on GitHub [1]. All models are trained using 4 NVIDIA 3080 Ti GPU. We fix a random seed, however, due to multi-gpu training, results could sligtly differ between different runs.

Baseline. As a baseline method we use EfficientNet-B0 trained with BCE loss on a target task. We also tried ResNet and MobileNet families, but EfficientNet showed the best result with ACER=5.343 % (see Tab. 1) on the test set. Meanwhile, on the validation set the result was 0.710 % ACER. This difference is caused by the presence of the unseen fake type in the test set. Also, this shows that validation set is not the best measure to select the best model. In the following we report results on the more challenging test set, and, if not mentioned explicitly, results are reported for this set. We also build up a learning rate strategy to drop to almost 0 by the end of the training and always select the last epoch as a checkpoint.

Pretraining. According to the terms of the challenge, the use of pretrained networks on external data was prohibited, however, we decided to evaluate the contribution of transfer learning to see if we can use any self-supervised training strategies. Pretraining the baseline network on ImageNet decreased the error ACER from 5.34 % to 4.47 %, however, the usage of various self-supervised methods such as jigsaw[23], simsiam[3], Moco[12] or simCLR[2] only increased the error to 5.72 % ACER. Presumably this is due to the fact that current self-learning methods require a large amount of training data, orders of magnitude larger than HiFiMask dataset size. Therefore, in all further experiments we trained our network with a random initialization.

Face crop preprocessing. All images are converted into centered crops in advance. During training, we extract face parts and apply a set of individual augmentations (Color Jitter, Blur, Random Crop, Horizontal Flip) for each face part. All face parts are then resized to 224×224 pixels to match the backbone input resolution.

4.2. Ablation study

Here we report results of ablation experiments (Tab. 1 to verify the contributions of each added feature. We examine the effect of the following modules:

- **Naive face part architecture**. Compared with Fig. 4 we do not use a shared convolutional block and do not aggregate face part scores at the end, using only p_{all} output. We also use the standard BCE for all parts.
- **Face part aggregation**. We transform final score using (3).
- **Shared convolution**. We make the weights of first EfficientNet-B0 block shared between different face part backbones.
- **Weighted BCE for face parts**. We substitute BCE with the weighted BCE loss for all face parts except the whole face branch.
- **Postprocessing**. We compute scores both for original and flipped images and report averaged results.

Results on the validation set contradict results on the test set. We believe that this is due to some portion of low-quality images that appear in HiFiMask. DFSD Face detector did not found any face in 1% of the dataset, therefore making any results that are close to same rate noisy. In further experiments we mostly analyze results on the test set.

Naive face part architecture. In this experiment we tested the proposed architecture but without shared convolution block and scores aggregation, e.g. using only $p_{\mathbf{all}}$ predictions. As a result, the model achieved 5.47 % ACER which is worse than the baseline.

Face part aggregation. We used the outputs from the previous experiment and aggregated them using (3). Formula coefficients were not finetuned and were selected by

	Face	Eyes	Nose	Ears	Mouth	All	hflip	APCER, %	BPCER, %	ACER, %
1	✓							1.89	3.70	2.80
2		✓						3.55	2.54	3.04
3			✓					6.24	3.42	4.83
4				✓				8.58	3.74	6.16
5					✓			4.23	4.67	4.45
6						✓		0.93	1.57	1.25
7	✓	✓	✓	✓	✓			1.17	1.62	1.39
8	✓	✓	✓	✓	✓	✓		0.85	1.57	1.21
9	✓	✓	✓	✓	✓	✓	✓	0.85	1.25	1.05

Table 2. Results on CASIA-SURF HiFiMask validation subset. Checkmark indicates the presence of the corresponding part/technique in final score.

rough analysis of individual face parts performance (Table 2). Aggregated score improved the resulting ACER by 0.5 %, proving that individual face part contribution should also be accounted.

Shared convolution. Challenge protocol bans the usage of any pretrained models, so all our experiments were trained from scratch. A strong pretrained baseline creates generic and robust features at the first network layers. We have tried to partially replicate this effect by sharing first convolution between all backbone instances. This slightly improved the final metric, achieving 4.9% ACER on the test set.

Weighted BCE for face parts. This experiment shows the importance of using weighted binary cross-entropy loss for partial face parts (eyes, nose, mouth, ears). Bona fide class weight increase leads to less penalty for the cases when the face part branch could not find any signs of a spoof. The model trained with new loss achieved 3.69% ACER, which is 1.21% better compared to the model trained with a standard loss.

Postprocessing. Using the model trained in the previous experiment, we additionally extract scores for horizontally flipped face images and average results for original and flipped images. This was the final addition of our method, which led to the score 3.05% ACER. The improvement by 0.64% with added horizontal flip augmentation is likely due to the fact, that face parts were not always cropped perfectly. We visually examined some examples with highest variation between original and flipped images and found that in most cases there were an error in face part cropping.

4.3. Face parts contribution

Here we examine the contribution of selected face parts into the final score. Table 2 shows ACER on the HiFiMask validation set. The whole face shows 2.796% ACER, which is the best individual result compared to other face parts. Among partial crops, the most meaningful result is demonstrated by eyes part, which achieved 3.044% ACER. We visually examined eye crops and in most cases we noticed clear visual artifacts corresponding to fake faces. However, the dataset contains both fake and bona fide examples with dark glasses, making eye classifier unreliable in such cases. Aggregated scores improved results to 1.21% ACER and the horizontal flip decreased the error further to 1.05% (see Tab. 2).

5. Summary

In this paper we introduced a novel method for 3D mask presentation attack detection bh combining analysis of the whole face and semantically meaningful face parts. We proposed a method that decomposes a face image into multiple face parts and processes them at a higher resolution, which results in better performance compared to baseline architectures. We also showed the importance of sharing first layers between different branches, since their mutual learning leads to a more general and robust convolution filters. Finally, we demonstrated that when using parts with partially missing discriminative features, the use of weighted loss is preferable. In our experimental validation we showed the effect of each of above-mentioned components on the final metric and also examined the importance of individual face parts for the PAD task. The proposed method resulted in the winning submission on the recently conducted 3D High-Fidelity Mask Face Presentation Attack Detection Challenge, reaching 3.05 % ACER.

References

[1] https://github.com/AlexanderParkin/chalearn_3d_hifi. 3, 5

[2] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. *arXiv preprint arXiv:2002.05709*, 2020. 5

[3] Xinlei Chen and Kaiming He. Exploring simple siamese representation learning. *CoRR*, abs/2011.10566, 2020. 5

[4] Ivana Chingovska, André Anjos, and Sébastien Marcel. On the effectiveness of local binary patterns in face anti-spoofing. pages 1–7, 2012. 1

[5] Gustavo Botelho de Souza, João Paulo Papa, and Aparecido Nilceu Marana. On the learning of deep local features

for robust face spoofing detection. *CoRR*, abs/1806.07492, 2018. 2

[6] Nesli Erdogmus and Sébastien Marcel. Spoofing in 2d face recognition with 3d masks and anti-spoofing with kinect. In *2013 IEEE Sixth International Conference on Biometrics: Theory, Applications and Systems (BTAS)*, pages 1–6, 2013. 1

[7] Javier Galbally and Riccardo Satta. Three-dimensional and two-and-a-half-dimensional face recognition spoofing using three-dimensional printed models. *IET Biometrics*, 5(2):83–91, 2016. 1, 2

[8] Anjith George and Sébastien Marcel. Cross modal focal loss for rgbd face anti-spoofing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7882–7891, 2021. 2

[9] 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. 2

[10] Huiling Hao, Mingtao Pei, and Meng Zhao. Face liveness detection based on client identity using siamese network. In *PRCV*. Springer, 2019. 2

[11] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *CVPR*, 2020. 2

[12] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020. 5

[13] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. pages 770–778, 2016. 2

[14] Andrew G Howard, Menglong Zhu, Bo Chen, Dmitry Kalenichenko, Weijun Wang, Tobias Weyand, Marco Andreetto, and Hartwig Adam. Mobilenets: Efficient convolutional neural networks for mobile vision applications. *arXiv preprint arXiv:1704.04861*, 2017. 2

[15] Yunpei Jia, Jie Zhang, Shiguang Shan, and Xilin Chen. Single-side domain generalization for face anti-spoofing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8484–8493, 2020. 1

[16] Jian Li, Yabiao Wang, Changan Wang, Ying Tai, Jianjun Qian, Jian Yang, Chengjie Wang, Jilin Li, and Feiyue Huang. Dsfd: dual shot face detector. pages 5060–5069, 2019. 3

[17] Lei Li, Zhaoqiang Xia, Xiaoyue Jiang, Fabio Roli, and Xiaoyi Feng. Compactnet: learning a compact space for face presentation attack detection. *Neurocomputing*, 2020. 2

[18] Ajian Liu, Jun Wan, Sergio Escalera, Hugo Jair Escalante, Zichang Tan, Qi Yuan, Kai Wang, Chi Lin, Guodong Guo, Isabelle Guyon, and Stan Z. Li. Multi-modal face anti-spoofing attack detection challenge at cvpr2019. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2019. 1

[19] 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. 1, 2, 5

[20] Siqi Liu, Baoyao Yang, Pong C Yuen, and Guoying Zhao. A 3d mask face anti-spoofing database with real world variations. pages 100–106, 2016. 1, 2

[21] Yaojie Liu, Joel Stehouwer, Amin Jourabloo, and Xiaoming Liu. Deep tree learning for zero-shot face anti-spoofing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4680–4689, 2019. 2

[22] Erdogmus Nesli and Sébastien Marcel. Spoofing in 2d face recognition with 3d masks and anti-spoofing with kinect. pages 1–8, 2013. 1, 2

[23] Mehdi Noroozi and Paolo Favaro. Unsupervised learning of visual representations by solving jigsaw puzzles. In Bastian Leibe, Jiri Matas, Nicu Sebe, and Max Welling, editors, *Computer Vision – ECCV 2016*, pages 69–84, Cham, 2016. Springer International Publishing. 5

[24] Aleksandr Parkin and Oleg Grinchuk. Creating artificial modalities to solve rgb liveness, 2020. 1

[25] P Jonathon Phillips, Amy N Yates, Ying Hu, Carina A Hahn, Eilidh Noyes, Kelsey Jackson, Jacqueline G Cavazos, Géraldine Jeckeln, Rajeev Ranjan, Swami Sankaranarayanan, et al. Face recognition accuracy of forensic examiners, superrecognizers, and face recognition algorithms. *Proceedings of the National Academy of Sciences*, 115(24):6171–6176, 2018. 2

[26] Rui Shao, Xiangyuan Lan, Jiawei Li, and Pong C Yuen. Multi-adversarial discriminative deep domain generalization for face presentation attack detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10023–10031, 2019. 1

[27] Rui Shao, Xiangyuan Lan, and Pong C Yuen. Regularized fine-grained meta face anti-spoofing. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 11974–11981, 2020. 1

[28] Tao Shen, Yuyu Huang, and Zhijun Tong. Facebagnet: Bag-of-local-features model for multi-modal face anti-spoofing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2019. 2

[29] Holger Steiner, Andreas Kolb, and Norbert Jung. Reliable face anti-spoofing using multispectral swir imaging. pages 1–8, 2016. 2

[30] Mingxing Tan and Quoc Le. Efficientnet: Rethinking model scaling for convolutional neural networks. pages 6105–6114, 2019. 2, 3

[31] Guoqing Wang, Hu Han, Shiguang Shan, and Xilin Chen. Improving cross-database face presentation attack detection via adversarial domain adaptation. In *ICB*. IEEE, 2019. 2

[32] Guoqing Wang, Hu Han, Shiguang Shan, and Xilin Chen. Cross-domain face presentation attack detection via multi-domain disentangled representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6678–6687, 2020. 1

[33] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020. 1

[34] Qing Yang, Xia Zhu, Jong-Kae Fwu, Yun Ye, Ganmei You, and Yuan Zhu. Pipenet: Selective modal pipeline of fusion network for multi-modal face anti-spoofing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2020. 2

[35] 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

[36] Zhiwei Zhang, Junjie Yan, Sifei Liu, Zhen Lei, Dong Yi, and Stan Z Li. A face antispoofing database with diverse attacks. pages 26–31, 2012. 1