Single Patch Based 3D High-Fidelity Mask Face Anti-Spoofing

Samuel Huang
CyberLink
samuel_huang@cyberlink.com

Wen-Huang Cheng
National Yang Ming Chiao Tung University
whcheng@nycu.edu.tw

Robert Cheng
CyberLink
robert_cheng@cyberlink.com

Abstract

Face anti-spoofing is rapidly increasing in importance as facial recognition systems have become common in the financial and security fields. Among all kinds of attack, 3D high-fidelity masks are especially hard to defend. Recently, CASIA introduced a large scale dataset CASIA-SURF HiFiMask, which comprises of 54,600 videos recorded from 75 subjects with 225 high-fidelity masks. In this paper, we design a lightweight network with single patch input on the basis of CDCN++, and supervise it by focal loss. The proposed method achieves the Average Classification Error Rate (ACER) of 3.215 on the Protocol 3 of CASIA-SURF HiFiMask dataset and ranks the third best model in the Chalearn 3D High-Fidelity Mask Face Presentation Attack Detection Challenge at ICCV 2021.

1. Introduction

Face recognition [5, 2, 7, 6] has been widely used in our life and brought convenience, e.g mobile payment, access control, and so forth. However, vulnerability to variable attacks curtails its reliable deployment. Plenty of malicious attacks such as printed photos, video replays, 3D masks and others could easily confuse the systems that make the wrong judgement. In order to protect the privacy and prevent the property from unauthorized use, various face anti-spoofing methods have been proposed over the past decades. Many approaches for 2D attacks such as photos and replay attacks have made great progress, where some of them use RGB images only and others use multiple inputs for higher accuracy like NIR or depth information [9]. However, ever-changing 3D printing technology makes it easier to fabricate masks with face. There are more realistic textures and structure of bona fide on these masks than the traditional 2D attacks.

Recently, a large-scale cross-ethnicity face anti-spoofing dataset built with all 3D mask attacks called CASIA-SURF HiFiMask [4] provides 25 subjects with different tone color and three kinds of high-fidelity mask of each subject. The dataset is utilized for Chalearn 3D High-Fidelity Mask Face Presentation Attack Detection Challenge at ICCV2021 [1]. The baseline results in CASIA-SURF HiFiMask [4] indicate that the best result is 10.5 ACER in the Protocol 3. Therefore, it is necessary to find out more effective methods for cross-environment and cross-ethnicity 3D high-fidelity mask attacks.

In this paper, we design a focal loss supervised network with single patch input on the basis of CDCN++ [10] for 3D mask anti-spoofing. The proposed method achieves 3.215 ACER on the Protocol 3 with CASIA-SURF HiFiMask dataset.

2. CASIA-SURF HifiMask dataset

CASIA-SURF HifiMask dataset [4] includes 75 subjects, each subject provides high-fidelity plaster, resin, and transparent masks. 6 different environments, 6 directional lighting, and 7 recording sensors are applied in the dataset. In total, the dataset provides 54,600 videos (13,650 live, 40,950 mask). For the Chalearn 3D High-Fidelity Mask Face Presentation Attack Detection Challenge at ICCV 2021, the organizers introduced the Protocol 3-'open set' [4] of the dataset, which is more general for real-world deployment because it considers both 'seen' and 'unseen' domains and masks. The training and development sets lack one mask type and some scenarios, while the testing set provides all. The training set contains 3,715 videos, development set contains 536 videos, and testing set contains 17,362 videos. In the challenge, videos are split into images so that only single image based methods can be applied.

3. Methodology

In this section, we will describe the proposed method and its training details for the Protocol 3 of CASIA-SURF HiFiMask dataset [4], which includes face detection, data augmentation, training configuration and architecture.

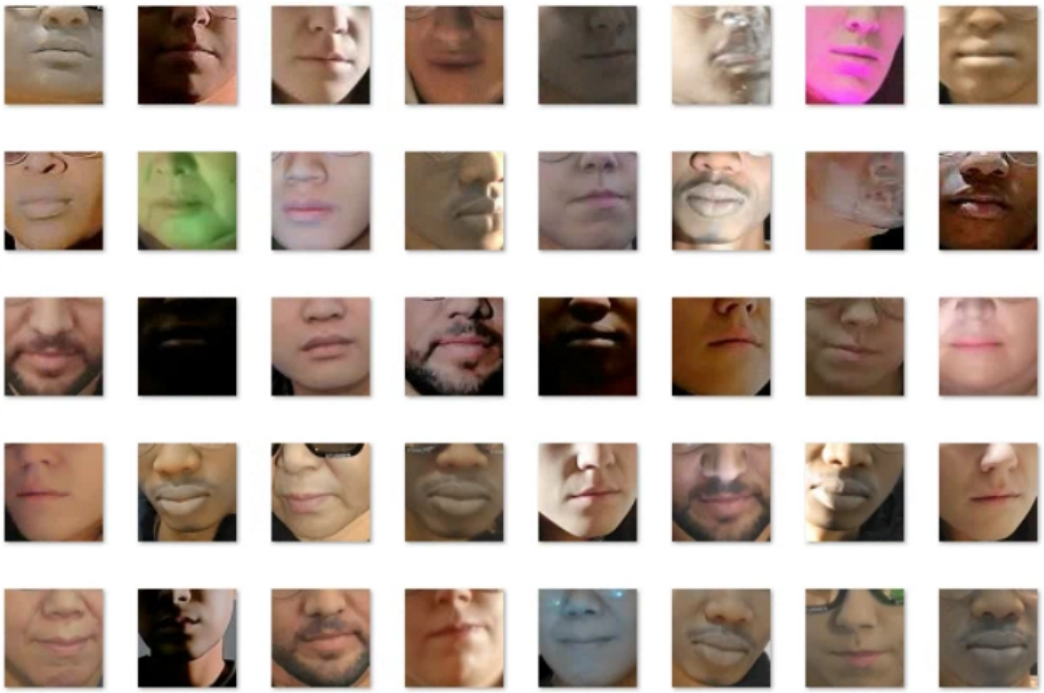

Figure 1. Square face patches which are randomly cropped in a small range around mouth.

3.1. Face Detection

We use CyberLink FaceMe® SDK[1] for face detection on the CASIA-SURF HiFiMask dataset [4]. Compared with the detection results of other modules like Dlib, the face bounding boxes detected by FaceMe® SDK is more close-fitting and accurate. The coordinates of face bounding boxes will be stored in a text file for the next step. If FaceMe® SDK detect more than one face in an image, the bounding box being stored will be the largest one. Also, if FaceMe® SDK didn't detect any face in a image, the bounding box being stored will be the whole image.

3.2. Data Augmentation

Due to the fact that the training and developing sets in the Protocol 3 of CASIA-SURF HiFiMask dataset [4] contain only parts of scenarios and mask types, several data augmentation strategies including random crop, cutout, random erase, and random flip are adopted to improve the generalization ability of the model. To increase robustness to new scenarios, we also randomly add pixel values and gamma value in a specific range.

3.3. Architecture

Our model is constructed based on CDCN++ [10]. The original CDCN++ is supervised by facial depth map, which is generated from PRNet [8]. Instead of facial depth map based loss, we find that focal loss [3] performs better when facing 3D mask attacks from CASIA-SURF HiFiMask dataset [4]. Focal loss is based on cross entropy (CE) loss for binary classification :

$$\mathrm{CE}(p, y) = \begin{cases} -\log(p) & \text{if } y = 1 \\ -\log(1-p) & \text{if } y = 0 \end{cases} \tag{1}$$

where $p \in [0, 1]$ is the model's probability for the class, and $y \in \{0, 1\}$ specifies class labels including attack and bona fide. We define p_t for notational convenience:

$$p_t = \begin{cases} p & \text{if } y = 1 \\ 1-p & \text{if } y = 0 \end{cases} \tag{2}$$

and define α_t analogously to how we defined p_t, so that we can write the α-balanced CE loss as:

$$\mathrm{CE}(p_t) = -\alpha_t \log(p_t). \tag{3}$$

With a modulating factor $(1-p_t)^\gamma$, the α-balanced focal loss (FL) [3] is defined as:

$$\mathrm{FL}(p_t) = -\alpha_t (1-p_t)^\gamma \log(p_t). \tag{4}$$

where the focusing parameter γ adjusts the rate at which easy examples are downweighted. We found $\alpha = 1, \gamma = 2$

[1] https://www.cyberlink.com/faceme

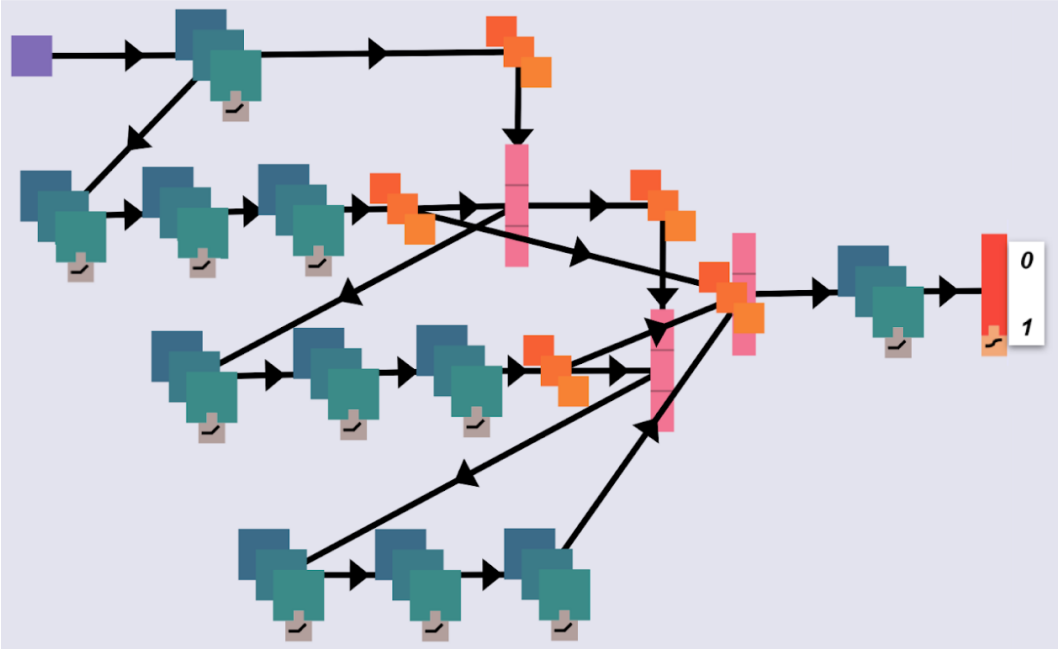

Figure 2. The architecture of our model. The purple blocks denote the input, green blocks denote convolution layers, and orange blocks denote average pooling layers. The pink stick represents the concat operation. Spatial attention block is applied before each concat operation. The orange stick denotes fully connected layer, and the white stick represent focal loss output.

to work best in the Protocol 3 of CASIA-SURF HiFiMask dataset [4].

Because all live faces and masks were randomly equipped with decorations like hat, sunglasses, wig, and glasses in the CASIA-SURF HiFiMask dataset [4], we use square face patch which is randomly cropped in a small range around mouth as the input, instead of the whole face to avoid possible side effects that may be caused by these appendages. The examples are shown in Figure 1. In the testing phase, we adopt a 10x10 sliding window self-voting strategy in the same range of the training phase to obtain the best performance on accuracy. We also found that there are a certain amount of images that FaceMe® SDK can't detect any face in the testset of Protocol 3, so we randomly apply a full image as the bounding box while training to make our model also capable of handling the situation mentioned above. The face patches are all resized to 56x56 as the final input of model.

We use the adam optimizer with weight decay to train the model. the model structure we contructed based on CDCN++ [10] is shown in Figure 2. The model's complexity is 1.73 GMac, which is fairly lightweight comparing to other common networks. We linearly warm-up the model in first 20 epochs, and then slowly reduce the learning rate to 0 until 100 epochs using cosine decay schedule.

4. Experiments

In this section, we will describe the performance of our proposed method on the Protocol 3 of CASIA-SURF HiFiMask dataset [4], and compare with the result of Chalearn 3D High-Fidelity Mask Face Presentation Attack Detection Challenge at ICCV 2021.

4.1. Evaluation Metrics

Evaluation Metrics are calculated based on the following values: True Positive (TP), False Positive (FP), True Negative (TN), and False Negative (FN). Attack Presentation Classification Error Rate (APCER) is denoted as:

$$APCER = \frac{FP}{TN + FP} \tag{5}$$

and Bona Fide Presentation Classification Error Rate (BPCER) is denoted as:

$$BPCER = \frac{FN}{TP + FN} \tag{6}$$

Finally, we denote Average Classification Error Rate (ACER) as:

$$ACER = \frac{APCER + BPCER}{2} \tag{7}$$

which is used to measure the performance on the test set and determine the final ranking of the competition (lower ACER value is better).

Table 1. Chalearn 3D High-Fidelity Mask Face Presentation Attack Detection Challenge at ICCV 2021

Team	APCER	BPCER	ACER
fscr	6.095	7.52	6.808
DXM-DI-AI-CV-TEAM	8.444	4.175	6.31
VIC_FACE	8.843	2.399	5.621
msxf_cvas	5.773	5.352	5.562
Piercing Eyes	3.669	7.313	5.491
inspire	5.834	4.06	4.947
Reconova-AI-LAB	2.126	6.367	4.247
oldiron666	4.944	2.653	3.798
We Only Look Once	1.858	4.452	3.155
VisionLabs	3.777	2.33	3.053
Ours	3.708	2.722	3.215

4.2. Performance

We achieve the score of 3.215 ACER (3.708 APCER, 2.722 BPCER) on the test set of the Protocol 3 of CASIA-SURF HiFiMask dataset [4], which is also the protocol of Chalearn 3D High-Fidelity Mask Face Presentation Attack Detection Challenge at ICCV 2021. The team scores of the competition and our score is shown in Table 1. Our ACER performance beats the baseline (10.5 ACER) and all teams except for the teams VisionLabs and We Only Look Once.

5. Conclusion

In this paper, we construct our focal loss supervised lightweight network based on CDCN++ for 3D high-fidelity mask anti-spoofing on Protocol 3 of CASIA-SURF HiFiMask dataset [4]. We use face patch around mouth instead of the whole face as input to avoid side effects caused by random decorations in the dataset, and apply several augmentations to improve generalizability. The proposed method achieves 3.215 ACER on Protocol 3 of CASIA-SURF HiFiMask dataset, and ranks the third best among all teams in the Chalearn 3D High-Fidelity Mask Face Presentation Attack Detection Challenge at ICCV 2021.

References

[1] Chalearn 3D High-Fidelity Mask Face Presentation Attack Detection Challenge@ICCV2021. `https://sites.google.com/qq.com/face-anti-spoofing/winners-results/challengeiccv2021`. Accessed: 2021-06-20.

[2] Fatma S. Abousaleh, Tekoing Lim, Wen-Huang Cheng, Neng-Hao Yu, M. Anwar Hossain, and Mohammed F. Alhamid. A novel comparative deep learning framework for facial age estimation. *No. 47, EURASIP Journal on Image and Video Processing, 2016*.

[3] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár. Focal loss for dense object detection, 2018.

[4] Ajian Liu, Chenxu Zhao, Zitong Yu, Jun Wan, Anyang Su, Xing Liu, Zichang Tan, Sergio Escalera, Junliang Xing, Yanyan Liang, Guodong Guo, Zhen Lei, Stan Z. Li, and Du Zhang. Contrastive context-aware learning for 3d high-fidelity mask face presentation attack detection, 2021.

[5] Ling Lo, Hong-Xia Xie, Hong-Han Shuai, and Wen-Huang Cheng. Facial chirality: Using self-face reflection to learn discriminative features for facial expression recognition. *The 2021 IEEE International Conference on Multimedia and Expo (ICME), 5-9 July, 2021, Shenzhen, China*.

[6] Jordi Sanchez-Riera, Kai-Lung Hua, Yuan-Sheng Hsiao, Tekoing Lim, Shintami C. Hidayati, and Wen-Huang Cheng. A comparative study of data fusion for rgb-d based visual recognition. *Pattern Recognition Letters, vol. 73, pp. 1-6, April 2016*.

[7] Hong-Xia Xie, Ling Lo, Hong-Han Shuai, and Wen-Huang Cheng. Au-assisted graph attention convolutional network for micro-expression recognition. *The 28th ACM International Conference on Multimedia (MM 2020), 12-16 October, 2020, Seattle, USA*.

[8] Fan Wu Yao Feng, Xiaohu Shao, Yanfeng Wang, and Xi Zhou. Joint 3d face reconstruction and dense alignment with position map regression network. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2017.

[9] Zitong Yu, Yunxiao Qin, Xiaobai Li, Zezheng Wang, Chenxu Zhao, Zhen Lei, and Guoying Zhao. Multi-modal face anti-spoofing based on central difference networks. In *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

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