# 3D High-Fidelity Mask Face Presentation Attack Detection Challenge

Ajian Liu[1], Chenxu Zhao[2], Zitong Yu[3], Anyang Su[2], Xing Liu[2], Zijian Kong[2]
Jun Wan[4,5,9*], Sergio Escalera[7,9], Hugo Jair Escalante[8,9], Zhen Lei[4,5,6], Guodong Guo[10]
[1]MUST, Macau [2]Mininglamp Academy of Sciences, Mininglamp Technology, China
[3]University of Oulu, Finland [4]NLPR, CASIA, China [5]SAI, UCAS, China [6]CAIR, HKISI, CAS
[7]CVC, UB, Spain [8]INAOE, CINVESTAV, Mexico [9] ChaLearn, USA [10]Baidu Research, China
ajianliu92@gmail.com, zhaochenxu@mininglamp.com, zitong.yu@oulu.fi
jun.wan@ia.ac.cn, sergio@maia.ub.es

## Abstract

*The threat of 3D masks to face recognition systems is increasingly serious and has been widely concerned by researchers. To facilitate the study of the algorithms, a large-scale High-Fidelity Mask dataset, namely CASIA-SURF HiFiMask (briefly HiFiMask) has been collected. Specifically, it consists of a total amount of* 54,600 *videos which are recorded from* 75 *subjects with* 225 *realistic masks under* 7 *new kinds of sensors [21]. Based on this dataset and Protocol 3 which evaluates both the discrimination and generalization ability of the algorithm under the open set scenarios, we organized a 3D High-Fidelity Mask Face Presentation Attack Detection Challenge to boost the research of 3D mask-based attack detection. It attracted* 195 *teams for the development phase with a total of* 18 *teams qualifying for the final round. All the results were verified and re-run by the organizing team, and the results were used for the final ranking. This paper presents an overview of the challenge, including the introduction of the dataset used, the definition of the protocol, the calculation of the evaluation criteria, and the summary and publication of the competition results. Finally, we focus on introducing and analyzing the top ranking algorithms, the conclusion summary, and the research ideas for mask attack detection provided by this competition.*

## 1. Introduction

Recently, Face Anti-Spoofing (FAS) has been attracted more and more attention [45, 43, 30] due to the wide application of face recognition in financial payment, access control, and phone unlocking. Therefore, face Presentation Attack Detection (PAD) technology is a critical stage to reinforce the face recognition systems by determining whether the captured face from an imaging sensor is real or fake. With the release of several high-quality 2D attack datasets [2, 27, 53, 18], the previous algorithms [46, 47, 39, 49, 31, 42, 51, 24, 6, 52, 50, 23] show better performance against Print attacks and video Replay attacks, but they are more vulnerable to mask attacks with more realistic color and structure. However, with the maturity of 3D printing technology, face mask has become a new type of Presentation Attack (PA), which can easily fool the FAS system based on coarse texture and facial depth information.

Although some works have been devoted to 3D mask attacks, including datasets [29, 22, 13, 9, 48] collection and algorithms [54, 12, 33, 22, 15, 25, 16] design, there are still some limitations that hinder the performance of the algorithm: (1) Lack of a high-quality and large-scale mask dataset for algorithm research, due to the limitation of high fidelity mask production cost. As far as we know, the existing mask datasets are insufficient in the aspects of the number of subjects and skin tones, mask quality and types, scene settings, lighting environments, and collection devices, which seriously limits the research of data-driven algorithms. (2) Lack of a challenging and public benchmark for performance comparison of different algorithms. As a result, the existing algorithms only work for specific mask types or in constrained environments. Several rPPG-based methods [15, 22, 25, 16, 26, 44] are proposed according to the evidence that periodic rPPG pulse cues could be recovered from the live faces but noisy for the mask attacks. However, they are vulnerable to the interference of illumination change and sensitive to detection distance. (3) Compared with 2D attacks, such as Print-Attack, Replay-Attack, high fidelity mask has realistic skin color and structure. It is difficult to distinguish between a live face and a mask from the visible spectrum.

In order to promote the community's research on mask attack detection, we solve the current difficulties from the

*Corresponding author

following three aspects based on the above analysis: (1) We collect and release a large-scale 3D high-fidelity mask face PAD dataset named HiFiMask. Compared with public 3D mask datasets, it has several advantages, such as high fidelity masks and amount of data in the term of identities, lightings, sensors, and videos. (2) We define a more general and valuable testing protocol for real-world deployment and provide a decent result as a benchmark. Our protocol evaluates both the discrimination and generalization ability of the algorithm under the open set scenarios. In other words, the training and developing sets contain only parts of common mask types and scenarios while there are more general mask types and scenarios on the testing set. (3) Based on the dataset and protocol, we successfully held a competition, *3D High-Fidelity Mask Face Presentation Attack Detection Challenge at ICCV2021* [1], attracting 195 teams from all over the world. The results of the top three teams are far better than our baseline results, which greatly pushes the current best performance of mask attack detection. A summary with the names and affiliations of teams that entered the final stage is shown in Tab. 1. Interestingly, compared with the previous challenges [20, 17, 1], the majority of the final participants of this competition come from the industrial community, which indicates the increased importance of the topic for daily life applications.

To sum up, the contributions of this paper are summarized as follows: (1) We describe the design of the *3D High-Fidelity Mask Face Presentation Attack Detection Challenge at ICCV2021* challenge. (2) We organize this challenge around the HiFiMask datsaset, proving the suitability of such a resource for boosting research on the topic. (3) We report and analyze the solutions developed by participants. (4) We conclude the effective scheme of mask attack detection from the top-ranked algorithms and point out the research direction through this competition.

## 2. Challenge Overview

In this section, we review the organized challenge, including a brief introduction of the HiFiMask dataset, the challenge process and timeline, the challenge protocol, and evaluation metrics.

**HiFiMask Dataset.** HiFiMask [21] is currently the largest 3D face mask PAD dataset, which contains $54,600$ videos captured from 75 subjects of three skin tones, including 25 subjects in yellow, white, and black, respectively. For mask types, it contains 3 high-fidelity masks for each identity, which are made of transparent, plaster, and resin materials, respectively. During the acquisition process, it considers 6 complex scenes for video recording, *i.e.*, White Light, Green Light, Periodic Three-color Light, Outdoor Sunshine, Outdoor Shadow, and Motion Blur. For each scene, there are 6 videos under different lighting directions (*i.e.*, NormalLight, DimLight, BrightLight, BackLight, SideLight, and TopLight) to explore the impact of directional lighting. Among them, there is periodic lighting within [0.7, 4]Hz for the first three scenarios to mimic the human heartbeat pulse, thus might interfere with the rPPG-based mask detection technology [15]. Finally, 7 mainstream imaging devices (*i.e.*, iPhone11, iPhone X, MI10, P40, S20, Vivo, and HJIM) are utilized for video recording to ensure high resolution and imaging quality.

In order to facilitate the participating teams to use the dataset, we have carried out some data preprocessing steps. We remove irrelevant background areas from original videos, such as the part below the neck. After face detection, we sample 10 frames at equal intervals from each video. Finally, we name the folder of this video according to the following rule: $Skin\_Subject\_Type\_Scene\_Light\_Sensor$.

**Challenge Protocol and Data Statistics.** In order to increase the challenge of the competition and meet the actual deployment requirements, we consider a protocol that can comprehensively evaluate the performance of algorithm discrimination and generalization. In other words, the training and developing sets contain only parts of common mask

Table 1. Team and affiliations name are listed in the final ranking of this challenge.

| Ranking | Team Name | Leader Name, Affiliation |
|---|---|---|
| 1 | VisionLabs | Oleg Grinchuk, visionlabs.ai |
| 2 | WeOnlyLookOnce | Ke-Yue Zhang, Tencent Youtu Lab |
| 3 | CLFM | Samuel Huang, FaceMe |
| 4 | oldiron666 | Zezheng Wang, Kuaishou Technology |
| 5 | Reconova-AI-LAB | Mingmu Chen, Reconova Technology |
| 6 | inspire | Jiang Hao, Bytedance Ltd. |
| 7 | Piercing Eyes | Hyokong, National University of Singapore |
| 8 | msxf_cvas | Liang Gao, MaShang Consumer Finance Co.,Ltd |
| 9 | VIC_FACE | Cheng Zhen, Meituan |
| 10 | DXM-DI-AI-CV-TEAM | Weitai Hu, Du Xiaoman Financial |
| 11 | fscr | Artem Petrov, Peter the Great St. Petersburg Polytechnic University |
| 12 | VIPAI | Yao Xiao, Zhejiang University |
| 13 | reconova-ZJU | Zhishan Li, Zhejiang University |
| 14 | sama_cmb | Yifan Chen, Chinese Merchants Bank(CMB) |
| 15 | Super | Yu He, Technische Universität München, mytum |
| 16 | ReadFace | Zhijun Tong, ReadFace |
| 17 | LsyL6 | Dongxiao Li, Zhejiang University |
| 18 | HighC | Minzhe Huang, Akuvox (Xiamen) Networks Co., Ltd. |

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

Table 2. Statistical information for Challenge Protocol. '#' means the number of videos. Note that 1, 2 and 3 in the third column mean Transparent, Plaster and Resin mask, respectively. Other columns refer to 2 in a similar way.

| Subset | subj. | Mask | Scene | Light | Sensor | # live | # mask | # all |
|---|---|---|---|---|---|---|---|---|
| Train | 45 | 1,3 | 1,4,6 | 1,3,4,6 | 1,2,3,4 | 1,610 | 2,105 | 3,715 |
| Dev | 6 | 1,3 | 1,4,6 | 1,3,4,6 | 1,2,3,4 | 210 | 320 | 536 |
| Test | 24 | 1∼3 | 1∼6 | 1∼6 | 1∼7 | 4,335 | 13,027 | 17,362 |

types and scenarios while there are more general mask types and scenarios on the testing set. Based on Protocol 1 [21], we define training and development sets with parts of representative samples while a full testing set is used. Thus, the distribution of testing sets is more complicated than the training and development sets in terms of mask types, scenes, lighting, and imaging devices. Different from Protocol 2 [21] with only 'unseen' mask types, the challenge protocol considers both 'seen' and 'unseen' domains as well as mask types, which are more general and valuable for real-world deployment.

In the challenge protocol, as shown in Tab. 2, all skin tones, part of mask types, such as transparent and resin materials (short for 1, 3), part of scenes, such as White Light, Outdoor Sunshine, and Motion Blur (short for 1, 4, 6), part of lightings, such as NormalLight, BrightLight, BackLight, and TopLight (short for 1, 3, 4, 6), and part of imaging devices, such as iPhone11, iPhone X, MI10, P40 (short for 1, 2, 3, 4) are presented in the training and development subsets. While all skin tones, mask types, scenes, lightings, and imaging devices are presented in the testing subset. For clarity, the dataset partition and video quantity of each subset of the challenge protocol are shown in Tab. 2.

**Challenge Process and Timeline.** The challenge was run in the CodaLab[2] platform, and comprised two stages as follows:

**Development Phase:** (*Started: April. 19, 2021 - Ended: in June 10, 2021*). During this phase, participants had access to labeled training data and unlabeled development data. Participants could use training data to train their models, and they could submit predictions on the development data. Training data was made available with samples labeled with the genuine, 2 types of the mask (short for 1, 3), 3 types of scenes (short for 1, 4, 6), 4 kinds of lightings (short for 1, 2, 4, 6) and 4 imaging sensors (short for 1, 2, 3, 4). Although the development data maintains the same data type as the training data, the label is not provided to the participants. Instead, participants could submit predictions on the development data and receive immediate feedback via the leader board.

**Final phase:** (*Started: June 10, 2021 - Ended: June 20, 2021*). During this phase, labels for the development set were made available to participants, so that they can have more labeled data for training their models. The unlabeled testing set was also released, participants had to make predictions for the testing data and upload their solutions to the challenge platform. The test set was formed by examples labeled with the genuine, and all skin tones, mask types (short for 1∼3), scenes (short for 1∼6), lightings (short for 1∼6), and imaging devices (short for 1∼7). Participants had the opportunity to make 3 submissions for the final phase, this was done with the goal of assessing the stability of their methods. Note that the CodaLab platform defaults to the result of the last submission.

The final ranking of participants was obtained from the performance of submissions in the testing sets. To be eligible for prizes, winners had to publicly release their code under a license of their choice and provide a fact sheet describing their solution.

**Evaluation Metrics.** In this challenge, we selected the recently standardized ISO/IEC 30107-3[3] metrics: Attack Presentation Classification Error Rate (APCER), Normal Presentation Classification Error Rate (NPCER) and Average Classification Error Rate (ACER) as the evaluation metrics. The ACER on the testing set is determined by the Equal Error Rate (EER) thresholds on the development set. Finally, The value ACER was the leading evaluation measure for this challenge, and Area Under Curve (AUC) was used as additional evaluation criteria.

## 3. Description of solutions

**VisionLabs** Due to the tiny fake features of 3D face masks and the complexity to distinguish, team VisionLabs proposed a pipeline based on high-resolution face parts cropped from the original image, as shown in Fig. 1. Those parts are used as additional information to classify the full images through the network.

During preparation state, centered face crops are created using the Dual Shot Face Detector (DSFD) detector [14]. The crop bounding box is expanded $1.3\times$ times around face detection bounding box. If the bounding box is out of the original image border, missing parts are filled with black. If no face is found, the original image is used instead of crop. Then, five face regions are cropped using prior information from face bounding box, as eyes, nose, chin, left, and right ear. Each part will be resized to $224 \times 224$ and input into the backbone after data augmentation (*i.e.*, rotation, random crop, color jitter). Additionally, as a regularization technique, they turned 10% of images into trash images by scaling random tiny parts of images. As shown in Fig. 1, team VisionLabs used a multi-branch network, including Face, Eyes, Nose, Chin, and Ears branches. Since any pre-trained weight is prohibited in this competition, they tried to replicate the generalization ability of convolutional filters by us-

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

[3] https://www.iso.org/obp/ui/iso

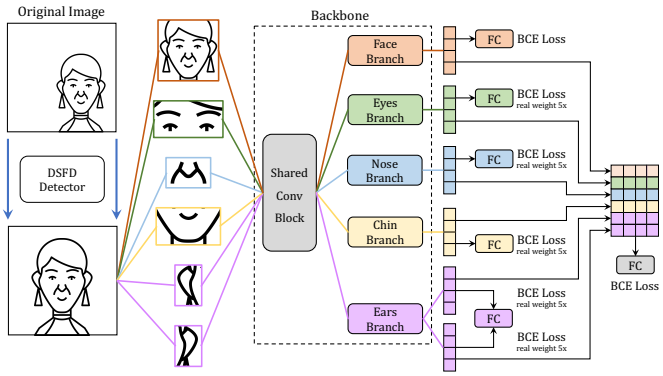


Figure 1. The pipeline of team VisionLabs. Original images are cropped by DSFD detector [14] and split into five regions by prior knowledge. Then, those parts are input into the backbone with one shared convolution block and five branches. Each branch will output a 320-dimensional vector (two vectors from the ears branch). All vectors are concatenated to one vector and calculated as $L_{total}$.

ing shared the first block of each branch. This made the first block filters learn more diverse features. Five branches all adopt EfficientNet-B0 [34] as the backbone, and the original descriptor size of each branch is reduced from 1280 to 320. Due to the presence of left and right ears, the Ears Branch outputs two vectors. Then, each loss and confidence of the current branch can be obtained through the fully connected layer, as $L_{face}$, $L_{eyes}$, $L_{nose}$, $L_{chin}$, $L_{Lear}$,$L_{Rear}$ ($Lear$ for left ear and $Rear$ for right ear). These six vectors are concatenated to obtain a 1920-dimensional vector, used to calculate the loss function $L_{total}$. All branches are trained simultaneously with the final loss:

$$L = 5 * L_{total} + 5 * L_{face} + L_{eyes} + L_{nose} + L_{chin} + 0.5 * L_{Lear} + 0.5 * L_{Rear} \quad (1)$$

where all losses are binary cross-entropy(BCE) losses. Since face parts do not always contain the tiny fake features, for eyes, nose, chin and ears, they increase positive class weight in BCE loss by a factor of 5. The partial face part descriptors will not be punished too hard if don't contain useful features.

They trained the model with Adam optimizer for 60 epochs using an initial learning rate of 0.0006 and decreasing it every 3 epochs by a factor of 0.9. During the inference phase, they chose 0.7 as the test set threshold when it is inaccurate to select a threshold from a validation set close to the full score. Based on average positions and prior information, some face parts of images may be cropped in the wrong way. So a test time augmentation is introduced. They flip an image and obtain the final results by averaging the scores of original and flipped faces.

**WeOnlyLookOnce** In this method, considering that there are irrelevant noises in the raw training data, a custom algorithm is used to detect black borders firstly. After

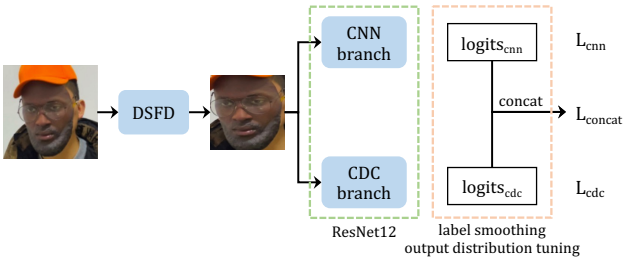


Figure 2. The framework of team WeOnlyLookOnce. The DSFD face detector is used to detect bounding box. Afterwards, a lightweight self-defined ResNet12 subsequently aims to classify the input into three categories. To be mentioned, Label smoothing and output distribution tuning are used as additional tricks.

that, the DSFD [14] is applied to detect potential faces for each image. To be specified, the training set is processed by wiping black borders merely, while the testing set and validation set are cropped with a ratio of 1.5 times bounding box further. What's more, the positive samples in the training set are much less than negative samples. The training augmentations include rotation, image crop, color jitter, etc.

As shown in Fig. 2, the framework [38, 3, 37, 41] consists of a CNN branch and a CDC branch. Both networks are self-designed lightweight ResNet12, and each of them is a three-class classification network aiming to detect real images and two kinds of mask. The realization of CNN branch is a vanilla convolution while the CDC branch used Central Difference Convolution [49]. To alleviate the overfitting problem, the team additionally adopted a label smoothing strategy and an output distribution tuning strategy inspired by temperature scaling [28]. After computing the cross-entropy loss by using the logits and label, the total loss is calculated by the following equation

$$L = L_{concat} + 0.5 * L_{cnn} + 0.5 * L_{cdc} \quad (2)$$

To minimize the distribution gap between validation and test sets, this team proposed an effective distribution tuner. They provide two strategies in this tuner both of which are proved to be effective. In the first strategy, they reform the three-class classification task into a binary classification task by adding the two attack-class logits into one uniformed value, then dividing the real logits by a factor of 3.6 and the fake logits by a factor of 5.0 before the softmax operation. In the second strategy, the task still remains a three-class classification problem while the real score on the validation set is subtracted by 0.07.

**CLFM** Team CLFM produced a model with only cross-entropy loss based on CDCN++ model but earn a good result. The central difference convolution was used to replace traditional convolution. Also, attention modules were introduced in each stage to make the model performed better.

Besides, they fuse three stages' output parameters as feature vectors before the fully connected layer.

For data pre-processing, they adopt their own face detection model to perform face detection and take patches of the face as input. Something should be the highlight that they play some brilliant and practical tricks both on train and test set. On the one hand, they find that there are hats/glasses that will most likely lead the model in the wrong direction. So they firstly crop the face according to the bounding box and then crop the region around the mouth. The face size is randomly set in a small range to ensure the generalization of the model. If the region is not enough to fill it, they tend to flip and mirror the region to keep the texture constant. The model input is square blocks resized to $56 \times 56$ and normalized with mean and standard deviation parameters which are summarized from ImageNet. On the other hand, they also notice that there are about 17% of the images in the test set that the face detection didn't detect any face, and in this kind of scenario the model will have no other choice but use the whole image as the bounding box. So they randomly make part of the training data's bounding box to be the whole image and make slight changes to the cropping process of the test set compared with the training set, so that at least the model won't be pure guessing when facing this situation. Finally, they use self voting by moving the patch in a small range and averaging them as the final score.

**Oldiron666** The team Oldiron666 proposed a self-dense regularization framework for face anti-spoofing. For data pre-processing, they expand an adaptive scale for the cropped face, which improves the performance. The input size of the images is 256, and the following data augmentations are performed to improve generalization, such as Random Crop, Cutout, and Patch Shuffle Augmentation, etc.

The team Oldiron666 used a representation learning framework similar to SimSiam [5], but introduces a multilayer perceptron(MLP) for supervised classification. During the training process, the face image $X$ will be randomly augmented to obtain two views for input, $X_1$ and $X_2$. The two views are processed by an encoder network $f$, which consists of the backbone and an MLP head called projector [4]. They found a more light network may bring better performance on the HiFiMask. Therefore, Resnet6 is utilized as the backbone, which contains a low computational complexity. The output of one view is transformed to match the other view by a dense predictor MLP head, denoted as $h$. The dense similarity loss, marked as $L_{contra}$, maximizes the similarity between both sides. To implement supervised learning, they perform the dense classifier $c$ at the end of the framework and use Mean Squared Error(MSE) to evaluate the output. MSE loss is calculated with the ground truth label on one side, denoted as $L_{cls}$, while $L_d$ is calculated as the difference between the category output of both two sides. The training loss can be defined as,

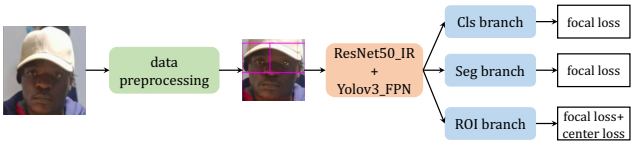


Figure 3. The application flow chart of team Reconova-AI-Lab. Raw images are firstly pre-processed by RetinaFace for face detection, crop and alignment in the upper flow. Then, they used a multi-task learning algorithm, which mainly includes three branches.

$$L = L_{contra} + L_{cls} + 0.1 * L_d \tag{3}$$

During the training process, they perform half-precision floating-point to obtain a faster training speed. The SGD optimizer is adopted, with an initial learning rate of 0.03, weight decay by 0.0005, and momentum by 0.9. During inference, only the $X_1$ side is executed to obtain the result of face anti-spoofing.

**Reconova-AI-Lab** Team Reconova-AI-LAB contributes a variety of models and generates many different results, the best of which is used for the competition. They proposed a multi-task learning algorithm, which mainly includes three branches, the direct classification branch, the real person learning Gaussian mask branch, and the Region of interest(ROI) classification branch. In the rest of this section, we take Cls, Seg, and ROI branches as abbreviations respectively. Cls branch takes a focal loss which combining Sigmoid and BCE Loss as the supervision information. It is annotated as $loss\_classi$. Seg branch adopts the same loss function as Cls and its loss annotation is $loss\_seg$. Concerning with ROI branch, it take three loss functions, which is $loss\_cls1$, $loss\_cls2$ and $loss\_center$, respectively. The effect of the first one is focal loss mentioned before. The second one aims to the alignment of ROI which is used to calibrate the operation of ROI pooling. Subsequently, the purpose of the last one is to reduce the distance between classes. Finally, the lost function of ROI branch $loss\_roi$ equals $loss\_cls1$ plus $loss\_center* \times 0.01$ plus $loss\_cls2$. All branches are trained synchronously with an SGD optimizer in 800 epochs, and the total loss function formulates as follows:

$$total\_loss = loss\_classi + loss\_seg + loss\_roi \tag{4}$$

Their application flow chart is shown in Fig. 3. First, the data pre-processing includes the use of RetinaFace to detect the face and generate 14 landmarks per face, including face coordinates and bounding boxes of left, right ear, and mouth. At that stage, they use some strategies to avoid large-angle posture and non-existence of face by constraining the size of the bounding box of ROI. Meanwhile, they

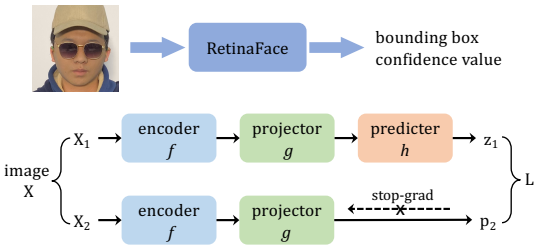


Figure 4. The framework of team inspire. Raw images are firstly processed in the upper flow. After that, this team used a Context Contrastive Learning framework to train, while the backbone is a SE-ResNeXt101 network.

take mirroring, random rotation, random color enhancement, random translation, and random scaling as treatments of data enhancement. Then they adopt a backbone called Res50_IR, which has stacked 3, 4, 14, and 3 blocks respectively in four stages. In order to enhance features, an improved Residual Bottleneck structure named Yolov3_FPN is connected to the different stages of the network. The slightly complicated network is followed by three branches mentioned before. All of the parameters are initialized by different methods according to different layers.

**inspire** The team firstly utilized a ResNet50 [11] based RetinaFace [7] to detect face bounding boxes for all images. To be noticed, three different threshold values of 0.8, 0.1, and 0.01 are used to record the different types of bounding boxes. If the detecting confidence value is above 0.1, the box label is set to be 2. If it is between 0.1 and 0.01, the box label is 1. While the value is less than 0.01, the box label remains to be 0. According to the box label depicted above, hard samples of the cropped images is partitive.

For the training stage, SE-ResNeXt101 [40] was selected as the backbone. Besides, the team applied the Context Contrastive Learning(CCL) [21] architecture as the framework, which is shown in Fig. 4. As a result, they used a sampling strategy the same as that in [21]. The MSE loss $L_{MSE}$, Cross Entropy loss $L_{CE}$ and Contrastive loss [10] $L_{Contra}$ are applied to calculate total loss by the following weights:

$$L = L_{MSE} + L_{CE} + 0.7 * L_{Contra} \quad (5)$$

Afterward, Ranger optimizer[4] is set as a learning strategy with an initial learning rate of 0.001. The total epoch is 70, and the learning rate decays by 0.1 at 20, 30, 60 epochs, respectively.

**Piercing Eye** The team Piercing Eye used the modified CDCN [49] as the basic framework, shown in Fig. 5. During data processing, the face regions are detected from the original images, which are resized to $256 \times 256$ and randomly cropped to $228 \times 228$. Same as other teams, some types of data augmentation like color jitter are used.

[4] https://github.com/lessw2020/Ranger-Deep-Learning-Optimizer

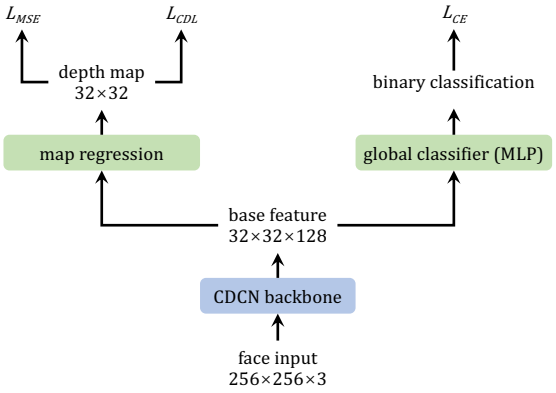


Figure 5. The framework of team Piercing Eye. Two branches are attached to the CDCN backbone, called map regression and global classifier, respectively.

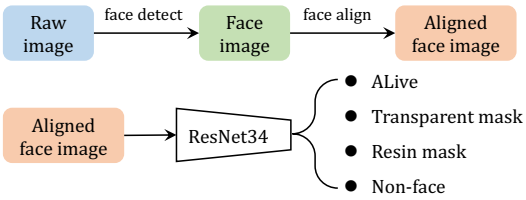


Figure 6. The framework of team msxf_cvas. Raw images are firstly processed by detecting face and alignment. After that, a ResNet34 network is utilized to classify the input image into four types including live, transparent mask, resin mask and no face.

In addition to the original output (depth map) of CDCN, a multi-layer perceptron (MLP) is attached to the backbone, implementing the global binary classification. The shape of the depth map is $32 \times 32$. The label of the real face region is set to 1, while the background and fake face region is set to 0. They trained the model with an SGD optimizer for 260 epochs using an initial learning rate of 0.002 and decreasing it by a factor of 0.5 with milestones. As in [49], both mean square error loss $L_{MSE}$ and contrastive depth loss $L_{CDL}$ are utilized for pixel-wise supervision. They also perform cross-entropy loss in a global branch, denoted as $L_{CE}$. So the overall loss function is formulated as

$$L = 0.5 * L_{MSE} + 0.5 * L_{CDL} + 0.8 * L_{CE} \quad (6)$$

**msxf_cvas** From the analysis of competition data, the team finds two different distributions of spoof masks which are transparent material and fidelity material. They consider two materials (plaster and resin) as one category as the features of these two types looks similar. Besides, there are small amounts of noisy data without a human face which do not contain spoof or live features. Therefore, the team try to classify them as one category called non-face. The fi-

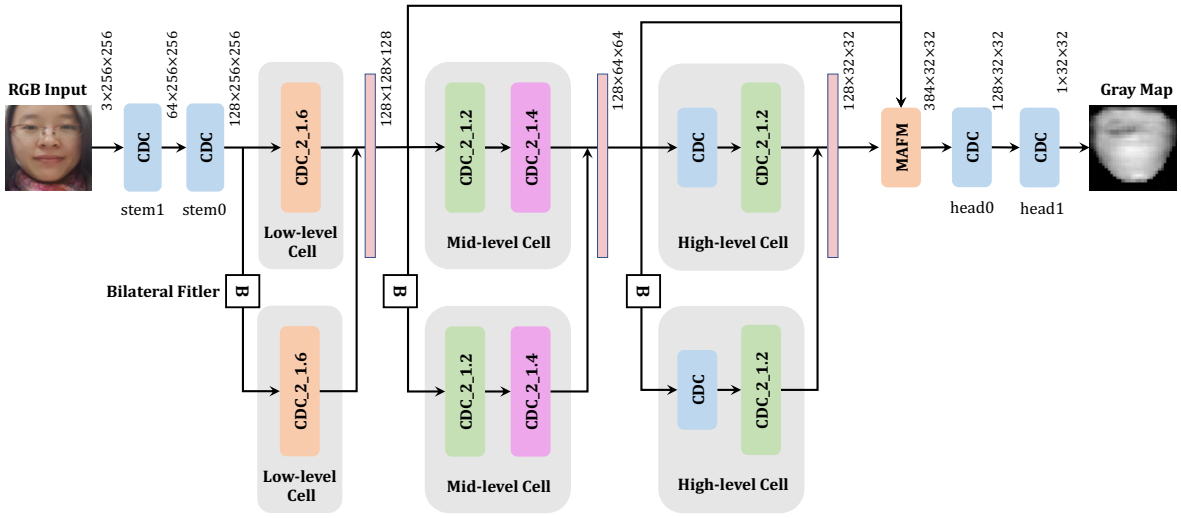


Figure 7. The framework of team VIC_FACE.

nal task is to classify all data into four categories which are the live, transparent mask, resin mask, and non-face. Considering that there are many extreme posture and light and low-quality data in the competition data, they focus more on data augmentation strategies during training including cutMix, ISONoise, randomSunFlare, randomFog, motionBlur, and imageCompression.

First of all, the team applied a face detector to detect faces and align faces by five points. After that, the mmclassification[5] project was used to train a face anti-spoofing model. To begin with, the team chose a ResNet34 [11] as the backbone and the cross-entropy loss was selected as the loss function. The whole framework is illustrated in Fig. 6.

**VIC_FACE** The prerequisites need to know is that deep bilateral has been successfully applied in convolutional networks to filter the deep features instead of original images. Inspired by this, team VIC_FACE proposed a novel method based on fusing the deep bilateral operator on the basis of original CDCN in order to learn more intrinsic features via aggregating multilevel bilateral macro- and micro- information. As shown in Fig. 7, the backbone model is an initial CDCN, which divides the backbone into multilevel (low-level, mid-level, and high-level) blocks to predict the gray-scale facial depth map with size $1 \times 32 \times 32$ from a single RGB facial image with size $3 \times 256 \times 256$. Besides, the DBO as a channel-wise deep bilateral filtering mimics a residual layer embedded in the network and replaces the original convolution layer by representing the aggregated bilateral base and residual features.

Specifically, at the first stage, they detect and crop the face area in the full image as input of the model. Secondly, they edit the images randomly with down-sampling and

[5] https://github.com/open-mmlab/mmclassification

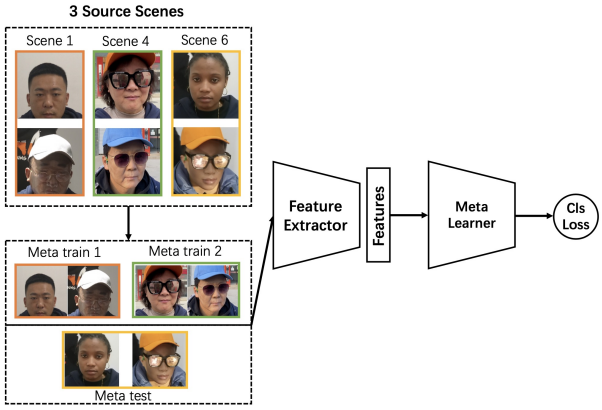


Figure 8. The framework of DXM-DI-AI-CV-TEAM.

jpeg compression, which often occur unintentionally when the images are captured from different devices. Moreover, it is worth mentioning that excepting some regular data augmentation methods including cutout, color jitter and erase to improve generalization of the model, affine transformation of brightness and color of random area based on OpenCV is applied to simulate light condition in training data. Finally, they design a contrastive loss function for controlling the contrast depth map of gray-scale output and a mean-square error loss function for reducing the difference between augmented input and binary mask, then combining them into one loss function with Adam optimizer.

**DXM-DI-AI-CV-TEAM** Due to the generalization performance of the challenge evaluation algorithm for unknown attack scenarios, this team casts faces anti-spoofing as a domain generalization (DG) problem. To let the model generalize well to unseen scenes, inspired by [32], the proposed framework trains their model to perform well in the simulated domain shift scenarios, which is achieved by find-

ing generalized learning directions in the meta-learning process. Different from [32], the team removed the branch of depth prior knowledge from the real face and mask, which contained similar depth information. Besides, a series of data augmentation and training strategies are used to achieve the best results.

In the challenge, the training data are collected in 3 scenes, namely White Light, Outdoor Sunshine, and Motion Blur (short for 1, 4, 6). Therefore, the objective of DG for this challenge is to make the model trained on the 3 source scenes can generalize well to unseen attacks from the target scene. To this end, as shown in Fig. 8, the framework in this team that composes of a feature extractor and a meta learner. At each training iteration, they divide the original 3 source scenes by randomly selecting 2 scenes as meta-train scenes and the remaining one as the meta-test scene. In each meta-train and meta-test scene, meta learner conducts the meta-learning in the feature space supervised by the image and label pairs denoted as $x$ and $y$, where $y$ are ground truth with binary class labels ($y = 0/1$ is the label of fake/real face). In this way, their model can learn how to perform well in the scene shift scenarios through many training iterations and thus learn to generalize well to unseen attacks.

# 4. Challenge Results

## 4.1. Challenge Results Report

We adopted four metrics to evaluate the performance of the solutions, which are APCER, NPCER, ACER, and AUC respectively. Please note that although we report performance for a variety of evaluation measures, the leading metric was ACER. See from the Tab. 3, which lists the results and ranking of the top 18 teams, we can draw three conclusions: (1) The ACER performance of the top 3 teams is relatively close, and the top 2 teams have the best results in all metrics. (2) The top 6 teams are from industry, which indicates that mask attack detection is no longer limited to academia, but also an urgent problem in practical application. (3) The ACER performance of all teams is evenly distributed between $3\%$ and $10\%$, which not only shows the rationality and selectivity of our challenge but also demonstrates the value of HiFiMask for further research.

## 4.2. Competition summary and Future Work

Through the challenge, we summarize the effective ideas for mask attack detection: (1) At the data level, data augmentation plays an important role in preventing the overfitting of the model and improving the stability of the algorithm. (2) The segmentation of the face region can not only enlarge the local information but also avoid the extraction of irrelevant features. (3) Multi-branch-based feature learning is a framework widely used by participating teams.

Table 3. Team and results are listed in the final ranking of this challenge.

| R. | Team Name | FP | FN | APCER | BPCER | ACER | AUC |
|---|---|---|---|---|---|---|---|
| 1 | VisionLabs | 492 | **101** | 3.777 | **2.330** | **3.053** | **0.995** |
| 2 | WeOnlyLookOnce | **242** | 193 | **1.858** | 4.452 | 3.155 | **0.995** |
| 3 | CLFM | 483 | 118 | 3.708 | 2.722 | 3.215 | 0.994 |
| 4 | oldiron666 | 644 | 115 | 4.944 | 2.653 | 3.798 | 0.992 |
| 5 | Reconova-AI-LAB | 277 | 276 | 2.126 | 6.367 | 4.247 | 0.991 |
| 6 | inspire | 760 | 176 | 5.834 | 4.060 | 4.947 | 0.986 |
| 7 | Piercing Eyes | 887 | 143 | 6.809 | 3.299 | 5.054 | 0.983 |
| 8 | msxf_cvas | 752 | 232 | 5.773 | 5.352 | 5.562 | 0.982 |
| 9 | VIC_FACE | 1152 | 104 | 8.843 | 2.399 | 5.621 | 0.965 |
| 10 | DXM-DI-AI-CV-TEAM | 1100 | 181 | 8.444 | 4.175 | 6.310 | 0.970 |
| 11 | fscr | 794 | 326 | 6.095 | 7.520 | 6.808 | 0.979 |
| 12 | VIPAI | 1038 | 268 | 7.968 | 6.182 | 7.075 | 0.976 |
| 13 | reconova-ZJU | 1330 | 183 | 10.210 | 4.221 | 7.216 | 0.974 |
| 14 | sama_cmb | 1549 | 188 | 11.891 | 4.337 | 8.114 | 0.969 |
| 15 | Super | 780 | 454 | 5.988 | 10.473 | 8.230 | 0.979 |
| 16 | ReadFace | 1556 | 202 | 11.944 | 4.660 | 8.302 | 0.965 |
| 17 | LsyL6 | 2031 | 138 | 15.591 | 3.183 | 9.387 | 0.951 |
| 18 | HighC | 1656 | 340 | 12.712 | 7.843 | 10.278 | 0.966 |

In the following work, we further improve the performance from the following aspects: (1) We will use additional or generate multi-modal data [19, 35] to assist mask attack detection. (2) Besides CNN, we will explore the effectiveness of recent vision transformer [8] and MLP-like [36] architectures. (3) As the HiFiMask contains challenging dynamic lighting and scenes, we will explore more reliable rPPG [26, 44] technology.

# 5. Conclusion

We organized the *3D High-Fidelity Mask Face Presentation Attack Detection Challenge at ICCV2021* based on the HiFiMask dataset and running on the CodaLab platform. 195 teams registered for the competition and 18 teams made it to the final stage. Among the latter, teams were formed by 12 companies and 6 academic institutes/universities. We first described the associated dataset, the challenge protocol, and the evaluation metrics. Then, we reviewed the top-ranked solutions and reported the results from the final phases. Finally, we summarized the relevant conclusions, and pointed out the effective methods against mask attacks explored by this challenge.

# 6. Acknowledgement

This work was supported by the Chinese National Natural Science Foundation Projects #61961160704, #61876179, the External cooperation key project of Chinese Academy Sciences # 173211KYSB20200002, the Key Project of the General Logistics Department Grant No.AWS17J001, Science and Technology Development Fund of Macau (No. 0010/2019/AFJ, 0008/2019/A1, 0025/2019/AKP, 0019/2018/ASC), by the Spanish project PID2019-105093GB-I00 (MINECO/FEDER, UE), and by ICREA under the ICREA Academia programme.

## References


[1] Zinelabdine Boulkenafet, Jukka Komulainen, Zahid Akhtar, Azeddine Benlamoudi, Djamel Samai, Salah Eddine Bekhouche, Abdelkrim Ouafi, Fadi Dornaika, Abdelmalik Taleb-Ahmed, Le Qin, et al. A competition on generalized software-based face presentation attack detection in mobile scenarios. In *IJCB*. IEEE, 2017. 2

[2] Zinelabinde Boulkenafet, Jukka Komulainen, Lei Li, Xiaoyi Feng, and Abdenour Hadid. Oulu-npu: A mobile face presentation attack database with real-world variations. In *FGR*, pages 612–618, 2017. 1

[3] Shen Chen, Taiping Yao, Yang Chen, Shouhong Ding, Jilin Li, and Rongrong Ji. Local relation learning for face forgery detection. 2021. 4

[4] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In Hal Daumé III and Aarti Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 1597–1607. PMLR, 13–18 Jul 2020. 5

[5] Xinlei Chen and Kaiming He. Exploring simple siamese representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 15750–15758, June 2021. 5

[6] 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. In *AAAI*, 2021. 1

[7] Jiankang Deng, Jia Guo, Yuxiang Zhou, Jinke Yu, Irene Kotsia, and Stefanos Zafeiriou. Retinaface: Single-stage dense face localisation in the wild. *arXiv preprint arXiv:1905.00641*, 2019. 6

[8] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale, 2021. 8

[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. 1

[10] Raia Hadsell, Sumit Chopra, and Yann LeCun. Dimensionality reduction by learning an invariant mapping. In *Proceedings of the 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition-Volume 2*, pages 1735–1742, 2006. 6

[11] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 6, 7

[12] Neslihan Kose and Jean-Luc Dugelay. Mask spoofing in face recognition and countermeasures. *Image and Vision Computing*, 2014. 1

[13] Haoliang Li, Wen Li, Hong Cao, Shiqi Wang, Feiyue Huang, and Alex C Kot. Unsupervised domain adaptation for face anti-spoofing. *TIFS*, 2018. 1

[14] 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 *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5055–5064, 2019. 3, 4

[15] Xiaobai Li, Jukka Komulainen, Guoying Zhao, Pong-Chi Yuen, and Matti Pietikäinen. Generalized face anti-spoofing by detecting pulse from face videos. In *ICPR*, 2016. 1, 2

[16] Bofan Lin, Xiaobai Li, Zitong Yu, and Guoying Zhao. Face liveness detection by rppg features and contextual patch-based cnn. In *ICBEA*. ACM, 2019. 1

[17] 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*, 2020. 2

[18] 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 *WACV*, 2021. 1

[19] 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 TIFS*, 2021. 8

[20] A. Liu, J. Wan, S. Escalera, H. J. Escalante, and S. Z. Li. Multi-modal face anti-spoofing attack detection challenge at cvpr2019. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2019. 2

[21] 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, 3, 6

[22] Siqi Liu, Pong C Yuen, Shengping Zhang, and Guoying Zhao. 3d mask face anti-spoofing with remote photoplethysmography. In *ECCV*. Springer, 2016. 1

[23] Shubao Liu, Ke-Yue Zhang, Taiping Yao, Mingwei Bi, Shouhong Ding, Jilin Li, Feiyue Huang, and Lizhuang Ma. Adaptive normalized representation learning for generalizable face anti-spoofing. In *ACM MM*, 2021. 1

[24] 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. In *IJCAI*, 2021. 1

[25] Si-Qi Liu, Xiangyuan Lan, and Pong C Yuen. Remote photoplethysmography correspondence feature for 3d mask face presentation attack detection. In *ECCV*, 2018. 1

[26] Si-Qi Liu, Xiangyuan Lan, and Pong C Yuen. Multi-channel remote photoplethysmography correspondence feature for 3d mask face presentation attack detection. *TIFS*, 2021. 1, 8

[27] Yaojie Liu, Amin Jourabloo, and Xiaoming Liu. Learning deep models for face anti-spoofing: Binary or auxiliary supervision. In *CVPR*, 2018. 1

[28] Rafael Müller, Simon Kornblith, and Geoffrey E Hinton. When does label smoothing help? In *Advances in Neural Information Processing Systems*, volume 32, 2019. 4

[29] Erdogmus Nesli and Sébastien Marcel. Spoofing in 2d face recognition with 3d masks and anti-spoofing with kinect. In *BTAS*, 2013. 1

[30] Yunxiao Qin, Zitong Yu, Longbin Yan, Zezheng Wang, Chenxu Zhao, and Zhen Lei. Meta-teacher for face anti-spoofing. *IEEE TPAMI*, 2021. 1

[31] 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. 1

[32] Rui Shao, Xiangyuan Lan, and Pong C. Yuen. Regularized fine-grained meta face anti-spoofing. In *Thirty-Fourth AAAI Conference on Artificial Intelligence (AAAI)*, 2020. 7, 8

[33] Holger Steiner, Andreas Kolb, and Norbert Jung. Reliable face anti-spoofing using multispectral swir imaging. In *ICB*. IEEE, 2016. 1

[34] Mingxing Tan and Quoc Le. EfficientNet: Rethinking model scaling for convolutional neural networks. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 6105–6114. PMLR, 09–15 Jun 2019. 4

[35] Hao Tang and Nicu Sebe. Total generate: Cycle in cycle generative adversarial networks for generating human faces, hands, bodies, and natural scenes. *IEEE TMM*, 2021. 8

[36] Ilya Tolstikhin, Neil Houlsby, Alexander Kolesnikov, Lucas Beyer, Xiaohua Zhai, Thomas Unterthiner, Jessica Yung, Daniel Keysers, Jakob Uszkoreit, Mario Lucic, et al. Mlp-mixer: An all-mlp architecture for vision. *arXiv preprint arXiv:2105.01601*, 2021. 8

[37] Wenxuan Wang, Bangjie Yin, Taiping Yao, Li Zhang, Yanwei Fu, Shouhong Ding, Jilin Li, Feiyue Huang, and Xiangyang Xue. Delving into data: Effectively substitute training for black-box attack. In *CVPR*, pages 4761–4770, 2021. 4

[38] Xinyao Wang, Taiping Yao, Shouhong Ding, and Lizhuang Ma. Face manipulation detection via auxiliary supervision. In *International Conference on Neural Information Processing*, pages 313–324. Springer, 2020. 4

[39] 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*, 2020. 1

[40] Saining Xie, Ross Girshick, Piotr Dollár, Zhuowen Tu, and Kaiming He. Aggregated residual transformations for deep neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1492–1500, 2017. 6

[41] Bangjie Yin, Wenxuan Wang, Taiping Yao, Junfeng Guo, Zelun Kong, Shouhong Ding, Jilin Li, and Cong Liu. Adv-makeup: A new imperceptible and transferable attack on face recognition. 2021. 4

[42] 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. 1

[43] Zitong Yu, Xiaobai Li, Jingang Shi, Zhaoqiang Xia, and Guoying Zhao. Revisiting pixel-wise supervision for face anti-spoofing. *IEEE TBIOM*, 2021. 1

[44] Zitong Yu, Xiaobai Li, Pichao Wang, and Guoying Zhao. Transrppg: Remote photoplethysmography transformer for 3d mask face presentation attack detection. *IEEE Signal Processing Letters*, 2021. 1, 8

[45] 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. 1

[46] Zitong Yu, Yunxiao Qin, Xiaqing Xu, Chenxu Zhao, Zezheng Wang, Zhen Lei, and Guoying Zhao. Auto-fas: Searching lightweight networks for face anti-spoofing. In *ICASSP*, pages 996–1000. IEEE, 2020. 1

[47] Zitong Yu, Yunxiao Qin, Hengshuang Zhao, Xiaobai Li, and Guoying Zhao. Dual-cross central difference network for face anti-spoofing. In *IJCAI*, 2021. 1

[48] Zitong Yu, Jun Wan, Yunxiao Qin, Xiaobai Li, Stan Z. Li, and Guoying Zhao. Nas-fas: Static-dynamic central difference network search for face anti-spoofing. In *TPAMI*, 2020. 1

[49] 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*, 2020. 1, 4, 6

[50] 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. 1

[51] 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 *IJCB*, pages 1–6, 2021. 1

[52] 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. In *ECCV*, 2020. 1

[53] Shifeng Zhang, Xiaobo Wang, Ajian Liu, Chenxu Zhao, Jun Wan, Sergio Escalera, Hailin Shi, Zezheng Wang, and Stan Z Li. A dataset and benchmark for large-scale multi-modal face anti-spoofing. In *CVPR*, 2019. 1

[54] Zhiwei Zhang, Dong Yi, Zhen Lei, and Stan Z Li. Face liveness detection by learning multispectral reflectance distributions. In *FG*. IEEE, 2011. 1