Deep Spatial Gradient and Temporal Depth Learning for Face Anti-spoofing

Zezheng Wang[1] Zitong Yu[2] Chenxu Zhao[3,*] Xiangyu Zhu[4] Yunxiao Qin[5] Qiusheng Zhou[6]
Feng Zhou[1] Zhen Lei[4]

[1]AIBEE [2]CMVS, University of Oulu [3]Academy of Sciences, Mininglamp Technology
[4]CBSR&NLPR, CASIA [5]Northwestern Polytechnical University [6]JD Digits

{zezhengwang, fzhou}@aibee.com zitong.yu@oulu.fi zhaochenxu@mininglamp.com
{xiangyu.zhu, zlei}@nlpr.ia.ac.cn qyxqyx@mail.nwpu.edu.cn zhouqiusheng3@jd.com

Abstract

Face anti-spoofing is critical to the security of face recognition systems. Depth supervised learning has been proven as one of the most effective methods for face anti-spoofing. Despite the great success, most previous works still formulate the problem as a single-frame multi-task one by simply augmenting the loss with depth, while neglecting the detailed fine-grained information and the interplay between facial depths and moving patterns. In contrast, we design a new approach to detect presentation attacks from multiple frames based on two insights: 1) detailed discriminative clues (e.g., spatial gradient magnitude) between living and spoofing face may be discarded through stacked vanilla convolutions, and 2) the dynamics of 3D moving faces provide important clues in detecting the spoofing faces. The proposed method is able to capture discriminative details via Residual Spatial Gradient Block (RSGB) and encode spatio-temporal information from Spatio-Temporal Propagation Module (STPM) efficiently. Moreover, a novel Contrastive Depth Loss is presented for more accurate depth supervision. To assess the efficacy of our method, we also collect a Double-modal Anti-spoofing Dataset (DMAD) which provides actual depth for each sample. The experiments demonstrate that the proposed approach achieves state-of-the-art results on five benchmark datasets including OULU-NPU, SiW, CASIA-MFSD, Replay-Attack, and the new DMAD. Codes will be available at `https://github.com/clks-wzz/FAS-SGTD`.

1. Introduction

Face recognition technology has become the most indispensable component in many interactive AI systems for their convenience and human-level accuracy. However, most of existing face recognition systems are easily to be spoofed through presentation attacks (PAs) ranging from printing a face on paper (print attack) to replaying a face on a digital device (replay attack) or bringing a 3D-mask (3D-mask attack). Therefore, not only the research community but also the industry has recognized face anti-spoofing [18, 19, 4, 33, 39, 11, 23, 55, 1, 29, 12, 49, 45, 54, 21] as a critical role in securing the face recognition system.

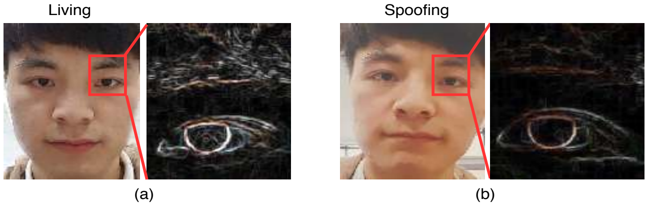

Figure 1. Spatial gradient magnitude difference between living (a) and spoofing (b) face. Notice that the large difference in gradient maps despite their similarities in the original RGB images.

In the past few years, both traditional methods [14, 42, 9] and CNN-based methods [35, 38, 20, 24, 46] have shown effectiveness in discriminating between the living and spoofing face. They often formalize face anti-spoofing as a binary classification between spoofing and living images. However, these approaches are challenging to explore the nature of spoofing patterns, such as the loss of skin details, color distortion, moiré pattern, and spoofing artifacts.

In order to overcome this issue, many auxiliary depth supervised face anti-spoofing methods have been developed. Intuitively, the images of living faces contain face-like depth, whereas the images of spoofing faces in print and by replaying carriers only have planar depth. Thus, Atoum *et al*. [2] and Liu *et al*. [34] propose single-frame depth supervised CNN architectures, and improve the presentation attack detection (PAD) accuracy.

By surveying the past face anti-spoofing methods, we

[*] denotes the corresponding author.

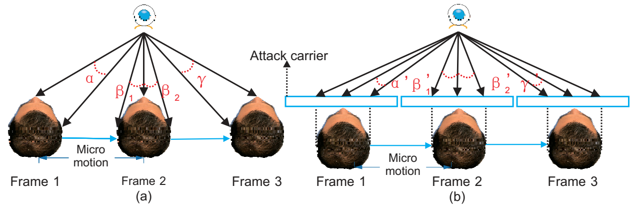

Figure 2. Temporal depth difference between live and spoof (print attack here) scenes. The change in camera viewpoint can result in facial motion among different keypoints. In the living scene (a), the angle α between nose and right ear is getting smaller, while the angle β_1 between left ear and nose is getting larger. However, in the spoofing scene (b), the observation could be different $\alpha' < \beta_2'$, and $\beta_1' > \gamma'$.

notice there are two problems that have not yet been fully solved: **1)** Traditional methods usually design local descriptors for solving PAD while modern deep learning methods can learn to extract relatively high-level semantic features instead. Despite their effectiveness, we argue that low-level fine-grained patterns can also play a vital role in distinguishing living and spoofing faces, e.g. the spatial gradient magnitude shown in Fig. 1. So how to aggregate local fine-grained information into convolutional networks is still unexplored for face anti-spoofing task. **2)** Recent depth supervised face anti-spoofing methods [2, 34] estimate facial depth based on a single frame and leverage depth as dense pixel-wise supervision in a direct manner. We argue that the *virtual* discrimination of depth between living and spoofing faces can be explored more adequately by multiple frames. A vivid and exaggerated example with assumed micro motion is illustrated in Fig. 2.

To address the problems, we present a novel depth supervised spatio-temporal network with Residual Spatial Gradient Block (RSGB) and Spatio-Temporal Propagation Module (STPM). Inspired by ResNet [22] , our RSGB aggregates learnable convolutional features with spatial gradient magnitude via shortcut connection. As a result, both local fine-grained patterns and traditional convolution features can be captured via stacked RSGB. To better utilize the information from multiple frames, STPM is designed for propagating short-term and long-term spatio-temporal features into depth reconstruction. To supervise the models with facial depth more effectively, we propose a Contrastive Depth Loss (CDL) to learn the topography of facial points.

We believe that the accuracy of facial depth directly affects the establishment of the relationship between temporal motion and facial depth. So we collect a double-modal anti-spoofing dataset named Double-modal Anti-spoofing Dataset (DMAD) which provides actual depth map for each sample. Extensive experiments are conducted to show that actual depth is more appropriate for monocular PAD than the generated depth. Note that this paper mainly focuses on the planar attack, which is the most common in practice.

We summarize the main contributions below.

- We propose a novel depth supervised architecture to capture discriminative details via Residual Spatial Gradient Block (RSGB) and encode spatio-temporal information from Spatio-Temporal Propagation Module (STPM) efficiently from monocular frame sequences.
- We develop a Contrastive Depth Loss to learn the topography of facial points for depth supervised PAD.
- We collect a double-modal dataset to verify that the actual depth is more appropriate for monocular PAD than the generated depth. This indicates an insight that collecting corresponding depth image to the RGB image brings benefit to the progress of the monocular PAD.
- We demonstrate the state-of-the-art performance by our method on widely used face anti-spoofing benchmarks.

2. Related Work

Roughly speaking, previous face anti-spoofing works generally fall into three categories: binary supervised, depth supervised, and temporal-based methods.

Binary supervised Methods Since face anti-spoofing is essentially a binary classification problem, most of previous anti-spoofing methods train a classifier under binary supervision, e.g., spoofing face as 0 and living face as 1. The early works usually rely on hand-crafted features, such as LBP [14, 15, 37], SIFT [42], SURF [9], HoG [28, 52], DoG [43, 48], and traditional classifiers, such as SVM and Random Forests. Because of the sensitiveness of manually-engineered features, traditional methods often generalize poorly across varied conditions such as camera devices, lighting conditions and presentation attack instruments (PAIs). Recently, CNN has emerged as a powerful tool in face anti-spoofing tasks with the help of hardware advancement and data abundance. For instance, in early works like [30, 41], pre-trained VGG-face model is fine-tuned to extract features in a binary-classification setting. However, most of them consider face anti-spoofing as a binary classification problem with cross-entropy loss, which easily learns the arbitrary patterns such as screen bezel.

Depth supervised Methods Compared with the binary setting, depth supervised methods aim to learn more faithful patterns. In [2], the depth map of a face is utilized as a supervisory signal for the first time. They propose a two-stream CNN-based approach for face anti-spoofing, by extracting both the patch features and holistic depth maps from the face images. It shows that depth estimation is beneficial for modeling face anti-spoofing to obtain promising results, especially on higher-resolution images. In another work [34], the authors propose a face anti-spoofing method by augmenting spatial facial depth as an auxiliary supervi-

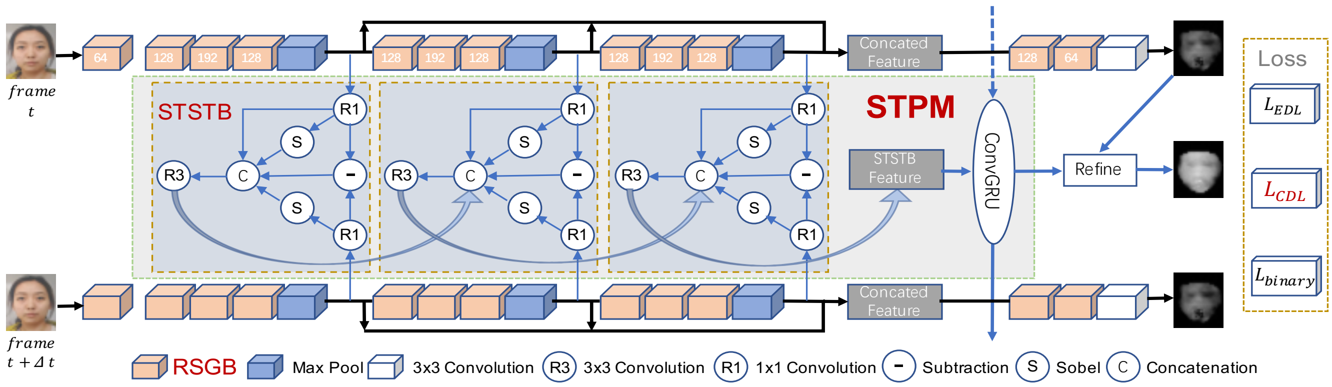

Figure 3. Illustration of the overall framework. The inputs are consecutive frames with a fixed interval. Each frame is processed by cascaded RSGB with a shared backbone which generates a corresponding coarse depth map. The number in RSGB cubes denotes the output channel number of RSGB. STPM is plugged between frames for estimating the temporal depth, which is used for refining the corresponding coarse depth map. The framework works well by learning with the overall loss functions.

sion along with temporal rPPG signals. More recently, [26] attempts to learn spoof noise and depth for generalized face anti-spoofing. However, these methods take stacked vanilla convolutional networks as the backbone and fail to capture the rich detailed patterns for depth estimation.

Temporal-based Methods Temporal information plays a vital role in face anti-spoofing tasks. Most of the prior works focus on the movement of key parts of the face. For example in [40, 41], the eye-blinking fact is used to predict spoofing. However, these methods are vulnerable to replay attacks since they heavily rely on some heuristic assumptions about the nature of these attacks. More general approaches like 3D convolution [20] or LSTM [50, 53] have recently been used to distinguish the live from spoof images. In addition, optical flow magnitude map and Shearlet feature have been taken as inputs in [16] to the CNN due to the obvious difference in flow patterns between living and spoofing faces. Based on the different color changes between the living and spoofing face videos, rPPG [31, 34, 32] features are also explored for PAD. To the best of our knowledge, no depth supervised temporal-based methods has ever been proposed for face anti-spoofing task.

3. The Proposed Approach

In this section, we first present our advanced depth-supervised spatio-temporal network structure, including Residual Spatial Gradient Block (RSGB) and Spatio-Temporal Propagation Module (STPM). Then our proposed novel Contrastive Depth Loss (CDL) and the overall loss would be demonstrated.

3.1. Network Structure

Designed in an end-to-end depth supervised fashion, our proposed framework takes N_f-frame face images as input and predicts the corresponding depth map directly. As shown in Fig. 3, the backbone is composed of cascaded RSGB followed by pooling layers, intending to extract fine-grained spatial features in low-level, mid-level and high-level, respectively. Then these multi-level features are concatenated to predict coarse depth map for each frame.

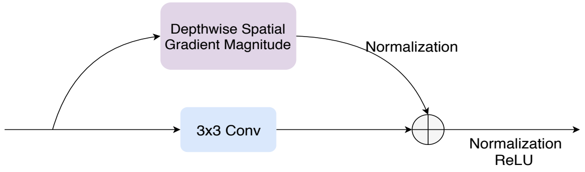

Figure 4. Residual spatial gradient block.

In order to capture rich dynamic information, STPM is plugged between frames. Short-term Spatio-Temporal Block (STSTB) picks up spatio-temporal features from adjacent frames while ConvGRU propagates these short-term features in a multi-frame long-term view. Finally, the temporal depth maps estimated from STPM are used to refine the coarse depth from the backbone.

3.1.1 Residual Spatial Gradient Block

Fine-grained spatial details are vital for distinguishing the bona fide and attack presentations. As illustrated in Fig. 1, the gradient magnitude response between the living (Fig. 1(a)) and spoofing (Fig. 1(b)) face is quite different, which gives the insight to design a residual spatial gradient block (RSGB) for capturing such discriminative clues. In this paper, we take the well-known Sobel [27] operation to compute gradient magnitude. In a nutshell, the horizontal and vertical gradients can be derived from the following convolutions respectively:

$$F_{hor}(x)=\begin{bmatrix}-1 & 0 & +1\\ -2 & 0 & +2\\ -1 & 0 & +1\end{bmatrix}\odot x, F_{ver}(x)=\begin{bmatrix}-1 & -2 & -1\\ 0 & 0 & 0\\ +1 & +2 & +1\end{bmatrix}\odot x, \tag{1}$$

where $\odot$ denotes the depthwise convolution operation, and x represents the input feature maps. As shown in Fig. 4, our RSGB adopts the advanced shortcut connection structure to aggregate the learnable convolutional features with gradient magnitude information, which intends to enhance representation ability of fine-grained spatial details. It can be formulated as

$$y = \phi(\mathcal{N}(F(x,\{W_i\}) + \mathcal{N}(F_{hor}(x^{'})^2 + F_{ver}(x^{'})^2))), \tag{2}$$

where x represents the input features maps while $x^{'}$ denotes the feature maps altered through 1x1 convolution, which intends to keep the consistent channel numbers for subsequent residual addition. y denotes the output feature maps. $\mathcal{N}$ and ϕ denote the normalization and Relu layer, respectively. The function $F(x,\{W_i\})$ represents the residual gradient magnitude mapping to be learned. Note that the proposed RSGB is able to plug in both image and feature levels, extracting rich spatial context for depth regression task.

3.1.2 Spatio-Temporal Propagation Module

Virtual discrimination of depth between living and spoofing faces can be explored adequately by multiple frames. Therefore, we design STPM to extract multi-frame spatio-temporal features for depth estimation, via Short-term Spatio-Temporal Block (STSTB) and ConvGRU.

STSTB. As illustrated in Fig. 3, STSTB extracts the generalized short-term spatio-temporal information by fusing five kinds of features: the current compressed features $F_l(t)$, the current spatial gradient features $F_l^S(t)$, the future spatial gradient features $F_l^S(t+\triangle t)$, the temporal gradient features $F_l^T(t)$, and the STSTB features from the previous level $STSTB_{l-1}(t)$. The fused features can provide weighted spatial and temporal information in a learnable/adaptive way. In this paper, the spatial and temporal gradients are implemented with Sobel-based depthwise convolution (similar to Eq. 1) and element-wise subtraction of temporal features, respectively. Note that the 1x1 convolutions intend to compress the channel number with more efficiency.

Different from the related OFF [47] work, we consider both spatial gradient of the current compressed features $F_l^S(t)$ and future spatial gradient features $F_l^S(t+\triangle t)$ while OFF only considers $F_l^S(t)$. Moreover, current compressed feature $F_l(t)$ itself also plays an important role in recovering the fine depth map, which is concatenated in STSTB as well. The detailed comparison between STSTB and OFF will be studied in Sec. 5.3, which shows the advancement of STSTB especially for depth-supervised face anti-spoofing task.

ConvGRU. As short-term information between two consecutive frames from STSTB has limited representation ability, it is natural to use the recurrent neural network to capture long-range spatio-temporal context. However, the classical LSTM and GRU [13] neglect the spatial information in hidden units. In consideration of the spatial neighbor relationship in the hidden layers, ConvGRU is conducted for propagating the long-range spatio-temporal information. ConvGRU can be described as below:

$$\begin{aligned}&R_t = \sigma(K_r \otimes [H_{t-1}, X_t]),\ U_t = \sigma(K_u \otimes [H_{t-1}, X_t]),\\ &\hat{H}_t = \tanh(K_{\hat{h}} \otimes [R_t * H_{t-1}, X_t]),\\ &H_t = (1 - U_t) * H_{t-1} + U_t * \hat{H}_t,\end{aligned} \tag{3}$$

where X_t, H_t, U_t and R_t are the matrix of input, output, update gate and reset gate, $K_r, K_u, K_{\hat{h}}$ are the kernels in the convolution layer, $\otimes$ is convolution operation, $*$ denotes element wise product, and σ denotes the sigmoid activation function.

3.1.3 Depth Map Refinement

Forwarding the RSGB based backbone and STPM for a given N_f-frame input, we could obtain the corresponding coarse depth maps D_{single}^t and temporal depth maps D_{multi}^t, respectively, where $t \in [1, N_f - 1]$ denotes the t-th frame. Then D_{multi}^t is utilized to refine D_{single}^t in a weighted summation manner:

$$\mathrm{D}_{refined}^t = (1-\alpha)\cdot \mathrm{D}_{single}^t + \alpha \cdot \mathrm{D}_{multi}^t, \alpha \in [0,1], \tag{4}$$

where α is the trade-off weight between D_{single}^t and D_{multi}^t. The higher value of α indicates the more importance about the multi-frame spatio-temporal features. Finally, $N_f - 1$ refined depth maps $\{\mathrm{D}_{refined}^t\}_{t=1}^{N_f-1}$ are obtained.

3.2. Loss Function

Besides designing the network architecture, we also need an appropriate loss function to guide the network training. One major step-forward of the current study is that we design a novel Contrastive Detph Loss, which is able to combine with classical loss, further boosting performance.

3.2.1 Contrastive Detph Loss

In the classical depth-based face anti-spoofing, Euclidean Distance Loss (EDL) is usually used for pixel-wise supervision, which is formulated:

$$L_{EDL} = ||\mathrm{D}_P - \mathrm{D}_G||_2^2, \tag{5}$$

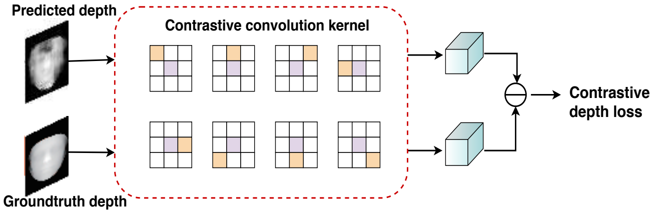

Figure 5. Contrastive Depth Loss. The purple, yellow, and white pieces indicate 1, -1, and 0, respectively. There are totally eight contrastive convolution kernels in CDL.

where D_P and D_G are the predicted depth and groundtruth depth, respectively. EDL applies supervision on the predicted depth based on pixel one by one, ignoring the depth difference among adjacent pixels. Intuitively, EDL merely assists the network to learn the absolute distance between the objects to the camera. However, the distance relationship of different objects is also important to be supervised for the depth learning. Therefore, as shown in Fig. 5, we propose the Contrastive Depth Loss (CDL) to offer extra strong supervision, which improves the generality of the depth-based face anti-spoofing model:

$$L_{CDL} = \sum_i ||\mathrm{K}_i^{CDL} \odot \mathrm{D}_P - \mathrm{K}_i^{CDL} \odot \mathrm{D}_G||_2^2, \tag{6}$$

where K_i^{CDL} is the ith contrastive convolution kernel, $i \in [0, 7]$. The details of the kernels can be found in Fig. 5.

3.2.2 Overall Loss

In view of the potentially unclear depth map, we hereby consider a binary loss when looking for the difference between living and spoofing depth map. Note that the depth supervision is decisive, whereas the binary supervision takes an assistant role to discriminate the different kinds of depth maps.

$$L_{binary} = -\mathrm{B}_G * log(fcs(\mathrm{D}_{avg})), \tag{7}$$

$$L_{overall} = \beta \cdot L_{binary} + (1 - \beta) \cdot (L_{EDL} + L_{CDL}), \tag{8}$$

where B_G is the binary groundtruth label, D_{avg} is the pool averaged map of $\{\mathrm{D}_{refined}^t\}_{t=1}^{N_f-1}$, and fcs denotes two fully connected layers and one softmax layer after the element-wise averaged depth maps, which outputs the logits of two classes, β is the hyper-parameter to trade-off binary loss and depth loss in the final overall loss $L_{overall}$.

4. Double-modal Anti-spoofing Dataset

In this work, we collect a real double-modal dataset (RGB and Depth). There are three kinds of display materials in replay attack: AMOLED screen, OLED screen, IPS/TFT screen. Meanwhile, three kinds of paper materials in print attacks are adopted: high-quality A4 paper, coated paper, and poster paper. The capture camera is RealSense SR300, which can offer corresponding RGB and Depth images. There are 300 subjects, each of which is recorded in three sessions and contains one real category and two attack categories (print and replay). Totally, we obtain 2700 samples (4˜12 seconds videos) in less than two months with two human workers. Tab. 1 demonstrates the details of DMAD, and Fig. 6 shows some corresponding examples.

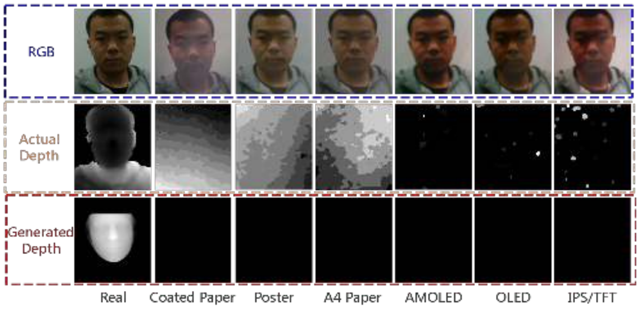

Figure 6. Some examples of DMAD. The actual depth is more precise than the generated depth.

Table 1. The details of our collected DMAD. This protocol of splitting subsets aims to evaluate the generalization of methods under unseen presentation materials.

Subset	Subject	Session	Modal Types	Presentation Material	# of live/attack vid.
Train	1˜100	1˜3	RGB, Depth	A4 Paper, AMOLED	900
	101˜200	1˜3	RGB, Depth	Coated Paper, OLED	900
Test	201˜300	1˜3	RGB, Depth	Poster Paper, IPS/TFT	900

5. Experiments

5.1. Databases and Metrics

5.1.1 Databases

Five databases - OULU-NPU [10, 5], SiW [34], CASIA-MFSD [56], Replay-Attack [12], DMAD are used in our experiment. OULU-NPU [10] is a high-resolution database, consisting of 4950 real access and spoofing videos and containing four protocols to validate the generalization of models. SiW [34] contains more live subjects and three protocols are used for testing. CASIA-MFSD [56] and Replay-Attack [12] both contain low-resolution videos.

5.1.2 Performance Metrics

In OULU-NPU and SiW dataset, we follow the original protocols and metrics for a fair comparison. OULU-NPU, SiW and DMAD utilize 1) Attack Presentation Classification Error Rate $APCER$, which evaluates the highest error among all PAIs (e.g. print or display), 2) Bona Fide Presentation Classification Error Rate $BPCER$, which evaluates the error of real access data, and 3) $ACER$ [25], which evaluates the mean of $APCER$ and $BPCER$:

$$ACER = \frac{APCER + BPCER}{2}. \tag{9}$$

HTER is adopted in the cross-database testing between CASIA-MFSD and Replay-Attack, evaluating the mean of False Rejection Rate (FRR) and False Acceptance Rate (FAR):

$$HTER = \frac{FRR + FAR}{2}. \tag{10}$$

5.2. Implementation Details

5.2.1 Depth Generation

Dense face alignment method PRNet [17] is adopted to estimate the 3D shape of the living face and generate the facial depth map $\mathrm{D}_G \in \mathbb{R}^{32\times 32}$. A typical sample can be found in the third row of Fig. 6. To distinguish living faces from spoofing faces, at the training stage, we normalize living depth map in a range of $[0, 1]$, while setting spoofing depth map to 0, which is similar to [34].

5.2.2 Training Strategy

The proposed method is trained with a two-stage strategy: *Stage 1:* We train the backbone with cascaded RSGB by the depth loss L_{EDL} and L_{CDL}, in order to learn a fundamental representation to predict coarse depth maps. *Stage 2:* We fix the parameters of the backbone, and train the STPM part by the overall loss $L_{overall}$ for refining depth maps. Our networks are fed by N_f frames, which are sampled by an interval of three frames. This sampling interval makes sampled frames maintain enough temporal information in the limited GPU memory.

5.2.3 Testing Strategy

For the final classification score, we feed the sequential frames into the network and obtain depth maps $\{\mathrm{D}^t_{refined}\}_{t=1}^{N_f-1}$ and the living logits $\hat{b}$ in $fcs(\mathrm{D}_{avg})$. The final living score can be obtained by:

$$score = \beta \cdot \hat{b} + (1-\beta) \cdot \frac{\sum_{t=1}^{N_f-1} ||\mathrm{D}^t_{refined} * \mathrm{M}^t||_1}{N_f - 1}, \tag{11}$$

where β is the same as that in equation 8, M^t is the mask of face at frame t , which can be generated by the dense face landmarks in PRNet [17], and the second module denotes that we compute the mean of depth values in the facial areas as one part of the score.

5.2.4 Hyper-parameter Setting

Our proposed method is implemented in Tensorflow, with a learning rate of 1e-4 for single-frame part and 1e-2 for multi-frame part. The batch size of single-frame part is 48, and that of multi-frame part is 2 with N_f being 5 in our experiment. Adadelta optimizer is used in our training procedure, with ρ as 0.95 and ϵ as 1e-8. We set $\alpha = 0.6$ and $\beta = 0.8$ by our experimental experience.

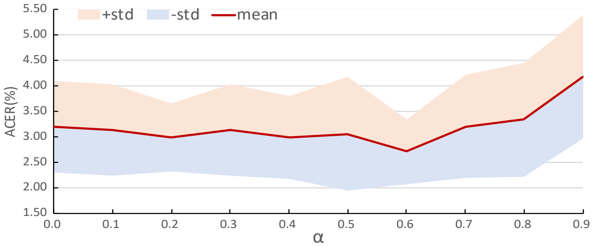

Figure 7. Ablation study of α in Eq. 4 on OULU-NPU Protocol 3. The red line denotes the mean ACER(%) value while the orange/blue area denotes the range of standard deviation.

Table 2. The results of ablation study on OULU-NPU Protocol 3.

Module	L_{CDL}	RSGB	STSTB	OFF	ConvGRU	L_{binary}	ACER(%)
Model 1							6.25±3.20
Model 2	√						5.07±1.83
Model 3	√	√					3.19±0.90
Model 4	√	√	√				2.99±0.72
Model 5	√	√	√		√		2.85±0.49
Model 6	√	√		√	√	√	3.20±1.00
Model 7	√	√	√		√	√	**2.71**±**0.63**

5.3. Experimental Comparison

5.3.1 Ablation Study

Seven architectures are implemented to demonstrate the efficacy of vital parts (i.e., RSGB, STPM and loss functions) in the proposed method. As shown in Tab. 2, Model 1 can be treated as a raw baseline, consisting of a backbone network with stacked vanilla convolutions. Model 2 is supervised with extra contrastive depth loss. Based on Model 2, vanilla convolutions are replaced by RSGB in Model 3. Moreover, Model 4 and Model 5 are designed for validating the effectiveness of STSTB and ConvGRU. In Model 6, STSTB is replaced by normal OFF [47]. Model 7 is our complete architecture with all modules and losses.

Efficacy of the Modules and Loss Functions. It can be seen from Tab. 2 that Model 2 outperforms Model 1, which means our proposed CDL helps to estimate more accurate depth maps. With the progressive lower ACER of Model 3, Model 4 and Model 5, it is clear that RSGB, STSTB and ConvGRU contribute to extract effective discriminative features respectively. Finally, in comparison between Model 5 and Model 7, binary supervision indeed assists to distinguish live vs. spoof.

STSTB vs. OFF. As illustrated in Tab. 2, Model 7 with STSTB surpasses Model 6 with OFF for a large margin, which implies that the current and future gradient information is valuable for spatio-temporal face anti-spoofing task. Model 6 even achieves inferior result compared with Model 3, indicating that it is challenging to design an effective temporal module for depth regression task.

Table 3. The results of intra-database testing on four protocols of OULU-NPU. For a fair comparison, here the results STASN [53] trained without extra private dataset are reported.

Prot.	Method	APCER(%)	BPCER(%)	ACER(%)
1	CPqD [6]	2.9	10.8	6.9
	GRADIANT [6]	1.3	12.5	6.9
	STASN [53]	1.2	2.5	1.9
	Auxiliary [34]	1.6	1.6	1.6
	FaceDs [26]	1.2	1.7	1.5
	OURs	2.0	0.0	**1.0**
2	MixedFASNet [6]	9.7	2.5	6.1
	FaceDs [26]	4.2	4.4	4.3
	Auxiliary [34]	2.7	2.7	2.7
	GRADIANT [6]	3.1	1.9	2.5
	STASN [53]	4.2	0.3	2.2
	OURs	2.5	1.3	**1.9**
3	MixedFASNet [6]	5.3±6.7	7.8±5.5	6.5±4.6
	GRADIANT [6]	2.6±3.9	5.0±5.3	3.8±2.4
	FaceDs [26]	4.0±1.8	3.8±1.2	3.6±1.6
	Auxiliary [34]	2.7±1.3	3.1±1.7	2.9±1.5
	STASN [53]	4.7±3.9	0.9±1.2	2.8±1.6
	OURs	3.2±2.0	2.2±1.4	**2.7±0.6**
4	Massy_HNU [6]	35.8±35.3	8.3±4.1	22.1±17.6
	GRADIANT [6]	5.0±4.5	15.0±7.1	10.0±5.0
	Auxiliary [34]	9.3±5.6	10.4±6.0	9.5±6.0
	STASN [53]	6.7±10.6	8.3±8.4	7.5±4.7
	FaceDs [26]	1.2±6.3	6.1±5.1	5.6±5.7
	OURs	6.7±7.5	3.3±4.1	**5.0±2.2**

Importance of Spatio-temporal Information for Depth Refinement. It can be seen from Eq. 4 that the depth map refinement is conducted in a weighted summation manner and hyperparameter α controls the contribution of the temporal depth maps predicted by STPM. As shown in Fig. 7, with appropriate valuel of α, the model can be benefited from spatio-temporal information and achieves better performance than that using only spatial information ($\alpha = 0.0$). And the best performance can be obtained when $\alpha = 0.6$.

Influence of Sampling Interval in Spatio-temporal Architecture. We conduct experiments on one sub-protocol of Protocol 3 with various sampling intervals ($\triangle t$). When $\triangle t$ euqals to $1, 3, 5$, and 7 frame(s), the ACER is 3.347%, 2.927%, 4.223%, and 2.934%, respectively. The ACER is the lowest when $\triangle t = 3$, which is used as the default setting for the following intra- and cross-database testing.

5.3.2 Intra-database Testing

We compare the performance of intra-database testing on OULU-NPU, SiW and DMAD datasets. There are four protocols in OULU-NPU for evaluating the generalization of the developed face presentation attack detection (PAD) methods. Protocol 1 and Protocol 2 are designed to evaluate the generalization of PAD methods under previously unseen illumination scene and under unseen attack medium (e.g., unseen printers or displays), respectively. Protocol 3 utilizes a Leave One Camera Out (LOCO) protocol, in order to study the effect of the input camera variation. Protocol 4 considers all the above factors and integrates all the constraints from protocols 1 to 3, so protocol 4 is the most challenging.

Table 4. The results of intra-database testing on three protocols of SiW [34].

Prot.	Method	APCER(%)	BPCER(%)	ACER(%)
1	Auxiliary [34]	3.58	3.58	3.58
	STASN [53]	–	–	1.00
	OURs	0.64	0.17	**0.40**
2	Auxiliary [34]	0.57±0.69	0.57±0.69	0.57±0.69
	STASN [53]	–	–	0.28±0.05
	OURs	0.00±0.00	0.04±0.08	**0.02±0.04**
3	STASN [53]	–	–	12.10±1.50
	Auxiliary [34]	8.31±3.81	8.31±3.80	8.31±3.81
	OURs	2.63±3.72	2.92±3.42	**2.78±3.57**

Table 5. The results of intra-database testing on DMAD.

Method	Depth Map	APCER(%)	BPCER(%)	ACER(%)
Model 7	Generated	9.17	3.48	6.33
	Actual	6.36	2.75	**4.55**

Results on OULU-NPU. As shown in Tab. 3, our proposed method ranks first on all 4 protocols, which indicates the proposed method performs well at the generalization of the external environment, attack mediums and input camera variation. It's worth noting that our model has the lowest mean and std of ACER in protocol 3 and 4, indicating its good accuracy and stablity.

Results on SiW. Tab. 4 compares the performance of our method with two state-of-the-art methods Auxiliary [34] and STASN [53] on SiW dataset. According to the purposes of three protocols on SiW and the results in Tab. 4, we can see that our method performs great advantages on the generalization of (a) variations of face pose and expression, (b) variations of different spoof mediums, (c) cross presentation attack instruments.

Results on DMAD. The results of intra-database testing on DMAD are shown in Tab. 5. In this experiment, we still set spoofing depth map to zero when training the actual depth model. Tab. 5 shows that the ACER(%) of multi-frame model (Model 7) supervised by actual depth obtains 1.78 lower than that supervised by generated depth. This demonstrates the actual depth map brings benefit to the improvement of monocular face anti-spoofing.

5.3.3 Cross-database Testing

We utilize four datasets (CASIA-MFSD, Replay-Attack, SiW and OULU-NPU) to perform cross-database testing for measuring the generalization potential of the models.

Results on CASIA-MFSD and Replay-Attack. In this experiment, there are two cross-database testing protocols. One is training on the CASIA-MFSD and testing on Replay-Attack, which we name protocol CR; the other is training on the Replay-Attack and testing on CASIA-

Table 6. The results of cross-database testing between CASIA-MFSD and Replay-Attack. The evaluation metric is HTER(%).

Method	Train	Test	Train	Test
	CASIA-MFSD	Replay-Attack	Replay-Attack	CASIA-MFSD
Motion [15]	50.2		47.9	
LBP-1 [15]	55.9		57.6	
LBP-TOP [15]	49.7		60.6	
Motion-Mag [3]	50.1		47.0	
Spectral cubes [44]	34.4		50.0	
CNN [51]	48.5		45.5	
LBP-2 [7]	47.0		39.6	
STASN [53]	31.5		30.9	
Colour Texture [8]	30.3		37.7	
FaceDs [26]	28.5		41.1	
Auxiliary [34]	27.6		28.4	
OURs	**17.0**		**22.8**	

Table 7. The results of cross-database testing from SiW to OULU-NPU dataset.

Prot.	Method	APCER(%)	BPCER(%)	ACER(%)
1	Auxiliary [34]	–	–	10.0
	OURs	1.7	13.3	**7.5**
2	Auxiliary [34]	–	–	14.1
	OURs	9.7	14.2	**11.9**
3	Auxiliary [34]	–	–	**13.8±5.7**
	OURs	17.5±4.6	11.7±12.0	14.6±4.8
4	Auxiliary [34]	–	–	10.0±8.8
	OURs	0.8±1.9	10.0±11.6	**5.4±5.7**

MFSD, which we name protocol RC. In Tab. 6, it is shown that HTER(%) of our proposed method is 17.0 on protocol CR and 22.8 on protocol RC, reducing 38.4% and 19.7% respectively compared with the previous state of the art. The improvement of performance on cross-database testing demonstrates the good generalization of proposed method.

Results from SiW to OULU-NPU. Here, It is shown that the cross-database testing results trained on SiW and tested on OULU-NPU in Tab. 7. It can be seen that our method outperforms Auxiliary [34] on three protocols (decrease 2.5%, 2.2% and 4.6% of ACER on protocol 1, protocol 2 and protocol 4, respectively). In protocol 3, ACER of our method is 14.6±4.8% and slightly higher than that of Auxiliary. Considering the rPPG used in Auxiliary method, it may also be a good choice combined with proposed method.

5.3.4 Visualization and Analysis

The predicted depth maps of hard samples in OULU-NPU Protocol 3 are partly visualized in Fig. 8. It can be seen that some samples are difficult for the single-frame PAD to be detected. In contrary, our multi-frame methods with STPM can estimate more precise depth maps than those of single-frame method. The difference of depth images from real and attack samples in third row is also more significant, indicating the good discriminative information with the results of STPM.

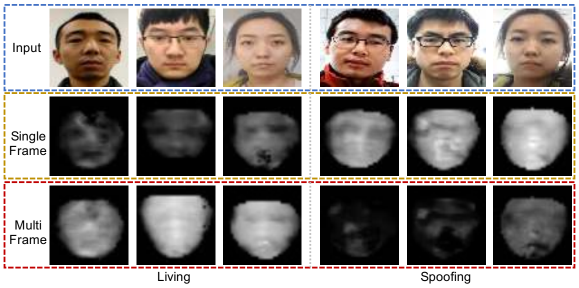

Figure 8. The generated results of hard samples in OULU-NPU. The predicted coarse depth maps from stacked RSGB backbone and temporal depth maps from STPM are illustrated in the second and third row, respectively.

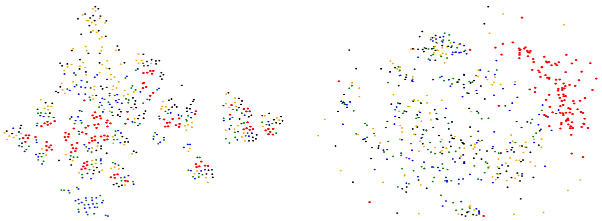

Figure 9. Feature distribution visualization of the testing videos on OULU-NPU Protocol 1 using t-SNE [36] . Left: features w/o RSGB, Right: features w/ RSGB. Color indicates red=live, $green$=printer1, $blue$=printer2, $orange$=display1, $black$=display2.

Feature distribution of the testing videos on OULU-NPU Protocol 1 is shown in Fig. 9. The right image (w/ RSGB) presents more well-clustered behavior than the left image (w/o RSGB), which demonstrates the excellent discrimination ability of our proposed RSGB for distinguishing the living and spoofing faces.

6. Conclusions

In this paper, we propose a novel face anti-spoofing method, which exploits fine-grained spatio-temporal information for facial depth estimation. In our method, Residual Spatial Gradient Block (RSGB) is utilized to detect more discriminative details while Spatio-Temporal Propagation Module (STPM) to encode spatio-temporal information. An extra Contrastive Depth Loss (CDL) is designed to improve the generality of depth-supervised PAD. We also investigate the effectiveness of actual depth map in face anti-spoofing. Extensive experiments demonstrate the superiority of our method.

Acknowledgment

This work has been partially supported by the Chinese National Natural Science Foundation Projects #61876178, #61806196, #61976229.

References

[1] Akshay Agarwal, Richa Singh, and Mayank Vatsa. Face anti-spoofing using haralick features. In *BTAS*, 2016.

[2] Yousef Atoum, Yaojie Liu, Amin Jourabloo, and Xiaoming Liu. Face anti-spoofing using patch and depth-based cnns. In *IJCB*, pages 319–328, 2017.

[3] Samarth Bharadwaj, Tejas I Dhamecha, Mayank Vatsa, and Richa Singh. Computationally efficient face spoofing detection with motion magnification. In *CVPRW*, pages 105–110, 2013.

[4] Samarth Bharadwaj, Tejas I. Dhamecha, Mayank Vatsa, and Richa Singh. Face anti-spoofing via motion magnification and multifeature videolet aggregation. 2014.

[5] Zinelabinde Boulkenafet. A competition on generalized software based face presentation attack detection in mobile scenarios. In *IJCB*, 2017.

[6] 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 *2017 IEEE International Joint Conference on Biometrics (IJCB)*, pages 688–696. IEEE, 2017.

[7] Zinelabidine Boulkenafet, Jukka Komulainen, and Abdenour Hadid. Face anti-spoofing based on color texture analysis. In *ICIP*, pages 2636–2640, 2015.

[8] Zinelabidine Boulkenafet, Jukka Komulainen, and Abdenour Hadid. Face spoofing detection using colour texture analysis. *IEEE Transactions on Information Forensics and Security*, 11(8):1818–1830, 2016.

[9] Zinelabidine Boulkenafet, Jukka Komulainen, and Abdenour Hadid. Face antispoofing using speeded-up robust features and fisher vector encoding. *IEEE Signal Processing Letters*, 24(2):141–145, 2017.

[10] 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.

[11] Girija Chetty and Michael Wagner. Multi-level liveness verification for face-voice biometric authentication. *CB*, 2006.

[12] Ivana Chingovska, André Anjos, and Sébastien Marcel. On the effectiveness of local binary patterns in face anti-spoofing. In *Biometrics Special Interest Group*, pages 1–7, 2012.

[13] Junyoung Chung, Caglar Gulcehre, KyungHyun Cho, and Yoshua Bengio. Empirical evaluation of gated recurrent neural networks on sequence modeling. *arXiv preprint arXiv:1412.3555*, 2014.

[14] Tiago de Freitas Pereira, André Anjos, José Mario De Martino, and Sébastien Marcel. Lbp- top based countermeasure against face spoofing attacks. In *ACCV*, pages 121–132, 2012.

[15] Tiago de Freitas Pereira, André Anjos, José Mario De Martino, and Sébastien Marcel. Can face anti-spoofing countermeasures work in a real world scenario? In *ICB*, pages 1–8, 2013.

[16] Litong Feng, Lai-Man Po, Yuming Li, Xuyuan Xu, Fang Yuan, Terence Chun-Ho Cheung, and Kwok-Wai Cheung. Integration of image quality and motion cues for face anti-spoofing: A neural network approach. *Journal of Visual Communication and Image Representation*, 38:451–460, 2016.

[17] Yao Feng, Fan Wu, Xiaohu Shao, Yanfeng Wang, and Xi Zhou. Joint 3d face reconstruction and dense alignment with position map regression network. In *CVPR*, 2017.

[18] Robert W. Frischholz and Ulrich Dieckmann. Biold: a multimodal biometric identification system. *Computer*, 33(2):64–68, 2000.

[19] Robert W Frischholz and Alexander Werner. Avoiding replay-attacks in a face recognition system using head-pose estimation. *AMFGW*, pages 234–235, 2003.

[20] Junying Gan, Shanlu Li, Yikui Zhai, and Chengyun Liu. 3d convolutional neural network based on face anti-spoofing. In *ICMIP*, pages 1–5, 2017.

[21] Jianzhu Guo, Xiangyu Zhu, Jinchuan Xiao, Zhen Lei, Genxun Wan, and Stan Z. Li. Improving face anti-spoofing by 3d virtual synthesis. In *ICB*, 2019.

[22] 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.

[23] Wei Hu, Gusi Te, Ju He, Dong Chen, and Zongming Guo. Exploring hypergraph representation on face anti-spoofing beyond 2d attacks. *arXiv preprint arXiv: 1811.11594v1*, 2018.

[24] Wei Hu, Gusi Te, Ju He, Dong Chen, and Zongming Guo. Aurora guard: Real-time face anti-spoofing via light reflection. *arXiv preprint arXiv: 1902.10311*, 2019.

[25] international organization for standardization. Iso/iec jtc 1/sc 37 biometrics: Information technology biometric presentation attack detection part 1: Framework. In *https://www.iso.org/obp/ui/iso*, 2016.

[26] Amin Jourabloo, Yaojie Liu, and Xiaoming Liu. Face despoofing: Anti-spoofing via noise modeling. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 290–306, 2018.

[27] Nick Kanopoulos, Nagesh Vasanthavada, and Robert L Baker. Design of an image edge detection filter using the sobel operator. *IEEE Journal of solid-state circuits*, 23(2):358–367, 1988.

[28] Jukka Komulainen, Abdenour Hadid, and Matti Pietikainen. Context based face anti-spoofing. In *BTAS*, pages 1–8, 2013.

[29] Jiangwei Li, Yunhong Wang, Tieniu Tan, and Anil K Jain. Live face detection based on the analysis of fourier spectra. *SPIE (BTHI)*, 5404:296–304, 2004.

[30] Lei Li, Xiaoyi Feng, Zinelabidine Boulkenafet, Zhaoqiang Xia, Mingming Li, and Abdenour Hadid. An original face anti-spoofing approach using partial convolutional neural network. In *IPTA*, pages 1–6, 2016.

[31] Xiaobai Li, Jukka Komulainen, Guoying Zhao, Pong-Chi Yuen, and Matti Pietikäinen. Generalized face anti-spoofing by detecting pulse from face videos. In *2016 23rd International Conference on Pattern Recognition (ICPR)*, pages 4244–4249. IEEE, 2016.

[32] Bofan Lin, Xiaobai Li, Zitong Yu, and Guoying Zhao. Face liveness detection by rppg features and contextual patch-based cnn. In *Proceedings of the 2019 3rd International Conference on Biometric Engineering and Applications*, pages 61–68. ACM, 2019.

[33] Siqi Liu, Pong Chi Yuen, Shengping Zhang, and Guoying Zhao. 3d mask face anti-spoofing with remote photoplethysmography. 2016.

[34] Yaojie Liu, Amin Jourabloo, and Xiaoming Liu. Learning deep models for face anti-spoofing: Binary or auxiliary supervision. In *CVPR*, pages 389–398, 2018.

[35] Oeslle Lucena, Amadeu Junior, Vitor Moia, Roberto Souza, Eduardo Valle, and Roberto Lotufo. Transfer learning using convolutional neural networks for face anti-spoofing. In *International Conference Image Analysis and Recognition*, pages 27–34, 2017.

[36] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(Nov):2579–2605, 2008.

[37] Jukka M*ää*tt*ä*, Abdenour Hadid, and Matti Pietik*ä*inen. Face spoofing detection from single images using micro-texture analysis. In *IJCB*, pages 1–7, 2011.

[38] Chaitanya Nagpal and Shiv Ram Dubey. A performance evaluation of convolutional neural networks for face anti spoofing. *arXiv preprint arXiv:1805.04176*, 2018.

[39] Ewa Magdalena Nowara, Ashutosh Sabharwal, and Ashok Veeraraghavan. Ppgsecure: Biometric presentation attack detection using photopletysmograms. 2017.

[40] Gang Pan, Lin Sun, Zhaohui Wu, and Shihong Lao. Eyeblink-based anti-spoofing in face recognition from a generic webcamera. In *ICCV*, pages 1–8, 2007.

[41] Keyurkumar Patel, Hu Han, and Anil K Jain. Cross-database face antispoofing with robust feature representation. In *Chinese Conference on Biometric Recognition*, pages 611–619, 2016.

[42] Keyurkumar Patel, Hu Han, and Anil K Jain. Secure face unlock: Spoof detection on smartphones. *IEEE transactions on information forensics and security*, 11(10):2268–2283, 2016.

[43] Bruno Peixoto, Carolina Michelassi, and Anderson Rocha. Face liveness detection under bad illumination conditions. In *ICIP*, pages 3557–3560. IEEE, 2011.

[44] Allan Pinto, Helio Pedrini, William Robson Schwartz, and Anderson Rocha. Face spoofing detection through visual codebooks of spectral temporal cubes. *IEEE Transactions on Image Processing*, 24(12):4726–4740, 2015.

[45] 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. *AAAI*, 2020.

[46] Xiao Song, Xu Zhao, Liangji Fang, and Tianwei Lin. Discriminative representation combinations for accurate face spoofing detection. *Pattern Recognition*, 85:220–231, 2019.

[47] Shuyang Sun, Zhanghui Kuang, Lu Sheng, Wanli Ouyang, and Wei Zhang. Optical flow guided feature: A fast and robust motion representation for video action recognition. In *CVPR*, pages 1390–1399, 2018.

[48] Xiaoyang Tan, Yi Li, Jun Liu, and Lin Jiang. Face liveness detection from a single image with sparse low rank bilinear discriminative model. In *ECCV*, pages 504–517, 2010.

[49] Di Wen, Hu Han, and Anil K Jain. Face spoof detection with image distortion analysis. *IEEE Transactions on Information Forensics and Security*, 10(4):746–761, 2015.

[50] Zhenqi Xu, Shan Li, and Weihong Deng. Learning temporal features using lstm-cnn architecture for face anti-spoofing. In *ACPR*, pages 141–145, 2015.

[51] Jianwei Yang, Zhen Lei, and Stan Z. Li. Learn convolutional neural network for face anti-spoofing. *Computer Science*, 9218:373–384, 2014.

[52] Jianwei Yang, Zhen Lei, Shengcai Liao, and Stan Z Li. Face liveness detection with component dependent descriptor. In *ICB*, page 2, 2013.

[53] Xiao Yang, Wenhan Luo, Linchao Bao, Yuan Gao, Dihong Gong, Shibao Zheng, Zhifeng Li, and Wei Liu. Face anti-spoofing: Model matters, so does data. In *CVPR*, 2019.

[54] Zitong Yu, Yunxiao Qin, Xiaqing Xu, Chenxu Zhao, Zezheng Wang, Zhen Lei, and Guoying Zhao. Autofas: Searching lightweight networks for face anti-spoofing. *ICASSP*, 2020.

[55] 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 multimodal face anti-spoofing. In *CVPR*, 2019.

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