# A Closer Look at Geometric Temporal Dynamics for Face Anti-Spoofing

Chih-Jung Chang[12]*, Yaw-Chern Lee[1], Shih-Hsuan Yao[1], Min-Hung Chen[1], Chien-Yi Wang[1]
Shang-Hong Lai[13], Trista Pei-Chun Chen[1]
[1]Microsoft AI R&D Center, Taiwan   [2]Stanford University   [3]National Tsing Hua University, Taiwan

## Abstract

*Face anti-spoofing (FAS) is indispensable for a face recognition system. Many texture-driven countermeasures were developed against presentation attacks (PAs), but the performance against unseen domains or unseen spoofing types is still unsatisfactory. Instead of exhaustively collecting all the spoofing variations and making binary decisions of live/spoof, we offer a new perspective on the FAS task to distinguish between normal and abnormal movements of live and spoof presentations. We propose **Geometry-Aware Interaction Network (GAIN)**, which exploits dense facial landmarks with spatio-temporal graph convolutional network (ST-GCN) to establish a more interpretable and modularized FAS model. Additionally, with our cross-attention feature interaction mechanism, GAIN can be easily integrated with other existing methods to significantly boost performance. Our approach achieves state-of-the-art performance in the standard intra- and cross-dataset evaluations. Moreover, our model outperforms state-of-the-art methods by a large margin in the cross-dataset cross-type protocol on CASIA-SURF 3DMask (+10.26 higher AUC score), exhibiting strong robustness against domain shifts and unseen spoofing types.*

## 1. Introduction

Face anti-spoofing (FAS) has played a critical role in securing face recognition systems from presentation attacks (*e.g.*, print, video-replay, and 3D-mask attacks). Most popular FAS approaches extract fine-grained texture to spot spoofing cues [1, 5, 44, 46, 48, 49]. These methods mainly adopt frame-level spoof detection, which aggregates the prediction of one or multiple frames to determine the liveness of a video instance. However, the extracted spoofing cues based on single-frame information might be insufficient to represent the characteristics of spoof faces.

One common way to address the above issue is to ex-

---

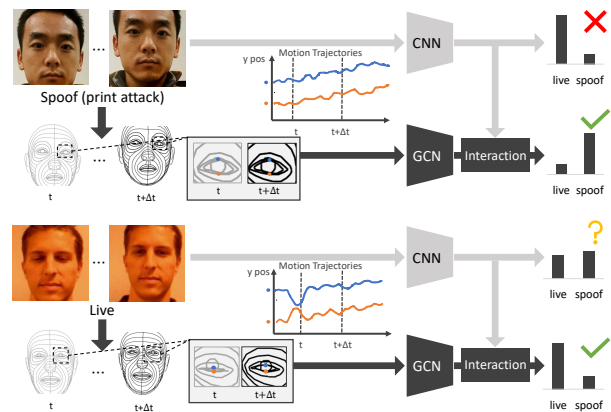*Work done during the internship at Microsoft AI R&D Center, Taiwan.



Figure 1. Instead of only focusing on photometric information by common CNN-based FAS methods, our proposed GAIN exploits geometric temporal dynamics based on dense facial landmarks to capture detailed facial motions, which is holistic and discriminative to enhance spoof detection. The motion trajectories correspond to the example facial landmarks, showing useful cues to distinguish live/spoof faces.

ploit the *temporal information* of videos. Live subjects differ from spoof attacks in motion (*i.e.*, facial movements), indicating that dynamics in consecutive frames are beneficial to discriminating presentation attacks from live faces. CNN-LSTM/GRU networks [22, 26, 38, 45] and 3D-CNNs [37, 42] have been used to extract temporal features from RGB frames. Despite the promising performance, focusing on RGB information may lead to overfitting to *photometric* properties of live/spoof instances (*e.g.*, material patterns) and under-exploring *geometric* information (*e.g.*, motion trajectories), which is important for robust temporal features (Fig. 1). Compared to a live subject's micro-movement, a paper-based attack commonly generates global motions such as translation or rotation, or abnormal transition movements of paper getting bent; for a 3D-mask attack, regardless of how realistic it appears, the movements introduced can be mostly treated as translation or rotation. All these motions appear relatively constrained in the geometric temporal modality, and some types of mo-

tions are even common across different attack types, indicating that *geometric temporal information* is the key to further improving the model robustness against domain variations and unseen spoofing types. Therefore, the question becomes: *How can we effectively exploit geometric temporal dynamics to benefit face anti-spoofing?* Inspired by the success of skeleton-based action recognition [6, 24, 29, 43]), which shows robustness to unseen domains by exploiting body joints, the use of *landmarks* is motivated to effectively extract geometric temporal dynamics on faces. In this work, we propose to improve the robustness of the temporal FAS features by extracting dynamic information on top of facial landmarks. Taking landmarks as inputs instead of raw RGB frames can explicitly extract geometric information that generalizes well to unseen attacking materials and is invariant under changes in cameras, lighting, and backgrounds. More specifically, we adopt dense facial landmark prediction [41] to better capture the detailed discriminative motion from live faces and spoof attacks. In addition, we adopt a Graph Convolutional Network (GCN) to extract spatio-temporal features from dense facial landmarks, providing robust and representative dynamic information and reducing the computational complexity compared to the commonly used CNN-based temporal FAS methods. Finally, we design a cross-attention feature interaction strategy to integrate our geometric temporal features with photometric features from other FAS methods. The overall framework is dubbed **Geometry-Aware Interaction Network (GAIN)**. Extensive experimental results on intra- and cross-dataset benchmarks demonstrate that the geometric feature of GAIN is robust to unseen domains and provides liveness information complementary to current FAS methods, significantly boosting the performance.

Our contributions are three-fold:

- To the best of our knowledge, we are the first to learn robust geometric temporal dynamics with the dense facial topology that captures fine-grained facial movements, which is critical to FAS but missing in previous works.

- In the proposed GAIN, our learned geometric temporal features can smoothly cooperate with photometric features from other existing FAS methods by our proposed cross-attention feature interaction strategy.

- The proposed GAIN has been evaluated on intra-dataset, cross-dataset, and domain generalization benchmarks, achieving state-of-the-art performance and outperforming other FAS methods on all protocols. This demonstrates the efficacy of our proposed usage of geometric temporal dynamics.

## 2. Related Works

**Face Anti-Spoofing:** Most of the recent FAS methods aim at frame-level spoofing detection. They extract fine-grained information to identify spoofing patterns (*e.g.*, lattice artifacts), using binary supervision [5, 44] or leveraging auxiliary tasks [1, 46, 48, 49]. In spite of promising results in intra-domain scenarios, these methods encounter degradation when evaluated in unseen domains, which motivates recent domain adaptation [17, 34] and domain generalization [14, 28, 33, 36] approaches. Another group of FAS works can be viewed as temporal-based methods, which make use of temporal information in consecutive frames to extract liveness cues. These methods generally adopt CNN-LSTM/GRU networks [22, 26, 38, 45] or 3D-CNNs [37, 42] to integrate spatial and temporal information. Nonetheless, by considering RGB information as input, such methods might run the risk of overfitting to photometric properties and neglecting geometric information that reveals facial motion patterns. In this work, we aim to extract robust geometric temporal features that focus on human facial dynamics to improve FAS performance.

**Node-Based Action Recognition:** In human action recognition, extracting dynamic information from key nodes (*i.e.*, skeleton, and joint) trajectories has attracted much attention due to its robustness to scene variation. For example, sequences of human joints are fed into temporal CNNs [15, 16, 30] or RNNs [19, 27, 50, 54] to predict human actions. However, these methods might overlook the inherent correlations of joints, thus having the limited capability of describing human dynamics. More recently, GCNs have demonstrated the power of leveraging graph topology to model joint correlations [6, 24, 29, 43]. ST-GCN [43] applies GCNs along with temporal convolutions to learn both spatial and temporal information from landmark sequences, achieving impressive results in action recognition. Inspired by such success, we use GCN in this paper to extract geometric information on top of human facial landmarks and model facial movements.

**Facial Landmark Prediction:** Landmarks are instrumental in several face-related computer vision tasks. While sparse landmark prediction [3, 10, 55] has been the mainstream approach, it fails to capture detailed facial characteristics and expressions since such information cannot be represented by a sparse set of landmarks. Recently, using dense landmarks has achieved remarkable results for 3D face reconstruction [41]. By training with synthetic data, which guarantees dense yet consistent landmark annotations, [40] shows their dense landmark prediction model can produce accurate and expressive facial performance capture. The rich information provided by dense landmarks

would be helpful to identifying fine-grained facial movements for spoof detection.

## 3. Proposed Method

With the aim of extracting robust temporal features to enhance FAS performance, we propose Geometry-Aware Interaction Network (GAIN), which is designed to extract detailed facial dynamics from the geometric information, as illustrated in Fig. 2. More specifically, we utilize GCN to extract features on top of dense facial landmarks and learn geometric information for fine-grained facial movements (Sec. 3.1). We then integrate our learned geometric features with photometric features by the proposed cross-attention mechanism, further boosting the performance of SOTA FAS methods (Sec. 3.2).

### 3.1. Dense Geometric Liveness Feature Learning

Facial movement is a discriminative component in describing live subjects. Formed by complex combinations of muscles, the detailed facial movements of live subjects are difficult to be produced by spoof attacks, such as print and multiple 3D-mask attacks, which usually produce *abnormal* movements instead of *normal* movements from live subjects. With the aim of robustly capturing such fine-grained movements, we propose to utilize GCN to learn geometric facial dynamics on top of dense facial landmarks.

#### 3.1.1 Dense Landmarks Detection

In order to invariantly describe motion patterns in different domains, we exploit landmarks as representative geometric features. More specifically, we adopt dense landmark prediction [41] to capture the fine-grained details of facial dynamics. As in [41], each facial landmark is predicted as a random variable $\{v_i, \sigma_i\}$, where $v_i = (x_i, y_i)$ represents the expected position of a landmark, and $\sigma_i$ is a measure of uncertainty. The landmark prediction network is trained with synthetic data of densely annotated facial landmarks using a Gaussian negative log-likelihood (GNLL) loss:

$$L_{GNLL}(v, \sigma) = \sum_{i=1}^{N} \lambda_i \left( \log(\sigma_i^2) + \frac{\|v_i - \hat{v}_i\|^2}{2\sigma^2} \right), \quad (1)$$

where $N$ is the total number of facial landmarks, $\hat{v}_i$ is the training label for $v_i$, and $\lambda_i$ is the loss weighting.

We then utilize the trained network to predict dense facial landmarks for live and spoof subjects. Concretely, given an input RGB sequence $I = \{I_t\}_{t=1}^{T}$ of length $T$, we predict $N$ facial landmarks $\{v_{ti}\}_{i=1}^{N}$ for each frame $I_t$, where $v_{ti} = (x_{ti}, y_{ti})$ represents the expected position of a landmark. To even better spot the micro-movements on a face, we align the landmarks in each frame using simple yet effective frame-wise min-max normalization. Namely, we obtain aligned landmarks $v'_{ti} = (x'_{ti}, y'_{ti})$ by:

$$x'_{ti} = \frac{x_{ti} - x_{t min}}{x_{t max} - x_{t min}}, \quad (2)$$

where $x_{t min}$ and $x_{t max}$ denote $\min_{i=1,...,N} x_{ti}$ and $\max_{i=1,...,N} x_{ti}$ respectively, and $y'_{ti}$ can be obtained likewise. The visualization of facial landmarks is shown in Fig. 3.

#### 3.1.2 Geometric Liveness Feature Learning

With a rich set of facial landmarks obtained, we can learn the fine-grained motion of facial movements. Inspired by the success of GCN-based methods in skeleton-based human action recognition, we adopt the spatio-temporal graph convolution operator in ST-GCN [43] to model the geometrical relationships of facial landmarks across the frames.

We first sub-sample $S$ frames out of $T$ aligned facial landmarks as the input of GCN. At each graph convolutional layer, a graph with $N \times S$ nodes is constructed, where each node is an input feature of size $C_{in}$. The input feature map can thus be represented by $f_{in} \in \mathbb{R}^{N \times S \times C_{in}}$. The edges of the graph are defined as the combination of: (1) the facial connectivity at each frame (Fig. 3), and (2) the same facial landmark in consecutive frames. We represent the facial connectivity by the adjacency matrix $\mathbf{A} \in \mathbb{R}^{N \times N}$, and additionally construct a learnable weight mask $\mathbf{M} \in \mathbb{R}^{N \times N}$ for $\mathbf{A}$ to adaptively learn edge weighting that helps capture detailed facial movements. We formulate the higher-level feature map $f_{out}$ that integrates both spatial and temporal information as:

$$f_{out} = conv(\Lambda^{-\frac{1}{2}}((\mathbf{A} + \mathbf{I}) \odot \mathbf{M})\Lambda^{-\frac{1}{2}} f_{in}\mathbf{W}), \quad (3)$$

where $\mathbf{I} \in \mathbb{R}^{N \times N}$ represents the identity matrix, $\Lambda^{ii} = \sum_j \mathbf{M}^{ij}(\mathbf{A}^{ij} + \mathbf{I}^{ij})$ is for normalization, $\mathbf{W} \in \mathbb{R}^{C_{in} \times C_{out}}$ is the weight for feature transformation, and $\odot$ and $conv(\cdot)$ denotes the element-wise product and a 1-D convolution along the temporal axis, respectively.

After the last spatio-temporal convolutional layer, we apply global pooling across all $N \times S$ nodes to obtain the final feature $f_g$. To learn $f_g$ that discriminates between live facial dynamics and abnormal motions, we distinctly define the class of normal and abnormal movements: We group live faces and video-replay attacks together to form the class of normal facial movements, denoted as $G_{nm}$, and define the class of abnormal movements $G_{am}$ as attacks that cannot produce realistic motions (*e.g.*, printed/displayed photos and plastic/plaster/resin masks). The objective of our geometric liveness learning $L_g$ is then formulated as:

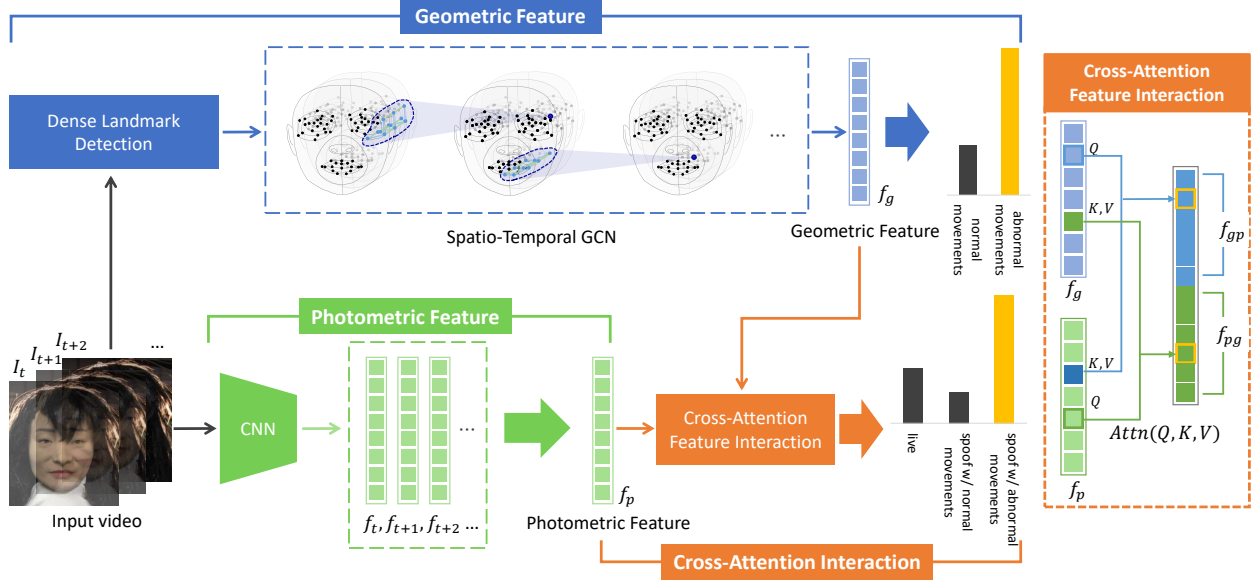$$L_g = BCE(l_g, \mathbf{W}f_g), \quad (4)$$

Figure 2. The proposed Geometry-Aware Interaction Network. Aiming at exploiting robust temporal dynamics that benefit spoof detection, we learn geometric features $f_g$ (blue) by utilizing ST-GCN on top of dense landmark predictions. To integrate the learned $f_g$ with photometric features $f_p$ from common FAS methods (green), we adopt a cross-attention mechanism (orange) that explores the interaction between geometric and photometric information, producing more discriminative features $f_{gp}$ and $f_{pg}$ for FAS.

where $BCE(\cdot)$ denotes the binary cross-entropy loss, $\mathbf{W}$ represents a linear projection for classification, and $l_g$ is the label of the input sequence $I$, defined by:

$$l_g = \begin{cases} 1, & \text{if } I \in G_{nm}. \\ 0, & \text{if } I \in G_{am}. \end{cases} \quad (5)$$

### 3.2. Cross-Attention Feature Interaction

Learned with the objective of identifying facial movements, the geometric feature obtained in Sec. 3.1 provides information complementary to common photometrics-based methods. To integrate the learned geometric features with these methods, it is important that the relationship between geometric and photometric features is highlighted and exploited. Fig. 2 illustrates our designed cross-attention strategy that fuses the learned geometric feature $f_g$ with feature extracted by common photometrics-based methods, $f_p$.

Given a chosen photometrics-based FAS network, we obtain the photometric feature $f_p$ by extracting the representation before the network's task-specific projection head. We then perform cross-attention on our geometric feature $f_g$ and the extracted $f_p$. Generally, the attention operation [32] is computed as:

$$Attn(Q, K, V) = Softmax(\frac{QK^\top}{\sqrt{d}})V, \quad (6)$$

where $Q$, $K$, $V$ represent the input query, key, and value, and $d$ is the feature dimension. The interaction of $f_g$ and $f_p$

can thus be modeled by:

$$\begin{aligned} f_{gp} &= Attn(\mathbf{W}_g^Q f_g, \mathbf{W}_p^K f_p, \mathbf{W}_p^V f_p), \\ f_{pg} &= Attn(\mathbf{W}_p^Q f_p, \mathbf{W}_g^K f_g, \mathbf{W}_g^V f_g), \end{aligned} \quad (7)$$

, where each of the $\mathbf{W}$s is a linear projection.

Finally, $f_{gp}$ and $f_{pg}$ are concatenated and fed into a linear classifier to predict the liveness probability. We train the cross-attention module using the cross-entropy loss, where live and spoof types are divided into 3 (instead of 2) classes: live faces, spoof attacks with normal facial movements, and spoof attacks with abnormal motions. This encourages the geometric information of $f_g$ to be further exploited in the final decision and leads to more accurate spoof detection, as later verified in Sec. 4.4.

## 4. Experiments

### 4.1. Datasets and Evaluation Metrics

We evaluate the proposed GAIN on five major public datasets: OULU-NPU [2], CASIA-FASD [51], Replay-Attack [7], MSU-MFSD [39], and CASIA-SURF 3DMask [47]. We utilize OULU-NPU for intra-dataset evaluation, and all five datasets for cross-dataset experiments. The performance is measured by the following evaluation metrics: Attack Presentation Classification Error Rate (APCER), Bona-fide Presentation Classification Error Rate (BPCER), Average Classification Error Rate
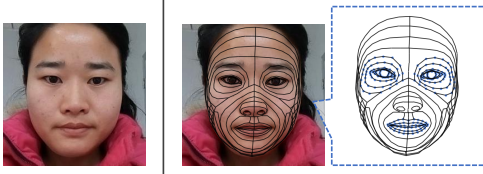
Figure 3. Visualization of dense landmark prediction. The prediction consists of 550 landmarks, in which 246 landmarks around eyes and mouth region (visualized by the blue dots) are used in our proposed framework.

| Prot. | Method | APCER(%) | BPCER(%) | ACER(%) |
|---|---|---|---|---|
| 1 | Auxiliary [22] | 1.6 | 1.6 | 1.6 |
| | ICLM [42] | <u>0.4</u> | <u>0.0</u> | <u>0.2</u> |
| | NAS-FAS [47] | <u>0.4</u> | <u>0.0</u> | <u>0.2</u> |
| | MTSS [13] | 2.0 | 1.0 | 1.5 |
| | CDCN++* [48] | **0.0** | 0.8 | 0.4 |
| | CDCN++* + **GAIN (Ours)** | **0.0** | **0.0** | **0.0** |
| 2 | Auxiliary [22] | 2.7 | 2.7 | 2.7 |
| | ICLM [42] | 2.5 | 1.9 | 1.6 |
| | NAS-FAS [47] | 1.5 | 0.8 | 1.2 |
| | MTSS [13] | <u>1.4</u> | **0.3** | <u>0.9</u> |
| | CDCN++* [48] | 1.9 | <u>0.6</u> | 1.3 |
| | CDCN++* + **GAIN (Ours)** | **0.8** | 0.8 | **0.8** |
| 3 | Auxiliary [22] | 2.7±1.3 | 3.1±1.7 | 2.9±1.5 |
| | ICLM [42] | 1.1±0.9 | 2.5±2.6 | 1.8±1.3 |
| | NAS-FAS [47] | 2.1±1.3 | 1.4±1.1 | 1.7±0.6 |
| | MTSS [13] | 2.1±1.3 | **0.5±0.4** | <u>1.3±0.8</u> |
| | CDCN++* [48] | **0.6±0.5** | 2.8±4.2 | 1.7±2.0 |
| | CDCN++* + **GAIN (Ours)** | <u>0.9±0.8</u> | <u>0.8±0.8</u> | **0.9±0.8** |
| 4 | Auxiliary [22] | 9.3±5.6 | 10.4±6.0 | 9.5±6.0 |
| | ICLM [42] | **3.3±3.9** | 4.1±4.4 | <u>3.7±2.8</u> |
| | NAS-FAS [47] | <u>4.2±5.3</u> | **1.7±2.6** | **2.9±2.8** |
| | MTSS [13] | 6.6±3.3 | <u>2.4±2.8</u> | 4.5±2.2 |
| | CDCN++* [48] | 5.8±5.9 | 4.2±6.1 | 5.0±2.4 |
| | CDCN++* + **GAIN (Ours)** | 4.6±2.2 | 4.2±1.9 | 4.4±2.0 |

Table 1. The results of the intra-dataset evaluation on OULU-NPU. We reproduce CDCN++ as our baseline method (noted as CDCN++*). The best results in each protocol are in **bold**, and the second-best ones are <u>underlined</u>.

(ACER), Half Total Error Rate (HTER), and Area Under Curve (AUC). APCER, BPCER, and ACER are utilized for intra-dataset protocols, and HTER and AUC are used for cross-dataset evaluation.

## 4.2. Implementation Details

Given an input RGB sequence, we first detect and crop the face in each frame using RetinaFace [8], and then resize it to $256 \times 256$ before performing dense landmark prediction. A total of $N = 246$ landmarks around the eyes and mouth region (as shown in Fig. 3) are chosen as the input of GCN. The sequence sub-sampling length $S$ is set to 64. Random sub-sampling is adopted during training. At the inference stage, we choose the 64 frames by selecting the landmarks with the highest variances of position from the input sequence. Mirror padding along the temporal axis is applied to any sequence of length less than 64.

The GCN is composed of 6 spatio-temporal graph convolutional units followed by a linear layer. We train the GCN for 65 epochs with a batch size of 16. The optimizer is SGD with a momentum of 0.9 and a weight decay of 0.0001. The initial learning rate is set to 0.1, and it is decayed with a factor of 0.1 after the 50-th epoch. We apply random rotation and horizontal flip to each landmark sequence for augmentation. For cross-attention feature interaction, we obtain $f_p$ using a chosen single-frame photometrics-based method by averaging features extracted from the frames in an input sequence. We specify the adopted photometrics-based method in each experiment in Sec. 4.3. The cross-attention module is trained for a maximum of 200 epochs with a batch size of 64. The optimizer is SGD with a momentum of 0.9, a learning rate of 0.1, and a weight decay of 0.0005.

## 4.3. Comparison to SOTA Methods

### 4.3.1 Intra-Dataset Evaluation

For intra-dataset evaluation, we follow the protocols in OULU-NPU [2] to validate GAIN's generalization capacity under varied environments and presentation mediums. We adopt CDCN++ [48] as our baseline single-
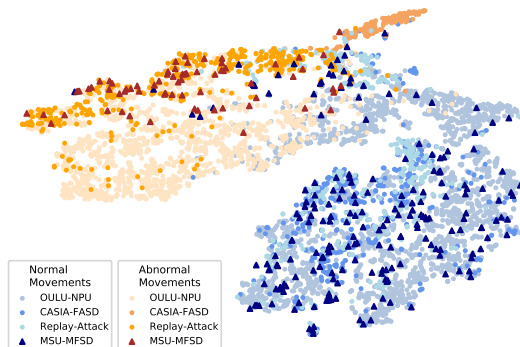


Figure 4. t-SNE visualization of the extracted geometric features by GAIN under the cross-dataset (O&C&I to M) setting. The distributions are well aligned between the testing dataset (M) and the training dataset (O&C&I).

frame photometrics-based method. Table 1 compares our CDCN++-GAIN fusion results with other FAS approaches, including those that also exploit temporal information. Combining learned geometric and extracted photometric features, GAIN significantly improves the performance of CDCN++ in all the four protocols of OULU-NPU. Moreover, our results outperform other FAS methods in most settings, demonstrating the efficacy of learning geometric facial dynamics for FAS.

| Method | O&C&I to M | | O&M&I to C | | O&C&M to I | | I&C&M to O | |
|---|---|---|---|---|---|---|---|---|
| | HTER(%) | AUC(%) | HTER(%) | AUC(%) | HTER(%) | AUC(%) | HTER(%) | AUC(%) |
| Auxiliary (Depth) [22] | 22.72 | 85.88 | 33.52 | 73.15 | 29.14 | 71.69 | 30.17 | 77.61 |
| NAS-FAS [47] | 19.53 | 88.63 | 16.54 | 90.18 | 14.51 | 93.84 | 13.80 | 93.43 |
| ANRL [20] | 10.83 | 96.75 | 17.85 | 89.26 | 16.03 | 91.04 | 15.67 | 91.90 |
| DRDG [21] | 12.43 | 95.81 | 19.05 | 88.79 | 15.56 | 91.79 | 15.63 | 91.75 |
| SDFANet [52] | <u>4.28</u> | 97.59 | 12.56 | 93.63 | **6.14** | 97.30 | 12.26 | 94.29 |
| HFN+MP [4] | 5.24 | 97.28 | 9.11 | <u>96.09</u> | 15.35 | 90.67 | 12.40 | 94.26 |
| VLAD-VSA (R) [35] | 4.29 | 98.25 | <u>8.76</u> | 95.89 | 7.79 | <u>97.79</u> | 12.64 | 94.00 |
| SSAN-R [36] | 6.67 | 98.75 | 10.00 | **96.67** | 8.88 | 96.79 | 13.72 | 93.63 |
| AMEL [53] | 10.23 | 96.62 | 11.88 | 94.39 | 18.60 | 88.79 | <u>11.31</u> | 93.96 |
| LMFD-PAD [9] | 10.48 | 94.55 | 12.50 | 94.17 | 18.49 | 84.72 | 12.41 | 94.95 |
| SSDG-R* [14] | 7.38 | 96.97 | 9.89 | 95.28 | 12.29 | 94.49 | 14.03 | 93.07 |
| SSDG-R* + **GAIN (Ours)** | **4.05** | <u>98.92</u> | **8.52** | 96.02 | 8.50 | 97.27 | 12.50 | <u>95.12</u> |
| PatchNet* [33] | 7.10 | 98.46 | 11.33 | 94.58 | 13.40 | 95.67 | 11.82 | 95.07 |
| PatchNet* + **GAIN (Ours)** | 4.29 | **99.45** | 10.67 | 95.13 | <u>7.29</u> | **98.01** | **9.74** | **96.75** |

Table 2. The results of the cross-dataset evaluation. We reproduce SSDG-R and PatchNet as our baseline methods (noted as SSDG-R* and PatchNet*). The best results in each protocol are in **bold**, and the second-best ones are <u>underlined</u>.

### 4.3.2 Cross-Dataset Evaluation

For cross-dataset evaluation, the following four datasets are used: OULU-NPU (denoted as O), CASIA-FASD (denoted as C), Replay-Attack (denoted as I), and MSU-MFSD (denoted as M). We follow the leave-one-out testing protocol in which one of the four datasets is selected for testing, and the remaining three are used for training (O&C&I to M, O&M&I to C, O&C&M to I, and I&C&M to O). The four datasets contain the same spoof types (live, print attack, and video-replay attack) while being collected under the variation of cameras, lighting conditions, resolutions, etc. This protocol indicates the models' generalization capability against unseen domains.

We adopt two baseline single-frame methods, SSDG-R [14] and PatchNet [33], and show the results in Table 2. GAIN leads to the substantial performance boost over both of the baseline methods and reaches state-of-the-art results in the four settings. This verifies that facial motion patterns provide information complementary to the RGB-photometric modality, contributing to performance improvement. The results also demonstrate the strong generalization capability of our geometric features. As shown as well in Fig. 4 by the t-SNE visualization [31], a close match can be observed between the distributions of features extracted from the training (O&C&I) and testing (M) datasets.

### 4.4. Ablation Studies

#### 4.4.1 Benefit of Dense Landmark Prediction

In our GAIN, dense landmark prediction is adopted to precisely capture fine-grained facial movements. In Fig. 5 we
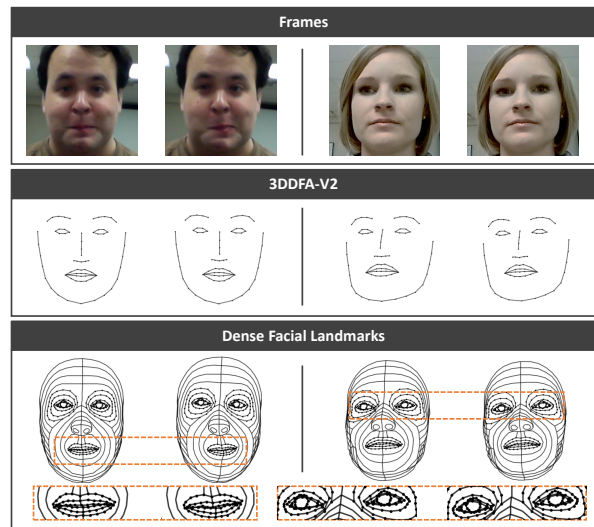


Figure 5. Comparison of sparse and dense landmark predictions. The sparse landmarks (68 on each face) are predicted by 3DDFA-V2 [10]. Details such as pursing lips (left) and rolling eyes (right) are better captured by the dense landmarks.

compare dense landmark predictions with sparse sets of 68 landmarks predicted by 3DDFA-V2 [10]. Dense landmarks show the superior ability to represent detailed movements such as pursing lips and rolling eyes. To quantitatively assess how using sparse and dense facial landmarks affect the learned geometric features, we integrate the features with the baseline photometrics-based method and report the performance in Table 3 (e - f). Geometric features learned from

| | Photometric | GCN Input | Attn | # of Cls | O&C&I to M | |
|---|---|---|---|---|---|---|
| | | | | | HTER(%) | AUC(%) |
| (a) | ✓ | ✗ | ✗ | 2 | 7.38 | 96.97 |
| (b) | ✗ | dense | ✗ | 2 | 35.95 | 65.92 |
| (c) | ✓ | dense | ✗ | 2 | 7.14 | 97.47 |
| (d) | ✓ | dense | ✗ | 3 | 4.29 | 98.76 |
| (e) | ✓ | dense | ✓ | 3 | **4.05** | **98.92** |
| (f) | ✓ | sparse | ✓ | 3 | 5.71 | 98.12 |

Table 3. Ablation study under the cross-dataset (O&C&I to M) setting. Note: *Photometric* indicates whether the baseline photometrics-based method (SSDG-R) is adopted; *GCN Input* specifies if the geometric features are learned from sparse or dense landmarks. *Attn* indicates whether cross-attention is adopted in the method; *# of Cls* indicates if common live/spoof binary classification or the proposed 3-class classification is performed.

| CASIA-SURF 3DMask | | |
|---|---|---|
| Method | HTER(%) | AUC(%) |
| ResNet50 [12] | 48.34 | 52.16 |
| Auxiliary [22] | 32.54 | 60.44 |
| DTN [23] | 38.97 | 69.24 |
| NAS-FAS [47] | 16.46 | 83.91 |
| NAS-FAS w/ T(Mask)-Meta [47] | 15.00 | 85.78 |
| **GAIN (ours)** | **6.23** | **96.04** |

Table 4. The results of the cross-dataset cross-type evaluation. All methods are first trained on OULU-NPU with live, paper attacks, and replay attacks, and then evaluated on 3DMask dataset with unseen 3D-mask attacks.

dense landmarks lead to higher performance in spoof detection.

### 4.4.2 Analysis of Geometric Features

It is important to note that the geometric features learned in our approach are not intended to address all possible liveness properties. Rather, their primary goal is to complement the common photometrics-based methods by providing critical temporal information. In Table 3 (b), we present the results obtained by using geometric features alone for detecting spoof attacks. A significant drop in performance can be observed as the geometric features alone are insufficient to identify certain types of spoof attacks that can produce realistic facial movements, such as video-replay attacks. In Table 3 (c), we combine geometric features with photometric features, enabling the model to leverage both sources of information for detecting spoof attacks. The result outperforms the photometrics-based baseline method (a).

While our learned geometric features alone may not be able to detect all types of spoof attacks, they are specifically suitable for spotting attacks that lack authentic fa-

cial dynamics and are quite robust. The CASIA-SURF 3DMask [47] (or 3DMask) dataset, which does not include any attacks featuring realistic facial movements, is particularly well-suited for evaluating the performance of our proposed geometric features. 3DMask consists of realistic 3D mask attacks produced with 3D printing and is collected under various challenging lighting conditions (indoor/outdoor, with back/front/side-light). The realistic presenting materials and complex environments make it hard for conventional FAS solutions to discriminate between live and spoof subjects based on single-frame information. We thus conduct the cross-dataset cross-type testing protocol proposed by the dataset to assess the ability of our geometric feature to detect unseen attacks with abnormal facial movements. Following the protocol, we first train GAIN using OULU-NPU dataset with live, paper attacks, and replay attacks, and then evaluate the model on the unseen dataset (3DMask) with unseen 3D-mask attacks. As shown in Table 4, our method significantly outperforms other photometrics-based FAS approaches. From this experiment, the robustness of our geometric feature learning is further justified.

### 4.4.3 Comparison with Standard Temporal-Based Methods

To exploit temporal information and focus on the *geometric* instead of photometric properties, GAIN utilizes GCN to extract geometric features on top of facial landmarks. In Table 5 we conduct two other experiments of abnormal movement detection, in which 3D-CNN (3D-ResNet-18 [11]) is adopted as the backbone, as a comparison to showcase the strength of extracting geometric information from facial landmarks. The first row shows the method using 64 raw RGB frames of aligned faces. The input is of shape $(S, C, H, W) = (64, 3, 64, 64)$, where $S$ is the sequence sub-sampling length, $C$ is the number of input channels, and $H$ and $W$ are the height and the width of a frame. The

| O&C&M to I Abnormal Movement Detection | |
|---|---|
| Method | AUC(%) |
| RGB + 3D-CNN | 65.66 |
| Flow + 3D-CNN | 92.28 |
| **GAIN (ours)** | **95.01** |

Table 5. Comparison with standard temporal-based methods under the cross-dataset (O&C&M to I) setting. The performance is evaluated in terms of abnormal movement detection. *RGB* and *Flow* refer to using raw RGB frames and optical flow as the input, respectively.

second row utilizes the frame-wise information extracted by the standard optical flow algorithm [18] to generate the input of shape $(64, 2, 64, 64)$. Specifically, the two channels represent the orientation and magnitude of the calculated flow field. The experiment shows that Flow 3D-CNN and GAIN provide significant improvements over the RGB 3D-CNN method, and GAIN reaches the best performance. This indicates that pre-defined dense landmarks dedicated to a face can be more suitable for the FAS task compared to the generic-purpose optical flow. Moreover, the geometrical relationships of GAIN bring stronger connectivity on the geometric and temporal domain.

#### 4.4.4 Feature Interaction Strategies

We examine the effectiveness of each design in our feature interaction strategy in Table 3 (c - e). The first row (c) shows the result of naively feeding concatenated features through linear layers and performing binary live/spoof classification. On the second row (d), by simply switching to the movement-aware 3-class classification (as described in Sec. 3.2), the performance improves remarkably (-3.09% in HTER and +1.26% in AUC). It shows that facial dynamic information plays a critical role in FAS, and exploiting such information during feature integration results in improved spoof detection. Replacing linear layers with a cross-attention module (e) further encourages the model to explore the relationship between geometric and photometric features, leading to the full performance of GAIN.

#### 4.4.5 GCN Visualization

Taking landmarks as inputs and training with the objective of movement classification, our GCN disentangles geometric temporal dynamics from texture information. In Fig. 6, GradCAM for GCN [25] is implemented to interpret the geometric feature learned by the GCN. We visualize the feature-node activations along the temporal axis from the last graph convolutional layer. In live subjects,
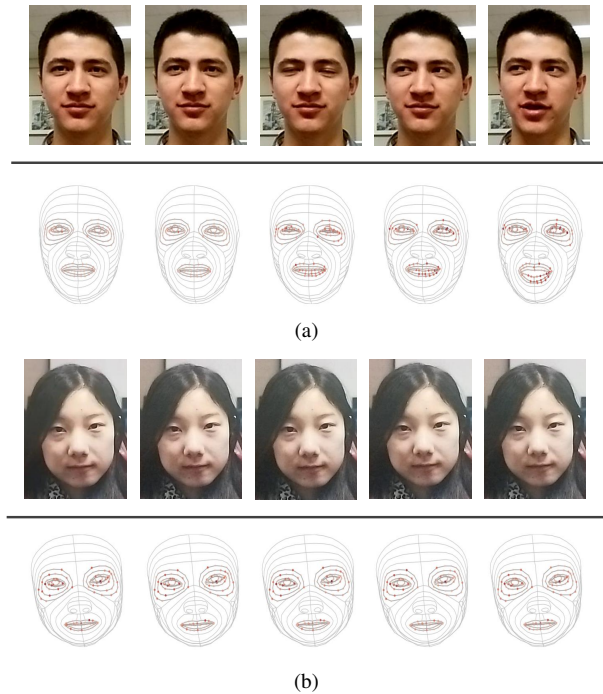


(a)



(b)

Figure 6. GradCAM Visualization for GCN. The feature-node activations are derived from the last layer of the GCN along the temporal axis from (a) a live subject, and (b) a print attack. The activation intensities are indicated by the tones of red color.

nodes tend to be more activated around the eye and mouth regions where facial motion can be observed at frames. Notably, activation intensities increase during movement transitions, such as raising eyebrows, moving eyeballs, and an open mouth. In contrast, similar intensities of attention on eye and mouth regions can be observed in the case of the print attack, in which the main motion is paper translation. The visualization once again verifies our proposed GAIN is capable of capturing facial movements discriminatively.

## 5. Conclusion

In this work, we aim to learn geometric temporal dynamics that capture discriminative facial movements for face anti-spoofing. To achieve this goal, we propose ***Geometry-Aware Interaction Network (GAIN)***, which adopts ST-GCN to extract fine-grained geometric facial dynamics from dense landmark predictions. The geometric features learned by GAIN can be easily integrated with other photometrics-based methods using our designed cross-attention feature interaction strategy. Extensive experiments on diverse testing protocols (intra-, cross-dataset, and cross-type evaluations) demonstrate that the geometric features of GAIN provide robust liveness information complementary to photometrics-based FAS methods, leading to a significant boost in performance.

# References

[1] Yousef Atoum, Yaojie Liu, Amin Jourabloo, and Xiaoming Liu. Face anti-spoofing using patch and depth-based cnns. In *2017 IEEE International Joint Conference on Biometrics (IJCB)*, pages 319–328. IEEE, 2017. 1, 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 *2017 12th IEEE international conference on automatic face & gesture recognition (FG 2017)*, pages 612–618. IEEE, 2017. 4, 5

[3] Adrian Bulat and Georgios Tzimiropoulos. How far are we from solving the 2d & 3d face alignment problem?(and a dataset of 230,000 3d facial landmarks). In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1021–1030, 2017. 2

[4] Rizhao Cai, Zhi Li, Renjie Wan, Haoliang Li, Yongjian Hu, and Alex C Kot. Learning meta pattern for face anti-spoofing. *IEEE Transactions on Information Forensics and Security*, 17:1201–1213, 2022. 6

[5] Haonan Chen, Guosheng Hu, Zhen Lei, Yaowu Chen, Neil M Robertson, and Stan Z Li. Attention-based two-stream convolutional networks for face spoofing detection. *IEEE Transactions on Information Forensics and Security*, 15:578–593, 2019. 1, 2

[6] Yuxin Chen, Ziqi Zhang, Chunfeng Yuan, Bing Li, Ying Deng, and Weiming Hu. Channel-wise topology refinement graph convolution for skeleton-based action recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13359–13368, 2021. 2

[7] Ivana Chingovska, André Anjos, and Sébastien Marcel. On the effectiveness of local binary patterns in face anti-spoofing. In *2012 BIOSIG-proceedings of the international conference of biometrics special interest group (BIOSIG)*, pages 1–7. IEEE, 2012. 4

[8] Jiankang Deng, Jia Guo, Evangelos Ververas, Irene Kotsia, and Stefanos Zafeiriou. Retinaface: Single-shot multi-level face localisation in the wild. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5203–5212, 2020. 5

[9] Meiling Fang, Naser Damer, Florian Kirchbuchner, and Arjan Kuijper. Learnable multi-level frequency decomposition and hierarchical attention mechanism for generalized face presentation attack detection. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 3722–3731, 2022. 6

[10] Jianzhu Guo, Xiangyu Zhu, Yang Yang, Fan Yang, Zhen Lei, and Stan Z Li. Towards fast, accurate and stable 3d dense face alignment. In *European Conference on Computer Vision*, pages 152–168. Springer, 2020. 2, 6

[11] Kensho Hara, Hirokatsu Kataoka, and Yutaka Satoh. Can spatiotemporal 3d cnns retrace the history of 2d cnns and imagenet? In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 6546–6555, 2018. 7

[12] 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. 7

[13] Yao-Hui Huang, Jun-Wei Hsieh, Ming-Ching Chang, Lipeng Ke, Siwei Lyu, and Arpita Samanta Santra. Multi-teacher single-student visual transformer with multi-level attention for face spoofing detection. In *BMVC*, 2021. 5

[14] Yunpei Jia, Jie Zhang, Shiguang Shan, and Xilin Chen. Single-side domain generalization for face anti-spoofing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8484–8493, 2020. 2, 6

[15] Qiuhong Ke, Mohammed Bennamoun, Senjian An, Ferdous Sohel, and Farid Boussaid. A new representation of skeleton sequences for 3d action recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3288–3297, 2017. 2

[16] Chao Li, Qiaoyong Zhong, Di Xie, and Shiliang Pu. Skeleton-based action recognition with convolutional neural networks. In *2017 IEEE International Conference on Multimedia & Expo Workshops (ICMEW)*, pages 597–600. IEEE, 2017. 2

[17] Haoliang Li, Wen Li, Hong Cao, Shiqi Wang, Feiyue Huang, and Alex C Kot. Unsupervised domain adaptation for face anti-spoofing. *IEEE Transactions on Information Forensics and Security*, 13(7):1794–1809, 2018. 2

[18] Ce Liu et al. *Beyond pixels: exploring new representations and applications for motion analysis*. PhD thesis, Massachusetts Institute of Technology, 2009. 8

[19] Jun Liu, Amir Shahroudy, Dong Xu, and Gang Wang. Spatio-temporal lstm with trust gates for 3d human action recognition. In *European conference on computer vision*, pages 816–833. Springer, 2016. 2

[20] 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 *Proceedings of the 29th ACM International Conference on Multimedia*, 2021. 6

[21] 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. *arXiv preprint arXiv:2106.16128*, 2021. 6

[22] Yaojie Liu, Amin Jourabloo, and Xiaoming Liu. Learning deep models for face anti-spoofing: Binary or auxiliary supervision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 389–398, 2018. 1, 2, 5, 6, 7

[23] Yaojie Liu, Joel Stehouwer, Amin Jourabloo, and Xiaoming Liu. Deep tree learning for zero-shot face anti-spoofing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4680–4689, 2019. 7

[24] Ziyu Liu, Hongwen Zhang, Zhenghao Chen, Zhiyong Wang, and Wanli Ouyang. Disentangling and unifying graph convolutions for skeleton-based action recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 143–152, 2020. 2

[25] Phillip E Pope, Soheil Kolouri, Mohammad Rostami, Charles E Martin, and Heiko Hoffmann. Explainability

methods for graph convolutional neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10772–10781, 2019. 8

[26] Suman Saha, Wenhao Xu, Menelaos Kanakis, Stamatios Georgoulis, Yuhua Chen, Danda Pani Paudel, and Luc Van Gool. Domain agnostic feature learning for image and video based face anti-spoofing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 802–803, 2020. 1, 2

[27] Amir Shahroudy, Jun Liu, Tian-Tsong Ng, and Gang Wang. Ntu rgb+ d: A large scale dataset for 3d human activity analysis. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1010–1019, 2016. 2

[28] Rui Shao, Xiangyuan Lan, Jiawei Li, and Pong C Yuen. Multi-adversarial discriminative deep domain generalization for face presentation attack detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10023–10031, 2019. 2

[29] Lei Shi, Yifan Zhang, Jian Cheng, and Hanqing Lu. Two-stream adaptive graph convolutional networks for skeleton-based action recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12026–12035, 2019. 2

[30] Tae Soo Kim and Austin Reiter. Interpretable 3d human action analysis with temporal convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 20–28, 2017. 2

[31] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008. 6

[32] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 4

[33] Chien-Yi Wang, Yu-Ding Lu, Shang-Ta Yang, and Shang-Hong Lai. Patchnet: A simple face anti-spoofing framework via fine-grained patch recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20281–20290, 2022. 2, 6

[34] Guoqing Wang, Hu Han, Shiguang Shan, and Xilin Chen. Improving cross-database face presentation attack detection via adversarial domain adaptation. In *2019 International Conference on Biometrics (ICB)*, pages 1–8. IEEE, 2019. 2

[35] Jiong Wang, Zhou Zhao, Weike Jin, Xinyu Duan, Zhen Lei, Baoxing Huai, Yiling Wu, and Xiaofei He. Vlad-vsa: Cross-domain face presentation attack detection with vocabulary separation and adaptation. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 1497–1506, 2021. 6

[36] Zhuo Wang, Zezheng Wang, Zitong Yu, Weihong Deng, Jiahong Li, Tingting Gao, and Zhongyuan Wang. Domain generalization via shuffled style assembly for face anti-spoofing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4123–4133, 2022. 2, 6

[37] Zhuming Wang, Yaowen Xu, Lifang Wu, Hu Han, Yukun Ma, and Guozhang Ma. Multi-perspective features learning for face anti-spoofing. In *Proceedings of the IEEE/CVF*

*International Conference on Computer Vision*, pages 4116–4122, 2021. 1, 2

[38] Zezheng Wang, Zitong Yu, Chenxu Zhao, Xiangyu Zhu, Yunxiao Qin, Qiusheng Zhou, Feng Zhou, and Zhen Lei. Deep spatial gradient and temporal depth learning for face anti-spoofing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5042–5051, 2020. 1, 2

[39] 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. 4

[40] Erroll Wood, Tadas Baltrušaitis, Charlie Hewitt, Sebastian Dziadzio, Thomas J Cashman, and Jamie Shotton. Fake it till you make it: face analysis in the wild using synthetic data alone. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 3681–3691, 2021. 2

[41] Erroll Wood, Tadas Baltrusaitis, Charlie Hewitt, Matthew Johnson, Jingjing Shen, Nikola Milosavljevic, Daniel Wilde, Stephan Garbin, Chirag Raman, Jamie Shotton, Toby Sharp, Ivan Stojiljkovic, Tom Cashman, and Julien Valentin. 3d face reconstruction with dense landmarks, 2022. 2, 3

[42] Yaowen Xu, Zhuming Wang, Hu Han, Lifang Wu, and Yongluo Liu. Exploiting non-uniform inherent cues to improve presentation attack detection. In *2021 IEEE International Joint Conference on Biometrics (IJCB)*, pages 1–8. IEEE, 2021. 1, 2, 5

[43] Sijie Yan, Yuanjun Xiong, and Dahua Lin. Spatial temporal graph convolutional networks for skeleton-based action recognition. In *Thirty-second AAAI conference on artificial intelligence*, 2018. 2, 3

[44] Jianwei Yang, Zhen Lei, and Stan Z Li. Learn convolutional neural network for face anti-spoofing. *arXiv preprint arXiv:1408.5601*, 2014. 1, 2

[45] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3507–3516, 2019. 1, 2

[46] Zitong Yu, Yunxiao Qin, Hengshuang Zhao, Xiaobai Li, and Guoying Zhao. Dual-cross central difference network for face anti-spoofing. *arXiv preprint arXiv:2105.01290*, 2021. 1, 2

[47] 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. *IEEE transactions on pattern analysis and machine intelligence*, 43(9):3005–3023, 2020. 4, 5, 6, 7

[48] Zitong Yu, Chenxu Zhao, Zezheng Wang, Yunxiao Qin, Zhuo Su, Xiaobai Li, Feng Zhou, and Guoying Zhao. Searching central difference convolutional networks for face anti-spoofing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5295–5305, 2020. 1, 2, 5

[49] 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 *European Conference on Computer Vision*, pages 641–657. Springer, 2020. 1, 2

[50] Songyang Zhang, Xiaoming Liu, and Jun Xiao. On geometric features for skeleton-based action recognition using multilayer lstm networks. In *2017 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 148–157. IEEE, 2017. 2

[51] Zhiwei Zhang, Junjie Yan, Sifei Liu, Zhen Lei, Dong Yi, and Stan Z Li. A face antispoofing database with diverse attacks. In *2012 5th IAPR international conference on Biometrics (ICB)*, pages 26–31. IEEE, 2012. 4

[52] Lifang Zhou, Jun Luo, Xinbo Gao, Weisheng Li, Bangjun Lei, and Jiaxu Leng. Selective domain-invariant feature alignment network for face anti-spoofing. *IEEE Transactions on Information Forensics and Security*, 16:5352–5365, 2021. 6

[53] Qianyu Zhou, Ke-Yue Zhang, Taiping Yao, Ran Yi, Shouhong Ding, and Lizhuang Ma. Adaptive mixture of experts learning for generalizable face anti-spoofing. *arXiv preprint arXiv:2207.09868*, 2022. 6

[54] Wentao Zhu, Cuiling Lan, Junliang Xing, Wenjun Zeng, Yanghao Li, Li Shen, and Xiaohui Xie. Co-occurrence feature learning for skeleton based action recognition using regularized deep lstm networks. In *Proceedings of the AAAI conference on artificial intelligence*, 2016. 2

[55] Xiangyu Zhu, Xiaoming Liu, Zhen Lei, and Stan Z Li. Face alignment in full pose range: A 3d total solution. *IEEE transactions on pattern analysis and machine intelligence*, 41(1):78–92, 2017. 2