Learning Multi-Granularity Temporal Characteristics for Face Anti-Spoofing

Zhuo Wang, Qiangchang Wang, Weihong Deng, *Member, IEEE*, and Guodong Guo, *Senior Member, IEEE*

***Abstract*— Face anti-spoofing (FAS) is essential for securing face recognition systems. Despite the decent performance, few existing works fully leverage temporal information. This would inevitably lead to inferior performance because real and fake faces tend to share highly similar spatial appearances, while important temporal features between consecutive frames are neglected. In this work, we propose a temporal transformer network (TTN) to learn multi-granularity temporal characteristics for FAS. It mainly consists of temporal difference attentions (TDA), a pyramid temporal aggregation (PTA), and a temporal depth difference loss (TDL). Firstly, the vision transformer (ViT) is used as the backbone where comprehensive local patches are utilized to provide subtle differences between live and spoof faces. Then, instead of learning temporal features on global faces which may miss some important local cues, the TDA is developed to extract motion-sensitive cues on each of the comprehensive local patches. Moreover, the TDA is inserted into different layers of the ViT, learning multi-scale motion-sensitive local cues to improve the FAS performance. Secondly, it is observed that different subjects may have different visual tempos in some actions, making it necessary to model different temporal speeds. Our PTA aggregates temporal features at various tempos, which could build short-range and long-range relations among multiple frames. Thirdly, depth maps for real parts may change continuously, while they remain zeros for spoof regions. In order to locate motion features on facial parts, the TDL is proposed to guide the network to locate spoof facial parts where motion patterns between neighboring frames are set as the ground truth. To the best of our knowledge, this work is the first attempt to learn temporal characteristics via transformers. Both qualitative and quantitative results on several challenging tasks demonstrate the usefulness and effectiveness of our proposed methods.**

Manuscript received October 4, 2021; revised January 12, 2022 and March 1, 2022; accepted March 2, 2022. Date of publication March 8, 2022; date of current version March 25, 2022. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Fernando Alonso-Fernandez. (Zhuo Wang and Qiangchang Wang are co-first authors.) *(Corresponding author: Guodong Guo.)*

Zhuo Wang was with the Institute of Deep Learning (IDL), Baidu Research, Beijing 100085, China. He is now with the Pattern Recognition and Intelligent System Laboratory, School of Artificial Intelligence, Beijing University of Posts and Telecommunications, Beijing 100876, China (e-mail: wz2019@bupt.edu.cn).

Weihong Deng is with the Pattern Recognition and Intelligent System Laboratory, School of Artificial Intelligence, Beijing University of Posts and Telecommunications, Beijing 100876, China (e-mail: whdeng@bupt.edu.cn).

Qiangchang Wang was with the Institute of Deep Learning (IDL), Baidu Research, Beijing 100085, China. He is now is with the Lane Department of Computer Science and Electrical Engineering, West Virginia University, Morgantown, WV 26506 USA (e-mail: qiangchang.wang@gmail.com).

Guodong Guo is with the Lane Department of Computer Science and Electrical Engineering, West Virginia University, Morgantown, WV 26506 USA, and also with the Institute of Deep Learning, Baidu Research, Beijing 100085, China (e-mail: guoguodong01@baidu.com).

Digital Object Identifier 10.1109/TIFS.2022.3158062

***Index Terms*— Face anti-spoofing, temporal difference attentions, pyramid temporal aggregation, temporal depth difference loss, transformer networks.**

I. Introduction

FACE recognition (FR) has achieved a series of breakthroughs [1]–[3] in recent years which has greatly benefited various applications, such as mobile phone login and airport check-in. Although great success has been achieved, FR models may be spoofed by various presentation attacks (PAs), including print photo, video replay, and 3D mask. Consequently, fields with high safety requirements are subject to attack risks which may lead to a dramatic loss. For example, an attacker could bypass a biometrics model in the financial institution to steal money. To secure FR systems, face anti-spoofing (FAS) [4] is developed to detect PAs which is receiving increasing attention from both academia and industry.

Recent CNN-based methods [5], [6] extract spatial features to discriminate between live and spoof faces. Although these approaches obtain significant improvements, they ignore temporal information which is important for FAS. Because real and fake faces tend to share very similar spatial appearances when presented only on a single frame. Consequently, complementary information between neighboring frames is neglected, resulting in the loss of important temporal information.

Previous methods utilize temporal cues to detect face liveness, such as eye-blinking, lip or mouth movements [7], [8]. These methods are robust to paper attacks, but vulnerable to replay attacks or print attacks with eye/mouth cut. Therefore, it is necessary to extract facial motion features on the whole face, instead of limited small areas. Some traditional methods achieve this by concatenating features from continuous frames [9], [10] or proposing temporal-specific features, such as Haralick features [11], HOOF [12], and optical flow [13]. However, these methods suffer from poor generalization due to the inferior representational ability. With the development of deep learning, several works apply it to model temporal features. The optical flow map and Shearlet image features are extracted [14]. Long-relation information is captured by using RNNs as temporal structures [15], [16]. The 3D convolution is used to learn spatiotemporal features of consecutive frames [17], [18]. Two-stream networks are proposed to model face temporal features [19], [20]. The rPPG signal is also explored [16], [21]. While these works can learn temporal information, they achieve this by directly learning on whole faces. However, many important cues may lay in local regions. To address this issue, some methods [22]–[26] explore

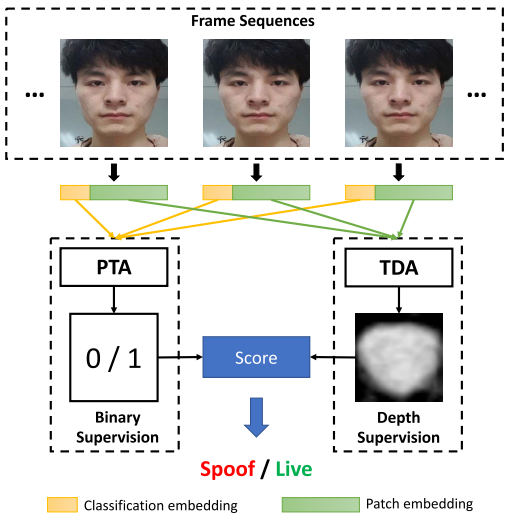


Fig. 1. The pipeline of learning multi-granularity temporal characteristics for FAS. Special information is extracted from single frames, including global classification embeddings and local patch embeddings. Then, those features are transmitted to the PTA and TDA for a more accurate binary and depth supervision. Thus, their mixed score is utilized for the final decision.

spatial features from partial patches for more local details. Nevertheless, we argue that some important local patches may appear at arbitrary locations for faces with pose variations, material changes, and different illumination. Therefore, it is expected to capture comprehensive local features from all patches along with the deepening of the network. Moreover, due to the subtle motion in local regions for FAS, it is expected that motion-sensitive cues to separate living and spoofing lie in local patches.

Combined with the abovementioned viewpoints, a temporal transformer network (TTN) is proposed to extract rich multi-granularity temporal information as illustrated in Fig. 1, which consists of temporal difference attentions (TDA), a pyramid temporal aggregation (PTA), and a temporal depth difference loss (TDL). Firstly, spatial information about each local patch in a frame is extracted via vision transformers (ViT) [27]. Different from the image-based structure CNNs, patch-based ViT utilizes the relation-based mechanism to exploit local fine-grained information for liveness detection. Then, to capture subtle motion-sensitive cues, our proposed TDA learns temporal attention on each local patch between neighboring frames, emphasizing motion-relative channels and suppressing motion-unrelated ones. Since every local patch is explored thoroughly, important motion-sensitive cues are less likely to be neglected. Besides, to explore complementary hierarchical information, our TDA is inserted into different transformer encoders. In such a way, fine-grained spatiotemporal information in low layers and abstract spatiotemporal representations in high layers are captured simultaneously to benefit the FAS. Secondly, when presented in facial videos, different subjects may have different temporal speeds because of various factors, such as mood. For example, an excited live face tends to blink faster than a sad live face. Since short-range and long-range relations among multiple frames are complementary, it is necessary to model variances in the temporal speed. To achieve this goal, a pyramid temporal aggregation (PTA) is developed to aggregate global face representations from multiple frames via shifted windows. Thirdly, depth maps may continuously change over time for live face regions, while they remain zero for spoof facial parts. Thus, a temporal depth difference loss (TDL) is proposed to supervise the depth changes between adjacent frames, which is also beneficial to distinguish the motion regions from static backgrounds. To the best of our knowledge, this work is the first effort to apply transformer networks to learn temporal information for FAS. Qualitative and quantitative results on several challenging tasks demonstrate the effectiveness of our proposed modules.

The main contributions of this work are four-fold:

1. Temporal difference attentions (TDA) are proposed to hierarchically learn temporal cues on comprehensive local patches, capturing coarse-to-fine motion-sensitive cues to distinguish the bona fide and presentation attacks.
2. Pyramid temporal aggregation (PTA) is developed to cover different temporal speeds on multiple frames, building rich short-range and long-range connections between different frames to utilize their complementarity for liveness detection.
3. With depth maps between adjacent frames as the ground truth, a temporal depth difference loss (TDL) is designed to guide networks to learn relative changes of depth information in temporal sequences, further boosting the FAS performance.
4. With these modules, our model outperforms the state-of-the-art methods on various benchmark tests.

The rest of this paper is organized as follows. In section II, we provide a brief review regarding the related works on face anti-spoofing and transformer networks. In section III, the whole framework of the temporal transformer network (TTN) is introduced. Then, experimental results and visual analysis are shown in Section IV. Section V concludes this paper.

II. Related Works

A. Face Anti-Spoofing

As the development of deep learning in image processing, several methods apply CNNs to learn discriminative features [5], [14], [16], [23], [28]–[30]. Recently, [31] employs the popular transformer network for zero-shot FAS. However, these works do not consider the discriminative temporal information.

Temporal-based methods in the early time focus on some easily observable movements [6], [7], such as eye-blinking, opening and closing the mouth. Those methods can effectively identify some print attacks, but become vulnerable to replay attacks, 3D mask attacks, or print attacks with eye/mouth cuts. Besides, some methods [11]–[13] also make a distinction by capturing temporal-specific features. For example, [12] adopts multifeature videolet aggregation of multi-LBP and HOOF for liveness detection. However, these traditional feature descriptors fail to exploit the comprehensive texture and motion cues, due to their limited representational abilities.

Recent temporal-based methods can be grouped into three categories. One common strategy is to capture temporal

information via recurrent neural networks. Feng *et al.* [15] adopts an LSTM to learn temporal features. Xu *et al.* [16] leverages the rPPG signal by a CNN-RNN. One alternative is to use the two-stream network. [19] introduces the temporal shift module for motion patterns. Besides, 3D CNNs are employed to extract spatiotemporal features on continuous frames [17], [18]. However, most of these methods extract temporal features on global images, but fail to fully explore motion-sensitive cues in patch levels.

To alleviate this issue, [22], [23] capture local features on random patches in face images. Atoum *et al.* [24] utilizes deep reinforcement learning to find suspicious sub-patches for FAS. Cai *et al.* [25] destroys the global structure of images into patches to make the network concentrate on local details. However, most existing patch-wise methods only explore spatial features from partial patches, which may miss some important patches due to the changes of external factors. Moreover, almost all these methods are conducted on a single frame, which may cause temporal relations to be not exploited thoroughly.

Therefore, in our work, the spatial transformer is employed to explore the comprehensive spatial information from all patches, while the temporal transformer is used to integrate continuous frames for temporal information extraction. Moreover, different from FAS-SGTD [32] with a two-stage training, our method utilizes patch embeddings for depth supervision, and classification embeddings for binary supervision simultaneously in an end-to-end fashion, which is more suitable for large-scale training.

B. Transformer Networks

Transformers [33] have been widely used in natural language processing (NLP). Recently, they have shown great potential in multiple computer vision (CV) tasks. A CNN and a transformer were combined for object detection [34]. A pure transformer network was applied for image classification [27]. The segmentation problem was reformulated from a sequence-to-sequence perspective [35]. A hierarchical architecture with shifted windows was utilized to reduce the calculation, obtaining excellent performances on several CV tasks [36].

Transformers are also utilized to extract temporal information in video-related tasks. A temporal transformer network was proposed for general time-series classification in [37]. A spatiotemporal transformer was utilized for visual tracking in [38]. Pure transformers were proposed for video classification, capturing spatial and temporal information simultaneously in [39]. The seminal idea of multi-scale feature hierarchies was combined with transformer models, outperforming concurrent vision transformers on video recognition [40]. To the best of our knowledge, this work is the first attempt to employ transformer networks to learn temporal information for FAS.

III. Methodology

In this section, for a comprehensive understanding, a single-frame structure based on transformers is firstly introduced to FAS. Then, to capture temporal information between frame sequences, the aforementioned single-frame structure is extended to a multi-frame one by adding temporal transformer layers. Furthermore, temporal difference attentions (TDA) are proposed to learn the temporal difference and promote depth estimation. A pyramid temporal aggregation (PTA) is developed to learn multi-scale temporal information based on the global information of each frame. Lastly, the proposed temporal depth difference loss (TDL) and overall loss are integrated for reliable training and great generalization capability.

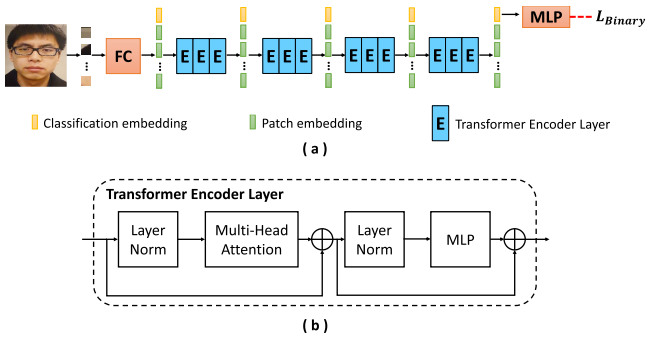


Fig. 2. (a) Illustration of vision transformer structure for single-frame FAS. Position embeddings are omitted in our diagram, because they have been added to patch and classification embeddings. (b) The detailed structure of the transformer encoder layer.

A. Transformer Network for Face Anti-Spoofing

To explore relations between local patches in bona fide and presentation attacks, vision transformer (ViT) [27] is utilized for spatial feature extraction. As shown in Fig. 2 (a), an input image x is firstly split into $N \times N$ non-overlapped patches. After linear projections, these patches are embedded into 1D patch embeddings $z_i \in R^{1\times C}, 1 \leq i \leq L$ and integrated as $z_p \in R^{L\times C}$, where $L = N^2$ is the number of split patches and C represents the embedding dimension. Then, learned classification embedding $z_{cls} \in R^{1\times C}$ and position embedding $z_{pos} \in R^{(L+1)\times C}$ are set to capture classification information and retain positional information, respectively. Therefore, the sequence of feature embeddings $z \in R^{(L+1)\times C}$ is input to the following transformer encoder layers as follows:

$$\begin{aligned} z &= concat(z_{cls}, z_p) + z_{pos}, \\ z_p &= [z_1, z_2, \ldots, z_L]. \end{aligned} \tag{1}$$

A transformer encoder layer consists of multi-head self-attention (MSA) and multi-layer perception (MLP), as shown in Fig. 2 (b). The relations between patch and classification embeddings are calculated continuously along with the cascaded transformer encoder layers. Therefore, feature embeddings of adjacent layers can be expressed as follows:

$$\begin{aligned} y^l &= MSA(LN(z^l)) + z^l, \\ z^{l+1} &= MLP(LN(y^l)) + y^l, \end{aligned} \tag{2}$$

where $z^l \in R^{(N+1)\times C}$ is the input of l-th encoder layer, y^l is the temporal variable, and MLP is composed of two linear projections with GELU as the activation function.

Finally, the classification embedding z_{cls} contains sufficient classification information. A linear classifier with a softmax layer is used to decode the classification information for the final decision between live and spoof faces.

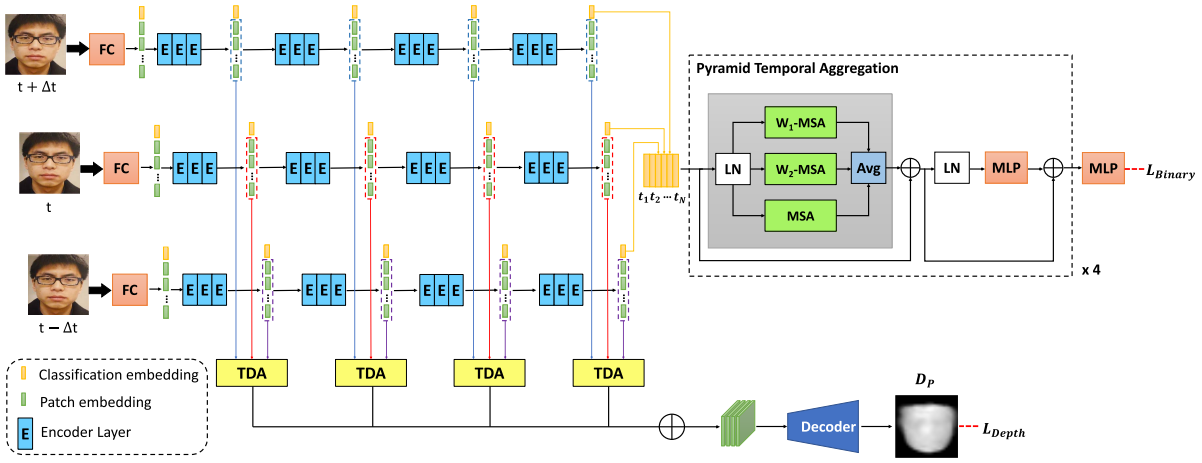


Fig. 3. The overall architecture of our temporal transformer network (TTN). Firstly, N_f successive frames are fed into transformer layers for extracting spatial features that contain global classification embeddings and local patch embeddings. Then, classification embeddings of N_f frames are transmitted into the pyramid temporal aggregation (PTA) to integrate spatiotemporal information for binary classification. Additionally, to promote the generalization capacity, temporal difference attentions (TDA) are proposed to enhance motion-related channels on each local patch for precise depth estimation. Lastly, the temporal depth difference loss (L_{TDL}) is used to model depth motion effectively. Thus, depth loss ($L_{\text{MSE}} + L_{\text{TDL}}$) and binary loss ($L_{\text{Binary}}$) are employed simultaneously to supervise the model learning.

Compared with CNNs, transformers have the following advantages for spatial feature extraction: 1) Different from directly encoding the whole images, transformers encode the separated patches. It can promote the network to describe fine-grained information and eliminate adverse effects of padding operation on subtle information; 2) Relation-aware mechanism is adopted to exploit intrinsic textures and capture long-range spatial features. Specifically, self-attention layers are applied to exploit relations between different patches, then weigh them continuously; 3) Transformers have global receptive fields through all layers, which provides a larger learning capability.

B. Temporal Transformer Network

Although the single-frame transformer can capture long-range spatial features by exploiting relations between local patches, temporal information in continuous frames is omitted, which contains sufficient fine-grained information based on short-term and long-term moving patterns. Therefore, we expand the single-frame model to the multi-frame one with some effective modules and build our new architecture, *i.e.*, the temporal transformer network (TTN), as shown in Fig. 3.

The TTN mainly consists of a spatial transformer network, a temporal transformer network, and a depth estimation. As introduced in III-A, continuous N_f face images $\{x(t_i)\}_{i=1}^{N_f}$ are fed into the network and their corresponding feature embeddings are extracted by a cascade of weight shared transformers. Feature embeddings $\{z(t_i)\}_{i=1}^{N_f}$ consist of classification embeddings $\{z_{cls}(t_i)\}_{i=1}^{N_f}$ and patch embeddings $\{z_p(t_i)\}_{i=1}^{N_f}$. Specifically, the former contains important global classification cues, while the latter describes local patches. Therefore, patch embeddings of each frame are utilized to predict depth maps with the assistance of temporal difference attentions (TDA) for depth supervision. Classification embeddings of each frame are passed into a pyramid-based temporal transformer structure, called pyramid temporal aggregation (PTA), to capture multi-scale temporal characteristics for binary classification. Depth supervision and binary supervision are used commonly to obtain an excellent generalization capability for FAS.

C. Temporal Difference Attentions

Temporal difference attentions (TDA) are designed to utilize temporal information in adjacent frames, enhancing motion-sensitive channels of patch embeddings for depth estimation.

Since depth supervision is proposed by [23], it has been widely used in FAS due to its great capability of extracting fine-grained details. These detailed cues in local regions are vital and effective for distinguishing between bona fide and presentation attacks. Specifically, pseudo depth maps are generated as supervisory signals to enlarge differences between living and spoofing. In training, these pseudo depth maps are set as the ground truth, which guides our networks to learn useful information. In testing, predicted depth maps constitute an important component of final decision scores.

An image is split into patches, then patch embeddings $\{z_p(t_i)\}_{i=1}^{N_f} \in R^{L\times C}$ about local patches are captured by a spatial transformer network. Patch embeddings can be utilized for depth estimation with the guidance of depth supervision. However, the dimension C of patch embedding z_p is usually too large to precisely focus on critical channels that are beneficial to depth estimation. Therefore, it is necessary to enhance vital channels of patch embeddings that encode motion information by utilizing the short-term temporal differences between adjacent frames. The above judgment is based on the following facts: 1) For a patch embedding, partial channels tend to encode motion information which is related to dynamic parts, while a part of ones encodes static information related to silent regions; 2) In living videos, dynamic regions such as face parts usually contain depth information, but silent regions such as background parts contain less; 3) For attack videos, subtle spoofing cues can be better exploited in channels that model motion information. To sum up, the

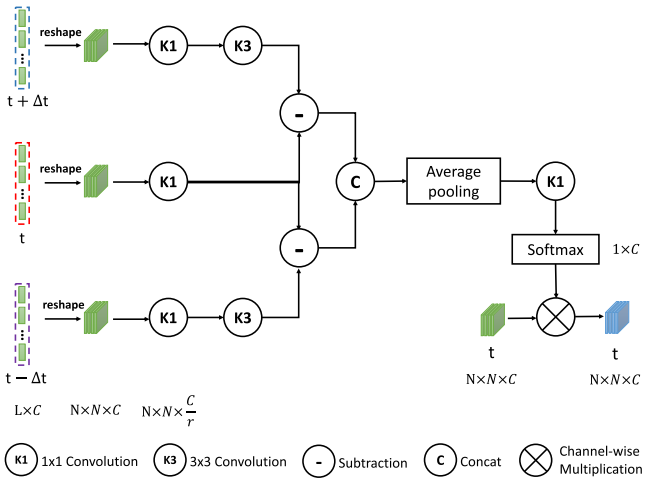


Fig. 4. Illustration of temporal difference attentions (TDA).

TDA is proposed to explore the temporal difference between local regions of adjacent frames and assist networks to utilize short-term motion information for precise depth estimation.

This TDA can be illustrated in Fig. 4. The patch embedding $z_p(t)$ at time t and that of two adjacent frames $\{z_p(t-\Delta t), z_p(t+\Delta t)\}$ are utilized to accurately estimate the depth map $D(t)$ at time t. Firstly, we convert the patch embeddings $\{z_p(t_i)\}_{i=1}^{N_f} \in R^{L\times C}$ to the 2D structure $\{f_p(t_i)\}_{i=1}^{N_f} \in R^{N\times N\times C}$. Next, a 1×1 2D convolution layer is adopted to reduce feature channels for efficient calculation and to obtain $f_p^r \in R^{N\times N\times C/r}$. Then, to capture the short-term differences between frame $x(t)$ and its two adjacent frames $\{x(t-\Delta t), x(t+\Delta t)\}$, the spatial features of extended receptive field in the two adjacent frames are extracted by a 3×3 2D convolution layer and a subtraction is conducted between them to obtain temporal differences. This process can be formally formulated as follows:

$$\begin{aligned} Diff(t, t-\Delta t) &= f_p^r(t) - K_3 \otimes f_p^r(t-\Delta t), \\ Diff(t, t+\Delta t) &= f_p^r(t) - K_3 \otimes f_p^r(t+\Delta t), \end{aligned} \tag{3}$$

where $\{Diff(t, t-\Delta t), Diff(t, t+\Delta t)\} \in R^{L\times C}$ are the temporal difference at time t. K_3 is a 3×3 2D convolution layer. $\otimes$ denotes the convolution operation.

Furthermore, we concatenate the temporal differences between frame $x(t)$ and adjacent frames $\{x(t-\Delta t), x(t+\Delta t)\}$:

$$Diff_{con}(t) = concat[Diff(t, t-\Delta t), Diff(t, t+\Delta t)], \tag{4}$$

where $Diff_{con}(t) \in R^{N\times N\times 2C/r}$. Then, a global average pooling layer is utilized to summarize the differences:

$$P(t) = avgPool[Diff_{con}(t)], \quad P(t) \in R^{1\times 2C/r}. \tag{5}$$

Another 1×1 2D convolution layer recovers the pooling difference $P(t)$ to the origin channel dimension C and the attention weights $A(t)$ can be obtained as follows:

$$\begin{aligned} P_{ori}(t) &= K_1 \otimes P(t), \quad P_{ori} \in R^{1\times C}, \\ A(t) &= \sigma(P_{ori}), \quad A(t) \in R^{1\times C}, \end{aligned} \tag{6}$$

where $\otimes$ is the convolution operation and σ denotes the sigmoid function.

Finally, to excite the motion-sensitive channels that contain vital depth information, we conduct a channel-wise multiplication between the original patches feature map $f_p(t)$ and the attention weight $A(t)$ at time t to obtain the enhanced patches feature map $f_o(t) \in R^{N\times N\times C}$, as follows:

$$f_o(t) = A(t) \cdot f_p(t). \tag{7}$$

On the other hand, there exists complementarity between patch embeddings from different layers. For example, in lower layers, most patches are correlated with surrounding areas, which would capture low-level location spatial details. In higher layers, most patches can correlate with larger regions, which would capture more high-level semantic information. Therefore, to fully exploit their complementarity as well as balance calculation costs, we collect patch embeddings of every three layers from the frame at t, represented as $\{z_p^{3i}(t)\}_{i=1}^4 \in R^{L\times C}$. Then after the channel weighting operation of TDA, we obtain the weighted patch feature maps $\{f_o^{3i}(t)\}_{i=1}^4 \in R^{N\times N\times C}$. Different from the concatenation in CNN-based methods [16], we fuse these features by using the addition operation according to the experimental validation. Therefore, the final patch features of frame $x(t)$ that are weighted by TDA and fused from different layers are represented as $f_{final}(t)$ as follows:

$$f_{final}(t) = \sum_{i=1}^{4} f_o^{3i}(t), \quad f_{final}(t) \in R^{N\times N\times C}. \tag{8}$$

Lastly, as shown in Fig. 3, the predicted depth map $D_p(t)$ of frame $x(t)$ is estimated after a Fully Convolutional Network (FCN) decoder. For our N_f-frame input $\{x(t_i)\}_{i=1}^{N_f}$, the corresponding predicted depth maps $\{D_p(t_i)\}_{i=1}^{N_f}$ are obtained, which can promote the reliability for depth supervision.

D. Pyramid Temporal Aggregation

Pyramid temporal aggregation (PTA) is developed with a pyramid structure to capture multi-scale temporal characteristics that contain short-term as well as long-term relations.

The classification embeddings $\{z_{cls}(t_i)\}_{i=1}^{N_f}$ contain global features of frames $\{x(t_i)\}_{i=1}^{N_f}$. Different from being directly used for classification, these N_f-embeddings are firstly integrated as $z_{cls} \in R^{N_f\times C}$, then go through the pyramid-like transformer structure PTA to further learn temporal information implied in consecutive frames with a binary supervision.

As shown in Fig.3, the major difference between the PTA and basic transformer layer in Fig. 2 (b), is that the basic transformer layer contains single multi-head self-attention (MSA), while our PTA has multiple MSAs. Specifically, besides the global attention module with a temporal receptive field, PTA contains two additional window attention modules with local receptive fields. Therefore, Eqn. (2) is replaced as follows:

$$\begin{aligned} y^l &= \{W_1\text{-}MSA, W_2\text{-}MSA, MSA\}(LN(z^l)) + z^l, \\ z^{l+1} &= MLP(LN(y^l)) + y^l, \end{aligned} \tag{9}$$

where $z^l \in R^{N_f\times C}$ is the input of l-th PTA layer. y^l is the intermediate variable. $\{W_1\text{-}MSA, W_2\text{-}MSA\}$ are MSAs whose receptive sizes are $w_1 = 2$ and $w_2 = \frac{N_f}{2}$.

The window-based self-attention module splits the classification embeddings of N_f-frames into $\lfloor \frac{N_f}{w} \rfloor$ non-overlapped

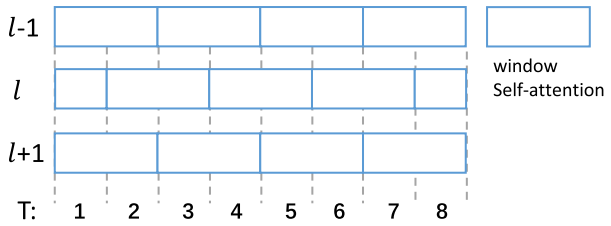


Fig. 5. Illustration of shifted window partitioning between adjacent temporal transformer layers. For example, when the sequence length N_f is 8 and the size of the window self-attention w is 2, the shifting step is $\lfloor \frac{N_f}{w} \rfloor = 1$ between the l-th layer and its adjacent ones.

windows. This partition is beneficial to separate the successive sequence into multiple independent slices for temporal information exploration. However, the lack of connections across windows may limit its information integration capability. Inspired by [36], a shifted window partitioning approach is utilized along with our temporal transformer layers. Specifically, the windows with the size of w constitute new windows by shifting $\frac{w}{2}$ step length across the adjacent temporal transformer layers, as shown in Fig. 5. Thus, the information interactions between different windows are obtained.

To sum up, the pyramid structure with multiple receptive fields is beneficial for multi-scale temporal information integration. Specifically, $\{W_1\text{-}MSA, W_2\text{-}MSA, MSA\}$ can independently concentrate on short-term, middle-term, and long-term temporal cues, respectively. Therefore, this pyramid structure provides rich modeling variances in temporal speed for FAS. Besides, transformers have the ability of parallel computing, thus are more efficient in processing temporal sequences, which is beneficial for temporal feature extraction.

E. Loss Function

Besides developing a novel temporal network architecture, we also propose a temporal loss function for network training. To guide the network to learn motion depth effectively, we develop the temporal depth difference loss (L_{TDL}) to supervise depth changes between adjacent frames.

1) Temporal Depth Difference Loss: Previous depth-based works usually use Euclidean Distance Loss (L_{EDL}) to supervise a single frame in a pixel-wise way. Intuitively, this approach merely assists the network to learn absolute distances from objects to the camera. However, the distance changes between adjacent frames are also important for depth supervision. Because for live faces, the depth maps may persistently change over time, while for spoof faces, such as print and replay attacks, depth maps always keep zero. On the other hand, the temporal difference depth can be taken to distinguish the motion regions from static backgrounds at the pixel level.

Therefore, we propose the temporal depth difference loss (L_{TDL}) to offer stronger supervision for multi-frame methods, which promotes a more accurate depth estimation by using temporal information. The formula is as follows:

$$L_{\text{TDL}} = \sum\nolimits_{t=2}^{N_f} \left\| \left(D_G^t - D_G^{t-1} \right) - \left(D_P^t - D_P^{t-1} \right) \right\|_2, \quad (10)$$

where $\{D_P^t,\ D_P^{t-1}\}$ and $\{D_G^t, D_G^{t-1}\}$ represent the predicted and ground-true depth maps at frames t and $t-1$, receptively.

TABLE I

THE SUMMARY OF DATASETS USED IN OUR EXPERIMENTS

Dataset	Subject	Attack Types	Num (Videos)	
			Live	Spoof
Replay-Attack [41]	50	Print, Replay	200	1000
CASIA-FASD [42]	50	Print, Replay	150	450
MSU-MFSD [43]	35	Print, Replay	70	210
OULU-NPU [44]	55	Print, Replay	720	2880
MLFP [45]	10	Mask	150	1200
SiW [17]	165	Print, Replay	1320	3300
CASIA-SURF [46]	1000	Print, Cut	3000	18000
WMCA [47]	72	Print, Replay, Partial, Mask	347	1332
CeFA [48]	1607	Print, Replay, Mask	6300	27900

2) Overall Loss: To make better use of global and local information, both binary and depth supervisions are employed jointly to supervise our network. The overall loss L_{Overall} is defined as follows:

$$L_{\text{Overall}} = \alpha \cdot L_{\text{Binary}} + (1-\alpha) \cdot L_{\text{Depth}}, \quad (11)$$

where L_{Binary} is a binary cross-entropy loss, and L_{Depth} measures the difference between the ground-true depth map (D_G) and predicted depth map (D_P). Specifically, L_{Depth} consists of two parts, $L_{\text{Depth}} = L_{\text{MSE}} + L_{\text{TDL}}$, where L_{MSE} measures Euclidean distances based on pixels, and L_{TDL} is the proposed temporal depth difference loss.

IV. EXPERIMENTS

A. Datasets and Metrics

1) Datasets: To evaluate our approach comprehensively, the following nine public datasets are used in our experiments. A summary of these datasets is given in Table I.

OULU-NPU [44] is a commonly used dataset for intra-dataset testing in FAS. Four protocols were designed to assess different performances. Protocol 1 evaluates the generalization capability under different environments. Protocol 2 evaluates the influence of different attack mediums. Protocol 3 evaluates the effect of different camera sensors. Protocol 4 is the most challenging, with all the above three factors considered simultaneously. Similarly, **SiW** [16] defines three protocols for a comprehensive assessment. Protocol 1 deals with the variations of face poses and expressions. Protocol 2 evaluates the performance under different attack mediums of the same spoof type. Protocol 3 evaluates the generalization of unknown presentation attacks. **CASIA-MFSD** [42], **Replay-Attack** [41], and **MSU-MFSD** [43] are three classical datasets in the FAS community. However, since they have been published for a long time, the number and quality of videos in these datasets may be outdated. Nowadays, these datasets are mainly used for cross-dataset and cross-type testings. **CASIA-SURF** [46] is a large-scale multi-modal dataset for FAS, including three modalities (*i.e.*, RGB, Depth, and NIR). Different attacks are divided into training, validation, and testing subsets for intra-dataset testing. **CeFA** [48] is a multi-modal FAS dataset, covering three modalities, different ethnicities, 1607 subjects, and 2D plus 3D attack types. Four protocols are reported in our experiments following the official definition. The above two datasets are used to investigate the performance in multi-modal datasets. **MLFP** [45] contains 1350 videos in visible, near-infrared, and thermal spectrums with presentation attacks using

TABLE II
DETAILED CONFIGURATIONS FOR TTN-T AND TTN-S

Model	Dimension	Heads	Spatial parts		Temporal parts	
			Input length	Layers	Input length	Layers
TTN-T	192	3	197	12	N_f	4
TTN-S	384	6	197	12	N_f	4

latex and paper masks. **WMCA** [47] consists of 1679 short videos, including multiple attack types shown in Table XIII. Two protocols are provided by this dataset: *seen* protocol and *unseen* attack protocol, respectively. The above two datasets are used to evaluate the performance against 3D mask attacks.

2) Evaluation Metrics: For a fair comparison with previous works, the original metrics are used in our experiments. Specifically, Attack Presentation Classification Error Rate (APCER), Bona Fide Presentation Classification Error Rate (BPCER), and Average Classification Error Rate (ACER) [49] are widely used in OULU-NPU, SiW, WMCA, CASIA-SURF, and CeFA. The ACER is the mean of APCER and BPCER as a whole assessment of bona Fide and attack presentation: $ACER = (APCER + BPCER)/2$. HTER is adopted for cross-dataset testing, which is the mean of False Rejection Rate (FRR) and False Acceptance Rate (FAR): $HTER = (FRR + FAR)/2$. Area Under Curve (AUC) is also utilized for cross-dataset testing and cross-type testing, which is the area under the ROC curve. Equal Error Rate (EER) is defined as the threshold point where False Positive Rate (FPR) is equal to False Rejection Rate (FRR), which is used in MLFP. Among the above metrics, a larger AUC value indicates better performance, while the other metrics are the opposite ($i.e.$, a small value means better performance).

B. Implementation Details

1) Data Preparation: All datasets we used are video data. Thus, we first extract all frames in every video. Next, if the datasets provide face locations, we crop and resize faces to 224×224 as the RGB format. Otherwise, we adopt MTCNN [50] for face detection. Then, a dense face alignment approach ($i.e.$, PRNet [51]) is used to generate the ground-truth depth maps with size 28×28 for genius faces, while spoof depth maps are set to zeros.

2) Networks Setting: Two network structures are tested, denoted as TTN-T and TTN-S. Their sizes of input images are both 224×224 and the patch sizes are set to 16×16. The main differences lie in the number of attention heads and embedding dimensions. More details are provided in Table II.

3) Training Setting: For the single-frame method, a single frame is extracted from a video as the input. For the multi-frame method, a video clip of length N_f is extracted from a video as the input, and the sampling interval is three to make sampled frames cover enough temporal information. To completely compute the depth maps of the first and last frame in the clip, we need to extend one extra frame at the place of beginning and end. However, the calculation of loss functions and decision scores cover the initial N_f merely.

4) Testing Setting: In testing, we calculate the final classification score to separate bona fide and presentation attacks. Specifically, N_f frames are fed into the network and the corresponding N_f depth maps are generated. The scoring formula is defined as follows:

$$Score = \alpha \cdot b_l + (1 - \alpha) \cdot \frac{1}{N_f} \sum\nolimits_{t=1}^{N_f} \left\| D^t_{mean} \right\|_1, \quad (12)$$

where α is the same as that in Eqn. (11). b_l is the living logit of binary classification. In the second term, D^t_{mean} represents the mean depth values on the pixel-wise level at frame t.

5) Hyper-Parameter Settings: The batch size is set to 256 for all single-frame methods and it is 56 for multi-frame ones with $N_f = 8$ due to the limited GPU memory. α is set to 0.4 through all experiments. Adam optimizer is used when the learning rate (lr) and weight decay are both set to 5e-5. The lr is halved every 50 epochs and training stops at 150^{th} epoch.

6) Pre-Training: Our networks consist of three parts: the spatial transformer, temporal transformer, and depth estimation. For better feature extraction, we utilize pre-trained weights on ImageNet [52] provided from [27] to initialize the transformer layers and input linear projection layers in the spatial transformer. Differently, the temporal transformer and depth estimation modules are trained from scratch.

C. Ablation Study

To verify the superiority of our TTN as well as the contributions of each component, multiple incomplete models are built up by controlling different variables. All ablation studies are conducted on TTN-S and the performance is measured on multiple testing scenarios, including intra-dataset testings on OULU-NPU, cross-dataset testings from SiW to OULU-NPU, and cross-type testings on CASIA-MFSD, Replay-Attack, and MSU-MFSD. Their quantitative results are shown in Table III.

1) Efficacy of Each Module and the Loss Function: As shown in Table III, Model 1 can be treated as a single-frame baseline, consisting of the backbone network (DeiT-S [53]). Model 2 utilizes single-scale temporal attention layers to integrate the features of multiple frames for classification. Model 3 adds the module of depth estimation whose cross-layer patch embeddings are fused in an additive manner. Model 4 introduces L_{TDL} to depth loss for forecasting temporal information of depth changes. Model 5 deploys TDA modules to enhance depth estimation by exploiting motion areas. Model 8 is our final model where multiple receptive fields are used to explore multi-scale temporal features independently. From the results, Model 2 is better than Model 1, which indicates the advantage of the multi-frame method compared with the single-frame one. Model 3 exceeds Model 2 by adding depth supervision which can illustrate its effectiveness for FAS. The usefulness of the proposed loss L_{TDL} is verified by comparing Model 3 and Model 4. The observation that Model 5 obtains a performance improvement over Model 4, demonstrates that motion information is beneficial for depth estimation. Finally, compared with Model 5, the improvement of Model 8 reflects the necessity of learning multi-scale temporal information.

2) Multi-Scale Temporal Features Analysis: As shown in Table III, different structures of pyramid temporal aggregation are compared by Models 5, 6, 7, and 8. Specifically, the size of attention windows W_1, W_2 is set to 2 and $\frac{N_f}{2}$, respectively,

TABLE III

THE ABLATION STUDY OF EACH PROPOSED MODULE AND THE LOSS FUNCTION

Model	L_{TDL}	Add	TDA	PTA	Intra-dataset (ACER%)↓				Cross-dataset (ACER%)↓				Cross-type (AUC%)↑	
					1	2	3	4	1	2	3	4	intra	inter
1				w/o PTA	2.8	1.2	1.3±0.9	4.0±1.5	7.9	9.8	9.1±3.6	4.6±4.4	97.5±4.6	86.9±12.1
2				w/ PTA (W)	1.7	1.2	1.4±1.3	3.8±3.4	7.5	9.8	8.9±3.2	4.6±2.5	97.5±4.2	87.1±12.4
3		√		w/ PTA (W)	1.2	1.1	1.4±0.6	3.5±3.4	7.1	9.6	8.3±2.4	4.2±2.6	97.7±3.8	88.6±10.3
4	√	√		w/ PTA (W)	1.0	1.3	1.2±0.9	3.3±5.2	6.9	9.3	7.8±2.2	4.4±3.7	97.6±3.9	88.3±10.8
5	√		√	w/ PTA (W)	0.8	1.0	1.2±0.8	3.3±3.0	6.4	8.9	7.5±2.8	4.0±3.1	97.9±4.4	88.8±10.5
6	√		√	w/ PTA (W_1, W)	0.7	0.8	1.2±1.0	3.1±2.9	5.9	8.7	7.1±1.9	3.7±2.8	98.0±4.2	89.3±10.2
7	√		√	w/ PTA (W_2, W)	0.4	0.9	1.0±1.3	2.7±2.8	5.7	**8.4**	7.4±2.4	3.5±2.3	**98.1±3.9**	89.6±10.3
8	√		√	w/ PTA (W_1, W_2, W)	**0.2**	**0.6**	**0.9±0.7**	**2.9±1.4**	**5.4**	8.5	**6.7±2.3**	**3.1±3.6**	97.9±4.2	**89.7±9.2**

Features	Models	Features	Models
Fea1	M1	Fea2, Fea3	M8
Fea2	M2	Fea2, Fea4	M9
Fea3	M3	Fea3, Fea4	M10
Fea4	M4	Fea1, Fea2, Fea3	M11
Fea1, Fea2	M5	Fea1, Fea2, Fea4	M12
Fea1, Fea3	M6	Fea2, Fea3, Fea4	M13
Fea1, Fea4	M7	Fea1, Fea2, Fea3, Fea4	M14

Fig. 6. The comparative experiments of models with different layers to fuse features for depth estimation on OULU-NPU Protocol 1. Different fusion methods are denoted by different colors, while different fusion strategies are distinguished by their corresponding symbols listed there.

which can concentrate on exploiting short-term and middle-term information for splitting living and spoofing. Attention window W represents the global receptive field. Experimental results indicate that exploring multi-scale attention regions can obtain better results. This mechanism is favorable to weakening the negative effects of abnormal frames by increasing the model variances in the temporal domain. More analysis will be provided in the latter section. Thus, for the trade-off between speed and accuracy, this pyramid structure which consists of three independent attention fields is adopted in our final model.

3) Multi-Scale Spatial Features Analysis: Different layers for feature fusion are investigated here. As shown in Fig. 6, Fea1, Fea2, Fea3, and Fea4 represent the patch features from layer 3, layer 6, layer 9, layer 12, respectively, which are distributed from low level to high level. Different fusion methods and strategies are evaluated on OULU-NPU Protocol 1. It can be observed: 1) For fusion, the result of directly adding is better than concatenation in general; 2) For specific fusion strategy, feature fusion from low layer to high layer progressively is beneficial for more accurate depth estimation. Thus, to make a trade-off between speed and accuracy, the feature fusion approach of M14 is applied in our final model.

4) Sampling Interval Analysis: We conduct several experiments by capturing input sequences with different sampling intervals (Δt), as shown in Fig. 7. Different sampling intervals imply different spans of temporal information. The ACER is the lowest when Δt is set to 3.

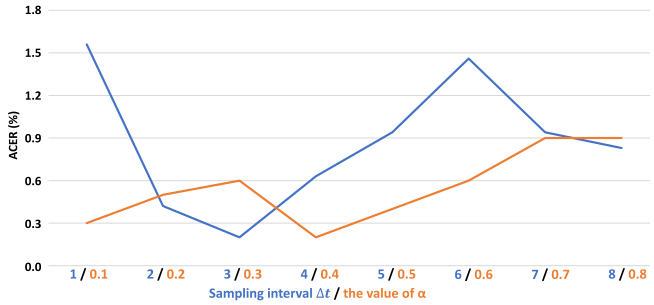


Fig. 7. Ablation study of different sampling intervals Δt and hyper-parameter α on OULU-NPU Protocol 1.

5) The Setting of Different α: Hyper-parameter α is to balance the proportion of binary and depth supervision by compositing loss function and evaluation score. Specifically, global information is used to make a judgment of binary classification, while local details are exploited by estimating depth maps. They work together to promote effectiveness and generalization simultaneously. As shown in Fig. 7, the model achieves the best performance when α is set to 0.4.

6) Channel Weighting Versus Spatial Weighting: The TDA is proposed to enhance relevant channels of patch embeddings for depth estimation, which is channel weighting. To verify its effectiveness, we replace it with other weighting methods for comparisons, including spatial weighting and CBAM [54]. Specifically, CBAM can be regarded as a cascaded structure of channel weighting and spatial weighting. Their ACERs on OULU-NPU Protocol 1 are 0.9 and 1.3 respectively, while channel weighting can achieve 0.2. These results can be explained by the following facts: 1) Different from convolutions, the spatial size of patch embeddings is equal to the number of sliced patches that are non-overlapped and encoded by channel-wise features. Thus, directly weighting the proportion of different patches may be less accurate than channel weighting; 2) For patch embedding, some channels encode motion, while some encode static information, which can be assigned to different weights by channel weighting. This is beneficial for utilizing temporal information. To sum up, the approach of channel weighting is adopted in our approach.

D. Intra-Dataset Testing

Intra-dataset testings are conducted on OULU-NPU and SiW datasets. We strictly follow their protocols designed for OULU-NPU and SiW to make a fair evaluation.

1) Results on OULU-NPU: In Table IV, we compare our model with recent SOTA methods. Our approach shows the best performance on all four protocols, which proves its generalization capability under different testing scenarios, including

TABLE IV
INTRA-DATASET TESTING ON FOUR PROTOCOLS OF OULU-NPU

Prot.	Method	APCER(%)↓	BPCER(%)↓	ACER(%)↓
1	DRL-FAS [25]	5.4	4.0	4.7
	Disentangled [55]	1.7	0.8	1.3
	STDN [56]	0.8	1.3	1.1
	CDCN [57]	0.4	1.7	1.0
	FAS-SGTD [31]	2.0	0.0	1.0
	CDCN-PS [58]	0.4	1.2	0.8
	DCN [26]	1.3	0.0	0.6
	DC-CDN [59]	0.5	0.3	0.4
	CDCN(MT) [60]	0.0	0.8	0.4
	CDCN++ [57]	0.4	0.0	0.2
	NAS-FAS [61]	0.4	0.0	0.2
	TTN-T (Ours)	1.2	0.0	0.6
	TTN-S (Ours)	0.4	0	**0.2**
2	Disentangled [55]	1.1	3.6	2.4
	DCN [26]	2.2	2.2	2.2
	DRL-FAS [25]	3.7	0.1	1.9
	FAS-SGTD [31]	2.5	1.3	1.9
	STDN [56]	2.3	1.6	1.9
	CDCN [57]	1.5	1.4	1.5
	CDCN-PS [58]	1.4	1.4	1.4
	CDCN(MT) [60]	1.3	1.4	1.4
	CDCN++ [57]	1.8	0.8	1.3
	DC-CDN [59]	0.7	1.9	1.3
	NAS-FAS [61]	1.5	0.8	1.2
	TTN-T (Ours)	0.8	0.8	0.8
	TTN-S (Ours)	0.4	0.8	**0.6**
3	DRL-FAS [25]	4.6±3.6	1.3±1.8	3.0±1.5
	STDN [56]	1.6±1.6	4.0±5.4	2.8±3.3
	FAS-SGTD [31]	3.2±2.0	2.2±1.4	2.7±0.6
	CDCN [57]	2.4±1.3	2.2±2.0	2.3±1.4
	Disentangled [55]	2.8±2.2	1.7±2.6	2.2±2.2
	CDCN(MT) [60]	2.3±1.5	1.9±1.8	2.1±1.7
	CDCN-PS [58]	1.9±1.7	2.0±1.8	2.0±1.7
	DCN [26]	2.3±2.7	1.4±2.6	1.9±1.6
	DC-CDN [59]	2.2±2.8	1.6±2.1	1.9±1.1
	CDCN++ [57]	1.7±1.5	2.0±1.2	1.8±0.7
	NAS-FAS [61]	2.1±1.3	1.4±1.1	1.7±0.6
	TTN-T (Ours)	0.8±0.9	1.4±1.8	1.1±0.9
	TTN-S (Ours)	1.0±1.1	0.8±1.3	**0.9±0.7**
4	DRL-FAS [25]	8.1±2.7	6.9±5.8	7.2±3.9
	CDCN [57]	4.6±4.6	9.2±8.0	6.9±2.9
	CDCN++ [57]	4.2±3.4	5.8±4.9	5.0±2.9
	FAS-SGTD [31]	6.7±7.5	3.3±4.1	5.0±2.2
	CDCN-PS [58]	2.9±4.0	5.8±4.9	4.8±1.8
	Disentangled [55]	5.4±2.9	3.3±6.0	4.4±3.0
	DC-CDN [59]	5.4±3.3	2.5±4.2	4.0±3.1
	STDN [56]	2.3±3.6	5.2±5.4	3.8±4.2
	CDCN(MT) [60]	0.9±2.0	6.4±4.9	3.7±2.9
	DCN [26]	6.7±6.8	0.0±0.0	3.3±3.4
	NAS-FAS [61]	4.2±5.3	1.7±2.6	2.9±2.8
	TTN-T (Ours)	4.2±2.4	3.8±4.0	4.0±2.3
	TTN-S (Ours)	3.3±2.8	2.5±2.0	**2.9±1.4**

external environment, attack mediums, and camera variations. It is worth noting that the excellent accuracy of ACER on protocols 2 and 3 is 0.6% and 0.9%, showing that our TTN model can grasp the interval cues by utilizing the relations of patches and frames. On protocol 4, our method has the same mean ACER value as NAS-FAS [61], but with a lower standard deviation, indicating great accuracy and stability.

2) Results on SiW: As shown in Table V, our method ranks first on protocols 1 and 2. This not only proves the effectiveness of our method toward poses and expressions changes, but also shows the generalization capability on cross mediums. Protocol 3 evaluates the performance of unknown PAs. Our method is better than FAS-SGTD [32] and DRL-FAS [24] which are multi-frame methods, but slightly worse than CDCN++ [57] which is a single-frame one. The reason could be that there exists a temporal difference between different attacks, such as print and replay attacks. Print attacks are in a static mode, while replay attacks are presented in a dynamic mode. Thus, this different attribute in sequences may cause difficulties to adapt to unknown PAs for multi-frame methods.

TABLE V
INTRA-DATASET TESTING ON THREE PROTOCOLS OF SIW

Prot.	Method	APCER(%)↓	BPCER(%)↓	ACER(%)↓
1	FAS-SGTD [31]	0.64	0.17	0.40
	CDCN++ [57]	0.07	0.17	0.12
	STDN [56]	0.00	0.00	0.00
	DRL-FAS [25]	-	-	0.00
	DCN [26]	0.00	0.00	0.00
	TTN-T (Ours)	0.00	0.00	0.00
	TTN-S (Ours)	0.00	0.00	**0.00**
2	CDCN++ [57]	0.00±0.00	0.09±0.10	0.04±0.05
	FAS-SGTD [31]	0.00±0.00	0.04±0.08	0.02±0.04
	STDN [56]	0.00±0.00	0.00±0.00	0.00±0.00
	DRL-FAS [25]	-	-	0.00±0.00
	DCN [26]	0.00±0.00	0.00±0.00	0.00±0.00
	TTN-T (Ours)	0.00±0.00	0.00±0.00	0.00±0.00
	TTN-S (Ours)	0.00±0.00	0.00±0.00	**0.00±0.00**
3	STDN [56]	8.30±3.30	7.50±3.30	7.90±3.30
	DRL-FAS [25]	-	-	4.51±0.00
	DCN [26]	3.80±4.30	3.00±2.60	3.40±0.90
	FAS-SGTD [31]	2.63±3.72	2.92±3.42	2.78±3.57
	CDCN++ [57]	1.97±0.33	1.77±0.10	**1.90±0.15**
	TTN-T (Ours)	3.66±3.02	3.51±3.18	3.58±3.09
	TTN-S (Ours)	2.69±2.05	2.67±2.00	2.68±2.03

TABLE VI
THE RESULTS OF CROSS-DATASET TESTING FROM SIW TO OULU-NPU

Prot.	Method	APCER(%)↓	BPCER(%)↓	ACER(%)↓
1	Auxiliary [17]	-	-	10.0
	FAS-SGTD [31]	1.7	13.3	7.5
	TTN-T (Ours)	3.8	10.0	6.9
	TTN-S (Ours)	0.8	10.0	**5.4**
2	Auxiliary [17]	-	-	14.1
	FAS-SGTD [31]	9.7	14.2	11.9
	TTN-T (Ours)	7.6	11.4	9.5
	TTN-S (Ours)	6.4	10.6	**8.5**
3	Auxiliary [17]	-	-	13.8±5.7
	FAS-SGTD [31]	17.5±4.6	11.7±12.0	14.6±4.8
	TTN-T (Ours)	6.1±3.3	8.1±5.1	7.1±1.9
	TTN-S (Ours)	6.9±4.3	6.4±5.3	**6.7±2.3**
4	Auxiliary [17]	-	-	10.0±8.8
	FAS-SGTD [31]	0.8±1.9	10.0±11.6	5.4±5.7
	TTN-T (Ours)	1.3±1.9	7.5±6.3	4.4±3.7
	TTN-S (Ours)	1.3±1.3	5.0±7.1	**3.1±3.6**

E. Cross-Dataset Testing

To evaluate the performance in cross domains, we utilize five datasets (CASIA-MFSD, Replay-Attack, MSU-MFSD, SiW, and OULU-NPU) to perform cross-dataset testings.

1) Results on OCIM: As shown in Table VII, for an overall evaluation, we conduct cross-dataset testing by using a leave-one-out (LOO) strategy. Specifically, three datasets are randomly selected for training and the rest one is used for testing. It can be observed that our method achieves the best performance among all models trained without domain information, which demonstrates its generalization capacity and robustness against unknown data distribution. Besides, when joined with SSDG [62], our method can gain a better performance, especially on O&C&I to M and O&C&M to I, which indicates the use of domain labels can further enhance the effectiveness of our approach. According to [27], training transformers requires a large-scale training data set, however, CASIA-MFSD, Replay-Attack, and MSU-MFSD are three low-resolution video datasets that contain few samples for sufficient training. Thus, we implement further cross-dataset testing experiments on relatively larger datasets (*i.e.*, SiW and OULU-NPU) to further verify the superiority of our method.

2) Results From SiW to OULU-NPU: Table VI shows the cross-dataset results where models are trained on SiW and tested on each protocol of the OULU-NPU dataset.

TABLE VII

THE RESULTS OF CROSS-DATASET TESTING ON OULU-NPU, CASIA-MFSD, REPLAY-ATTACK, AND MSU-MFSD

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(%)↑
Binary CNN [5]	29.25	82.87	34.88	71.94	34.47	65.88	29.61	77.54
IDA [43]	66.67	27.86	55.17	39.05	28.35	78.25	54.20	44.59
Color Texture [63]	28.09	78.47	30.58	76.89	40.40	62.78	63.59	32.71
Auxiliary (Depth) [17]	22.72	85.88	33.52	73.15	29.14	71.69	30.17	77.61
NAS-FAS [61]	19.53	88.63	16.54	90.18	14.51	93.84	13.80	93.43
TTN-T (Ours)	11.25	95.08	11.30	**95.33**	15.75	91.25	14.44	93.50
TTN-S (Ours)	**9.58**	**95.79**	**9.81**	95.07	**14.15**	**94.06**	**12.64**	**94.20**
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(%)↑
MMD-AAE [64]	27.08	83.19	44.59	58.29	31.58	75.18	40.98	63.08
MADDG [29]	17.69	88.06	24.50	84.51	22.19	84.99	27.98	80.02
RFMeta [65]	13.89	93.98	20.27	88.16	17.30	90.48	16.45	91.16
SSDG-R [62]	7.38	97.17	10.44	95.94	11.71	96.59	15.61	91.54
SDFANet [66]	**4.28**	97.59	12.56	93.63	**6.14**	97.30	**12.26**	94.29
VLAD-VSA [67]	4.29	**98.25**	**8.76**	95.89	7.79	**97.79**	12.64	94.00
TTN-T-SSDG (Ours)	6.25	97.27	11.11	95.33	13.62	94.70	14.69	92.48
TTN-S-SSDG (Ours)	5.42	98.08	10.00	**96.11**	10.12	95.73	12.47	**94.58**

TABLE VIII

AUC (%) OF THE INTRA-DATASET CROSS-TYPE AND INTER-DATASET CROSS-TYPE TESTING ON CASIA-MFSD, REPLAY-ATTACK, AND MSU-MFSD

Method	Protocol	CASIA-MFSD			Replay-Attack			MSU-MFSD			Overall
		Video	Cut photo	Warpped Photo	Video	Digital Photo	Printed Photo	Printed Photo	HR Video	Mobile Video	
DTN [68]	Intra	90.00	97.30	97.50	99.90	99.90	99.60	81.60	99.90	97.50	95.90±6.20
CDCN [57]		98.48	99.90	99.80	100.00	99.43	99.92	70.82	100.00	99.99	96.48±9.64
CDCN++ [57]		98.07	99.90	99.60	99.98	99.89	99.98	72.29	100.00	99.98	96.63±9.15
BCN [69]		99.62	100.00	100.00	99.99	99.74	99.91	71.64	100.00	**99.99**	96.77±9.99
NAS-FAS [61]		**99.62**	100.00	100.00	99.99	99.89	99.98	74.62	100.00	99.98	97.12±8.94
TTN-T (Ours)		99.06	99.89	**100.00**	**100.00**	**100.00**	**100.00**	**87.25**	99.81	96.75	**98.08±3.96**
TTN-S (Ours)		99.57	**100.00**	**100.00**	**100.00**	**100.00**	**100.00**	87.06	**100.00**	94.50	97.90±4.19
SVM1+IMQ [70]	Inter	88.41	75.14	75.23	88.21	71.20	56.41	56.62	71.12	49.75	70.23±12.69
CDCN [57]		72.20	79.31	84.22	**97.73**	94.89	96.70	74.25	98.88	100.00	88.69±10.56
CDCN++ [57]		73.12	76.64	78.36	96.66	92.92	97.67	74.25	98.13	**100.00**	87.53±10.90
TTN-T (Ours)		88.90	**90.12**	91.93	85.50	**98.16**	**99.80**	74.44	**99.19**	99.94	**92.00±8.01**
TTN-S (Ours)		**90.26**	79.60	**95.17**	68.81	93.82	95.88	**88.87**	95.19	99.82	89.71±9.17

OULU-NPU and SiW are two high-resolution video datasets that contain a relatively larger number of samples for training and testing. Our method outperforms the multi-frame method FAS-SGTD [32] and Auxiliary [16] on all four protocols (5.4%, 8.5%, 6.7%, and 3.1% ACER, respectively). The above experimental results and phenomena demonstrate the inductive learning capability of our method in cross-domain scenarios.

F. Cross-Type Testing

Following the protocol proposed in [70], we use CASIA-MFSD, Replay-Attack, and MSU-MFSD to perform intra-dataset cross-type testing and inter-dataset cross-type testing.

1) Intra-Dataset Cross-Type Testing: As shown in Table VIII, we adopt the Leave-One-Out (LOO) strategy for different attack types in the same dataset to evaluate the robustness of encountering unknown attacks. Five SOTA methods are listed for comparison. Our proposed methods achieve the best overall performance (98.08±3.96% AUC), which indicates the capability to process unknown presentation attacks.

2) Inter-Dataset Cross-Type Testing: The data distribution in the same dataset is similar. However, in reality, unknown presentation attacks usually appear in different domains. Therefore, to make a comprehensive evaluation, we further estimate the performance of our method with inter-dataset protocols in [70], as shown in Table VIII. Specifically, SVM1+IMQ is proposed in [70], which consists of one-class SVM with a Gaussian kernel based on image quality features. To compare comprehensively, we implement the popular methods, *i.e.*, CDCN and CDCN++ [57], and evaluate their performance under the inter-dataset setting. In this testing, our method retains the best performance compared with other methods, demonstrating that our method has a great adaption capability towards unseen domains and unknown attacks.

On the other hand, it is observed that the performance of TTN-T is better than TTN-S on both intra- and inter-datasets cross-type testing even though TTN-T has fewer parameters than TTN-S. This indicates that small models may have advantageous in cross-type scenarios, especially when only small datasets are available for training.

G. Comparison With Pre-Trained Baselines

To prove the effectiveness of our proposed multi-frame methods, several single-frame networks are implemented as baselines. Specifically, we utilize DeiT-T [53] and Deit-S [53] as the backbone of ViTranZFAS [31] for a fair comparison. To explore the superiority of transformer structures, we compare different network structures as the backbone for FAS, including ResNet [71], DenseNet [72], and CDCN [57]. Meanwhile, all baseline models are pre-trained on ImageNet. Specifically, for ResNet, DenseNet, and Deit based structures, we take the standard pre-trained models and replace the final layers with a fully connected layer for binary classification. For CDCN based structures, we first flatten the generated map in an element-wise way, then connect a suitable head classifier for ImageNet pre-training. Lastly, the head classifier is removed when training on FAS with depth supervision.

1) Evaluation Performance: To make an overall analysis, we compare our method with baseline methods on

TABLE IX
THE OVERALL EVALUATION OF OUR TTNS AND PRE-TRAINED BASELINES

Testing condition	Intra-dataset (ACER%)↓				Cross-dataset (ACER%)↓				Cross-type (AUC%)↑		Params(M)	FLOPs(G)	Run-time(ms)
Protocol	1	2	3	4	1	2	3	4	intra	inter			
ResNet18-TF [71]	5.3	1.9	2.5±2.9	8.1±4.9	11.4	11.9	8.9±3.2	8.3±5.7	93.3±9.6	83.1±12.9	11.18	14.6	17.8
ResNet50-TF [71]	2.7	1.2	2.0±2.8	5.6±1.9	6.5	9.9	7.5±2.8	4.8±4.6	90.0±13.2	77.7±12.7	23.51	32.9	35.7
DenseNet121-TF [72]	2.8	1.6	3.0±3.5	5.8±2.0	7.4	11.3	9.6±3.8	4.0±3.1	95.1±12.3	84.5±12.4	6.96	22.9	78.9
DenseNet161-TF [72]	1.3	1.2	1.7±2.3	5.0±2.4	8.2	10.5	7.8±2.2	4.6±2.5	96.8±6.4	86.9±11.3	26.48	62.3	86.9
CDCN-TF [57]	1.0	1.0	1.6±1.2	4.4±1.0	11.2	12.8	11.7±3.6	8.3±10.0	96.6±8.4	90.0±9.7	2.63	291.5	5203.3
CDCNpp-TF [57]	0.2	1.0	1.4±1.1	3.3±0.9	14.8	15.5	11.7±5.6	5.4±7.5	96.8±6.3	90.8±10.3	2.26	312.5	5416.2
ViTranZFAS-T [30]	2.2	1.1	1.6±1.7	4.2±2.5	7.7	11.5	9.8±3.5	6.7±6.1	96.9±6.3	85.2±11.4	5.72	8.6	16.8
ViTranZFAS-S [30]	1.8	1.2	1.3±0.9	4.0±1.5	7.9	9.8	9.1±3.6	4.6±4.4	97.5±4.6	86.9±12.1	22.05	33.9	17.5
TTN-T (Ours)	0.6	0.8	1.1±0.9	4.0±2.3	6.9	9.5	7.1±1.9	4.4±3.7	**98.1±4.0**	**92.0±8.0**	8.99	10.1	36.4
TTN-S (Ours)	**0.2**	**0.6**	**0.9±0.7**	**2.9±1.4**	**5.4**	**8.5**	**6.7±2.3**	**3.1±3.6**	97.9±4.2	89.7±9.2	34.44	37.0	37.4

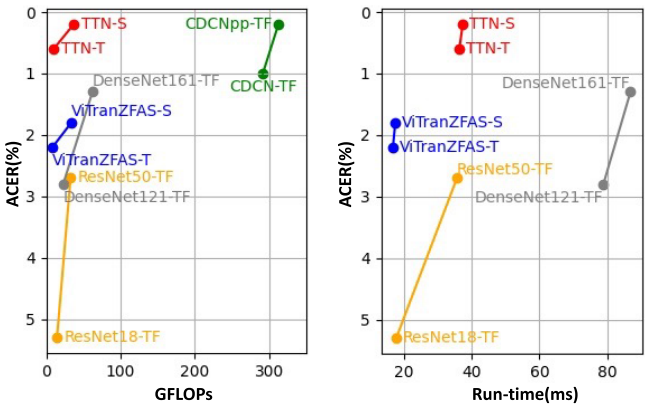


Fig. 8. The comparison between different models on ACER(%) of OULU-NPU protocol 1, GFLOPs and run-time (ms).

multiple evaluation scenarios, including intra-dataset testings on OULU-NPU, cross-dataset testings from SiW to OULU-NPU, and cross-type testings on CASIA-MFSD, Replay-Attack, and MSU-MFSD. Their quantitative results are shown in Table IX.

Compared with single-frame baselines (*i.e.*, ViTranZFAS-T and ViTranZFAS-S), our multi-frame methods obtain better performance, which proves that the proposed temporal modules can explore additional information to distinguish between living and spoofing. Meanwhile, the performance of the other compared networks is inferior to TTNs, which indicates the advantages of our transformer-based network to integrate long-range spatial information and multi-scale temporal information, simultaneously. On the other hand, it is worth noting that some models show unbalanced performances in different evaluation scenarios. For example, ResNet networks have a poor performance in cross-type conditions, compared to the performance on intra-dataset and cross-dataset scenarios. However, our TTNs retain great results under all protocols. This phenomenon demonstrates the great generalization capacity of our method toward different scenarios.

2) Model Efficiency: In Table IX, we list the number of parameters, FLOPs, and Run-time to compare the model size and computation efficiency between different methods. Specifically, the run-time is the inference time for one video on a single 2080Ti GPU. Although TTNs have a large learning capacity with a large number of parameters (8.99M and 34.44M), they have relatively low FLOPs (10.1 GFLOPs and 37.0 GFLOPs) and run-time (36.4ms and 37.4ms). In contrast, CDCN and CDCN++ have higher FLOPs (291.5 GFLOPs and 312.5 GFLOPs) and run-time (5203.3ms and 5416.2ms) despite fewer parameters (2.63M and 2.26M). The inconsistency between parameters and computational cost is due to the parameter sharing mechanism of the convolution operator. However, computational cost can affect the efficiency of the whole system, thus it is usually considered more important in practice. Furthermore, we analyze the trade-off between the performance ACERs, FLOPs and run-time, as shown in Fig. 8. It can be noted that our TTNs achieve the best balance between these factors. Therefore, our TTNs are expected to work well in large-scale learning and real-world applications.

3) The Importance of Pre-Training: It is difficult to train transformers from scratch, due to the lack of sufficient training data. Thus, following [27], pre-trained weights on ImageNet are utilized in our spatial transformers for better feature extraction. Due to the shortage of inductive bias, vision transformers are usually pre-trained on a large number of samples and transferred to medium-scale or small-scale datasets, which are different from CNN-based methods. This strategy is widely used in many other tasks when using the vision transformer, such as classification [27], detection [73], segmentation [35], and so on. To sum up, using pre-trained weights in vision transformers is a vital characteristic and indispensable.

H. Experiments on Multi-Modal Datasets

Besides single-modal RGB input, multi-modal data (*i.e.*, RGB, Depth, and NIR) can also be beneficial for performance improvement. To investigate such gains on our network, we implement TTNs on multi-modal datasets CASIA-SURF [46] and CeFA [48]. To utilize the multi-modal input, we adopt a halfway fusion approach [46] to assemble the spatial and temporal information from different modalities, which can be marked as TTN-T-NHF and TTN-S-NHF, respectively.

1) Experiments on CASIA-SURF: As shown in Table X, we measure the performance of our networks on the intra-testing protocol of CASIA-SURF. When only using single-modal RGB input, our method can obtain the lowest ACER of 3.5%, which demonstrates the effectiveness of our method on subtle feature extraction. When joining multi-modal inputs, the ACERs can even be further minimized to 1.0%, which shows a competitive performance compared to the SOTA multi-modal fusion methods SEF [46] and PSMM-Net [48].

2) Experiments on CeFA: Besides the intra-testing on CASIA-SURF, we conduct experiments on protocols 1, 2, 3, and 4 of CeFA, respectively. The results are reported in Table XI. It can be observed that our method can

TABLE X
THE RESULTS OF INTRA-DATASET TESTING ON CASIA-SURF

Modality	Method	APCER(%)↓	BPCER(%)↓	ACER(%)↓
Single	ResNet18 [71]	40.3	1.6	21.0
	Single-scale SEF [74]	8.0	14.5	11.3
	TTN-T (Ours)	4.0	3.3	3.7
	TTN-S (Ours)	3.8	3.2	**3.5**
Multi	NHF [46]	5.6	3.8	4.7
	Single-scale SEF [74]	3.8	1.0	2.4
	Multi-scale SEF [74]	1.6	0.08	0.8
	PSMM-Net [48]	0.7	0.06	**0.4**
	TTN-T-NHF (Ours)	0.8	2.2	1.5
	TTN-S-NHF (Ours)	0.4	1.6	1.0

TABLE XI
THE RESULTS ON FOUR PROTOCOLS OF CEFA

Prot.	Method	APCER(%)↓	BPCER(%)↓	ACER(%)↓
1	PSMM-Net [48]	2.4±0.6	4.6±2.3	3.5±1.3
	MA-Net [75]	16.7±5.6	12.4±8.1	14.6±6.6
	TTN-T (Ours)	2.4±1.0	8.7±5.2	5.6±3.0
	TTN-S (Ours)	5.4±4.6	4.1±2.0	4.7±2.6
	TTN-T-NHF (Ours)	0.9±1.2	3.4±0.9	2.3±0.9
	TTN-S-NHF (Ours)	0.0±0.0	4.2±1.6	**2.1±0.8**
2	PSMM-Net [48]	7.7±9.0	3.1±1.6	5.4±5.3
	MA-Net [75]	20.9±6.8	1.2±1.7	11.1±4.4
	TTN-T (Ours)	3.7±2.0	0.4±0.1	2.1±1.1
	TTN-S (Ours)	5.6±5.7	0.6±0.4	3.1±3.0
	TTN-T-NHF (Ours)	2.9±0.8	0.5±0.0	1.7±0.4
	TTN-S-NHF (Ours)	1.6±0.6	0.3±0.2	**1.0±0.4**
3	PSMM-Net [48]	19.4±8.7	5.0±1.8	12.2±5.2
	MA-Net [75]	-	-	-
	TTN-T (Ours)	-	-	-
	TTN-S (Ours)	-	-	-
	TTN-T-NHF (Ours)	2.0±0.8	0.8±0.3	1.4±0.4
	TTN-S-NHF (Ours)	1.1±0.5	0.6±0.1	**0.9±0.2**
4	PSMM-Net [48]	7.8±2.9	5.5±3.0	6.7±2.2
	MA-Net [75]	30.4±13.6	18.1±3.3	24.3±5.9
	TTN-T (Ours)	7.8±1.6	8.8±3.4	8.3±2.4
	TTN-S (Ours)	6.4±2.3	6.5±3.6	6.5±2.1
	TTN-T-NHF (Ours)	2.4±2.2	1.9±0.4	2.2±1.0
	TTN-S-NHF (Ours)	1.4±1.9	2.4±0.7	**1.9±1.3**

achieve a competitive performance despite the single-modal RGB input. Moreover, when utilizing multi-modal inputs, the performance is further improved and ranks the first on all four protocols (2.1%, 1.0%, 0.9%, and 1.9% ACER, respectively). The above results prove the generalization capacity of our method on multi-modal cross-domain testing scenarios.

I. Experiments on 3D Mask Datasets

To prove the effectiveness of TTN against 3D mask attacks, experiments on MLFP [45] and WMCA [47] are conducted.

1) Experiments on MLFP: Dataset MLFP contains three different data types: visible, near-infrared, and thermal spectrums. The training-testing protocol based on subjects and masks unseen is obeyed in our experiments, thus their average EERs are reported in Table XII. For near-infrared and thermal spectrums, it is difficult to generate pseudo depth maps, thus we use binary maps to replace depth maps in these two testing scenarios, which are labeled as TTN-T-BM and TTN-S-BM. As shown in Table XII, it can be observed: 1) The results of thermal spectrums arrive at the minimum EER of 1.2%, which proves the viewpoint in [45] that the thermal imaging spectrum is most effective in detecting mask attacks; 2) Compared with the results of using different ground-truths in visible, we can observe that pseudo depth maps can further reduce the EER, thus are more suitable for map supervision in our network; 3) There exist substantial reductions (25.5%, 36.9%, and 9.6%, respectively) in EER, compared with the other methods, which proves the effectiveness of our method against mask attacks.

TABLE XII
EER (%) OF FACE PRESENTATION ATTACK DETECTION ON MLFP

Method	VIS↓	NIR↓	Thermal↓
HOG [76]	34.9	45.5	24.5
ULBP [77]	38.7	49.0	21.3
BSIF [78]	29.2	43.3	15.0
LPQ [79]	40.0	43.8	13.7
RDWT+Haralick [12]	32.9	42.0	10.8
TTN-T (Ours)	7.1	-	-
TTN-S (Ours)	**3.7**	-	-
TTN-T-BM (Ours)	7.6	6.0	2.7
TTN-S-BM (Ours)	5.0	**5.1**	**1.2**

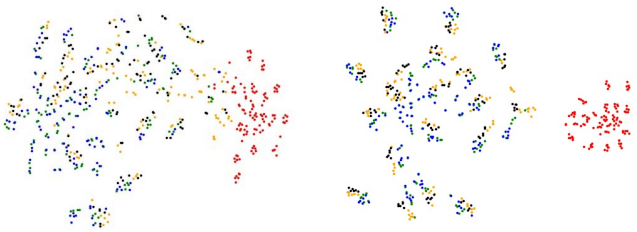


Fig. 9. Feature distributions on OULU-NPU Protocol 1. Left: features of single-frame baseline. Right: features of our TTN. Color code means: *red* = live, *green* = printer1, *blue* = printer2, *orange* = replay1, *black* = replay2.

2) Experiments on WMCA: The results of protocols *seen* and *unseen* on WMCA are shown in Table XIII. Specifically, on protocol *unseen*, the Leave-One-Out (LOO) strategy is adopted to measure the generalization capacity when encountering different attacks. With only RGB input, our method can achieve the best performance on protocols *seen* (2.6% ACER) and *unseen* (8.8% ACER). It can be observed that our method has poor performance for unseen attacks *replay* and *glasses*. For *replay*, our video-based method may be confused by the similar temporal characteristics between replay attacks and live faces. For *glasses*, it may cause mistakes in depth estimation due to the similarities between the appearance of the glass attacks and bone fide wearing medical glasses. However, the above shortages can be effectively relieved when using RGB and depth inputs together. The performance can be further improved and achieve the first rank among all methods.

J. Visualization and Analysis

1) Feature Distribution Visualization: As shown in Fig. 9, the feature distribution of testing videos on OULU-NPU Protocol 1 is visualized in the 2D plane via t-SNE [84]. The right image represents features of our multi-frame method TTN-S which presents a well-clustered characteristic, compared with the left image representing features of single-frame baseline ViTranZFAS-S [31]. This observation indicates that the temporal information embedded between frames is beneficial to distinguish living from spoofing more accurately.

2) Depth Estimation Visualization: As shown in Fig. 10, original images represent input RGB images; Original depth maps are generated by PRNet [51]; Depth estimation results of w/o TDA and final model are also shown for comparison. It can be observed that our TTN achieves more accurate depth estimation compared with the model w/o TDA. Because TDA modules utilize feature differences of adjacent frames to

TABLE XIII

COMPARISON OF THE RESULTS OF PROTOCOLS 'SEEN' AND 'UNSEEN' ON WMCA. THE VALUES ACER(%) REPORTED ON TESTING SETS ARE OBTAINED WITH THRESHOLDS COMPUTED FOR BPCER=1% ON DEVELOPMENT SETS. 'RGB-D' DENOTES USING BOTH RGB AND DEPTH INPUTS

Modality	Method	Seen	Unseen							
			Flexiblemask	Replay	Fakehead	Prints	Glasses	Papermask	Rigidmask	Mean±Std
RGB	ResNet50 [71]	40.9	14.5	15.7	38.0	32.7	27.3	20.1	30.2	25.5±9.0
	CDCN [57]	38.4	12.1	**8.7**	42.7	30.1	**11.7**	11.9	30.4	21.1±13.2
	Auxiliary (Depth) [17]	42.7	13.2	12.5	47.3	32.2	23.7	13.9	40.4	26.2±14.1
	TTN-T (Ours)	3.0	15.1	33.8	1.3	0.4	40.4	3.0	6.0	14.3±16.4
	TTN-S (Ours)	**2.6**	**10.7**	21.9	**1.3**	**0.0**	25.4	**0.0**	**2.0**	**8.8±10.9**
RGB-D	MC-PixBiS [80]	1.8	49.7	3.7	0.7	0.1	16.0	**0.2**	3.4	10.5±16.7
	MCCBB-OCCL-GMM [81]	3.3	22.8	31.4	1.9	30.0	50.0	4.8	18.3	22.7±15.3
	MC-ResNetDLAS [82]	4.2	33.3	38.5	49.6	3.8	41.0	47.0	20.6	33.4±14.9
	CMFL [83]	-	**12.4**	1.0	2.5	0.7	33.5	1.8	**1.7**	7.6±11.2
	TTN-T-NHF (Ours)	0.8	26.4	**0.0**	**0.0**	0.0	**15.9**	1.8	8.0	7.4±10.2
	TTN-S-NHF (Ours)	**0.3**	21.7	1.7	1.7	**0.0**	21.3	0.7	2.3	**7.1±9.9**

Fig. 10. The depth estimation and attention visualization results of hard samples in OULU-NPU. For attention visualization, bright regions represent activation areas for spoof cues. (a) Original images. (b) Original depth. (c) Depth estimation w/o TDA. (d) Depth estimation of final network. (e) Attention visualization w/o TDA. (f) Attention visualization w/o PTA. (g) Attention visualization of the final network.

enhance the motion-related channels that play an important role in depth estimation. Specifically, some local cues are easily distinguished in dynamic mode, but difficultly in static mode. Thus, TDA strengthens this based on motion and promotes the ability of our model for depth estimation.

3) Spatial Attention Visualization: To locate spatial regions that distinguish between live and spoof faces, we use the Transformer Attribution Method [85] to realize spatial attention visualization on the partial and final networks, as shown in Fig. 10 (e)-(g). Specifically, brighter areas in attention maps represent higher weights for spoofing cues during classification. It is observed that the attention values are evenly scattered on live faces, while biased distributions are shown on spoof faces. Specifically, in spoof faces, there exist large attention values in facial parts, especially corners of the eye, and sides of the nose, indicating that these regions may contain more spoofing cues. Furthermore, compared with (e) and (g), attention activation in spoofing of (e) is merely focused on the mouth and its surrounding area while that of (g) can cover more valid areas. Besides, attention activation in living of (e) includes many noise activation regions compared with that of (e). The above phenomenon proves the effectiveness of our TDA on extracting the inherent characteristics by weighing motion-sensitive feature channels. Compared with (f) and (g), attention in spoofing of (f) shows an incomplete covering when missing some valid areas, which demonstrates the use of our PTA is beneficial to capture more potential spoofing cues by adding multiple time windows for rich short-range and long-range connections between different frames.

4) Temporal Attention Illustration: The temporal transformer also utilizes multiple self-attention layers to integrate classification information from different frames. Different

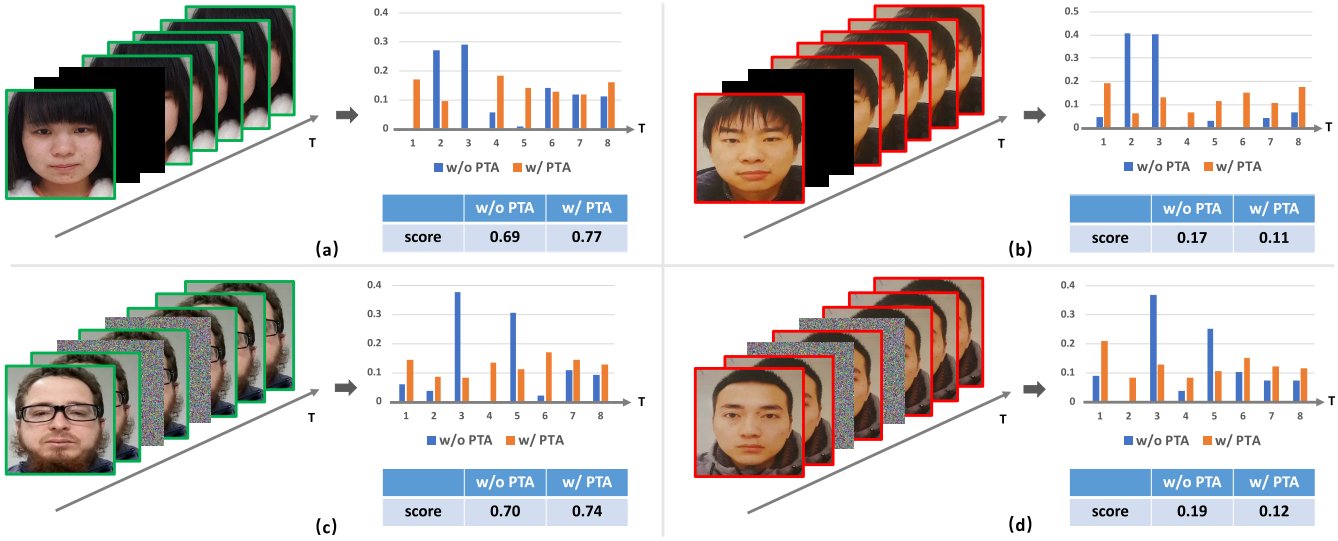


Fig. 11. Illustration of temporal attention distribution upon clips containing abnormal frames. The clips in (a) and (b) contain black frames, while clips in (c) and (d) contain noise frames. The green border represents a living video clip and the red border represents a spoofing video clip. Temporal attention weights are illustrated in histograms. The score is calculated according to Eqn. (12), which is normalized between 0 and 1. Specifically, the closer the score is to 1, the more likely to be judged as live faces. The closer the score is to 0, the more likely to be judged as spoof faces.

from the basic transformer layer, we propose a pyramid structure to capture multi-scale temporal characteristics by using multiple window attention of different sizes parallelly. Thus, each attention layer can independently concentrate on different scales of temporal information. This not only extracts the overall temporal cues, but also enhances the robustness against abnormal frames in video clips. Specifically, for the single global attention layer, abnormal frames may draw high weights due to their abnormality compared with other frames, which may damage the reasonable temporal information extraction. However, our PTA with multiple window attention of different sizes can weaken this influence by collecting local and global temporal features independently. The ablation illustration is performed in Fig. 11. The temporal attention distribution is more scattered for the model w/ PTA, which can fully utilize the classification embedding from different frames for the final decision. On the contrary, the model w/o PTA assigns large weights on abnormal frames, which may lead to missing some important classification clues that are concealed by high-weight abnormal features. The score changes before and after adding PTA prove our point of view experimentally.

V. Conclusion

In this paper, we propose a temporal transformer network (TTN) to learn rich multi-granularity temporal characteristics for face anti-spoofing (FAS). It consists of temporal difference attentions (TDA), pyramid temporal aggregation (PTA), and a temporal depth difference loss (TDL). Firstly, unlike most existing works that learn temporal features on global images, the TDA captures motion-sensitive local cues on comprehensive local patches. Secondly, the PTA aggregates features on multiple tempo speeds, learning short-range and long-range relations among different frames. Thirdly, with the motion-sensitive spoof patterns as the ground truth, the TDL can supervise networks to locate spoof facial parts accurately. To the best of our knowledge, learning temporal information via transformers for FAS has not been studied before. Experimental results on several benchmarks have demonstrated the superiority of our proposed methods.

References

[1] J. Deng, J. Guo, N. Xue, and S. Zafeiriou, "ArcFace: Additive angular margin loss for deep face recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 4690–4699.

[2] Y. Huang *et al.*, "CurricularFace: Adaptive curriculum learning loss for deep face recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 5901–5910.

[3] Q. Wang and G. Guo, "LS-CNN: Characterizing local patches at multiple scales for face recognition," *IEEE Trans. Inf. Forensics Security*, vol. 15, pp. 1640–1653, 2020.

[4] Z. Yu, Y. Qin, X. Li, C. Zhao, Z. Lei, and G. Zhao, "Deep learning for face anti-spoofing: A survey," 2021, *arXiv:2106.14948*.

[5] J. Yang, Z. Lei, and S. Z. Li, "Learn convolutional neural network for face anti-spoofing," 2014, *arXiv:1408.5601*.

[6] K. Patel, H. Han, and A. K. Jain, "Cross-database face antispoofing with robust feature representation," in *Proc. Chin. Conf. Biometric Recognit.* Cham, Switzerland: Springer, 2016, pp. 611–619.

[7] G. Pan, L. Sun, Z. Wu, and S. Lao, "Eyeblink-based anti-spoofing in face recognition from a generic webcamera," in *Proc. IEEE 11th Int. Conf. Comput. Vis.*, Oct. 2007, pp. 1–8.

[8] G. Pan, L. Sun, Z. Wu, and Y. Wang, "Monocular camera-based face liveness detection by combining eyeblink and scene context," *Telecommun. Syst.*, vol. 47, pp. 215–225, Aug. 2011.

[9] T. de Freitas Pereira, A. Anjos, J. M. De Martino, and S. Marcel, "LBP-TOP based countermeasure against face spoofing attacks," in *Proc. ACCV*. Berlin, Germany: Springer, 2012, pp. 121–132.

[10] J. Komulainen, A. Hadid, M. Pietikäinen, A. Anjos, and S. Marcel, "Complementary countermeasures for detecting scenic face spoofing attacks," in *Proc. Int. Conf. Biometrics (ICB)*, Jun. 2013, pp. 1–7.

[11] A. Agarwal, R. Singh, and M. Vatsa, "Face anti-spoofing using Haralick features," in *Proc. IEEE 8th Int. Conf. Biometrics Theory, Appl. Syst. (BTAS)*, Sep. 2016, pp. 1–6.

[12] T. A. Siddiqui *et al.*, "Face anti-spoofing with multifeature videolet aggregation," in *Proc. 23rd Int. Conf. Pattern Recognit. (ICPR)*, Dec. 2016, pp. 1035–1040.

[13] W. Bao, H. Li, N. Li, and W. Jiang, "A liveness detection method for face recognition based on optical flow field," in *Proc. Int. Conf. Image Anal. Signal Process.*, 2009, pp. 233–236.

[14] L. Feng, L. Po, and Y. Li, "Integration of image quality and motion cues for face anti-spoofing: A neural network approach," *J. Vis. Commun. Image Represent.*, vol. 38, no. 1, pp. 451–460, Jul. 2016.

[15] Z. Xu, S. Li, and W. Deng, "Learning temporal features using LSTM-CNN architecture for face anti-spoofing," in *Proc. 3rd IAPR Asian Conf. Pattern Recognit. (ACPR)*, Nov. 2015, pp. 141–145.

[16] Y. Liu, A. Jourabloo, and X. Liu, "Learning deep models for face anti-spoofing: Binary or auxiliary supervision," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 389–398.

[17] J. Gan, S. Li, Y. Zhai, and C. Liu, "3D convolutional neural network based on face anti-spoofing," in *Proc. 2nd Int. Conf. Multimedia Image Process. (ICMIP)*, Mar. 2017, pp. 1–5.

[18] D. Li, Y. Yang, Y.-Z. Song, and T. M. Hospedales, "Learning to generalize: Meta-learning for domain generalization," in *Proc. AAAI*, 2018, pp. 3490–3497.

[19] W. Zheng, M. Yue, S. Zhao, and S. Liu, "Attention-based spatial-temporal multi-scale network for face anti-spoofing," *IEEE Trans. Biometrics, Behav., Identity Sci.*, vol. 3, no. 3, pp. 296–307, Jul. 2021.

[20] S. Liu, P. C. Yuen, S. Zhang, and G. Zhao, "3D mask face anti-spoofing with remote photoplethysmography," in *Proc. ECCV*. Cham, Switzerland: Springer, 2016, pp. 85–100.

[21] Z. Yu, X. Li, P. Wang, and G. Zhao, "TransRPPG: Remote photoplethysmography transformer for 3D mask face presentation attack detection," *IEEE Signal Process. Lett.*, vol. 28, pp. 1290–1294, 2021.

[22] T. Shen, Y. Huang, and Z. Tong, "FaceBagNet: Bag-of-local-features model for multi-modal face anti-spoofing," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2019, pp. 1–6.

[23] Y. Atoum, Y. Liu, A. Jourabloo, and X. Liu, "Face anti-spoofing using patch and depth-based CNNs," in *Proc. IEEE Int. Joint Conf. Biometrics (IJCB)*, Oct. 2017, pp. 319–328.

[24] R. Cai, H. Li, S. Wang, C. Chen, and A. C. Kot, "DRL-FAS: A novel framework based on deep reinforcement learning for face anti-spoofing," *IEEE Trans. Inf. Forensics Security*, vol. 16, pp. 937–951, 2021.

[25] K.-Y. Zhang *et al.*, "Structure destruction and content combination for face anti-spoofing," in *Proc. IEEE Int. Joint Conf. Biometrics (IJCB)*, Aug. 2021, pp. 1–6.

[26] X. Yang *et al.*, "Face anti-spoofing: Model matters, so does data," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 3507–3516.

[27] A. Dosovitskiy *et al.*, "An image is worth 16×16 words: Transformers for image recognition at scale," in *Proc. Int. Conf. Learn. Represent.*, 2020, pp. 1–12.

[28] R. Shao, X. Lan, J. Li, and P. C. Yuen, "Multi-adversarial discriminative deep domain generalization for face presentation attack detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 10023–10031.

[29] Y. Qin *et al.*, "Learning meta model for zero-and few-shot face anti-spoofing," in *Proc. AAAI*, 2020, pp. 11916–11923.

[30] Z. Yu *et al.*, "Auto-fas: Searching lightweight networks for face anti-spoofing," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2020, pp. 996–1000.

[31] A. George and S. Marcel, "On the effectiveness of vision transformers for zero-shot face anti-spoofing," in *Proc. IEEE Int. Joint Conf. Biometrics (IJCB)*, Aug. 2021, pp. 1–8.

[32] Z. Wang *et al.*, "Deep spatial gradient and temporal depth learning for face anti-spoofing," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 5042–5051.

[33] A. Vaswani *et al.*, "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 5998–6008.

[34] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-end object detection with transformers," in *Proc. ECCV*. Cham, Switzerland: Springer, 2020, pp. 213–229.

[35] S. Zheng *et al.*, "Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 6881–6890.

[36] Z. Liu *et al.*, "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 10012–10022.

[37] S. Lohit, Q. Wang, and P. Turaga, "Temporal transformer networks: Joint learning of invariant and discriminative time warping," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 12426–12435.

[38] B. Yan, H. Peng, J. Fu, D. Wang, and H. Lu, "Learning spatio-temporal transformer for visual tracking," 2021, *arXiv:2103.17154*.

[39] A. Arnab, M. Dehghani, G. Heigold, C. Sun, M. Lučić, and C. Schmid, "ViViT: A video vision transformer," 2021, *arXiv:2103.15691*.

[40] H. Fan *et al.*, "Multiscale vision transformers," 2021, *arXiv:2104.11227*.

[41] I. Chingovska, A. Anjos, and S. Marcel, "On the effectiveness of local binary patterns in face anti-spoofing," in *Proc. Int. Conf. Biometrics Special Interest Group (BIOSIG)*, 2012, pp. 1–7.

[42] Z. Zhang, J. Yan, S. Liu, Z. Lei, D. Yi, and S. Z. Li, "A face antispoofing database with diverse attacks," in *Proc. 5th IAPR Int. Conf. Biometrics (ICB)*, Mar. 2012, pp. 26–31.

[43] D. Wen, H. Han, and A. K. Jain, "Face spoof detection with image distortion analysis," *IEEE Trans. Inf. Forensics Security*, vol. 10, no. 4, pp. 746–761, Apr. 2015.

[44] Z. Boulkenafet, J. Komulainen, L. Li, X. Feng, and A. Hadid, "OULU-NPU: A mobile face presentation attack database with real-world variations," in *Proc. 12th IEEE Int. Conf. Autom. Face Gesture Recognit. (FG)*, May 2017, pp. 612–618.

[45] A. Agarwal, D. Yadav, N. Kohli, R. Singh, M. Vatsa, and A. Noore, "Face presentation attack with latex masks in multispectral videos," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jul. 2017, pp. 81–89.

[46] S. Zhang *et al.*, "A dataset and benchmark for large-scale multi-modal face anti-spoofing," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 919–928.

[47] A. George, Z. Mostaani, D. Geissenbuhler, O. Nikisins, A. Anjos, and S. Marcel, "Biometric face presentation attack detection with multi-channel convolutional neural network," *IEEE Trans. Inf. Forensics Security*, vol. 15, pp. 42–55, 2020.

[48] A. Liu, Z. Tan, J. Wan, S. Escalera, G. Guo, and S. Z. Li, "CASIA-SURF CeFA: A benchmark for multi-modal cross-ethnicity face anti-spoofing," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Jan. 2021, pp. 1179–1187.

[49] *Information Technology—Biometric Presentation Attack Detection—Part 1: Framework*, ISO Standard ISO/IEC 30107-1:2016, ISO, Geneva, Switzerland, 2016. [Online]. Available: https://www.iso.org/standard/53227.html

[50] K. Zhang, Z. Zhang, Z. Li, and Y. Qiao, "Joint face detection and alignment using multitask cascaded convolutional networks," *IEEE Signal Process. Lett.*, vol. 23, no. 10, pp. 1499–1503, Oct. 2016.

[51] Y. Feng, F. Wu, X. Shao, Y. Wang, and X. Zhou, "Joint 3D face reconstruction and dense alignment with position map regression network," in *Proc. ECCV*, 2018, pp. 534–551.

[52] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2009, pp. 248–255.

[53] H. Touvron, M. Cord, M. Douze, F. Massa, A. Sablayrolles, and H. Jégou, "Training data-efficient image transformers & distillation through attention," in *Proc. Int. Conf. Mach. Learn.*, 2021, pp. 10347–10357.

[54] S. Woo, J. Park, J. Y. Lee, and I. S. Kweon, "CBAM: Convolutional block attention module," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 3–19.

[55] K.-Y. Zhang *et al.*, "Face anti-spoofing via disentangled representation learning," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2020, pp. 641–657.

[56] Y. Liu, J. Stehouwer, and X. Liu, "On disentangling spoof trace for generic face anti-spoofing," in *Proc. ECCV*. Cham, Switzerland: Springer, 2020, pp. 406–422.

[57] Z. Yu *et al.*, "Searching central difference convolutional networks for face anti-spoofing," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 5295–5305.

[58] Z. Yu, X. Li, J. Shi, Z. Xia, and G. Zhao, "Revisiting pixel-wise supervision for face anti-spoofing," *IEEE Trans. Biometrics, Behav., Identity Sci.*, vol. 3, no. 3, pp. 285–295, Jul. 2021.

[59] Z. Yu, Y. Qin, H. Zhao, X. Li, and G. Zhao, "Dual-cross central difference network for face anti-spoofing," in *Proc. 30th Int. Joint Conf. Artif. Intell.*, Aug. 2021, pp. 1–9.

[60] Y. Qin, Z. Yu, L. Yan, Z. Wang, C. Zhao, and Z. Lei, "Meta-teacher for face anti-spoofing," *IEEE Trans. Pattern Anal. Mach. Intell.*, early access, Jun. 22, 2021, doi: 10.1109/TPAMI.2021.3091167.

[61] Z. Yu, J. Wan, Y. Qin, X. Li, S. Z. Li, and G. Zhao, "NAS-FAS: Static-dynamic central difference network search for face anti-spoofing," 2020, *arXiv:2011.02062*.

[62] Y. Jia, J. Zhang, S. Shan, and X. Chen, "Single-side domain generalization for face anti-spoofing," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 8484–8493.

[63] Z. Boulkenafet, J. Komulainen, and A. Hadid, "Face antispoofing using speeded-up robust features and Fisher vector encoding," *IEEE Signal Process. Lett.*, vol. 24, no. 2, pp. 141–145, Feb. 2017.

[64] H. Li, S. J. Pan, S. Wang, and A. C. Kot, "Domain generalization with adversarial feature learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 5400–5409.

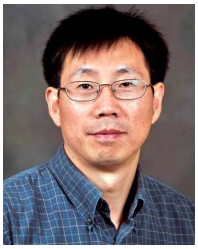
[65] R. Shao, X. Lan, and P. C. Yuen, "Regularized fine-grained meta face anti-spoofing," in *Proc. AAAI*, 2020, vol. 34, no. 7, pp. 11974–11981.

[66] L. Zhou, J. Luo, X. Gao, W. Li, B. Lei, and J. Leng, "Selective domain-invariant feature alignment network for face anti-spoofing," *IEEE Trans. Inf. Forensics Security*, vol. 16, pp. 5352–5365, 2021.
[67] J. Wang *et al.*, "VLAD-VSA: Cross-domain face presentation attack detection with vocabulary separation and adaptation," in *Proc. 29th ACM Int. Conf. Multimedia*, Oct. 2021, pp. 1497–1506.
[68] Y. Liu, J. Stehouwer, A. Jourabloo, and X. Liu, "Deep tree learning for zero-shot face anti-spoofing," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 4680–4689.
[69] Z. Yu, X. Li, X. Niu, J. Shi, and G. Zhao, "Face anti-spoofing with human material perception," in *Proc. ECCV*. Cham, Switzerland: Springer, 2020, pp. 557–575.
[70] S. R. Arashloo, J. Kittler, and W. Christmas, "An anomaly detection approach to face spoofing detection: A new formulation and evaluation protocol," *IEEE Access*, vol. 5, pp. 13868–13882, 2017.
[71] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.
[72] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, vol. 1, no. 2, pp. 4700–4708.
[73] J. Beal, E. Kim, E. Tzeng, D. H. Park, A. Zhai, and D. Kislyuk, "Toward transformer-based object detection," 2020, *arXiv:2012.09958*.
[74] S. Zhang *et al.*, "CASIA-SURF: A large-scale multi-modal benchmark for face anti-spoofing," *IEEE Trans. Biometrics, Behav., Identity Sci.*, vol. 2, no. 2, pp. 182–193, Apr. 2020.
[75] A. Liu *et al.*, "Face anti-spoofing via adversarial cross-modality translation," *IEEE Trans. Inf. Forensics Security*, vol. 16, pp. 2759–2772, 2021.
[76] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, vol. 1, Jun. 2005, pp. 886–893.
[77] T. Ojala, M. Pietikäinen, and T. Mäenpää, "Multiresolution gray-scale and rotation invariant texture classification with local binary patterns," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 24, no. 7, pp. 971–987, Jul. 2002.
[78] J. Kannala and E. Rahtu, "BSIF: Binarized statistical image features," in *Proc. 21st Int. Conf. Pattern Recognit. (ICPR)*, 2012, pp. 1363–1366.
[79] V. Ojansivu and J. Heikkilä, "Blur insensitive texture classification using local phase quantization," in *Proc. Int. Conf. Image Signal Process.* Berlin, Germany: Springer, 2008, pp. 236–243.
[80] A. George and S. Marcel, "Deep pixel-wise binary supervision for face presentation attack detection," in *Proc. Int. Conf. Biometrics (ICB)*, Jun. 2019, pp. 1–8.
[81] A. George and S. Marcel, "Learning one class representations for face presentation attack detection using multi-channel convolutional neural networks," *IEEE Trans. Inf. Forensics Security*, vol. 16, pp. 361–375, 2021.
[82] A. Parkin and O. Grinchuk, "Recognizing multi-modal face spoofing with face recognition networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2019, pp. 1–7.
[83] A. George and S. Marcel, "Cross modal focal loss for RGBD face anti-spoofing," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 7882–7891.
[84] L. Van der Maaten and G. Hinton, "Visualizing data using t-SNE," *J. Mach. Learn. Res.*, vol. 9, no. 11, pp. 1–27, 2008.
[85] H. Chefer, S. Gur, and L. Wolf, "Transformer interpretability beyond attention visualization," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 782–791.

Zhuo Wang received the B.S. degree from the Beijing University of Posts and Telecommunications (BUPT), Beijing, China, in 2019, where he is currently pursuing the M.E. degree with the School of Artificial Intelligence. His research interests include pattern recognition and computer vision, with a particular emphasis on face anti-spoofing.

Qiangchang Wang is currently pursuing the Ph.D. degree with West Virginia University. His research interests lie in the areas of computer vision and deep learning.

Weihong Deng (Member, IEEE) is currently a Professor with the School of Artificial Intelligence, Beijing University of Posts and Telecommunications. He has published over 150 technical papers in international journals and conferences, such as IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE, IEEE TRANSACTIONS ON IMAGE PROCESSING, IJCV, CVPR, and ICCV. His research interests include computer vision and affective computing, with a particular emphasis in face recognition and expression analysis. His dissertation titled "Highly accurate face recognition algorithms" was awarded the Outstanding Doctoral Dissertation Award by the Beijing Municipal Commission of Education in 2011. He has been supported by the program of New Century Excellent Talents in 2014, Beijing Nova in 2016, Young Chang Jiang Scholar, and Elsevier Highly Cited Chinese Researcher in 2020. He serves as the Area Chair for major international conferences, such as IJCAI, ACMMM, IJCB, FG, and ICME; a Guest Editor for IEEE TRANSACTIONS ON BIOMETRICS, BEHAVIOR, AND IDENTITY SCIENCE and *Image and Vision Computing* journal; and a Reviewer for dozens of international journals, such as IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE, IEEE TRANSACTIONS ON IMAGE PROCESSING, IEEE TRANSACTIONS ON INFORMATION FORENSICS AND SECURITY, IEEE TRANSACTIONS ON NEURAL NETWORKS AND LEARNING SYSTEMS, IEEE TRANSACTIONS ON AFFECTIVE COMPUTING, IEEE TRANSACTIONS ON MULTIMEDIA, IJCV, and PR.



Guodong Guo (Senior Member, IEEE) received the B.E. degree in automation from Tsinghua University, Beijing, China, and the Ph.D. degree in computer science from the University of Wisconsin-Madison, Madison, WI, USA. He is currently the Head of the Institute of Deep Learning (IDL), Baidu Research, and also affiliated with the Department of Computer Science and Electrical Engineering, West Virginia University (WVU), USA. In the past, he studied, visited or worked in several places, including the Institute of Automation, Chinese Academy of Sciences; INRIA, Sophia Antipolis, France; Ritsumeikan University, Kyoto, Japan; and Microsoft Research, Beijing, China. He has authored a book "*Face, Expression, and Iris Recognition Using Learning-Based Approaches*" (2008), co-edited two books "*Support Vector Machines Applications*" (2014) and "*Mobile Biometrics*" (2017), and coauthored a book "*Multi-Modal Face Presentation Attack Detection*" (2020). He published over 180 technical papers, and he is the creator of the visual BMI (body mass index) estimator. His research interests include computer vision, biometrics, machine learning, and multimedia. He received the North Carolina State Award for Excellence in Innovation in 2008, the New Researcher of the Year (2010–2011), and the Outstanding Researcher (2017–2018 and 2013–2014) at CEMR, WVU. He was selected as the "People's Hero of the Week" by BSJB under Minority Media and Telecommunications Council (MMTC) in 2013. Two of his papers were selected as "The Best of FG'13" and "The Best of FG'15," respectively. He is an AE of several journals, including IEEE TRANSACTIONS ON AFFECTIVE COMPUTING.