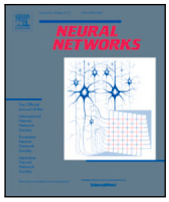
Neural Networks 175 (2024) 106275

Contents lists available at ScienceDirect

Neural Networks

journal homepage: www.elsevier.com/locate/neunet



Full Length Article

Face anti-spoofing with cross-stage relation enhancement and spoof material perception

Daiyuan Li [a,d,e], Guo Chen [a,f], Xixian Wu [b], Zitong Yu [c,*], Mingkui Tan [a,*]

[a] South China University of Technology, Guangzhou, 510006, Guangdong, China
[b] HuNan Gmax Intelligent Technology, Changsha, 410000, Hunan, China
[c] Great Bay University, Dongguan, 523000, Guangdong, China
[d] Pazhou Laboratory, Guangzhou, 510000, Guangdong, China
[e] Key Laboratory of Big Data and Intelligent Robot, Ministry of Education, Guangzhou, 510000, Guangdong, China
[f] CSSC Systems Engineering Research Institute, Beijing, 100000, Beijing, China

ARTICLE INFO

Keywords:
Face anti-spoofing
Presentation attack
Transformer
Dataset

ABSTRACT

Face Anti-Spoofing (FAS) seeks to protect face recognition systems from spoofing attacks, which is applied extensively in scenarios such as access control, electronic payment, and security surveillance systems. Face anti-spoofing requires the integration of local details and global semantic information. Existing CNN-based methods rely on small stride or image patch-based feature extraction structures, which struggle to capture spatial and cross-layer feature correlations effectively. Meanwhile, Transformer-based methods have limitations in extracting discriminative detailed features. To address the aforementioned issues, we introduce a multi-stage CNN-Transformer-based framework, which extracts local features through the convolutional layer and long-distance feature relationships via self-attention. Based on this, we proposed a cross-attention multi-stage feature fusion, employing semantically high-stage features to query task-relevant features in low-stage features for further cross-stage feature fusion. To enhance the discrimination of local features for subtle differences, we design pixel-wise material classification supervision and add a auxiliary branch in the intermediate layers of the model. Moreover, to address the limitations of a single acquisition environment and scarcity of acquisition devices in the existing Near-Infrared dataset, we create a large-scale Near-Infrared Face Anti-Spoofing dataset with 380k pictures of 1040 identities. The proposed method could achieve the state-of-the-art in OULU-NPU and our proposed Near-Infrared dataset at just 1.3GFlops and 3.2M parameter numbers, which demonstrate the effective of the proposed method.

1. Introduction

Face recognition (Deng, Guo, Xue, & Zafeiriou, 2019; *Face recognition: A literature survey*, 2003; Guo, Zhang, Hu, He, & Gao, 2016), due to its convenience and accuracy, is widely used in fields such as access control, electronic payment, and security surveillance systems. However, face recognition can be easily tricked by Presentation Attacks (PAs), where attackers copy authorized faces by various material physical media such as photos, videos, masks or 3d models, leading to security risks for face recognition-based identity verification systems. To address these issues, Face Anti-Spoofing (FAS) technology (Liu, Tan et al., 2023; Wang, Lu, Yang & Lai, 2022; Yu et al., 2022; Yu, Zhao et al., 2020; Yue et al., 2023) has attracted wide attention from both of the academic and industrial communities. This technology is proposed to determine whether a face in a video or image is real or a spoofing version presented through a physical medium. It typically serves as an initial step in face recognition procedures.

Traditional face anti-spoofing methods (Boulkenafet, Komulainen, & Hadid, 2016; Boulkenafet, Komulainen, Li, Feng, & Hadid, 2017; Peixoto, Michelassi, & Rocha, 2011) manually designed local descriptors to extract recapture cues (Komulainen, Hadid, & Pietikäinen, 2013) (*e.g.*, texture, color) from images. However, their limited presentation capability makes it challenging to implement them in practical scenarios. Deep learning-based methods learn face anti-spoofing features from data. Previous deep learning methods (Li et al., 2016; Yang, Lei, & Li, 2014) based on binary classification supervision only applied general deep learning frameworks to the FAS task, may lead the model to overfit on FAS-irrelevant facial attribution (gender, age, *etc*), resulting in poor generalization ability. Therefore, recent studies (Bekhouche,

* Corresponding authors.
E-mail addresses: selidaiyuan@mail.scut.edu.cn (D. Li), qwead134@gmail.com (G. Chen), wuxixian@gmax-ai.com (X. Wu), zitong.yu@ieee.org (Z. Yu), mingkuitan@scut.edu.cn (M. Tan).

https://doi.org/10.1016/j.neunet.2024.106275
Received 19 June 2023; Received in revised form 7 December 2023; Accepted 25 March 2024
Available online 27 March 2024

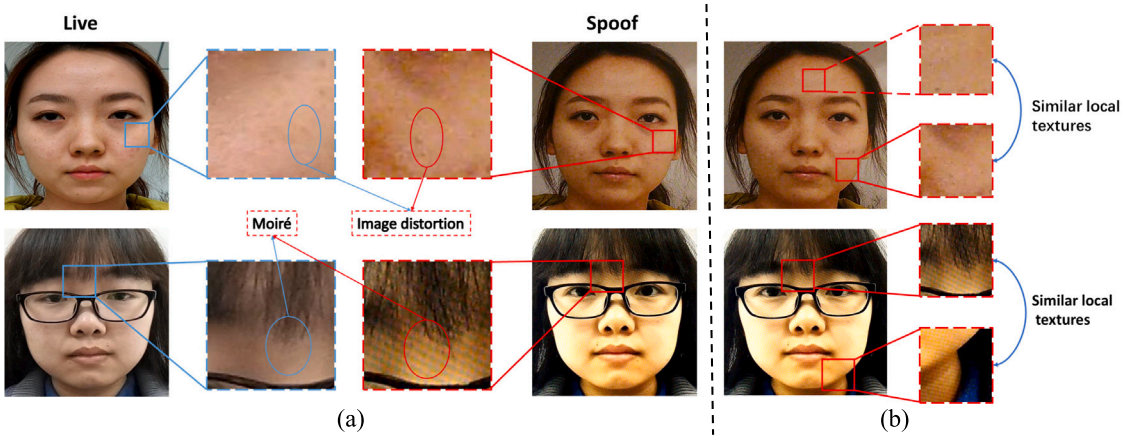


Fig. 1. (a) Capturing differences between live and spoof samples can be challenging, but by closely examining localized areas of the face, we can observe unique artificial features that are indicative of spoofing (*e.g.*, stripe, rough textures). (b) Local areas of the same type of attack samples should have similar texture patterns.

Kajo, Ruichek, & Dornaika, 2022; Liu, Stehouwer, & Liu, 2020; Liu, Zhang et al., 2021; Sun, Song, Chen, Huang, & Kot, 2020; Yu et al., 2022) tend to design more complex vision auxiliary tasks (*e.g.*, dense prediction, texture generation) to facilitate the model to extract more intrinsic face anti-spoofing features.

Recent methods (Yu, Qin, Zhao, Li & Zhao, 2021; Yu, Zhao et al., 2020) have highlighted that local features play a crucial role in face anti-spoofing tasks. Intuitively, as illustrated in Fig. 1(a), subtle texture patterns in image also reflect the differences between live and spoofing samples. In order to effectively capture detailed information, Lin et al. (2021), Wang, Yu and Zhou (2023), Yu, Li, Niu, Shi and Zhao (2020), Yu, Wan et al. (2020) devise a convolutional neural network with small strides to extract detailed information and introduce Squeeze and Extraction mechanism to re-weight features from different layers for multi-layer fusion. Some other works (Almeida et al., 2020; Cai, Li, Wang, Chen, & Kot, 2020; Wang, Lu et al., 2022) propose patch-based methods, which crop patches from the face image as inputs, and then integrate predictions of local patches to obtain the final result. However, excessive focus on local cue may lead to ambiguous predictions. The CNN-based methods adequately extract detailed pattern features but overlook the relationships between local features and cross-layer features. Also, employing small strides in structure and an ensemble inference strategy lead to substantial computational complexity and space occupation. Some recent works (Liu & Liang, 2022; Wang, Wang, Deng & Guo, 2022; Yu, Li, Wang & Zhao, 2021) apply the Vision Transformers (ViT) (Dosovitskiy et al., 2020; Han et al., 2021; Liu, Lin et al., 2021) to overcome the limitation associated with long-distance dependencies of CNN-based face anti-spoofing methods. However, ViT is ineffective in capturing fine-grained local information due to weak inductive bias (Deininger et al., 2022). Although most recent works introduce convolution (Ming et al., 2022) or local inductive bias (Lee, Kwak, & Shin, 2023) to alleviate the limitations of ViT-based methods in local feature extraction. These approaches neglect the relationship between high-level semantic features and low-level detailed information, and they fail to extract sufficient discriminative information for local details. Moreover, these methods generally involve a large number of parameters and high computational costs. However, face anti-spoofing is typically applied in computation- and memory-constrained scenarios. Therefore, a primary challenge we aim to address in this study is how to extract discriminative local image features and capture the relationships between local features and multi-level features in constrained computation resources. Furthermore, most works focus on solving the problem of visible light FAS (Boulkenafet et al., 2017). Near-Infrared (NIR) sensor is broadly applicable in some practical application scenarios (*e.g.*, dark illumination). Although multi-modal datasets (Shao, Lan, Li, & Yuen, 2019; Zhang, Liu et al., 2020) and methods (Liu, Tan et al., 2023; Yu et al., 2023) have been proposed. However, the amount of NIR data is still lacking, the diversity of Near-Infrared images in these datasets is not enough to cover complex real-world scenarios. Consequently, training the FAS model with existing NIR datasets results in poor generalization performance.

In this paper, we aim to address the aforementioned limitations. To capture both local details and their relationships in constrained computing resources, we propose a face anti-spoofing framework based on the lightweight CNN-Transformer unit STDA Decoder (Maaz et al., 2022). To efficiently fuse high and low-stage features, we aim to filter out task-irrelevant information from the low-level features before proceeding with further fusion. Unlike generally fusion methods (Chang, Chang, Hsiao, & Fu, 2020; Vandenhende, Georgoulis, & Van Gool, 2020) learning a fixed module to predict weight for features from different layer, we introduce underlying relationships between high and low-stage to guide cross-stage feature fusion. Specifically, we design a cross-stage multi-head attention feature fusion module. Given a pair of high- and low-stage features, the module generates the correlation map between the query of the high-stage feature and the key of the low-stage feature. Subsequently, the correlation map is used to reweight the low-stage features, ensuring that the weighted low-stage features preserve more task-relevant information. Furthermore, inspired by Wang, Lu et al. (2022), Yu, Li et al. (2020), we introduce the material information to improve the subtle discrimination of features. Motivated by the observation that: (1) low-stage features primarily capture local information (Child, Gray, Radford, & Sutskever, 2019), and (2) as illustrated in Fig. 1(b), the image should exhibit a consistent texture pattern across distinct local regions, we insert a pixel-wise material classification branch into the low and middle stage of our model. Moreover, we collect a large-scale NIR dataset to address small NIR data in current public datasets and the lack of diversity in capture devices and environments.

We summarize the main contributions in the following:

- To tackle the issue that existing FAS methods fail to simultaneously capture the detailed representation and cross-stage feature relationship. We introduce a lightweight CNN-Transformer-based and propose a cross-stage fusion scheme. By leveraging semantically rich high-stage features to query task-relevant information in low-stage features, the proposed method facilitates a more efficient cross-stage feature fusion.
- To improve the discriminative of detailed features, we design a pixel-wise multi-class supervision map based on material categories and insert an auxiliary pixel-wise supervision branch in the intermediate stage of our backbone accordingly. This enables us to improve the sublet discriminative of feature.

- To address the limitations of existing Near-Infrared Dataset in the diversity of identities acquisition environment and acquisition devices, we collect a comprehensive Near-Infrared face anti-spoofing Dataset, which incorporates six illumination conditions and comprises 380,000 images from 1,040 distinct identities.

2. Related works

2.1. Face anti-spoofing

Face Anti-Spoofing (FAS) plays a crucial role in ensuring the security and reliability of face recognition. Choudhury, Clarkson, Jebara, and Pentland (1999) first highlight the importance of face anti-spoofing for face recognition tasks and conducted systematic research. Traditional methods always extract local texture patterns (*e.g.*, SURF Boulkenafet et al., 2016, LBP Boulkenafet et al., 2017, HOG Peixoto et al., 2011) or capture micro-movement traces (blinking Li, 2008, head bobbing Wang, Ding, & Fang, 2009, mouth Singh, Joshi, & Nandi, 2014) from key parts of the human body for classification. These methods require lots of task-specific knowledge and exhibit limited generalization and robustness.

Binary classification face anti-spoofing. With the remarkable achievements of Convolution Neural Networks (CNN) in vision tasks, convolution neural networks be applied to solve the challenges in face anti-spoofing tasks. Early CNN-based methods treat face anti-spoofing as a binary classification task. Yang et al. (2014) first devise an end-to-end convolutional neural network for face anti-spoofing. Li et al. (2016) proposed a hybrid face anti-spoofing framework by combining pretrained VGGNet and Support Vector Machine. However, face images contain various semantic information unrelated to face anti-spoofing, such as identity, accessories, pose, expression, *etc.* Binary classification-based methods may overfit to this information, potentially compromising their generalization. Therefore, more efforts focus on creating complex auxiliary tasks or specialized convolution structures to learn more intrinsic features for face anti-spoofing tasks.

Auxiliary task design for face anti-spoofing. Some recent methods focus on learning face anti-spoofing features by designing complex tasks (*e.g.*, pixel-wise supervision Liu, Jourabloo, & Liu, 2018; Yu, Li et al., 2020; Yu, Zhao et al., 2020, image generation Liu et al., 2020; Liu, Zhang et al., 2021, others Wang, Lu et al., 2022). Liu et al. (2018) propose to replace binary classification task with a face depth regression task. In subsequent research, various pixel-wise supervision (*e.g.*, binary mask Yu, Zhao et al., 2020, reflective surface Yu, Li et al., 2020, rPPG Yu, Li et al., 2021) were introduced into the face anti-spoofing task. Liu et al. (2020) develop a fine-grained predictor by introducing a generation task that disentangles the spoofing trajectory map from the spoofing face. Wang, Lu et al. (2022) propose a patch-wise angular-margin softmax loss for fine-grained local spoofing cues mining. However, generation-based methods often lack stability and supervised approaches generally require the introduction of additional models or operations. In contrast, we introduce a straightforward pixel-wise multi-classification task based on material categories, such as human faces, paper, and screens, which enables model to conduct more nuanced discrimination of various surface types and textures pertinent to face anti-spoofing, thereby enhancing the discrimination of feature to subtle difference.

Task-specific architecture design in face anti-spoofing. Some works carefully design task-specific model architecture. Yu, Zhao et al. (2020) devise a Central Difference Convolution (CDC) module to extract invariant local features, and further developed the dual cross version (Yu, Qin et al., 2021) and the CDC-based neural network structure search scheme (Yu, Wan et al., 2020) to improve the efficiency of feature extraction. Wang, Yu et al. (2023) design a learnable gradient operator in a data-driven manner for adaptively feature extraction. These CNN-based methods adequately extract detailed pattern features but wake in capturing the relationships between local features and cross-layer features.

Face anti-spoofing for unseen scenario. Moreover, Some studies concentrate on addressing the face anti-spoofing challenges in unseen scenario that arise due to variations in environments, devices, etc. Jia, Zhang, Shan, and Chen (2020) perform single-side adversarial learning to learn a domain invariants FAS feature space. Wang, Wang et al. (2022) further extract general features by disentangling and shuffling the FAS-related and irrelated information in face image. Yue et al. (2023) devise a progressive pseudo label to generate method for sample from the unseen domain. Lin et al. introduces meta-learning for training face anti-spoofing model capable of quick adaptation (Lin et al., 2023) or lightweight framework (Lin et al., 2021) with strong generalization capabilities. To tackle the catastrophic forgetting and unseen domain generalization problems, Cai et al. (2023) propose a central difference convolutional adapter for a continual learning session.

2.2. Transformer for vision tasks

Recently, Transformers (Vaswani et al., 2017) have achieved superior performance in computer vision tasks (Chen, You, Zhang, Xi, & Le, 2022; Dai, Cai, Lin, & Chen, 2021; Dai, Zhang, Wang, Du, Yu, Liu, & Huang, 2023; Touvron et al., 2021) and natural language processing (NLP) tasks. Furthermore, DeiT (Touvron et al., 2021) leverage Transformer in a more effective way to model the local and global dependencies for image classification tasks. To better model contextual information in Transformer, some works have designed new variants. For example, Chu et al. (2021) use conditional position encoding instead of normal position encoding in ViT. Han et al. (2021) established patch-level and pixel-level representation information simultaneously through two-level encoding. PVT (Wang et al., 2021) obtains multi-level features by replacing convolutional modules in common CNNs with Transformer modules to achieve high-resolution dense prediction. Liu, Lin et al. (2021) introduce a hierarchical Transformer and sliding window to improve both efficiency and performance. Recently, convolution is widely used in Transformers (Gulati et al., 2020; Wu, Liu, Lin, Lin, & Han, 2020), such as Wu, Xiao et al. (2021) by introducing convolution token encoding and convolutional projection, which taking advantage of CNNs and Transformers in image recognition tasks.

Transformer for face anti-spoofing. Due to the advantages of Transformer in visual tasks, some recent methods have been proposed to solve the challenges in FAS using ViT. For example, Yu, Li et al. (2021) propose a lightweight ViT that can extract pure rPPG information. TransFAS (Wang, Wang et al., 2022) propose a multi-level face anti-spoofing framework based on Tiny-Deit. Liu and Liang (2022) propose a multimodal face anti-spoofing framework based on attention mechanisms. Yu et al. (2023) Masked Autoencoders-based pretrain task for multi-model face ant-spoofing. However, transformer-based face anti-spoofing methods often struggle to extract local detailed information due to their weak inductive bias. Consequently, most recent works have turned to incorporating convolution as a solution to this challenge of extracting local details. Wang, Wen, Zheng, Ying and Liu (2022) apply convolution to extract local features at the patch level, followed by employing MLP (Multi-Layer Perceptron) to facilitate the interaction of these patch features. Ming et al. (2022) employ multiple scales of attention head to learn fine-grained feature representation further. Lee et al. (2023) introduce a ConViT for local feature extraction and spatial correlations. Compared to these methods. In contrast, we introduce a lightweight CNN-Transformer-based and propose a cross-stage fusion scheme. By leveraging semantically rich high-stage features to query task-relevant information in low-stage features, the proposed method facilitates a more efficient cross-stage feature fusion.

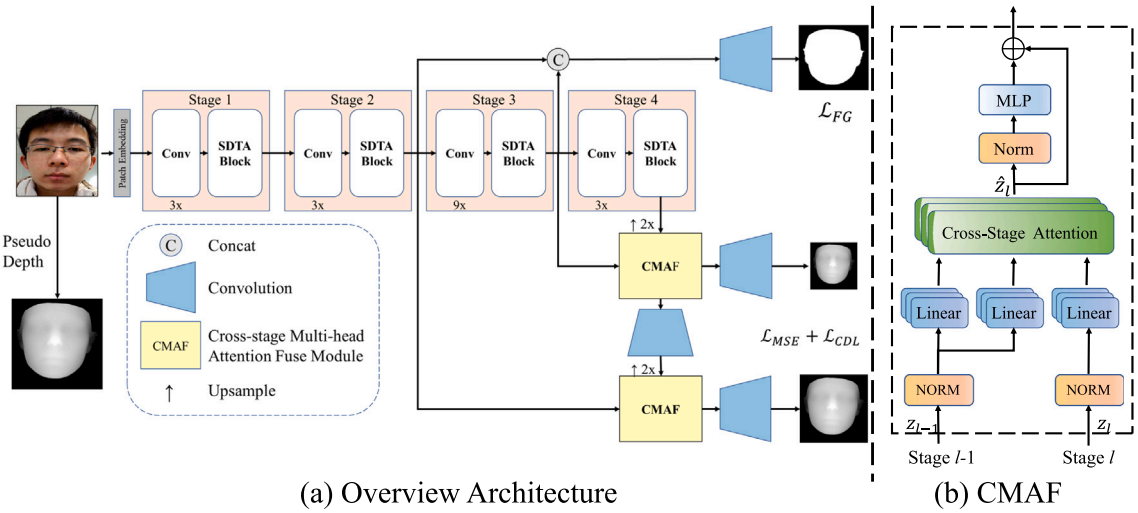


Fig. 2. Overview of the proposed framework. (a) This framework utilizes a backbone composed of four SDTA modules to extract image features. The CMAF module is used to fuse the features from adjacent stages. The model is trained with material classification loss $\mathcal{L}_{FG}$ and pseudo-depth regression loss $\mathcal{L}_{MSE} + \mathcal{L}_{CDL}$. (b) The CMAF module uses high-level features as keys to fuse the highly correlated low-level features.

3. Approach

3.1. Problem definition

Face anti-spoofing is a binary classification task that aims to identify whether a given face image $\mathbf{X} \in \mathbb{R}^{H \times W \times 3}$, is a genuine or fake face. The face anti-spoofing model takes $\mathbf{X}$ as input and outputs a liveness score. If the liveness score is greater than a predefined threshold θ, then the face is considered to be genuine, otherwise a fake face.

3.2. Architecture overview

Combining the superior local feature extraction capability of convolutional neural networks and the capability of the Transformer to model long-range local feature relationships, we propose a lightweight CNN-Transformer-based face anti-spoofing framework. As shown in Fig. 2, the framework is comprised of three components: backbone, cross-stage multi-head attention fusion block, and pixel-wise material auxiliary classification branch. The backbone deploys depthwise convolution to extract local features and employs the SDTA module (Maaz et al., 2022), which leverages a self-attentive mechanism to capture the interdependencies among these local features. Moreover, we design a Cross-Stage Multi-head Attention Fuse (CMAF) module to fully integrate features from different stages of the image. In addition, in order to enhance the model's ability to distinguish local details, we introduce a pixel-wise material classification module in the model. We will introduce each module in the following.

3.3. Cross-stage multi-head attention fusion (CMAF) module

The fusion of multi-stage features is essential for face anti-spoofing tasks (*e.g.*, CDCN Yu, Zhao et al., 2020, DC-CDN Yu, Qin et al., 2021, DSGD Wu, Zeng, Hu, Shi & Mei, 2021). However, simply concatenating or adding low-level and high-level features may disregard the direct relationship between features of different levels, which may lead to suboptimal results. As the higher-level features contain richer task-related semantic information, we hope to fuse the low-level feature maps that have a high correlation to high-level feature maps. Thus, we design a feature fusion module based on multi-head attention mechanism.

The Cross-Stage Multi-head Attention Fusion module consists of a Cross-Stage Multi-head Attention (CMA)(as shown in Fig. 3) module and an MLP module. We take $\mathbf{X}$ as input, and obtain the feature map $\mathbf{Z}_l$ of the lth stage with size $(\lfloor H/2^l \rfloor, \lfloor W/2^l \rfloor)$. The $\mathbf{Z}_{l-1}$ is used to calculate

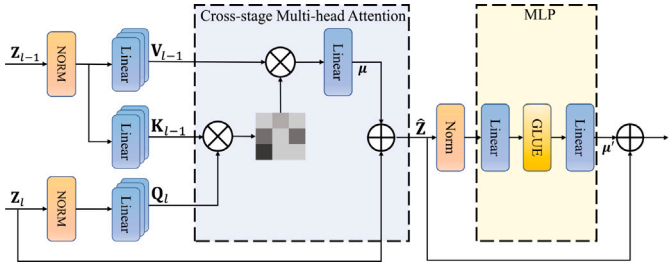


Fig. 3. An illustration of Cross-Stage Multi-head Attention Fusion Module (CMAF).

the key $\mathbf{K}_{l-1}$ and value $\mathbf{V}_{l-1}$, and the $\mathbf{Z}_l$ is used to calculate query $\mathbf{Q}_l$, which can be formulated as:

$$\mathbf{Q}_l = \text{Linear}(\text{Norm}(\mathbf{Z}_l)), \tag{1}$$

$$\mathbf{V}_{l-1}, \mathbf{K}_{l-1} = \text{Linear}\left(\text{Norm}(\mathbf{Z}_{l-1})\right), \tag{2}$$

where the Norm is LayerNorm, and the Linear is linear projection transform layers. The Cross-stage Multi-head Attention can be defined as:

$$\bar{\mathbf{Z}} = Softmax\left(\frac{\mathbf{Q}_l \mathbf{K}_{l-1}^T}{\sqrt{d}}\right)\mathbf{V}_{l-1}, \tag{3}$$

$$\hat{\mathbf{Z}} = \text{Linear}\left(\boldsymbol{\mu}\bar{\mathbf{Z}} + \mathbf{Z}_l\right), \tag{4}$$

where the Norm is LayerNorm, and the μ is a learnable scale parameters. Finally, the output of the CMA will be input into the MLP to obtain the fused feature $\mathbf{Z}'$.

$$\mathbf{Z}' = \text{MLP}\left(\text{Norm}\left(\hat{\mathbf{Z}}\right)\right) + \hat{\mathbf{Z}}, \tag{5}$$

Two step fusion strategy. Inspired by Ronneberger, Fischer, and Brox (2015), we adopt a two-step fusion strategy. Specifically, for a model with 4 stages, we first upsample the output feature map and fuse it with the feature map from stage 3. Then, after passing through a convolution layer, we upsample the feature map again and fuse it with the feature map from stage 2. The final depth prediction is obtained through the convolution operation. Specifically, we set the head number for both multi-head attention fusion modules to 4.

3.4. Pixel-wise material auxiliary supervision

To improve the detail discrimination of local features, we introduce a pixel-wise material auxiliary classification branch into our model and leverage material labels as supervision information to guide the model training. Specifically, we insert a pixel-wise material classification module in the intermediate stage of the backbone. As shown in Fig. 2, this module is comprised of two convolution layers and ReLU activation functions. It takes the concatenated intermediate feature maps of the backbone as inputs and generates pixel-wise material classification output $\mathcal{P} \in \mathbb{R}^{M\times N\times C}$, where (M, N) denotes the output size and C represents the number of material classes.

Construction of pixel-wise classification supervision. Given a face anti-spoofing dataset consisting of live faces and K types of spoofing materials. First, we construct a fine-grained class label $y \in \{l, s_1, s_2, \ldots, s_K\}$, where l denotes the live face label and s denotes the spoofing label. Then, we create a supervision map and assign the corresponding material labels y to all entries of the map. Finally, to eliminate interference from background pixels of the real face. Unlike in Sun et al. (2020), where background supervision is filtered out, we employ the mask as a supervisory signal to guide the model in learning the texture differences between the background and live faces. Specifically, we generate a binary mask by thresholding the pseudo depth map $\hat{D} \in \mathbb{R}^{N\times M\times 1}$, and set background entries of the supervision map to be 0 by conducting a AND operation on binary-mask and the supervision map. Let $\mathcal{F}$ be the masked supervision map, and let t be the threshold of the pseudo-depth. The masked supervision can then be formulated as follows:

$$\mathcal{F}_{ij} = \begin{cases} 0 & \text{if } \hat{D}_{ij} \le t \text{ and } y = L, \\ y & otherwise. \end{cases} \tag{6}$$

Let $\hat{\mathcal{F}} \in \{0,1\}^{M\times N\times C}$ be the one-hot encoding of the material supervision map $\mathcal{F}$. The pixel-wise material classification loss function $\mathcal{L}_{FG}$ is defined as follows.

$$\mathcal{L}_{FG} = -\frac{1}{MN}\sum_{i=0}^{M}\sum_{j=0}^{N}\sum_{k=0}^{C}\hat{\mathcal{F}}_{ijk}\log(\mathcal{P}_{ijk}), \tag{7}$$

where the $C = K + 2$ (C denote the total number of categories, which includes all spoofing material categories K, as well as the background and live face categories.)

3.5. Training and testing

In the training phase, we optimize both pixel-wise material classification loss and pseudo-depth estimation loss. Specifically, the pseudo depth estimation loss consists of mean square error loss $\mathcal{L}_{MSE}$ (Yu, Zhao et al., 2020) and contrast depth loss $\mathcal{L}_{CDL}$ (Wang et al., 2018). The overall loss can be formulated as:

$$\mathcal{L}_{Overall} = \mathcal{L}_{MSE} + \mathcal{L}_{CDL} + \lambda\mathcal{L}_{FG}, \tag{8}$$

where λ denote the hyper-parameter of $\mathcal{L}_{FG}$.

In the testing phase, we use the pixel-wise material classification results to refine the pseudo depth prediction $\mathcal{D} \in \mathcal{R}^{N\times M\times 1}$. Specifically, we perform element-wise multiplication between the material prediction map of the live face channel and the pseudo-depth prediction map. Following previous works (Yu, Li et al., 2020; Yu, Qin et al., 2021; Yu, Zhao et al., 2020), the average of the refined pseudo depth prediction is used for classification, which can be formulated as follows:

$$LivenessScore = \frac{\sum_{i=0}^{M}\sum_{j=0}^{N}\mathcal{P}_{ijl}D_{ij}}{NM}. \tag{9}$$

4. Face anti-spoofing in near-infrared dataset

4.1. Construction of the face anti-spoofing in near-infrared dataset

Near-Infrared sensors are commonly used image capture devices for edge-based face recognition applications. Despite the availability of a growing number of face anti-spoofing datasets that incorporate Near-Infrared data (Agarwal et al., 2017; Bhattacharjee et al., 2018; Chingovska et al., 2016; George et al., 2019; Heusch et al., 2020; Xiao et al., 2019) (see Table 1) . But, several shortcomings within the existing datasets: (1) The acquisition environment is relatively simplistic, and the scene lacks the necessary complexity. (2) The variety of Near-Infrared devices is limited and multiple Near-Infrared band sensors is not taken into account. To address the aforementioned drawbacks, we introduce a large-scale NIR face anti-spoofing dataset called FASN[1], which consists of 1043 subjects and 380k images. Compared to existing datasets, our dataset has the following characteristics: (1) We use a variety of Near-Infrared devices (containing 2 types of Near-Infrared bands) for data acquisition. To the best of our knowledge, the proposed dataset has a richer variety of Near-Infrared devices than existing datasets. (2) Unlike other datasets, the collection environment is artificially controlled during the collection process, our dataset is captured from real industrial applications. The variable environmental factors in the real application scenarios are included to make the data contain richer lighting and background environment.

4.2. Data collection

We acquire image data using three different Near-Infrared cameras: (1) 850 nm Near-Infrared USB camera with a resolution of 640 × 480. (2) 850 nm edge-device Near-Infrared camera with resolution of 640 × 480, and (3) 940 nm edge-device Near-Infrared camera with resolution of 960 × 1280. 1032 people participated, two capture scenarios (indoor, outdoor), and three lighting directions (front light, lateral light, back light) are considered. We also randomly selected multiple lighting intensities for data collection. To increase the variety of attack materials, we select 7 different materials (*e.g.*, ordinary A4 paper, copper plate paper, sulfate paper, glossy paper, white cardboard, and suede) to print real face photos. Our proposed dataset consists of four different types of photos as they can be a threat to face recognition systems under Near-Infrared sensing. This dataset contains 380k images, including 237,780 training images and 145,047 testing samples (The samples of NIR dataset are visualized in Fig. 4).

4.3. Evaluation protocol

We find that changes in capture devices and presentation materials can negatively impact the performance of face anti-spoofing models in industrial applications. Based on this, we design two types of evaluation protocol. *Protocol 1*: The first protocol is designed for material generalization ability evaluation. In this setting, the spoofing samples in the training set and testing set are printed on different materials. *Protocol 2*: The second protocol is designed to evaluate the model's generalization capability in handling device variations. Specifically, We use the data acquired by the 850 nm Near-Infrared camera for training and the data acquired by the 940 nm Near-Infrared camera for testing. The data scale of the two protocols is illustrated in Table 2.

[1] The Face Anti-spoofing in Near-Infrared Dataset available in: https://github.com/SCUT-AILab/FASN

Table 1
Comparison with other face anti-spoofing datasets containing Near-Infrared data.

Dataset	Identity	Number of image/video	Type of NIR camera	Attack type
MSSPOOF (Chingovska, Erdogmus, Anjos, & Marcel, 2016)	21	4704 (video)	1	Print
MLFP (Agarwal et al., 2017)	10	1350 (video)	1	Mask
CSMAD (Bhattacharjee, Mohammadi, & Marcel, 2018)	14	308 (video)	1	Mask
3DMA (Xiao et al., 2019)	67	920 (video)	1	Mask
CASIA-SURF (Zhang, Liu et al., 2020)	1000	21000 (video)	1	Print, Cut
WMCA (George et al., 2019)	72	1941 (video)	2	Print, Mask
WMCA-HQ (Heusch, George, Geissbühler, Mostaani, & Marcel, 2020)	51	2904 (video)	1	Print, Mask, Replay, MakeUp
FASN (Ours)	1032	380k (images)	3	Print, Cut

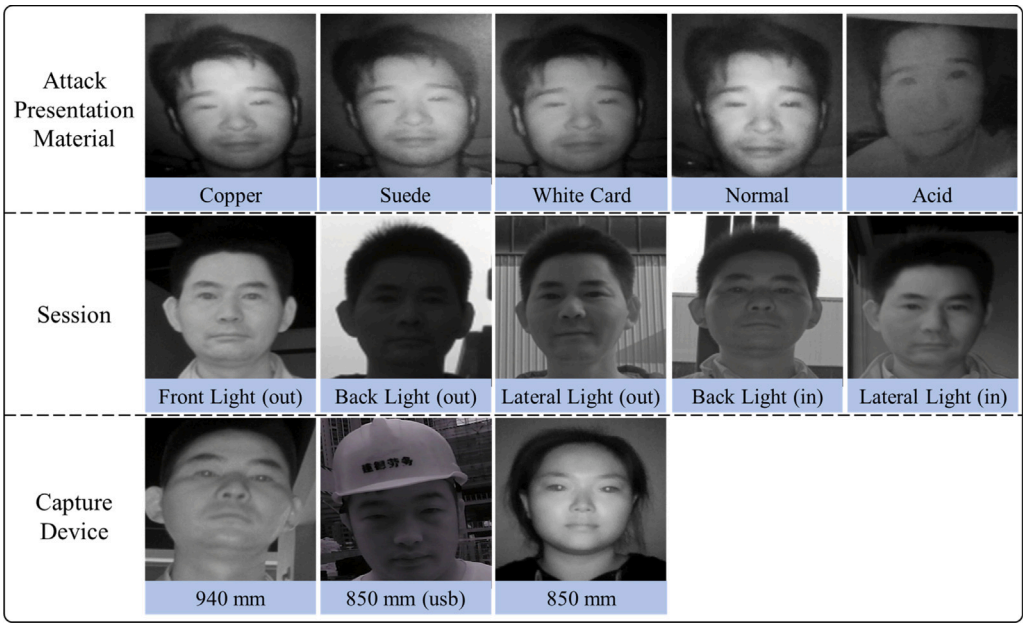


Fig. 4. Sample images of different capture settings in FASN. The first row showcases samples of spoofing attacks performed using different materials. The second row displays samples captured under varying lighting conditions (The terms "in" and "out" represent indoor and outdoor environments, respectively). The third row presents faces captured using different NIR cameras.

Table 2
The number of samples included in two testing protocols of the FASN dataset.

Protocol	Number of TrainSet		Number of TestSet	
	Live	Spoof	Live	Spoof
Protocol 1	72 609	165 171	30 362	65 840
Protocol 2	72 609	165 171	10 903	37 942

5. Experiments and results

5.1. Dataset and evaluation metric

We will conduct experiments on 5 visible FAS datasets, including OULU-NPU (Boulkenafet et al., 2017), SiW (Liu et al., 2018), CASIA-FASD (Zhang, Liu et al., 2020), Replay-Attack (Costa-Pazo, Bhattacharjee, Vazquez-Fernandez, & Marcel, 2016), MSU-MFSD (Chingovska, Anjos, & Marcel, 2012), and our NIR dataset.

The OULU-NPU dataset contains 6 cameras and 3 sessions, as well as two types of printed spoof face and two types of replayed spoof face. Protocols 1, 2, and 3 assess the model performance in cross-camera, cross-session, and cross-spoof-type scenarios, respectively. Protocol 4 is the most difficult, as it evaluates the model performance in a simultaneous cross-camera, cross-session, and cross-spoof-type scenario.

The SIW dataset captures 160 targets, each with 8 live videos and 11 prosthetic videos. The dataset was designed with three protocols. Protocol 1, 2 and 3 evaluate the generalization capability on pose, cross-spoof medium, and cross-spoof material, respectively.

The CASIA-FASD dataset contains live and spoof faces captured from 50 genuine subjects. Three attack manners are used to create spoof faces, each of which is recorded with three imaging qualities.

The MSU-MFSD dataset utilizes two different cameras to record all live and spoof faces from 35 genuine subjects. Three types of spoof faces are included, comprising two types of replayed faces and one type of printed face. Consequently, each subject has 2 kinds of live faces and 6 kinds of spoof faces captured with the two cameras.

The Replay-Attack dataset captures all live and spoof faces from 50 genuine subjects under two different lighting conditions. Five attack manners, including four types of replayed faces and one type of printed face, are used to capture spoof faces.

Evaluation Metrics. Referring to previous work, we evaluate the performance of the face anti-spoofing model using Attack Presentation Classification Error Rate (APCER), Bona Fide Presentation Classification Error Rate (BPCER), and Average Classification Error Rate (ACER) metrics in within-dataset experiments, where APCER denotes the proportion of misclassified spoof attack samples to total spoof samples, BPCER denotes the proportion of misclassified live face samples to total live samples, and ACER = (APCER + BPCER) / 2. The Area Under Curve (AUC) metric is introduced in the cross-dataset experiments.

5.2. Implementation details

Image and ground-truth depth map processing. We use the MTCNN (Zhang, Zhang, Li, & Qiao, 2016) to perform face detection and resize the face area to 256×256. Additionally, for live face images,

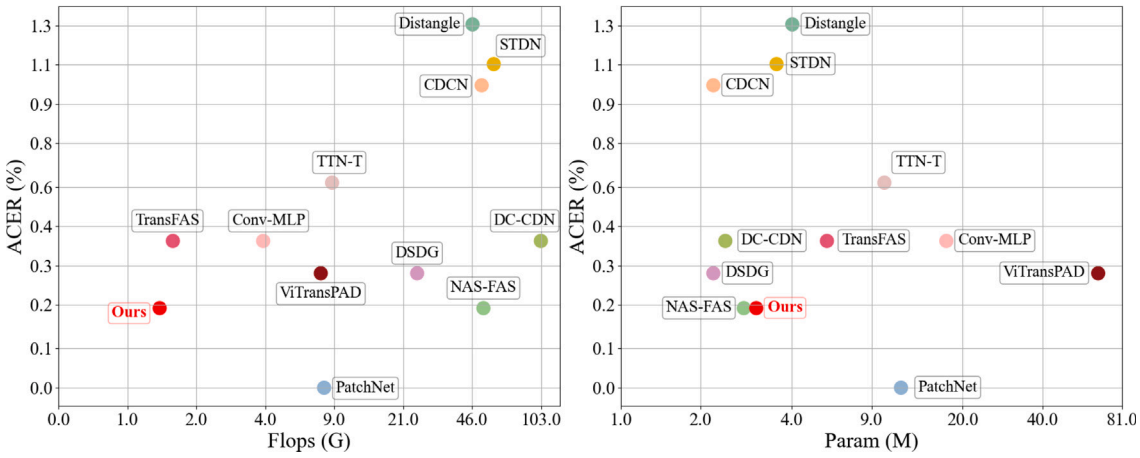


Fig. 5. Visualization of ACER (%) in OULU-NPU p1, versus number of Flops (G) (left) and parameters (M) (right).

we generate pseudo-depth faces using PRNet (Feng, Wu, Shao, Wang, & Zhou, 2018) and normalized its depth value to the range of [0, 1], the depth map is also resized to 32 × 32 as the face anti-spoofing ground-truth. For spoof images, an all-0 image is generated as the ground-truth.

Training and testing setting. We implement our methods based on the PyTorch. In the training phase, we train the model for 600 epochs, with an initial learning rate of 1e-4 and weight decay of 1e-5. The learning rate is decreased to 1e-5 using the cosine decay method. Moreover, we employ 20 epoch warm-up at the start of the training process. In the testing phase, we calculate the mean value of the dot product result of predicted depth map and material map as the final score.

5.3. Comparison experiments

Results on Intra-Dataset Experiment. Here we first evaluate our method in 4 protocols of OULU-NPU. As shown in Table 3. the proposed method obtained the second-lowest ACER in protocol 1 of OULU-NPU, which is only **0.2%** higher than PatchNet (Wang, Lu et al., 2022). Moreover, our method achieve the best ACER of **0.9%** and **2.3%** in protocol 3 and 4, respectively, and rank second with an ACER of **1.0%** in protocol 2. Moreover, as shown in Fig. 5, Among all methods, our approach achieves the optimal trade-off between computational efficiency(Flops) and performance (ACER), and it also benefits from having a smaller number of parameters. Overall, we achieve optimal performance in two protocols of this dataset at a relatively low computational cost and with a small number of parameters. We further compare with competitors on SiW. As shown in Table 4. we achieved ACER of 0.08%, 0.08%, and 2.46% in protocol 1, 2, and 3 of the SiW dataset, which achieve a comparable result in all of methods.

Results on Cross-dataset Experiment. The proposed method is evaluated for its generalizability across different FAS datasets with larger domain gap such as presentation material, acquisition device, and illumination, *etc*. We evaluate domain generalization performance on cross-dataset benchmarks, which include OULU-NPU (O), CASIA-FASD (C), MSU-MFSD (M), and Replay-Attack (I). We select three as the source domains for training, and one remaining database as the target domain for testing. Thus, there are four experimental result in total: O&C&I to M, O&M&I to C, O&C&M to I, and I&C&M to O. As illustrated in Table 5, we divide the methods into two groups for comparison. In the first 9 lines, we list the method focusing on the FAS in cross-domain scenarios, and in the last 7 lines, we compare with general FAS frameworks such as Auxiliary (Liu et al., 2018), CDCN (Yu, Zhao et al., 2020), NAS-FAS (Yu, Wan et al., 2020), PatchNet (Wang, Lu et al., 2022), TransFAS (Wang, Wang et al., 2022), *etc*. The results are reported as the HTER and AUC scores. The proposed method achieve the best HTER and AUC in the O&M&I to C experiment, and take the 3rd place in the O&C&M to I experiment. These competitive results demonstrate that the proposed method has a generalization ability across different domains. Furthermore, comparing with other general face anti-spoofing methods in last 7 line in Table 5, the proposed method is no less generalizable than PatchNet, with fewer computations and parameters.

Results on FASN. To validate the effectiveness of the proposed method on NIR device, we conducted experiments on our Near-Infrared face anti-spoofing dataset FASN. The proposed method is compared with a general deep convolutional neural network model Resnet18 and state-of-the-art FAS method in Visible spectrum domain such as CDCN, DC-CDN and PatchNet. All re-implements in Table 6 follow the default hyper-parameters setting in original papers. As shown in Table 6, the proposed method achieved the best results with ACERs of **0.46**% and **2.7**% on the two tested protocols (across attack materials and across NIR sensor devices) in the NIR dataset. The proposed method outperforms Resnet18, CDCN, and DC-CDN in two protocols of the NIR dataset, which demonstrates the general applicability of our method in both visible and Near-Infrared sensors.

5.4. Ablation study

Effectiveness of each module. We conduct ablation experiments on the easiest and most difficult settings of the OULU-NPU dataset, protocol 1 and protocol 4 (referred to as P1 and P4) respectively. First, we verify the importance of each module proposed in this paper. As shown in Table 7, we compared the following schemes: (1) Feature fusion method: The Concat indicates that the features of multiple stages are concatenated in the channel dimension and then fused by a convolution operation; the Attention indicates that the cross-stage multi-head attention scheme proposed in this paper. (2) Pixel-wise material classification indicates whether to use pixel-wise material classification supervision branches. P1@ACER and P4@ACER respectively indicate the ACER metrics at P1 and P4. Comparing the results in rows 1 to 3 of Table 7, it can be seen that the best ACER results can be achieved by introducing the Attention-based multi-stage fusion method, while simply concatenating the features and then fusing them using convolution performs worse than Cross-Stage Attention method on P1, and even achieves the opposite effect on P4. This proves that simply concatenating multi-layer features cannot consistently bring superior results. Comparing the first three rows of the table with the last three rows, it can be seen that there is a consistent improvement in the overall performance of the model after introducing the intermediate layer pixel-wise classification supervision.

Effectiveness of different feature interaction module. We conduct comparison experiments on both OULU-NPU and our Near-Infrared

Table 3
Comparison with state-of-the-art methods on four protocols of OULU-NPU.

	Method	APCER (%)↓	BPCER (%)↓	ACER (%)↓
Protocol 1	Distangle (Zhang, Yao et al., 2020)	1.7	0.8	1.3
	STDN (Liu et al., 2020)	0.8	1.3	1.1
	CDCN (Yu, Wan et al., 2020)	0.4	1.7	1.0
	LGON (Wang, Yu et al., 2023)	1.5	0.0	0.8
	Conv-MLP (Wang, Wen et al., 2022)	2.5	3.2	0.8
	DC-CDN (Yu, Qin et al., 2021)	0.5	0.3	0.4
	TTN-T (Wang, Wang, Deng & Guo, 2022)	1.2	0.0	0.6
	TransFAS (Wang, Wang et al., 2022)	0.8	0.0	0.4
	DSCE (Liu, Wu, Li & Wang, 2023)	1.3	0.8	0.3
	ViTransPAD (Ming et al., 2022)	0.4	0.2	0.3
	DSDG (Wu, Zeng et al., 2021)	0.6	0.0	0.3
	NAS-FAS (Yu, Wan et al., 2020)	0.4	0.0	0.2
	PatchNet (Wang, Lu et al., 2022)	**0.0**	**0.0**	**0.0**
	Ours	0.4	0.0	0.2
Protocol 2	Distangle (Zhang, Yao et al., 2020)	1.1	3.6	2.4
	STDN (Liu et al., 2020)	2.3	1.6	1.9
	LGON (Wang, Yu et al., 2023)	2.3	1.4	1.9
	CDCN (Yu, Wan et al., 2020)	1.5	1.4	1.5
	DC-CDN (Yu, Qin et al., 2021)	0.7	1.9	1.3
	NAS-FAS (Yu, Wan et al., 2020)	1.5	0.8	1.2
	DSDG (Wu, Zeng et al., 2021)	1.5	0.8	1.2
	PatchNet (Wang, Lu et al., 2022)	1.1	1.2	1.2
	ViTransPAD (Ming et al., 2022)	2.0	**0.4**	1.2
	TransFAS (Wang, Wang et al., 2022)	1.5	0.5	1.0
	DSCE (Liu, Wu et al., 2023)	0.7	1.4	1.1
	TTN-T (Wang, Wang et al., 2022)	1.5	0.5	1.0
	Conv-MLP (Wang, Wen et al., 2022)	**0.0**	1.6	**0.8**
	Ours	1.5	**0.5**	1.0
Protocol 3	Distangle (Zhang, Yao et al., 2020)	2.8 ± 2.2	1.7 ± 2.6	2.2 ± 2.2
	STDN (Liu et al., 2020)	1.6 ± 1.6	4.0 ± 5.4	2.8 ± 3.3
	CDCN (Yu, Wan et al., 2020)	2.4 ± 1.3	2.2 ± 2.0	2.3 ± 1.4
	Conv-MLP (Wang, Wen et al., 2022)	2.5 ± 1.0	2.0 ± 0.8	2.2 ± 0.6
	ViTransPAD (Ming et al., 2022)	3.1 ± 3.0	1.0 ± 1.3	2.0 ± 1.5
	DC-CDN (Yu, Qin et al., 2021)	2.2 ± 2.8	1.6 ± 2.1	1.9 ± 1.1
	NAS-FAS (Yu, Wan et al., 2020)	2.1 ± 1.3	1.4 ± 1.1	1.7 ± 0.6
	DSCE (Liu, Wu et al., 2023)	1.5 ± 1.3	1.4 ± 1.3	1.5 ± 1.1
	DSDG (Wu, Zeng et al., 2021)	1.2 ± 0.8	1.7 ± 3.3	1.4 ± 1.5
	PatchNet (Wang, Lu et al., 2022)	1.8 ± 1.5	**0.6 ± 1.2**	1.2 ± 1.3
	TTN-T (Wang, Wang et al., 2022)	0.8 ± 0.9	1.4 ± 1.8	1.1 ± 0.9
	LGON (Wang, Yu et al., 2023)	1.3 ± 0.9	0.8 ± 0.9	1.0 ± 0.6
	TransFAS (Wang, Wang et al., 2022)	**0.6 ± 0.7**	1.1 ± 2.5	**0.9 ± 1.1**
	Ours	0.7 ± 0.9	1.1 ± 2.7	0.9 ± 1.3
Protocol 4	CDCN (Yu, Wan et al., 2020)	4.6 ± 4.6	9.2 ± 8.0	6.9 ± 2.9
	Distangle (Zhang, Yao et al., 2020)	5.4 ± 2.9	3.3 ± 6.0	4.4 ± 3.0
	DC-CDN (Yu, Qin et al., 2021)	5.4 ± 3.3	2.5 ± 4.2	4.0 ± 3.1
	TTN-T (Wang, Wang et al., 2022)	4.2 ± 2.4	3.8 ± 4.0	4.0 ± 2.3
	Conv-MLP (Wang, Wen et al., 2022)	6.4 ± 4.5	3.4 ± 5.1	4.9 ± 4.8
	ViTransPAD (Ming et al., 2022)	2.3 ± 3.6	5.2 ± 5.4	3.8 ± 4.2
	DSCE (Liu, Wu et al., 2023)	4.2 ± 3.0	1.7 ± 2.6	3.0 ± 1.9
	NAS-FAS (Yu, Wan et al., 2020)	4.2 ± 5.3	1.7 ± 2.6	2.9 ± 2.8
	PatchNet (Wang, Lu et al., 2022)	2.5 ± 3.8	3.3 ± 3.7	2.9 ± 3.0
	TransFAS (Wang, Wang et al., 2022)	2.1 ± 2.2	3.8 ± 3.5	2.9 ± 2.4
	STDN (Liu et al., 2020)	2.3 ± 3.6	5.2 ± 5.4	3.8 ± 4.2
	LGON (Wang, Yu et al., 2023)	3.3 ± 2.6	3.3 ± 4.1	3.3 ± 2.6
	DSDG (Wu, Zeng et al., 2021)	**2.1 ± 1.0**	2.5 ± 4.2	2.3 ± 2.3
	Ours	2.9 ± 2.9	**1.7 ± 2.6**	**2.3 ± 2.2**

Face Antis-Spoofing datasets to evaluate the cross-stage feature fusion module. Specifically, we replaced our cross-stage multi-head attention module with feature fusion modules MTAN following Liu et al. (2019), FPM following Vandenhende et al. (2020), and CLAM following Chang et al. (2020) and report the ACER. (all of the above modules were implemented based on the authors' code). As shown in Table 8. Our cross-attention module has achieved the best ACER in 4 protocols of OULU-NPU and protocol 1 of our Near-Infrared Face Anti-Spoofing dataset (FSAN). Moreover, our cross-attention fusion method has achieved state-of-the-art performance without introducing significant additional computational or memory costs. The experimental results demonstrate the effectiveness of our fusion method in face anti-spoofing tasks.

Effectiveness of head number in cross-attention module. We supplement the ablation experiment in OULU-NPU P1 and P4 and respectively report the ACER of the model when number of heads = 1, 2, 4 (Our proposed module requires a consistent head count in its inputs. However, the maximum common divisor for the dimensions of backbone features is 4. Thus, we set the maximum number of heads to 4 in ablation experiments). As shown in Table 9, the model achieves the best performance when the number of heads in the cross-stage module is set to 4. Moreover, the performance consistently improves with an increase in the number of heads.

Effectiveness of hyper-parameter λ. In this ablation experiment, we explore how the hyper-parameter λ of the pixel-wise material classification branch affects the face anti-spoofing performance. We conduct hyper-parameter experiments on Protocol 1 and Protocol 4 of the OULU-NPU dataset with different values of λ. Specifically, we set $\lambda = 0.0, 0.2, 0.4, 0.6, 0.8, 1$ and report the ACER of the model. As shown

Table 4
Comparison with state-of-the-art methods on three protocol of SiW.

	Method	APCER (%)↓	BPCER (%)↓	ACER (%)↓
Protocol 1	NAS-FAS-base (Yu, Wan et al., 2020)	0.34	1.58	0.96
	Distangle (Zhang, Yao et al., 2020)	0.07	0.50	0.28
	NAS-FAS (Yu, Wan et al., 2020)	0.07	0.17	0.12
	CDCN (Yu, Wan et al., 2020)	0.07	0.17	0.12
	LGON (Wang, Yu et al., 2023)	0.06	0.00	0.03
	DSDG (Wu, Zeng et al., 2021)	**0.00**	**0.00**	**0.00**
	STDN (Liu et al., 2020)	**0.00**	**0.00**	**0.00**
	PatchNet (Wang, Lu et al., 2022)	**0.00**	**0.00**	**0.00**
	TransFAS (Wang, Wang et al., 2022)	**0.00**	**0.00**	**0.00**
	Ours	0.00	0.16	0.08
Protocol 2	NAS-FAS-base (Yu, Wan et al., 2020)	0.18 ± 0.24	0.28 ± 0.07	0.23 ± 0.18
	Distangle (Zhang, Yao et al., 2020)	0.08 ± 0.17	0.13 ± 0.09	0.10 ± 0.04
	CDCN (Yu, Wan et al., 2020)	0.00 ± 0.00	0.13 ± 0.09	0.06 ± 0.04
	LGON (Wang, Yu et al., 2023)	0.00 ± 0.00	0.12 ± 0.08	0.06 ± 0.04
	NAS-FAS (Yu, Wan et al., 2020)	0.00 ± 0.00	0.09 ± 0.10	0.04 ± 0.05
	DSDG (Wu, Zeng et al., 2021)	**0.00 ± 0.00**	**0.00 ± 0.00**	**0.00 ± 0.00**
	STDN (Liu et al., 2020)	**0.00 ± 0.00**	**0.00 ± 0.00**	**0.00 ± 0.00**
	PatchNet(Wang, Lu et al., 2022)	**0.00 ± 0.00**	**0.00 ± 0.00**	**0.00 ± 0.00**
	TransFAS (Wang, Wang et al., 2022)	**0.00 ± 0.00**	**0.00 ± 0.00**	**0.00 ± 0.00**
	Ours	0.00 ± 0.00	0.16 ± 0.00	0.08 ± 0.00
Protocol 3	STDN (Liu et al., 2020)	8.30 ± 3.30	7.50 ± 3.30	7.90 ± 3.30
	Distangle (Zhang, Yao et al., 2020)	9.35 ± 6.14	1.84 ± 2.60	5.59 ± 4.37
	NAS-FAS-base (Yu, Wan et al., 2020)	3.67 ± 1.04	7.35 ± 1.56	5.51 ± 1.23
	DSDG (Wu, Zeng et al., 2021)	3.75 ± 1.46	3.85 ± 1.42	3.80 ± 1.44
	PatchNet (Wang, Lu et al., 2022)	3.06 ± 1.10	1.83 ± 0.83	2.45 ± 0.45
	TransFAS (Wang, Wang et al., 2022)	1.95 ± 0.40	1.92 ± 0.11	1.94 ± 0.26
	LGON (Wang, Yu et al., 2023)	3.00 ± 2.40	**0.52 ± 0.24**	1.78 ± 1.08
	CDCN (Yu, Wan et al., 2020)	1.67 ± 0.11	1.76 ± 0.12	1.71 ± 0.11
	NAS-FAS (Yu, Wan et al., 2020)	**1.58 ± 0.23**	1.46 ± 0.08	**1.52 ± 0.13**
	Ours	2.13 ± 1.22	2.25 ± 1.06	2.19 ± 1.14

Table 5
Comparison with SOTA methods on Cross-dataset Setting.

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↑
MADDG (Shao et al., 2019)	17.69	88.06	24.50	84.51	22.19	84.99	27.89	80.02
SSDG-M (Jia et al., 2020)	16.67	90.47	23.11	85.45	18.21	94.61	25.17	81.83
ANRL (Liu, Zhang et al., 2021)	16.03	91.04	10.83	96.75	17.85	89.26	15.67	91.90
DRDG (Liu, Zhang et al., 2021)	15.56	91.79	12.43	95.81	19.05	88.79	15.63	91.75
RFM (Shao, Lan, & Yuen, 2020)	13.89	93.98	20.27	88.16	17.30	90.48	16.45	91.16
SSAN-M (Wang, Wang et al., 2022)	10.42	94.76	16.47	90.81	14.00	94.58	19.51	88.17
AMEL (Zhou et al., 2022)	10.23	96.62	11.88	94.39	18.60	88.79	15.67	91.90
EBDG (Du, Li, Zuo, Zhu, & Lu, 2022)	9.56	97.17	18.34	90.01	18.69	92.28	15.66	92.02
DRDN (He, Peng, & Long, 2023)	12.9	92.1	22.40	84.50	12.40	94.50	17.1	89.20
NDA-FAS (Wang, Liu, Zheng, Ying & Wen, 2023)	**4.29**	**99.18**	12.67	94.21	**7.50**	**96.79**	20.21	87.26
DiVT-V(Tiny) (Liao et al., 2023)	7.14	98.27	11.89	95.17	11.43	**97.00**	15.42	92.97
SSDG-R (Jia et al., 2020)	7.38	97.17	10.44	95.94	11.71	96.59	15.61	91.54
SSAN-R (Wang, Wang et al., 2022)	6.67	98.75	10.00	96.67	8.88	96.79	**13.72**	93.63
Auxiliary (Liu et al., 2018)	22.72	85.88	33.52	73.15	29.14	71.69	30.17	77.61
NAS-FAS-base (Yu, Wan et al., 2020)	19.53	88.63	16.54	90.18	14.51	93.84	13.80	93.43
DC-CDN (Yu, Qin et al., 2021)	17.08	89.04	33.12	73.88	23.88	83.77	22.35	85.36
CDCN (Yu, Wan et al., 2020)	15.47	91.19	33.12	73.24	23.38	83.43	21.20	85.87
NAS-FAS (Yu, Wan et al., 2020)	14.63	94.26	17.24	87.48	19.73	88.52	19.81	86.80
ConViT (Lee et al., 2023)	12.92	93.29	17.78	88.10	18.75	91.92	15.90	90.54
PBMS-GSAL (Huang & Wang, 2023)	12.92	93.29	17.78	88.10	18.75	91.92	15.90	90.90
PatchNet (Wang, Lu et al., 2022)	7.10	98.46	11.33	94.58	13.40	95.67	11.82	**95.07**
TransFAS (Wang, Wang et al., 2022)	7.08	96.69	9.81	96.13	10.12	95.53	15.53	91.10
Ours	12.92	94.33	**9.26**	**96.98**	10.87	95.46	15.13	91.43

Table 6
Comparison with SOTA methods on two protocol of FASN.

	Method	AUC (%)↓	ACER (%)↓	FLops(G)↓
Protocol 1	Resnet18	99.71	2.61	2.3
	CDCN	99.75	1.60	51.6
	DC-CDC	99.87	0.57	103.0
	PatchNet (9 patch)	99.91	0.49	8.2
	Ours	**99.91**	**0.46**	**1.2**
Protocol 2	Resnet18	90.03	17.18	2.3
	CDCN	72.54	32.93	51.6
	DC-CDC	72.96	32.88	103.0
	PatchNet (9 patch)	98.43	4.50	8.2
	Ours	**98.93**	**2.77**	**1.2**

Table 7
Ablation studise on OULU-NPU protocol 1 and protocol 4.

Feature fusion methods		Pixel-wise material classification	P1@ACER (%)↓	P4@ACER (%)↓
Concat	Attention			
			1.4	4.16 ± 4.44
✓			0.7	5.43 ± 4.33
	✓		**0.6**	**4.10 ± 3.67**
		✓	0.7	3.12 ± 3.60
✓		✓	0.7	3.70 ± 2.00
	✓	✓	**0.2**	**2.29 ± 2.20**

Table 8
Performance comparisons (ACER (%)) of the proposed cross-stage fusion method against other cross-layer feature fusion on OULU-NPU and our proposed NIR face anti-spoofing dataset.

Fusion methods	OULU-NPU				Near-Infrared		Flops (G)	Params (M)
	P1	P2	P3	P4	P1	P2		
MTAN (Liu, Johns, & Davison, 2019)	0.8	1.3	1.2 ± 2.0	2.5 ± 1.3	0.68	3.69	**1.1**	3.1
FPM (Vandenhende et al., 2020)	1.7	1.8	1.8 ± 3.6	3.9 ± 2.2	0.49	3.32	1.7	4.2
CLAM (Chang et al., 2020)	1.1	1.7	1.8 ± 2.5	2.5 ± 2.4	0.58	**2.02**	1.2	**2.8**
Ours	**0.2**	**1.1**	**0.9 ± 1.1**	**2.3 ± 2.2**	**0.49**	2.77	1.3	3.1

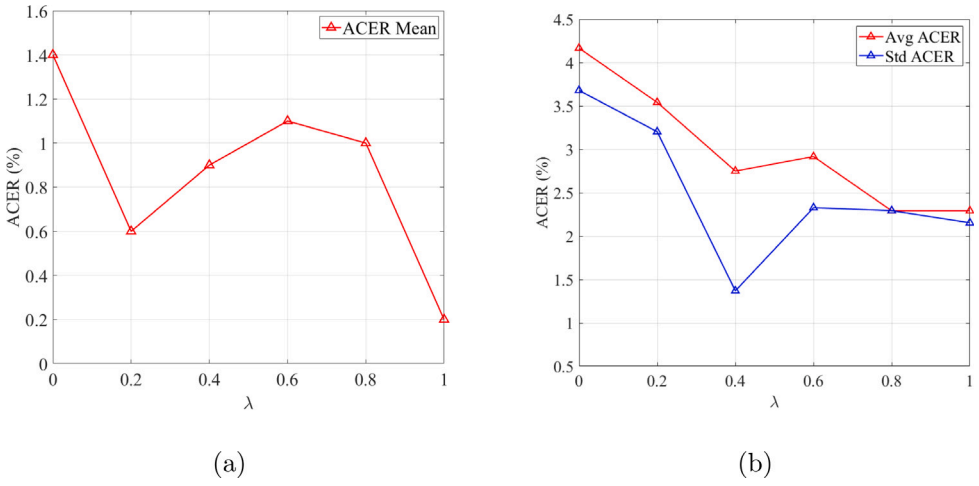


Fig. 6. ACER (%) with different λ.

Table 9
Ablation experiments on the number of cross-attention heads. We report ACER(%) in Protocol 1 and Protocol 4 of OULU-NPU.

	Head number		
	1	2	4
Protocol 1	0.9	0.4	0.2
Protocol 4	2.5 ± 1.3	2.3 ± 2.7	2.3 ± 2.0

Table 10
Impact of different stages feature combinations for pixel-wise material supervisions.

Stage combination	2	3	4	2@4	2@3	2@3@4
Protocol 1	3.9	1.6	1.9	1.7	0.2	0.7
Protocol 4	6.5 ± 3.5	3.5 ± 2.8	4.58 ± 2.19	6.6 ± 4.2	2.3 ± 2.2	4.8 ± 2.7

in Fig. 6, the model achieves the best performance in both Protocol 1 and Protocol 4 of the OULU-NPU dataset when $\lambda = 1$.

Effectiveness of Stage selection for fusion. We study the influence of feature fusion strategy on performance. In this experiment, we compare 6 different stages fusion combinations and report the ACER in Table 10. We denote "@" as an operation that concatenates the features of multiple stages. Additionally, introducing material classification supervision in combination 2@3 achieves the best performance on both protocol 1 and protocol 4 of the OULU-NPU dataset, which indicates that introducing material supervision in the medium and low features will achieve the best performance improvement.

Effectiveness of auxiliary supervision. To further study the effectiveness of pixel-wise material auxiliary classification supervision, we compare the performance of binary classification supervision, pixel-wise binary classification supervision, and pixel-wise material classification supervision. As shown in Table 11, the proposed pixel-wise material classification supervision achieves the best performance on

Table 11
Impact of different auxiliary supervisions. We report the ACER(%) in protocol 1 and protocol 4 of OULU-NPU.

Supervision	Binary classification	Binary pixel-wise classification	Pixel-wise material classification
Protocol 1	1.6	1.6	0.2
Protocol 4	3.8 ± 2.8	2.9 ± 1.7	2.3 ± 2.2

both protocol 1 and protocol 4, proving the effectiveness of introducing auxiliary material classification.

6. Conclusion

In this paper, we have investigated how to extract both local details and global semantics for effective anti-spoofing. Specifically, we propose a lightweight CNN-Transformer-based face anti-spoofing framework with a cross-stage multi-head attention module based on the multi-head attention mechanism. Moreover, to improve discrimination of local features for subtle differences, we design a pixel-wise attack material classification auxiliary supervision task for local feature learning. Furthermore, to overcome the limited data volume, lack of environmental variety, and facial identities in existing Near-Infrared face anti-spoofing datasets, we collect a large-scale NIR dataset containing 380k images from 1031 identities will be released in the community later. We made several findings in this study: (1) From the ablation studies on the pixel-wise material classification task, we find that applying attack material supervision at intermediate stages has the best effect on face anti-spoofing task. (2) Ablation experiments involving multi-head attention fusion indicate that a higher number of heads consistently improves performance. (3) The ablation study in different fusion blocks indicates that merely incorporating multi-stage features fusion mechanisms might have negative impacts on face anti-spoofing. Moreover, our method weak in processing samples with significant domain shifts. In the future, we focus on overcoming the challenges in face anti-spoofing under unseen domain and more challenge attacks (e.g. Adversarial Attack (Li, Zhang, Cao, & Tan, 2023) , Face Synthesis (Zixiong, Qi, Libo, Yifan, Naizhou, Mingkui, & Qi, 2024)).

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

Acknowledgments

This work was partially supported by National Natural Science Foundation of China (NSFC) 62072190, Program for Guangdong Introducing Innovative and Entrepreneurial Teams 2017ZT07X183, National Natural Science Foundation of China under Grant 62306061, Guangdong Provincial Regional Joint Foundation, China under Grant 2023A1515140037.

References

Agarwal, A., Yadav, D., Kohli, N., Singh, R., Vatsa, M., & Noore, A. (2017). Face presentation attack with latex masks in multispectral videos. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops* (pp. 81–89).

Almeida, W. R., Andaló, F. A., Padilha, R., Bertocco, G., Dias, W., Torres, R. d. S., et al. (2020). Detecting face presentation attacks in mobile devices with a patch-based CNN and a sensor-aware loss function. *PLoS One*, *15*(9), Article e0238058.

Bekhouche, S. E., Kajo, I., Ruichek, Y., & Dornaika, F. (2022). Spatiotemporal CNN with pyramid bottleneck blocks: Application to eye blinking detection. *Neural Networks*, *152*, 150–159.

Bhattacharjee, S., Mohammadi, A., & Marcel, S. (2018). Spoofing deep face recognition with custom silicone masks. In *2018 IEEE 9th international conference on biometrics theory, applications and systems* (pp. 1–7). IEEE.

Boulkenafet, Z., Komulainen, J., & Hadid, A. (2016). Face anti-spoofing using speeded-up robust features and fisher vector encoding. *IEEE Signal Processing Letters*, *24*(2), 141–145.

Boulkenafet, Z., Komulainen, J., Li, L., Feng, X., & Hadid, A. (2017). OULU-NPU: A mobile face presentation attack database with real-world variations. In *2017 12th IEEE international conference on automatic face & gesture recognition (FG 2017)* (pp. 612–618). IEEE.

Cai, R., Cui, Y., Li, Z., Yu, Z., Li, H., Hu, Y., et al. (2023). Rehearsal-free domain continual face anti-spoofing: Generalize more and forget less. arXiv preprint arXiv:2303.09914.

Cai, R., Li, H., Wang, S., Chen, C., & Kot, A. C. (2020). DRL-FAS: A novel framework based on deep reinforcement learning for face anti-spoofing. *IEEE Transactions on Information Forensics and Security*, *16*, 937–951.

Chang, C.-Y., Chang, S.-E., Hsiao, P.-Y., & Fu, L.-C. (2020). Epsnet: efficient panoptic segmentation network with cross-layer attention fusion. In *Proceedings of the Asian conference on computer vision*.

Chen, L., You, Z., Zhang, N., Xi, J., & Le, X. (2022). Utrad: Anomaly detection and localization with u-transformer. *Neural Networks*, *147*, 53–62.

Child, R., Gray, S., Radford, A., & Sutskever, I. (2019). Generating long sequences with sparse transformers. arXiv preprint arXiv:1904.10509.

Chingovska, I., Anjos, A., & Marcel, S. (2012). On the effectiveness of local binary patterns in face anti-spoofing. In *2012 BIOSIG-proceedings of the international conference of biometrics special interest group* (pp. 1–7). IEEE.

Chingovska, I., Erdogmus, N., Anjos, A., & Marcel, S. (2016). Face recognition systems under spoofing attacks. *Face Recognition Across the Imaging Spectrum*, 165–194.

Choudhury, T., Clarkson, B., Jebara, T., & Pentland, A. (1999). Multimodal person recognition using unconstrained audio and video. In *Proceedings, international conference on audio-and video-based person authentication* (pp. 176–181).

Chu, X., Tian, Z., Zhang, B., Wang, X., Wei, X., Xia, H., et al. (2021). Conditional positional encodings for vision transformers. arXiv preprint arXiv:2102.10882.

Costa-Pazo, A., Bhattacharjee, S., Vazquez-Fernandez, E., & Marcel, S. (2016). The replay-mobile face presentation-attack database. In *2016 international conference of the biometrics special interest group* (pp. 1–7). IEEE.

Dai, Z., Cai, B., Lin, Y., & Chen, J. (2021). Up-detr: Unsupervised pre-training for object detection with transformers. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 1601–1610).

Dai, G., Zhang, Y., Wang, Q., Du, Q., Yu, Z., Liu, Z., et al. (2023). Disentangling writer and character styles for handwriting generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 5977–5986).

Deininger, L., Stimpel, B., Yuce, A., Abbasi-Sureshjani, S., Schönenberger, S., Ocampo, P., et al. (2022). A comparative study between vision transformers and CNNs in digital pathology. arXiv preprint arXiv:2206.00389.

Deng, J., Guo, J., Xue, N., & Zafeiriou, S. (2019). Arcface: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 4690–4699).

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., et al. (2020). An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929.

Du, Z., Li, J., Zuo, L., Zhu, L., & Lu, K. (2022). Energy-based domain generalization for face anti-spoofing. In *Proceedings of the 30th ACM international conference on multimedia* (pp. 1749–1757).

(2003). Face recognition: A literature survey. *ACM Computing Surveys (CSUR)*, *35*(4), 399–458.

Feng, Y., Wu, F., Shao, X., Wang, Y., & Zhou, X. (2018). Joint 3d face reconstruction and dense alignment with position map regression network. In *Proceedings of the European conference on computer vision (ECCV)* (pp. 534–551).

George, A., Mostaani, Z., Geissenbuhler, D., Nikisins, O., Anjos, A., & Marcel, S. (2019). Biometric face presentation attack detection with multi-channel convolutional neural network. *IEEE Transactions on Information Forensics and Security*, *15*, 42–55.

Gulati, A., Qin, J., Chiu, C.-C., Parmar, N., Zhang, Y., Yu, J., et al. (2020). Conformer: Convolution-augmented transformer for speech recognition. arXiv preprint arXiv:2005.08100.

Guo, Y., Zhang, L., Hu, Y., He, X., & Gao, J. (2016). Ms-celeb-1m: A dataset and benchmark for large-scale face recognition. In *Computer vision–ECCV 2016: 14th European conference, amsterdam, the netherlands, October 11-14, 2016, proceedings, part III 14* (pp. 87–102). Springer.

Han, K., Xiao, A., Wu, E., Guo, J., Xu, C., & Wang, Y. (2021). Transformer in transformer. *Advances in Neural Information Processing Systems*, *34*, 15908–15919.

He, Y., Peng, F., & Long, M. (2023). Dynamic residual distillation network for face anti-spoofing with feature attention learning. *IEEE Transactions on Biometrics, Behavior, and Identity Science*.

Heusch, G., George, A., Geissbühler, D., Mostaani, Z., & Marcel, S. (2020). Deep models and shortwave infrared information to detect face presentation attacks. *IEEE Transactions on Biometrics, Behavior, and Identity Science*, *2*(4), 399–409.

Huang, R., & Wang, X. (2023). Face anti-spoofing using feature distilling and global attention learning. *Pattern Recognition*, *135*, Article 109147.

Jia, Y., Zhang, J., Shan, S., & Chen, X. (2020). Single-side domain generalization for face anti-spoofing. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 8484–8493).

Komulainen, J., Hadid, A., & Pietikäinen, M. (2013). Context based face anti-spoofing. In *2013 IEEE sixth international conference on biometrics: theory, applications and systems* (pp. 1–8). IEEE.

Lee, Y., Kwak, Y., & Shin, J. (2023). Robust face anti-spoofing framework with convolutional vision transformer. In *2023 IEEE international conference on image processing* (pp. 1015–1019). IEEE.

Li, J.-W. (2008). Eye blink detection based on multiple gabor response waves. *vol. 5*, In *2008 international conference on machine learning and cybernetics* (pp. 2852–2856). IEEE.

Li, L., Feng, X., Boulkenafet, Z., Xia, Z., Li, M., & Hadid, A. (2016). An original face anti-spoofing approach using partial convolutional neural network. In *2016 sixth international conference on image processing theory, tools and applications* (pp. 1–6). IEEE.

Li, J., Zhang, S., Cao, J., & Tan, M. (2023). Learning defense transformations for counterattacking adversarial examples. *Neural Networks*, *164*, 177–185.

Liao, C.-H., Chen, W.-C., Liu, H.-T., Yeh, Y.-R., Hu, M.-C., & Chen, C.-S. (2023). Domain invariant vision transformer learning for face anti-spoofing. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision* (pp. 6098–6107).

Lin, J.-D., Han, Y.-H., Huang, P.-H., Tan, J., Chen, J.-C., Tanveer, M., et al. (2023). DEFAEK: Domain effective fast adaptive network for face anti-spoofing. *Neural Networks*, *161*, 83–92.

Lin, J.-D., Lin, H.-H., Dy, J., Chen, J.-C., Tanveer, M., Razzak, I., et al. (2021). Lightweight face anti-spoofing network for telehealth applications. *IEEE Journal of Biomedical and Health Informatics*, *26*(5), 1987–1996.

Liu, S., Johns, E., & Davison, A. J. (2019). End-to-end multi-task learning with attention. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 1871–1880).

Liu, Y., Jourabloo, A., & Liu, X. (2018). Learning deep models for face anti-spoofing: Binary or auxiliary supervision. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 389–398).

Liu, A., & Liang, Y. (2022). MA-ViT: Modality-agnostic vision transformers for face anti-spoofing. In L. D. Raedt (Ed.), *Proceedings of the thirty-first international joint conference on artificial intelligence, IJCAI-22* (pp. 1180–1186). International Joint Conferences on Artificial Intelligence Organization, http://dx.doi.org/10.24963/ijcai.2022/165, Main Track.

Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., et al. (2021). Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 10012–10022).

Liu, Y., Stehouwer, J., & Liu, X. (2020). On disentangling spoof trace for generic face anti-spoofing. In *European conference on computer vision* (pp. 406–422). Springer.

Liu, A., Tan, Z., Yu, Z., Zhao, C., Wan, J., Lei, Y. L. Z., et al. (2023). Fm-vit: Flexible modal vision transformers for face anti-spoofing. *IEEE Transactions on Information Forensics and Security*.

Liu, Y., Wu, L., Li, Z., & Wang, Z. (2023). Dual-stream correlation exploration for face anti-spoofing. *Pattern Recognition Letters*, *170*, 17–23.

Liu, S., Zhang, K.-Y., Yao, T., Bi, M., Ding, S., Li, J., et al. (2021). Adaptive normalized representation learning for generalizable face anti-spoofing. In *Proceedings of the 29th ACM international conference on multimedia* (pp. 1469–1477).

Liu, S., Zhang, K.-Y., Yao, T., Sheng, K., Ding, S., Tai, Y., et al. (2021). Dual reweighting domain generalization for face presentation attack detection. arXiv preprint arXiv:2106.16128.

Maaz, M., Shaker, A., Cholakkal, H., Khan, S., Zamir, S. W., Anwer, R. M., et al. (2022). Edgenext: efficiently amalgamated cnn-transformer architecture for mobile vision applications. arXiv preprint arXiv:2206.10589.

Ming, Z., Yu, Z., Al-Ghadi, M., Visani, M., Luqman, M. M., & Burie, J.-C. (2022). Vitranspad: video transformer using convolution and self-attention for face presentation attack detection. In *2022 IEEE international conference on image processing* (pp. 4248–4252). IEEE.

Peixoto, B., Michelassi, C., & Rocha, A. (2011). Face liveness detection under bad illumination conditions. In *2011 18th IEEE international conference on image processing* (pp. 3557–3560). IEEE.

Ronneberger, O., Fischer, P., & Brox, T. (2015). U-net: Convolutional networks for biomedical image segmentation. In *Medical image computing and computer-assisted intervention–mICCAI 2015: 18th international conference, munich, Germany, October 5-9, 2015, proceedings, part III 18* (pp. 234–241). Springer.

Shao, R., Lan, X., Li, J., & Yuen, P. C. (2019). Multi-adversarial discriminative deep domain generalization for face presentation attack detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 10023–10031).

Shao, R., Lan, X., & Yuen, P. C. (2020). Regularized fine-grained meta face anti-spoofing. *vol. 34*, In *Proceedings of the AAAI conference on artificial intelligence* (07), (pp. 11974–11981).

Singh, A. K., Joshi, P., & Nandi, G. C. (2014). Face recognition with liveness detection using eye and mouth movement. In *2014 international conference on signal propagation and computer technology (ICSPCT 2014)* (pp. 592–597). IEEE.

Sun, W., Song, Y., Chen, C., Huang, J., & Kot, A. C. (2020). Face spoofing detection based on local ternary label supervision in fully convolutional networks. *IEEE Transactions on Information Forensics and Security*, *15*, 3181–3196.

Touvron, H., Cord, M., Douze, M., Massa, F., Sablayrolles, A., & Jégou, H. (2021). Training data-efficient image transformers distillation through attention. In *International conference on machine learning* (pp. 10347–10357). PMLR.

Vandenhende, S., Georgoulis, S., & Van Gool, L. (2020). Mti-net: Multi-scale task interaction networks for multi-task learning. In *Computer vision–ECCV 2020: 16th European conference, glasgow, UK, August 23–28, 2020, proceedings, part IV 16* (pp. 527–543). Springer.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., et al. (2017). Attention is all you need. *Advances in Neural Information Processing Systems*, *30*.

Wang, L., Ding, X., & Fang, C. (2009). Face live detection method based on physiological motion analysis. *Tsinghua Science & Technology*, *14*(6), 685–690.

Wang, W., Liu, P., Zheng, H., Ying, R., & Wen, F. (2023). Domain generalization for face anti-spoofing via negative data augmentation. *IEEE Transactions on Information Forensics and Security*.

Wang, C.-Y., Lu, Y.-D., Yang, S.-T., & Lai, S.-H. (2022). PatchNet: A simple face anti-spoofing framework via fine-grained patch recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 20281–20290).

Wang, Z., Wang, Q., Deng, W., & Guo, G. (2022). Face anti-spoofing using transformers with relation-aware mechanism. *IEEE Transactions on Biometrics, Behavior, and Identity Science*, *4*(3), 439–450.

Wang, Z., Wang, Q., Deng, W., & Guo, G. (2022). Learning multi-granularity temporal characteristics for face anti-spoofing. *IEEE Transactions on Information Forensics and Security*, *17*, 1254–1269.

Wang, Z., Wang, Z., Yu, Z., Deng, W., Li, J., Gao, T., et al. (2022). Domain generalization via shuffled style assembly for face anti-spoofing. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 4123–4133).

Wang, W., Wen, F., Zheng, H., Ying, R., & Liu, P. (2022). Conv-mlp: A convolution and mlp mixed model for multimodal face anti-spoofing. *IEEE Transactions on Information Forensics and Security*, *17*, 2284–2297.

Wang, W., Xie, E., Li, X., Fan, D.-P., Song, K., Liang, D., et al. (2021). Pyramid vision transformer: A versatile backbone for dense prediction without convolutions. In *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 568–578).

Wang, C., Yu, B., & Zhou, J. (2023). A learnable gradient operator for face presentation attack detection. *Pattern Recognition*, *135*, Article 109146.

Wang, Z., Zhao, C., Qin, Y., Zhou, Q., Qi, G., Wan, J., et al. (2018). Exploiting temporal and depth information for multi-frame face anti-spoofing. arXiv preprint arXiv:1811.05118.

Wu, Z., Liu, Z., Lin, J., Lin, Y., & Han, S. (2020). Lite transformer with long-short range attention. arXiv preprint arXiv:2004.11886.

Wu, H., Xiao, B., Codella, N., Liu, M., Dai, X., Yuan, L., et al. (2021). Cvt: Introducing convolutions to vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 22–31).

Wu, H., Zeng, D., Hu, Y., Shi, H., & Mei, T. (2021). Dual spoof disentanglement generation for face anti-spoofing with depth uncertainty learning. *IEEE Transactions on Circuits and Systems for Video Technology*, *32*(7), 4626–4638.

Xiao, J., Tang, Y., Guo, J., Yang, Y., Zhu, X., Lei, Z., et al. (2019). 3DMA: A multi-modality 3D mask face anti-spoofing database. In *2019 16th IEEE international conference on advanced video and signal based surveillance* (pp. 1–8). IEEE.

Yang, J., Lei, Z., & Li, S. Z. (2014). Learn convolutional neural network for face anti-spoofing. arXiv preprint arXiv:1408.5601.

Yu, Z., Cai, R., Cui, Y., Liu, X., Hu, Y., & Kot, A. (2023). Rethinking vision transformer and masked autoencoder in multimodal face anti-spoofing. arXiv preprint arXiv:2302.05744.

Yu, Z., Cai, R., Li, Z., Yang, W., Shi, J., & Kot, A. C. (2022). Benchmarking joint face spoofing and forgery detection with visual and physiological cues. arXiv preprint arXiv:2208.05401.

Yu, Z., Li, X., Niu, X., Shi, J., & Zhao, G. (2020). Face anti-spoofing with human material perception. In *European conference on computer vision* (pp. 557–575). Springer.

Yu, Z., Li, X., Wang, P., & Zhao, G. (2021). Transrppg: Remote photoplethysmography transformer for 3d mask face presentation attack detection. *IEEE Signal Processing Letters*, *28*, 1290–1294.

Yu, Z., Qin, Y., Zhao, H., Li, X., & Zhao, G. (2021). Dual-cross central difference network for face anti-spoofing. arXiv preprint arXiv:2105.01290.

Yu, Z., Wan, J., Qin, Y., Li, X., Li, S. Z., & Zhao, G. (2020). NAS-FAS: Static-dynamic central difference network search for face anti-spoofing. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *43*(9), 3005–3023.

Yu, Z., Zhao, C., Wang, Z., Qin, Y., Su, Z., Li, X., et al. (2020). Searching central difference convolutional networks for face anti-spoofing. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 5295–5305).

Yue, H., Wang, K., Zhang, G., Feng, H., Han, J., Ding, E., et al. (2023). Cyclically disentangled feature translation for face anti-spoofing. *vol. 37*, In *Proceedings of the AAAI conference on artificial intelligence* (3), (pp. 3358–3366).

Zhang, S., Liu, A., Wan, J., Liang, Y., Guo, G., Escalera, S., et al. (2020). Casia-surf: A large-scale multi-modal benchmark for face anti-spoofing. *IEEE Transactions on Biometrics, Behavior, and Identity Science*, *2*(2), 182–193.

Zhang, K.-Y., Yao, T., Zhang, J., Tai, Y., Ding, S., Li, J., et al. (2020). Face anti-spoofing via disentangled representation learning. In *European conference on computer vision* (pp. 641–657). Springer.

Zhang, K., Zhang, Z., Li, Z., & Qiao, Y. (2016). Joint face detection and alignment using multitask cascaded convolutional networks. *IEEE Signal Processing Letters*, *23*(10), 1499–1503.

Zhou, Q., Zhang, K.-Y., Yao, T., Yi, R., Ding, S., & Ma, L. (2022). Adaptive mixture of experts learning for generalizable face anti-spoofing. In *Proceedings of the 30th ACM international conference on multimedia* (pp. 6009–6018).

Zixiong, H., Qi, C., Libo, S., Yifan, Y., Naizhou, W., Mingkui, T., et al. (2024). G-nerf: geometry-enhanced novel view synthesis from single-view images. *arXiv preprint arXiv:2404.07474*, arXiv:2404.07474.