mmFAS: Multimodal Face Anti-Spoofing Using Multi-Level Alignment and Switch-Attention Fusion

Geng Chen[1,2], Wuyuan Xie[1], Di Lin[3], Ye Liu[4], Miaohui Wang[1*]

[1]College of CSSE, Guangdong Key Laboratory of Intelligent Information Processing, Shenzhen University
[2]State Key Laboratory of Networking and Switching Technology, Beijing University of Posts and Telecommunications
[3]College of Intelligence and Computing, Tianjin University
[4]School of Automation, Nanjing University of Posts and Telecommunications
chengeng@bupt.edu.cn, wuyuan.xie@gmail.com, wang.miaohui@gmail.com

Abstract

The increasing number of presentation attacks on reliable face matching has raised concerns and garnered attention towards face anti-spoofing (FAS). However, existing methods for FAS modeling commonly fuse multiple visual modalities (*e.g.*, RGB, Depth, and Infrared) in a straightforward manner, disregarding latent feature gaps that can hinder representation learning. To address this challenge, we propose a novel multimodal FAS framework (*mmFAS*) that focuses on explicit alignment and fusion of latent features across different modalities. Specifically, we develop a multimodal alignment module to alleviate the latent feature gap by using instance-level contrastive learning and class-level matching simultaneously. Further, we explore a new switch-attention based fusion module to automatically aggregate complementary information and control model complexity. To evaluate the anti-spoofing performance more effectively, we adopt a challenging yet meaningful cross-database protocol involving four benchmark multimodal FAS datasets to simulate real-world scenarios. Extensive experimental results demonstrate the effectiveness of *mmFAS* in improving the accuracy of FAS systems, outperforming 10 representative methods.

1 Introduction

Face recognition (Xiao et al. 2022; Gao et al. 2022) has become a popular means of providing certification services in many multimedia applications. However, the increasing prevalence of malicious face *presentation attack* (Zhang et al. 2020b; George et al. 2019; Heusch et al. 2020), such as *print attack*, *replay attack* and *3D mask attack*, has made this technology vulnerable and unreliable, resulting in significant threats to personal, financial, and even national security. To tackle this problem, many face presentation attack detection methods (Zhou et al. 2023; Sun et al. 2023; Srivatsan, Naseer, and Nandakumar 2023) have been put forward aiming to better distinguish between real and fake face images. Existing face anti-spoofing (FAS) methods (Yu et al. 2022a) are mainly based on feature descriptors. These methods have gained considerable attention but their performance is still unsatisfactory when facing diverse and complex presentation attacks.

Recently, deep learning-based FAS methods (Guo et al. 2022; Zhou et al. 2022a; Huang et al. 2022; Zhou et al. 2022b) have been developed and proven to be powerful. These methods have achieved state-of-the-art performance on several benchmark datasets. However, poor generalization limits their applications in many real-world scenarios while new types of face attacks are emerging all the time (Dong et al. 2021). One of the main reasons for the cross-dataset performance is that they might easily overfit the training data and lack the ability to learn intrinsic representative features. On the other hand, the emergence of multimodal data provides a new perspective for solving the generalization ability (Kuang et al. 2019; Zhang et al. 2020a; Liu et al. 2021; Heusch et al. 2020). Therefore, how to use multimodal learning to solve this challenging problem is the fundamental motivation of this paper.

As the RGB modality contains redundant and forensically irrelevant information (Yu et al. 2022b) (*e.g.*, clothing, accessories, and skin color), other easily acquired modality data (*i.e.*, depth and infrared) have been explored in various face-based security tasks (Liu et al. 2021; Heusch et al. 2020; Zhang et al. 2020a). In addition, there is a growing number of multimodality-based FAS datasets (Agarwal et al. 2017; Zhang et al. 2020a; Liu et al. 2021; George et al. 2019). These multimodal datasets highlight the anisotropy between different data sources. However, existing multimodal FAS schemes (Kuang et al. 2019; Yang et al. 2020; Wang et al. 2022) either implicitly align or directly fuse these different modalities, which can negatively impact both training efficiency and generalization ability. Therefore, we investigate an explicit alignment mechanism for the FAS task, exploring the performance effect of multimodal feature alignment on face presentation attacks for the first time.

In this paper, we propose a novel and effective multimodal-based FAS framework (called *mmFAS*), which consists of multiple-level alignment and switch-attention fusion. Specifically, we utilize three feature extractors to obtain shallow features from three modalities (*e.g.*, RGB, depth, and infrared). To alleviate the latent feature gap, we jointly adopt instance contrastive learning and spoof class matching, which explicitly align the three different modalities while enlarging their inter-instance and inter-class discrepancies. In addition, we propose the use of switch-

*Corresponding author: Miaohui Wang

attention to effectively fuse the aligned features, facilitating the aggregation of complementary information. Finally, the fused features are fed to a fully-connected layer for face anti-spoofing classification, where the binary cross-entropy loss is optimized.

In summary, our main contributions are threefold:

- We propose a novel multimodal-driven FAS framework (*i.e.*, *mmFAS*), which incorporates multiple-level alignment and switch-attention fusion, delivering better representation learning and generalization ability. To the best of our survey, *mmFAS* is one of the earliest studies to explicitly consider latent feature gap and perform multimodal alignment for FAS.
- To better understand the relationships between different modalities, we propose a new multiple-level feature alignment approach that combines instance contrastive learning and spoof class matching. Further, we also employ hard negative and hard positive mining to accelerate the learning convergence.
- To improve performance and remove redundant features, we develop a plug-and-play, parameter-free, and computation-free switch attention module to facilitate the aggregation of complementary information from three modalities. Experimental results validate the effectiveness of *mmFAS* in improving the accuracy of FAS systems and outperforming 10 representative methods.

2 Related Work

In this section, we provide a brief overview of representative algorithms that cover unimodal and multimodal FASs.

Unimodal Face Anti-spoofing

Binary Supervision Methods. Inspired by the success of convolutional neural network (CNN), various network architectures (Yang, Lei, and Li 2014; Kong et al. 2023; Yang et al. 2019) have been taken into account for better feature learning. On the other hand, multiple loss functions or novel network components (Cai, Rizhao and Cui, Yawen and Li, Zhi and Yu, Zitong and Li, Haoliang and Hu, Yongjian and Kot, Alex 2023; Huang et al. 2022; Jia et al. 2021; Xu, Xiong, and Xia 2021) have been designed to provide discriminative supervision signals to enforce a model to learn intrinsic spoofing clues against various attack types.

Multimodal Face Anti-spoofing

Recently, multiple modalities have also been studied in FAS, including modality choice and fusion strategy.

Modality Choice. Initially, FAS methods relied solely on image modality (Kose and Dugelay 2013; Patel, Han, and Jain 2016) or video modality (Wen, Han, and Jain 2015; Pinto et al. 2015). However, the development of different camera sensors has made it possible to acquire multimodal data, leading to increased interest in multimodal FAS methods due to their effectiveness. The most widely-used modalities in FAS were depth and infrared (George et al. 2019; Heusch et al. 2020; Liu et al. 2021; Zhang et al. 2020a). Existing schemes primarily explored the interaction among the RGB, depth, and infrared modalities. However, more novel modalities (Heusch et al. 2020; Li et al. 2020; Kong et al. 2022; Srivatsan, Naseer, and Nandakumar 2023)have also been introduced to explore the potential for better performance. Those modalities can be classified as whether having strong correlations with attributes of a living human face. In recent years, flexible-modal (Liu et al. 2024; Liu and Liang 2023; Liu et al. 2023; Yu et al. 2023) approaches have gained traction as all modalities may not be available during the testing phase.

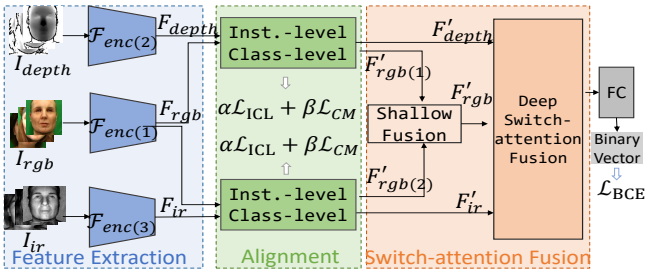


Figure 1: Overview of the proposed *mmFAS* network. It mainly consists of three feature extraction modules, two modality alignment modules, and two fusion modules. The final binary vectors are used for forgery classification. '$\mathcal{L}_{ICL}$', '$\mathcal{L}_{CM}$', 'FC', and '$\mathcal{L}_{BCE}$' represent instance contrastive learning loss, class matching loss, fully-connected layer, and binary cross entropy loss, respectively.

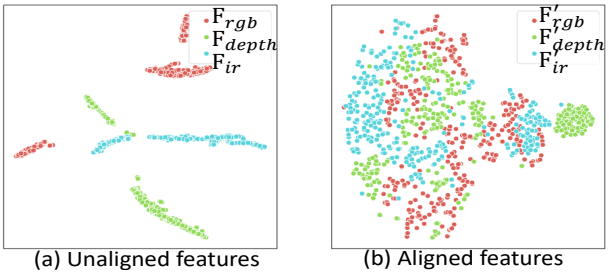
Fusion Strategy. Existing multimodal FAS methods can be categorized into 1) input-level fusion, 2) decision-level fusion, and 3) feature-level fusion.

Input-level fusion (Jiang et al. 2020; Wang et al. 2022) usually involves lighter network architectures and lower computations compared to the feature-level and decision-level fusions. These methods primarily explored the relevancy of low-level features among different modalities, helping mitigate the effects of missing or noisy data. Decision-level fusion (Zhang et al. 2019) typically involves ensembling multiple diverse modalities. The cross-modality interaction is shallow and even absent, and each modality requires its own classifier. Feature-level fusion (Yang et al. 2020; George and Marcel 2021; Shen, Huang, and Tong 2019) involves a trade-off between shallow modality interaction, high computation cost, and weak classification ability.

Therefore, our *mmFAS* utilizes a feature-level fusion approach, employing a novel strategy that retains the individual modality information while integrating other modalities as context in the form of queries, keys, or values within the attention mechanism, thereby enhancing the current modality.

3 Proposed *mmFAS* Method

As illustrated in Figure 1, our *mmFAS* method consists of three feature extraction modules, two modality alignment modules, two fusion modules, and a final classification head. Firstly, we train three independent feature extractors $\mathcal{F}_{enc(1,2,3)}$ for the RGB modality, depth modality, and infrared modality. The extracted features are denoted



(a) Unaligned features (b) Aligned features

Figure 2: Illustration of multimodal feature with and without alignment: (a) unaligned features from three modalities, and (b) aligned features from our multiple-level alignment method.

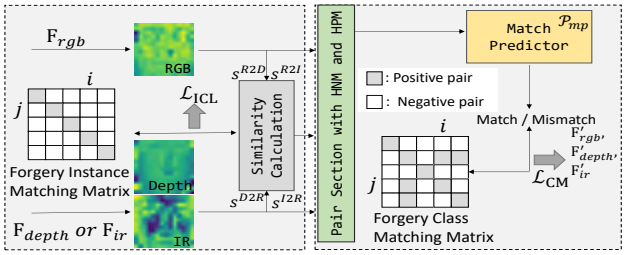


Figure 3: Overview of the proposed instance-level and class-level alignment modules from two modalities. 'HNM' and 'HPM' are short for hard negative mining and hard positive mining, respectively. i and j indicate the i-th and j-th instances in one batch.

as $\boldsymbol{F}_{rgb} = \mathcal{F}_{enc(1)}(\boldsymbol{I}_{rgb})$, $\boldsymbol{F}_{depth} = \mathcal{F}_{enc(2)}(\boldsymbol{I}_{depth})$, and $\boldsymbol{F}_{ir} = \mathcal{F}_{enc(3)}(\boldsymbol{I}_{ir})$, respectively. Subsequently, we employ two multi-level alignment modules, which includes the instance contrastive learning and class matching tasks, to align $\boldsymbol{F}_{rgb}$ with $\boldsymbol{F}_{depth}$ and $\boldsymbol{F}_{rgb}$ with $\boldsymbol{F}_{ir}$, respectively. Later, latent features from three modalities are fed into our proposed switch-attention fusion module to aggregate complementary information across different modalities. Finally, the fused feature $\boldsymbol{F}_{fuse}$ is passed to a fully-connected (FC) module.

Multiple-level Alignment

Previous FAS methods have often overlooked the importance of aligning multimodal features. Instead, they have primarily focused on how to more efficiently fuse these features. However, it is crucial to note that unaligned multimodal features exhibit significant distribution differences, as shown in Figure 2. These differences hinder the efficient fusion of subsequent modalities (Wang et al. 2024).

To illustrate this point, let us consider the feature differences between the $\boldsymbol{I}_{rgb}$ and $\boldsymbol{I}_{depth}$ modalities. $\boldsymbol{I}_{rgb}$ typically contains semantic information such as color and texture, while $\boldsymbol{I}_{depth}$ represents object distance information. An RGB face photograph usually exhibits clear contours and vibrant colors, while the resolution of a depth photograph is relatively low, resulting in a lack of detailed information in the face region. Therefore, it is essential to emphasize the necessity of cross-modality alignment before fusing the features.

To address this issue, we propose a multiple-level alignment module as shown in Figure 3. Given that $\boldsymbol{I}_{rgb}$ contains the most information compared to $\boldsymbol{I}_{depth}$ and $\boldsymbol{I}_{ir}$, we leverage the RGB modality as a bridge and incorporate two alignment modules to align the depth and infrared modalities. In *mmFAS*, the alignment module takes inputs from three unimodal feature extractors and produces aligned features, namely $\boldsymbol{F}'_{rgb}$, $\boldsymbol{F}'_{depth}$, and $\boldsymbol{F}'_{ir}$. In *mmFAS*, we employ instance-level contrastive learning (ICL) and class-level matching (CM) simultaneously in the multiple-level alignment module to increase the inter-instance and inter-class discrepancy. Further, we utilize online hard negative-positive mining to improve the CM model. Since the alignment process between RGB-depth and RGB-infrared is nearly identical, we only describe the specific alignment details of RGB-depth for the sake of conciseness.

Instance-level Contrastive Learning (ICL). We consider different modality representations of a face photograph as the same sample instance, but photographs of the same person captured at different time are not considered as the same instance. Therefore, there are only three different modalities belonging to the same instance. To align a paired $\boldsymbol{F}_{rgb}$ and $\boldsymbol{F}_{depth}$, we first need to measure the similarity between them using the following method:

$$s^{R2D}_{i,j} = \boldsymbol{F}^T_{rgb(i)}\boldsymbol{F}_{depth(j)}, s^{D2R}_{i,j} = \boldsymbol{F}^T_{depth(i)}\boldsymbol{F}_{rgb(j)}, \quad (1)$$

where $s^{R2D}_{i,j}$ represents the similarity score between the i-th sample of $\boldsymbol{F}_{rgb}$ and the j-th sample of $\boldsymbol{F}_{depth}$. In contrast, $s^{D2R}_{i,j}$ represents the similarity score between the i-th sample of $\boldsymbol{F}_{depth}$ and the j-th sample of $\boldsymbol{F}_{rgb}$. Then, the softmax-normalized similarity scores between s^{R2D} and s^{D2R} are defined as:

$$p^{R2D}_{i,j} = \frac{\exp\left(s^{R2D}_{i,j}/\tau\right)}{\sum_{k=1}^{N}\exp\left(s^{R2D}_{i,k}/\tau\right)}, p^{D2R}_{i,j} = \frac{\exp\left(s^{D2R}_{i,j})/\tau\right)}{\sum_{k=1}^{N}\exp\left(s^{D2R}_{i,k}/\tau\right)}, \quad (2)$$

where τ denotes a learnable temperature parameter and N denotes the batch size.

Similarly, we can obtain the similarity scores between $\boldsymbol{F}_{rgb}$ and $\boldsymbol{F}_{ir}$ as in Eq. (1), *i.e.*, $s^{R2I}_{i,j}$ and $s^{I2R}_{i,j}$. Then, the softmax-normalized similarity scores can also be calculated as in Eq. (2), *i.e.*, $p^{R2I}_{i,j}$ and $p^{I2R}_{i,j}$.

Finally, we use $y^{R2D(\boldsymbol{F}_{rgb(i)},\boldsymbol{F}_{depth(j)})}$ to represent whether the i-th RGB feature and the j-th depth feature match. Meanwhile, we use $y^{R2I(\boldsymbol{F}_{rgb(i)},\boldsymbol{F}_{ir(j)})}$ to represent that of the infrared feature. Taking into account the alignment costs of RGB-depth and RGB-infrared, we calculate the overall ICL loss as follows.

$$\begin{aligned}\mathcal{L}_{\text{ICL}} = &\frac{1}{2}\mathbb{E}_{(\boldsymbol{I}_{rgb},\boldsymbol{I}_{depth})\sim\mathcal{D}_1}\left[\text{H}\left(\boldsymbol{y}^{R2D},\boldsymbol{p}^{R2D}\right)+\text{H}\left(\boldsymbol{y}^{D2R},\boldsymbol{p}^{D2R}\right)\right] \\ &+\frac{1}{2}\mathbb{E}_{(\boldsymbol{I}_{rgb},\boldsymbol{I}_{ir})\sim\mathcal{D}_2}\left[\text{H}\left(\boldsymbol{y}^{R2I},\boldsymbol{p}^{R2I}\right)+\text{H}\left(\boldsymbol{y}^{I2R},\boldsymbol{p}^{I2R}\right)\right],\end{aligned} \quad (3)$$

where, $\mathcal{D}_1$ and $\mathcal{D}_2$ refer to the entire training set that contains all RGB-depth pairs as well as RGB-infrared pairs, respectively. $\text{H}(\cdot)$ represents the cross-entropy loss function.

Class-level Matching (CM). In this section, we define a class as the forgery label of a face photograph on the FAS

dataset, where 0 and 1 represent bonafide and spoof, respectively. The proposed CM model is used to predict whether modality pairs belong to the same class, thereby pulling closer intra-samples in the feature space and enhancing the intra-class feature diversity.

Specifically, $\boldsymbol{F}_{rgb}$ and $\boldsymbol{F}_{depth}$ are jointly fed to a multi-modal encoder to obtain a joint embedding. In our implementation, we concatenate features along the feature dimension for simplicity. The embedding is then fed to a 540-wide fully-connected (FC) module $\mathcal{P}_{mp}$ as a matching predictor, where the softmax is used to obtain the matching probability $\boldsymbol{p}^{CM}(\boldsymbol{F}_{rgb}, \boldsymbol{F}_{depth}) = \mathsf{softmax}(\mathcal{P}_{mp}(\boldsymbol{F}_{rgb}, \boldsymbol{F}_{depth}))$. The ground-truth CM label is denoted as $\boldsymbol{y}^{CM}(\boldsymbol{F}_{rgb}, \boldsymbol{F}_{depth})$, which represents whether $\boldsymbol{F}_{rgb}$ and $\boldsymbol{F}_{depth}$ belong to the same class.

Considering both matched RGB-depth and RGB-infrared pairs, the total CM loss is formulated as:

$$\begin{aligned}\mathcal{L}_{CM} = \mathbb{E}_{(\boldsymbol{I}_{rgb},\boldsymbol{I}_{depth})\sim\mathcal{D}_1} \mathrm{H}\left(\boldsymbol{y}^{CM}(\boldsymbol{F}_{rgb},\boldsymbol{F}_{depth}), \boldsymbol{p}^{CM}(\boldsymbol{F}_{rgb},\boldsymbol{F}_{depth})\right) \\ + \mathbb{E}_{(\boldsymbol{I}_{rgb},\boldsymbol{I}_{ir})\sim\mathcal{D}_2} \mathrm{H}\left(\boldsymbol{y}^{CM}(\boldsymbol{F}_{rgb},\boldsymbol{F}_{ir}), \boldsymbol{p}^{CM}(\boldsymbol{F}_{rgb},\boldsymbol{F}_{ir})\right)\end{aligned}. \quad (4)$$

Hard Negative Mining (HNM) and Hard Positive Mining (HPM). To accelerate computation, we further adopt HNM and HPM strategies to select pairs for faster convergence. For instance, there are N×N possible sample pairs in two batches. We select sample pairs based on the contrastive similarity from Eq. (1). Unlike (Li et al. 2021) using only hard negative samples, we use both hard negative and hard positive samples. A hard negative pair refers to two samples that do not belong to the same class but have high similarity, while a hard positive pair refers to two samples that belong to the same class but have low similarity. In total, we select 2N sample pairs. One of the main benefits of this negative-positive mining strategy is that it increases the training speed while maintaining performance.

Switch-attention Fusion

In this section, we fuse three aligned features through shallow and deep fusion modules. The fusion modules are used to enhance the interaction of multimodal information between different modalities, which promotes the aggregation of complementary information and mitigates the influence of irrelevant information.

Shallow Fusion In the shallow fusion module, we obtain two sets of aligned RGB features as the inputs: $\boldsymbol{F}'_{rgb(1)}$ denotes the alignment feature between RGB and depth modalities, and $\boldsymbol{F}'_{rgb(2)}$ denotes the alignment feature between RGB and infrared modalities, respectively. Although RGB is aligned with two different modalities, $\boldsymbol{F}'_{rgb(1)}$ and $\boldsymbol{F}'_{rgb(2)}$ essentially share a common RGB feature space and align well with each other. Therefore, a complex fusion module is not necessary and a shallow additive fusion is adopted, *i.e.*, $\boldsymbol{F}'_{rgb} = \frac{\boldsymbol{F}'_{rgb(1)} + \boldsymbol{F}'_{rgb(2)}}{2}$.

Deep Switch-attention Fusion In this section, we provide a detailed description of our proposed switch-attention method. Our switch-attention approach is more efficient than merge-attention (Zheng et al. 2021) and cross-attention (Lu et al. 2019), as it rarely adds new parameters. The inputs

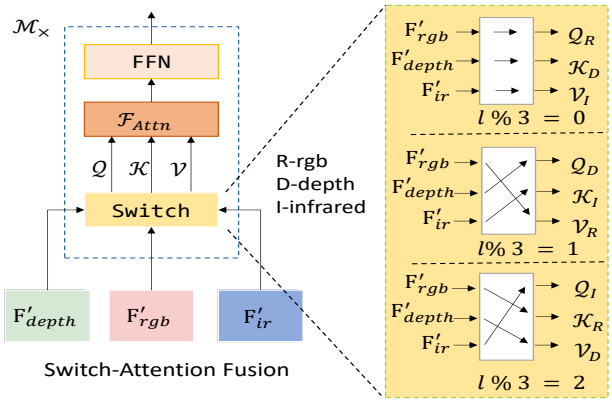


Figure 4: Illustration of the proposed switch-attention fusion. 'FFN' represents a feed forward network module with the neuron layer structure of 270×1080×270.

to our method are the aligned features from three modalities (*e.g.*, $\boldsymbol{F}'_{rgb}$, $\boldsymbol{F}'_{depth}$, and $\boldsymbol{F}'_{ir}$), while the output is the fused feature $\boldsymbol{F}_{fuse}$.

The self-attention mechanism is a powerful framework, which obtains learned weights for a sequence of values based on their relevance to given queries and keys. The attention function $\mathcal{F}_{Attn}$ is commonly expressed as:

$$\mathcal{F}_{Attn}(Q, K, V) = \mathtt{softmax}\left(QK^T \Big/ \sqrt{d_k}\right) V, \quad (5)$$

where Q, K, and V denote queries, keys, and values, respectively. d_k is the dimensionality of keys.

As presented in Figure 4, we propose a new switch-attention structure that combines the advantages of merge-attention and cross-attention, aiming to maintain only a single modality branch while efficiently and effectively integrating information from the other two modalities. Our approach differs from merge-attention and cross-attention. For instance, merge-attention uses a single branch and simply stacks multimodal features for coarse information fusion. Cross-attention maintains multiple branches but cannot directly accommodate tri-modal inputs. This structure is designed to improve computational efficiency and promote rich inter-modality interactions. Specifically, we introduce a plug-and-play switch operation before the attention module, which takes three modalities as the inputs and generates switched queries, keys, and values in a cyclic manner. This switch operation is parameter-free and can dynamically allocate attention weights to different modalities, thus encouraging the aggregation of useful information and the exclusion of irrelevant clues. The switch operation $\mathcal{F}_{sw}$ can be represented by the following formula:

$$\mathcal{F}^l_{sw} = \begin{cases} \mathcal{F}_{Attn}(\boldsymbol{F}'_{rgb}\boldsymbol{W}^Q_l, \boldsymbol{F}'_{depth}\boldsymbol{W}^K_l, \boldsymbol{F}'_{ir}\boldsymbol{W}^V_l), & \text{if } l \bmod 3 = 0, \\ \mathcal{F}_{Attn}(\boldsymbol{F}'_{depth}\boldsymbol{W}^Q_l, \boldsymbol{F}'_{ir}\boldsymbol{W}^K_l, \boldsymbol{F}'_{rgb}\boldsymbol{W}^V_l), & \text{if } l \bmod 3 = 1, \\ \mathcal{F}_{Attn}(\boldsymbol{F}'_{ir}\boldsymbol{W}^Q_l, \boldsymbol{F}'_{rgb}\boldsymbol{W}^K_l, \boldsymbol{F}'_{depth}\boldsymbol{W}^V_l), & \text{if } l \bmod 3 = 2, \end{cases} \quad (6)$$

where l represents the depth of the current transformer block. The switched three modalities represent the queries, keys, and values that will be subsequently passed to the attention module.

Taking the aligned feature $\boldsymbol{F}'$ as an example, we first calculate the query, key, and value matrices: $Q = \boldsymbol{F}'\boldsymbol{W}^Q$, $K = \boldsymbol{F}'\boldsymbol{W}^K$, $V = \boldsymbol{F}'\boldsymbol{W}^V$, where $\boldsymbol{W}^Q$, $\boldsymbol{W}^K$, and $\boldsymbol{W}^V$ represent learned projection matrices corresponding to Q, K, and V, respectively. Assuming $\boldsymbol{F}' = \texttt{Concat}(\boldsymbol{F}'_{rgb}, \boldsymbol{F}'_{depth}, \boldsymbol{F}'_{ir})$, we use the left-shift operation to change the order of aligned RGB, depth, and infrared features as shown in Figure 4: $\boldsymbol{F}'_1, \boldsymbol{F}'_2, \boldsymbol{F}'_3 = \boldsymbol{F}' \ll (l\%3)$. For instance, when $l\%3$=1, $\boldsymbol{F}'_1 = \boldsymbol{F}'_{depth}$, $\boldsymbol{F}'_2 = \boldsymbol{F}'_{ir}$, and $\boldsymbol{F}'_3 = \boldsymbol{F}'_{rgb}$, respectively. Consequently, we can formulate the i-th head $\boldsymbol{h}_i$ at the l-th depth as

$$\boldsymbol{h}_i^l = \mathcal{F}_{sw}^l \left(\boldsymbol{F}'_1 \boldsymbol{W}_i^Q, \boldsymbol{F}'_2 \boldsymbol{W}_i^K, \boldsymbol{F}'_3 \boldsymbol{W}_i^V \right). \tag{7}$$

Switch-attention can be achieved by several single transformer encoders, which fuses the multimodal features together. We denote the fused features by the concatenation $\texttt{Concat}(\cdot)$. Finally, the calculation of a single layer of switch-attention is:

$$\mathcal{F}_{mhsa}^l = \texttt{Concat} \left(\boldsymbol{h}_1^l, \boldsymbol{h}_2^l, ... \right) \boldsymbol{W}^L, \tag{8}$$

where $\mathcal{F}_{mhsa}^l$ represents a multi-head switch-attention module and $\boldsymbol{W}^L$ represents a linear projection matrix. Switch-attention has the benefits of simple implementation and sufficient modality interaction, though it comes with higher parameters. Note that we omit all layer normalization modules, subsequent feed forward networks, and residual connections here, as they are the same as the original ViT. We update the output of each layer to the color feature while keeping the other two modal features unchanged.

The final fused feature, which is the mean of three modal features, is subsequently fed into a 270-wide fully connected (FC) module for binary classification to predict the forgery result. With binary cross-entropy loss, the overall loss function is given by:

$$\mathcal{L}_{switch-attention} = \mathcal{L}_{BCE} + \alpha \mathcal{L}_{ICL} + \beta \mathcal{L}_{CM}. \tag{9}$$

For the sake of convenience, α and β are set to 1 based on our experiments.

4 Experimental Validation and Analysis

Experiment Settings

Datasets. In this section, we evaluate the performance of representative FAS models on four commonly-used multimodal benchmark datasets. (i) MmFA (Zhang et al. 2020a) is composed of a vast collection of 1000 subjects and 21000 video clips, incorporating 3 modalities including *RGB*, *depth*, and *infrared*. Two attacks are included: *print* and *cut*. (ii) CeFa (Liu et al. 2021) is the largest database in our experiment, including 23346 videos from 1607 subjects, 4 attack types and three modalities: *RGB*, *depth*, and *infrared*. (iii) WMCA (George et al. 2019) represents the wide multi-channel presentation attack (WMCA) dataset, consisting of 1941 short video recordings under varied conditions and covers 4 modalities (*i.e.*, *RGB*, *depth*, *infrared*, and *thermal*). (iv) HQ-WMCA (Heusch et al. 2020) contains 2904 recordings from 51 participants with 5 modalities (*i.e.*, *RGB*, *depth*, *infrared*, *thermal*, and *short-wave infrared*).

Modality Completion. In HQ-WMCA, the depth modality is inaccessible, so we generate a high-quality depth map using PRNet (Feng et al. 2018) for modality completion. Meanwhile, PRNet generates a depth map, and we manually remove the depth modality for flat presentation attack types (*i.e.*, print and replay) to more accurately represent real-world conditions. Finally, the first three datasets use RGB, depth, and near-infrared modalities, while HQ-WMCA uses RGB, depth, and short-wave infrared for evaluation.

Evaluation Protocols. To ensure fair evaluation, all comparison models are trained, validated, and tested on CeFa. The main reason is that CeFa is the largest in these four databases. Then, we further perform the cross-dataset testing on the other three complete datasets. The cross-dataset testing protocol is challenging and effectively assesses the generalization ability and robustness of FAS models.

Three commonly used metrics, namely *Half Total Error Rate (HTER)*, *Area Under Curve (AUC)*, and *Accuracy*, are utilized to evaluate the overall performance of all FAS methods. HTER computes the average of False Rejection Rate (FRR) and False Acceptance Rate (FAR). AUC is computed based on the Receiver Operating Characteristic (ROC) curve, which plots True Positive Rate (TPR) against False Positive Rate (FPR) under different classification thresholds. AUC quantifies the ability to distinguish between bonafide and spoofing. Accuracy refers to the percentage of images correctly classified by the FAS system as true or false. Particularly, HTER is measured under the *equal error rate* (EER) threshold and Accuracy is measured under the BPCER=1%.

Implementation details. Our method is implemented in *PyTorch*, where all images are resized to 224×224. We use a balance sampler to randomly sample data from a dataset while ensuring that the number of *bonafide* and *spoofing* images is roughly equal within the same batch. We use three identical independent ViTs with a depth of 3 and a dimension of 270. The matching predictor, $\mathcal{F}_{mp}$, in class-level matching task consists of a 540-wide FC. The fusion modules use transformer blocks with 6 attention heads, a depth of 6, and a dimension of 45 for each head, respectively. The feature dimension is 270. The fused features are passed through a classification head, consisting of two 270-wide FC modules with the *ReLU* activation.

In the training stage, we train our *mmFAS* model using the *Adam* optimizer with a weight decay setting to *1e-2*, and use a cosine annealing scheduler with a maximum number of warmup steps set to ten percent of total steps and a maximum learning rate of *1e-4*. *mmFAS* is trained for 30 epochs on 1 *NVIDIA-A100* GPU with a batch size of 128. In the testing stage, the alignment module is disabled and we report the HTER, AUC, and Accuracy results, respectively. The ratio of image pairs used for training, validating, and testing are 6:1:13.

Comparison Methods. To make a reasonable comparison, we compare our *mmFAS* with several unimodal and multimodal FAS models. We also include three general methods, namely ResNet (He et al. 2016), MLP-Mixer (Tolstikhin et al. 2021), and ViT (Dosovitskiy et al. 2021). Specifically, only the RGB modality is used in the unimodal FAS models, and three modalities (*i.e.*, RGB, depth, and infrared) are used

Method	CeFa			MmFA			WMCA			HQ-WMCA			Weighted		
	HTER↓	AUC↑	Acc↑	HTER↓	AUC↑	Acc↑	HTER↓	AUC↑	Acc↑	HTER↓	AUC↑	Acc↑	HTER↓	AUC↑	Acc↑
DeepPixBiS	**0.0001**	99.99	99.67	49.66	54.47	30.15	45.49	55.42	31.70	39.69	62.27	25.14	36.86	65.07	46.21
ADeepPixBiS	0.1183	99.98	99.36	59.17	41.56	39.18	45.76	52.52	49.64	37.97	45.24	31.71	40.60	58.87	56.59
ResNet50	0.5944	99.97	99.23	41.80	62.45	52.75	44.49	56.23	76.28	45.34	55.95	76.33	33.68	68.35	72.45
MLP-Mixer	1.392	99.99	98.77	42.22	61.21	56.29	44.79	56.19	78.61	47.52	54.60	72.62	34.15	67.86	74.57
ConvMLP	0.0120	99.99	99.29	10.17	96.09	30.29	68.98	23.87	78.86	48.02	53.45	26.72	31.04	68.58	64.47
PipeNet	0.3896	99.99	99.19	12.90	93.11	38.44	**37.02**	59.77	74.62	49.94	51.60	26.43	19.79	81.34	65.90
FeatherNet	0.0011	99.99	**99.99**	39.22	62.52	**69.65**	49.71	52.34	43.94	57.70	38.54	23.32	34.70	66.71	66.03
FlexModal-FAS	0.0011	99.99	99.98	39.22	65.76	30.36	41.40	59.03	40.55	57.70	66.17	23.32	31.48	70.79	49.77
FaceBagNet	0.0954	99.99	99.27	28.06	78.73	40.93	56.19	41.61	79.25	47.79	53.47	65.16	32.90	68.86	69.03
ViT-S/16	0.0546	99.99	99.27	10.30	95.49	67.79	58.05	38.52	79.22	49.00	52.33	38.58	26.87	74.02	78.97
Ours	0.0183	**99.99**	99.93	**9.24**	**97.80**	67.14	38.68	**64.79**	**79.26**	**36.92**	**69.85**	**76.38**	**18.84**	**85.24**	**79.24**

Table 1: Overall performance comparison of the proposed method and 10 representative methods in terms of HTER(%), AUC(%), and Accuracy(%) on four multimodal datasets. The best results are highlighted in **bold**. "↑" represents the higher the better, and "↓" represents the lower the better.

in multimodal FAS models, except for *FeatherNet* (Zhang et al. 2019), which uses only depth and infrared. All models are trained from scratch with the same settings.

For unimodal methods, two FAS-specific methods (*e.g.*, *DeepPixBiS* (George and Marcel 2019) and *ADeepPixBiS* (Hossain et al. 2020)) are supervised by auxiliary tasks and two generic methods (*e.g.*, *ResNet* (He et al. 2016) and *MLP-Mixer* (Tolstikhin et al. 2021)) are supervised by the binary cross-entropy loss. It is worth noting that for *ResNet* and *MLP-Mixer*, only the last fully-connected layers are modified to suit the unimodal FAS tasks.

For multimodal methods, five FAS-specific methods and one generic method are chosen for comparison. Specifically, three modalities (*i.e.*, RGB, depth, infrared) are used in *ConvMLP* (Wang et al. 2022), *PipeNet* (Yang et al. 2020), *FlexModal-FAS* (Yu et al. 2023), *FaceBagNet* (Shen, Huang, and Tong 2019), *ViT* (Dosovitskiy et al. 2021), and *mmFAS*, while two modalities (*i.e.*, only depth and infrared) are used in *FeatherNet* (Zhang et al. 2019).

Comparison Results

In the experiments, each FAS model is only trained on CeFa and tested on the rest of CeFa and other three datasets (*i.e.*, MmFA, WMCA, and HQ-WMCA). The weighted average results (*i.e.*, in terms of image numbers) are also reported, where the weight is the ratio of the number of images in one dataset to the total number of images in all four datasets.

Table 1 reveals that all methods achieve impressive performance on the CeFa dataset, where they are trained and tested on the same dataset. However, when transferring to the other three datasets in a zero-shot manner, different levels of performance degradation are observed. The qualitative metrics (*i.e.*, HTER, AUC, and Accuracy) of all models drop significantly, and some models even perform worse than random prediction, such as DeepPixBis (George and Marcel 2019) and ADeepPixBis (Hossain et al. 2020) on MmFA. For example, FeatherNet (Zhang et al. 2019) achieves the best Accuracy (69.65%) on the MmFA dataset, but ranks last on the HQ-WMCA dataset in terms of AUC (38.54%). As seen, our *mmFAS* outperforms all generic models and FAS-specific methods, achieving the best weighted average results.

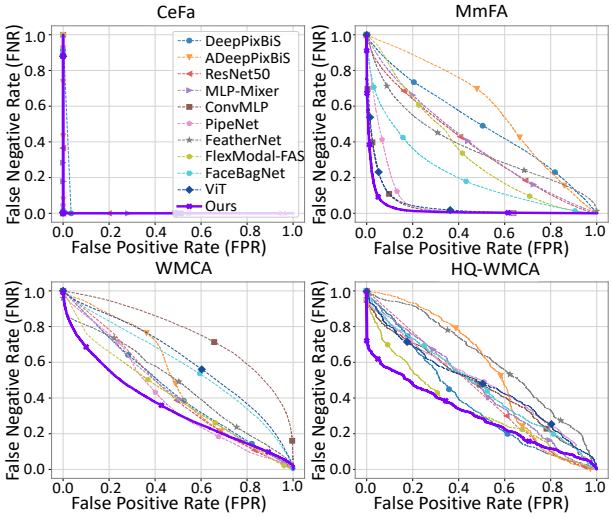


Figure 5: FPR-FNR curves of four datasets in inter-dataset testing. The closer the curve is to the origin, the better.

Figure 5 further highlights the superior performance of our *mmFAS*, demonstrating its ability to leverage cross-modal information and validate the effectiveness of the proposed alignment and fusion modules.

Ablation Study

Our proposed *mmFAS* framework consists of two innovative modules: the multiple-level alignment and the switch-attention fusion modules. To evaluate the effectiveness of these modules, we carry out 18 independent ablation experiments on the above mentioned cross-dataset settings, as shown in Table 2. The symbol '↑' indicates the higher the better, while '↓' means that the lower the better.The abbreviations 'ICL', 'CM', 'HCM', 'MA', 'CA', 'SA' and 'wAcc' represent instance contrastive learning, class matching, class matching with hard negative mining, hard positive mining, merge-attention (Zheng et al. 2021), cross-attention (Lu et al. 2019), switch-attention and weighted accuracy respectively. The checkmark symbol '✓' indicates the usage of the related module or modality in the experiments.

Modality			Alignment Loss			Attention Choice			Metrics		
R	D	I	ICL	CM	CM+HNM +HPM	MA	CA	SA	Params (M)↓	MmFA Acc(%)↑	Weighted Acc(%)↑
√			Modality Ablation						2.92	52.31	70.72
	√								2.92	58.88	68.68
		√							2.92	45.67	71.01
√	√								5.76	49.09	71.51
√		√							5.76	41.69	69.34
	√	√							5.76	46.44	60.41
√	√	√							8.60	55.18	74.10
√	√	√	√			Alignment Ablation			8.60	33.29	65.05
√	√	√		√					8.60	34.22	56.36
√	√	√			√				8.60	35.41	56.67
√	√	√	√	√					8.60	34.07	65.35
√	√	√	√		√				8.60	30.26	61.09
√	√	√	Attention without Alignment			√			56.11	34.81	58.74
√	√	√					√		29.66	46.90	71.31
√	√	√						√	15.63	48.47	61.81
√	√	√	√	Attention with Alignment	√	√			56.11	51.75	72.94
√	√	√	√		√		√		29.66	50.80	72.98
√	√	√	√		√			√	15.63	67.14	79.24

Table 2: Ablation experiments on choices of modalities, alignment losses, and attentions.

RDI	*RID*	*DRI*	*IDR*	*DIR*	*IRD*
85.24	82.28	82.18	84.74	82.76	82.32

Table 3: Comparison results of different switching modes.

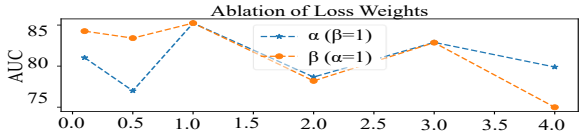


Figure 6: Parameter sensitivity experiments of α and β.

Impact of modality choice. We evaluate the performance of *mmFAS* under three different modality configurations: unimodal, bimodal, and trimodal. In the first 7 rows of Table 2, the trimodal configuration achieves the best weighted accuracy, demonstrating the importance of all three modalities. Therefore, it is important to consider multimodal feature alignment and fusion.

Effect of alignment setting. In the ablation study of the alignment loss function, there is a varying reduction in weighted accuracy between 8.75% and 17.74%. This performance decline may be attributed to the use of an alignment module without the accompaniment of a fusion module. The optimization goal of the alignment, which may not align perfectly with the ultimate classification objectives, can result in the model astray if overly emphasized.

Effect of attention selection. Based on the difference between merge-attention, cross-attention, and switch-attention, we have conducted the ablation studies under two settings: attention with and without alignment. To ensure fairness, we keep the parameter counts comparable. As seen, our switch-attention outperforms merge-attention and cross-attention while maintaining similar parameter costs.

Ablation of switch mode. As shown in Figure 4, there are six possible Q-K-V sequences under a fixed switching direction. For example, we use RDI to denote the switch mode Q-K-V sequence as $RDI \rightarrow DIR \rightarrow IRD \rightarrow \cdots$. Table 3 shows that this sequence order RDI achieves the best performance. Notably, even our least effective configuration surpasses all other methods in Table 1. RDI is adopted for both training and inference in the experiments.

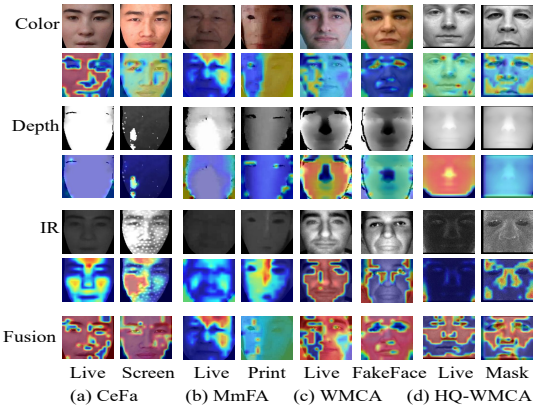


Figure 7: Visualization of presentation attacks.

Hyper-parameter sensitivity of α and β. In our experiments, α and β are the weights for the instance-level and class-level alignment loss, respectively. We have varied α while fixing $\beta = 1$, and then varied β while fixing $\alpha = 1$, as shown in Figure 6. The results reveal that either under- or over-emphasis on any alignment loss degrades performance. Therefore, we set both α and β to 1, as this configuration yields the highest AUC for *mmFAS*.

Visualization Comparison

To further demonstrate the cross-dataset testing results, Figure 7 illustrates the original RGB, depth, infrared images, and gradient maps of representative samples from four datasets. The visualization is computed using Grad-CAM (Selvaraju et al. 2017). It demonstrates that *mmFAS* is resilient to noise by adaptively weighting and aggregating complementary information, making it more reliable and robust in real-world scenarios with complex environment.

5 Conclusion

In this paper, we propose a novel framework called *mmFAS* for the multimodal face anti-spoofing task, which incorporates multi-level alignment and switch-attention fusion modules. We conduct extensive experiments to demonstrate the effectiveness of the proposed framework, and the results reveal that our method outperforms both generic and FAS-specific approaches, thereby confirming the significance of aligning multimodal features prior to fusion. Furthermore, we find that multimodal models without alignment may result in even poorer generalization and significant performance degradation in cross-dataset experiments compared to unimodal models.

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China under Grants 62472290 and 62372306, and in part by the Natural Science Foundation of Guangdong Province under Grants 2024A1515011972, 2023A1515011197 and 2022A1515011245.

References

Agarwal, A.; Yadav, D.; Kohli, N.; Singh, R.; Vatsa, M.; and Noore, A. 2017. Face Presentation Attack with Latex Masks in Multispectral Videos. In *IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 275–283.

Cai, Rizhao and Cui, Yawen and Li, Zhi and Yu, Zitong and Li, Haoliang and Hu, Yongjian and Kot, Alex. 2023. Rehearsal-Free Domain Continual Face Anti-Spoofing: Generalize More and Forget Less. In *IEEE International Conference on Computer Vision (ICCV)*, 8037–8048.

Dong, X.; Liu, H.; Cai, W.; Lv, P.; and Yu, Z. 2021. Open Set Face Anti-Spoofing in Unseen Attacks. In *ACM International Conference on Multimedia (ACM MM)*, 4082–4090.

Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; Uszkoreit, J.; and Houlsby, N. 2021. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In *International Conference on Learning Representations (ICLR)*.

Feng, Y.; Wu, F.; Shao, X.; Wang, Y.; and Zhou, X. 2018. Joint 3d face reconstruction and dense alignment with position map regression network. In *European Conference on Computer Vision (ECCV)*, 534–551.

Gao, G.; Yu, Y.; Yang, J.; Qi, G.-J.; and Yang, M. 2022. Hierarchical Deep CNN Feature Set-Based Representation Learning for Robust Cross-Resolution Face Recognition. *IEEE Transactions on Circuits and Systems for Video Technology (TCSVT)*, 32(5): 2550–2560.

George, A.; and Marcel, S. 2019. Deep Pixel-Wise Binary Supervision for Face Presentation Attack Detection. In *International Conference on Biometrics (ICB)*, 1–8.

George, A.; and Marcel, S. 2021. Cross modal focal loss for rgbd face anti-spoofing. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 7882–7891.

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

Guo, X.; Liu, Y.; Jain, A.; and Liu, X. 2022. Multi-domain Learning for Updating Face Anti-spoofing Models. In *European Conference on Computer Vision (ECCV)*, 230–249.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 770–778.

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

Hossain, M. S.; Rupty, L.; Roy, K.; Hasan, M.; Sengupta, S.; and Mohammed, N. 2020. A-DeepPixBis: Attentional Angular Margin for Face Anti-Spoofing. In *Digital Image Computing: Techniques and Applications (DICTA)*, 1–8.

Huang, H.-P.; Sun, D.; Liu, Y.; Chu, W.-S.; Xiao, T.; Yuan, J.; Adam, H.; and Yang, M.-H. 2022. Adaptive transformers for robust few-shot cross-domain face anti-spoofing. In *European Conference on Computer Vision (ECCV)*, 37–54.

Jia, S.; Li, X.; Hu, C.; Guo, G.; and Xu, Z. 2021. 3D face anti-spoofing with factorized bilinear coding. *IEEE Transactions on Circuits and Systems for Video Technology*, 31(10): 4031–4045.

Jiang, F.; Liu, P.; Shao, X.; and Zhou, X. 2020. Face anti-spoofing with generated near-infrared images. *Multimedia Tools and Applications*, 79: 21299–21323.

Kong, C.; Zheng, K.; Wang, S.; Rocha, A.; and Li, H. 2022. Beyond the Pixel World: A Novel Acoustic-Based Face Anti-Spoofing System for Smartphones. *IEEE Transactions on Information Forensics and Security (TIFS)*, 17: 3238–3253.

Kong, Z.; Zhang, W.; Liu, F.; Luo, W.; Liu, H.; Shen, L.; and Ramachandra, R. 2023. Taming Self-Supervised Learning for Presentation Attack Detection: De-Folding and De-Mixing. *IEEE Transactions on Neural Networks and Learning Systems (TNNLS)*.

Kose, N.; and Dugelay, J.-L. 2013. Shape and Texture Based Countermeasure to Protect Face Recognition Systems against Mask Attacks. In *IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 111–116.

Kuang, H.; Ji, R.; Liu, H.; Zhang, S.; Sun, X.; Huang, F.; and Zhang, B. 2019. Multi-modal multi-layer fusion network with average binary center loss for face anti-spoofing. In *ACM International Conference on Multimedia (ACM MM)*, 48–56.

Li, J.; Selvaraju, R.; Gotmare, A.; Joty, S.; Xiong, C.; and Hoi, S. C. H. 2021. Align before fuse: Vision and language representation learning with momentum distillation. *Advances in neural information processing systems (NeurIPS)*, 34: 9694–9705.

Li, X.; Wan, J.; Jin, Y.; Liu, A.; Guo, G.; and Li, S. Z. 2020. 3DPC-Net: 3D point cloud network for face anti-spoofing. In *IEEE International Joint Conference on Biometrics (IJCB)*, 1–8.

Liu, A.; and Liang, Y. 2023. Ma-vit: Modality-agnostic vision transformers for face anti-spoofing. *arXiv preprint arXiv:2304.07549*.

Liu, A.; Ma, H.; Zheng, J.; Yuan, H.; Yu, X.; Liang, Y.; Escalera, S.; Wan, J.; and Lei, Z. 2024. FM-CLIP: Flexible Modal CLIP for Face Anti-Spoofing. In *ACM International Conference on Multimedia (ACM MM)*, 8228–8237.

Liu, A.; Tan, Z.; Wan, J.; Escalera, S.; Guo, G.; and Li, S. Z. 2021. Casia-surf cefa: A benchmark for multi-modal cross-ethnicity face anti-spoofing. In *IEEE Winter Conference on Applications of Computer Vision (WACV)*, 1179–1187.

Liu, A.; Tan, Z.; Yu, Z.; Zhao, C.; Wan, J.; Liang, Y.; Lei, Z.; Zhang, D.; Li, S. Z.; and Guo, G. 2023. Fm-vit: Flexible modal vision transformers for face anti-spoofing. *IEEE Transactions on Information Forensics and Security (TIFS)*, 18: 4775–4786.

Lu, J.; Batra, D.; Parikh, D.; and Lee, S. 2019. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. *Advances in Neural Information Processing Systems (NeurIPS)*, 32.

Patel, K.; Han, H.; and Jain, A. K. 2016. Secure face unlock: Spoof detection on smartphones. *IEEE Transactions on Information Forensics and Security (TIFS)*, 11(10): 2268–2283.

Pinto, A.; Schwartz, W. R.; Pedrini, H.; and Rocha, A. d. R. 2015. Using Visual Rhythms for Detecting Video-Based Facial Spoof Attacks. *IEEE Transactions on Information Forensics and Security (TIFS)*, 10(5): 1025–1038.

Selvaraju, R. R.; Cogswell, M.; Das, A.; Vedantam, R.; Parikh, D.; and Batra, D. 2017. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *IEEE International Conference on Computer Vision (ICCV)*, 618–626.

Shen, T.; Huang, Y.; and Tong, Z. 2019. FaceBagNet: Bag-Of-Local-Features Model for Multi-Modal Face Anti-Spoofing. In *IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 1611–1616.

Srivatsan, K.; Naseer, M.; and Nandakumar, K. 2023. FLIP: Cross-domain Face Anti-spoofing with Language Guidance. In *IEEE International Conference on Computer Vision (ICCV)*, 19685–19696.

Sun, Y.; Liu, Y.; Liu, X.; Li, Y.; and Chu, W.-S. 2023. Rethinking Domain Generalization for Face Anti-spoofing: Separability and Alignment. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 24563–24574.

Tolstikhin, I. O.; Houlsby, N.; Kolesnikov, A.; Beyer, L.; Zhai, X.; Unterthiner, T.; Yung, J.; Steiner, A.; Keysers, D.; Uszkoreit, J.; et al. 2021. Mlp-mixer: An all-mlp architecture for vision. *Advances in Neural Information Processing Systems (NeurIPS)*, 34: 24261–24272.

Wang, M.; Xu, Z.; Xu, M.; and Lin, W. 2024. Blind Multimodal Quality Assessment of Low-light Images. *Springer International Journal of Computer Vision*, 1–24.

Wang, W.; Wen, F.; Zheng, H.; Ying, R.; and Liu, P. 2022. Conv-MLP: A Convolution and MLP Mixed Model for Multimodal Face Anti-Spoofing. *IEEE Transactions on Information Forensics and Security (TIFS)*, 17: 2284–2297.

Wen, D.; Han, H.; and Jain, A. K. 2015. Face spoof detection with image distortion analysis. *IEEE Transactions on Information Forensics and Security (TIFS)*, 10(4): 746–761.

Xiao, D.; Li, J.; Li, J.; Dong, S.; and Lu, T. 2022. IHEM Loss: Intra-Class Hard Example Mining Loss for Robust Face Recognition. *IEEE Transactions on Circuits and Systems for Video Technology (TCSVT)*, 32(11): 7821–7831.

Xu, X.; Xiong, Y.; and Xia, W. 2021. On improving temporal consistency for online face liveness detection system. In *IEEE International Conference on Computer Vision Workshops (ICCVW)*, 824–833.

Yang, J.; Lei, Z.; and Li, S. Z. 2014. Learn convolutional neural network for face anti-spoofing. *arXiv preprint arXiv:1408.5601*.

Yang, Q.; Zhu, X.; Fwu, J.-K.; Ye, Y.; You, G.; and Zhu, Y. 2020. PipeNet: Selective modal pipeline of fusion network for multi-modal face anti-spoofing. In *IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 644–645.

Yang, X.; Luo, W.; Bao, L.; Gao, Y.; Gong, D.; Zheng, S.; Li, Z.; and Liu, W. 2019. Face anti-spoofing: Model matters, so does data. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 3507–3516.

Yu, Z.; Li, X.; Shi, J.; Xia, Z.; and Zhao, G. 2022a. Revisiting pixel-wise supervision for face anti-spoofing. *IEEE Transactions on Biometrics, Behavior, and Identity Science*, 3(3): 285–295.

Yu, Z.; Liu, A.; Zhao, C.; Cheng, K. H.; Cheng, X.; and Zhao, G. 2023. Flexible-modal face anti-spoofing: A benchmark. In *IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 6345–6350.

Yu, Z.; Qin, Y.; Li, X.; Zhao, C.; Lei, Z.; and Zhao, G. 2022b. Deep learning for face anti-spoofing: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 45(5): 5609–5631.

Zhang, P.; Zou, F.; Wu, Z.; Dai, N.; Mark, S.; Fu, M.; Zhao, J.; and Li, K. 2019. FeatherNets: Convolutional Neural Networks as Light as Feather for Face Anti-Spoofing. In *IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 1574–1583.

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

Zhang, Y.; Yin, Z.; Li, Y.; Yin, G.; Yan, J.; Shao, J.; and Liu, Z. 2020b. Celeba-spoof: Large-scale face anti-spoofing dataset with rich annotations. In *European Conference on Computer Vision (ECCV)*, 70–85.

Zheng, R.; Chen, J.; Ma, M.; and Huang, L. 2021. Fused acoustic and text encoding for multimodal bilingual pretraining and speech translation. In *International Conference on Machine Learning (ICML)*, 12736–12746.

Zhou, Q.; Zhang, K.-Y.; Yao, T.; Lu, X.; Yi, R.; Ding, S.; and Ma, L. 2023. Instance-Aware Domain Generalization for Face Anti-Spoofing. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 20453–20463.

Zhou, Q.; Zhang, K.-Y.; Yao, T.; Yi, R.; Ding, S.; and Ma, L. 2022a. Adaptive mixture of experts learning for generalizable face anti-spoofing. In *ACM International Conference on Multimedia (ACM MM)*, 6009–6018.

Zhou, Q.; Zhang, K.-Y.; Yao, T.; Yi, R.; Sheng, K.; Ding, S.; and Ma, L. 2022b. Generative domain adaptation for face anti-spoofing. In *European Conference on Computer Vision (ECCV)*, 335–356.