Towards Unified Representation of Invariant-Specific Features in Missing Modality Face Anti-Spoofing

Guanghao Zheng[1†], Yuchen Liu[2†], Wenrui Dai[1⋆], Chenglin Li[2], Junni Zou[1], and Hongkai Xiong[2]

[1] Department of Computer Science and Engineering, Shanghai Jiao Tong University
[2] Department of Electronic Engineering, Shanghai Jiao Tong University
{zgh990318,liuyuchen6666, daiwenrui, lcl1985, zoujunni, xionghongkai}@sjtu.edu.cn

Abstract. The effectiveness of Vision Transformers (ViTs) diminishes considerably in multi-modal face anti-spoofing (FAS) under missing modality scenarios. Existing approaches rely on modality-invariant features to alleviate this issue but ignore modality-specific features. To solve this issue, we propose a **M**issing **M**odality **A**dapter framework for **F**ace **A**nti-**S**poofing (MMA-FAS), which leverages modality-disentangle adapters and LBP-guided contrastive loss for explicit combination of modality-invariant and modality-specific features. Modality-disentangle adapters disentangle features into modality-invariant and -specific features from the view of frequency decomposition. LBP-guided contrastive loss, together with batch-level and sample-level modality masking strategies, forces the model to cluster samples according to attack types and modal combinations, which further enhances modality-specific and -specific features. Moreover, we propose an adaptively modal combination sampling strategy, which dynamically adjusts the sample probability in masking strategies to balance the training process of different modal combinations. Extensive experiments demonstrate that our proposed method achieves state-of-the-art intra-dataset and cross-dataset performance in all the missing modality scenarios.

Keywords: Missing Modality · Face Anti Spoofing · Adapters

1 Introduction

Face recognition (FR) systems are widely used in many security fields. However, there are plenty of Presentation Attacks that deceive face recognition systems and steal privacy, including 2D attacks (e.g., print and replay attack) and 3D attacks (e.g., rigidmask and papermask attack). To protect FR systems from being attacked, face anti-spoofing (FAS) techniques have been widely concerned. Most existing FAS methods focus on RGB images [5,7,9,16–19,23,25,26,31,35–

⋆ Correspondence to Wenrui Dai. † Equal contribution.

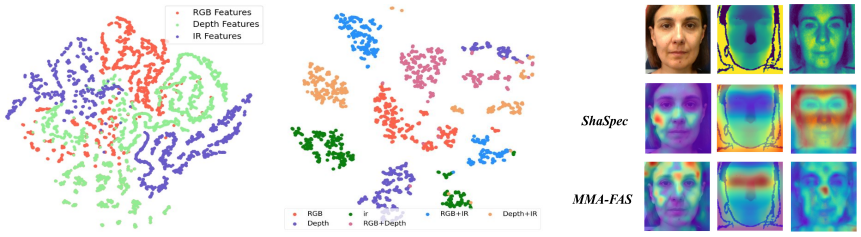


Fig. 1: (a) Visualization of features of different modalities extracted by AMA [29] (b) Visualization of feature extracted by MMANet [27]. (c) Visualization of attention in ShaSpec [24] and our MMA-FAS. We visualize the attention regions in different specific encoders in ShaSpec.

42]. Deep learning techniques like convolutional neural networks (CNNs) and vision transformers (ViTs) have been widely adopted to facilitate FAS using RGB images but cannot generalize well to unknown attacks and unseen deployment scenarios due to limited information in RGB images.

With the development of acquisition devices, images of various modalities (e.g., depth images, infrared (IR) images, and thermal images) are considered to provide complementary discriminative information beyond RGB images. Recent multi-modal FAS methods leverage ViT backbones to extract spoofing features for cross-modality fusion to improve generalization ability. However, these methods are trained and deployed under the requirement of inputs with complete modality [12, 22] and could fail in missing modality scenarios for FAS due to unmatched modalities in training and deployments.

Missing modality scenarios are common in practical applications of FAS, since we cannot guarantee each modality will be available at deployment. For example, the quality of RGB images could be poor in dark environments or the sensors of IR modality might be broken. Existing methods for missing modality FAS focus on extracting modality-invariant features for robust prediction across various modal combinations [13, 15, 29]. On the contrary, existing general methods only leverage modality-specific features for missing modality prediction [11, 24, 27]. The detailed clarification is provided in Section 3.1. Furthermore, these methods ignore the specific problem for FAS where different attack types have different spoofing cues and are significantly degraded when directly employed in missing-modality FAS.

To address the above problems, we propose to enhance the modality-specific information and integrate it with modality-invariant features during missing modal training in FAS. Different from existing methods that extract modality-invariant and modality-specific features with multiple encoders and loss functions [24], we explicitly extract them from the view of frequency based on the observation that modality-invariant cues are mainly represented in the low-frequency components, while modality-specific cues are in the high-frequency component in multi-modal FAS. As shown in Fig. 2(a), in the low-frequency components, the forehead regions (marked by red boxes) exhibit similar traits across modalities, whereas in the high-frequency components, the inherent spoofing

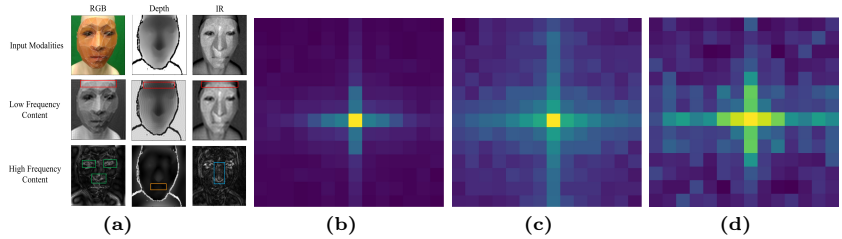


Fig. 2: (a) Spoofing cues for FAS of three modalities in low- and high-frequency components. The red boxes indicate the modality-invariant features, while other color boxes refer to modality-specific features. (b) Visualization of frequency maps for features extracted by AMA [29]. (c) Visualization of frequency maps for features extracted by Missing-aware Prompts [11]. (d) Visualization of frequency maps for features extracted by our MMA-FAS.

cues (marked by other color boxes) differ among different modalities. Visualization of the frequency maps of features extracted by AMA [29] and Missing-aware Prompt [11] shows that AMA [29] extracts modality-invariant features and focuses more on low-frequency regions, while Missing-aware Prompt [11] learns modality-specific features and focuses more on low-frequency regions.

Motivated by this observation, we propose a new framework named MMA-FAS. With a multi-modal vision transformer as the backbone, we develop a modality-disentangle adapter (MDA) to separate the fusion feature into modality-invariant and modality-specific ones by frequency decomposition and extract local fine-grained invariant and specific features via convolutions. Besides, we propose a LBP-guided missing modal contrastive loss, which leverages batch-level and sample-level masking strategy with LBP-guided (local binary pattern [1]) contrastive loss to cluster the images according to their attack types and modal combinations for further modality-specific feature enhancement and separation of spoofing features in different attack types. Furthermore, we introduce an Adaptively Modal combination Sampling (AMS) into Masking to dynamically adjust the sampling probabilities, ensuring balanced training for different modal combinations. In our experiment, we provide a comprehensive benchmark in all the missing modality cases. All the experiments demonstrate that MMA-FAS enhances modality-specific information and achieves SOTA performance in all the missing modality scenarios. To sum up,

- We propose a new framework named MMA-FAS to extract both modality-invariant and modality-specific information in FAS and solve the missing modality problem.
- To extract modality-specific features accurately, we propose a modality-disentangle adapter that decomposes the frequency information and extracts fine-grained spoofing cues from both modality-invariant and modality-specific features.
- We design batch-level and sample-level masking strategies with LBP-guided contrastive loss for further enhancing modality-specific features. Adaptive sampling strategy on modal combination is further developed to balance the training process of different modal combinations.

– We provide an extensive evaluation across all scenarios with missing modalities, and achieve SOTA performance on various datasets, e.g., WMCA, CASIA-SURF and CASIA-SURF-CeFA, under several settings.

2 Related Works

2.1 Multi-modal Face Anti-Spoofing

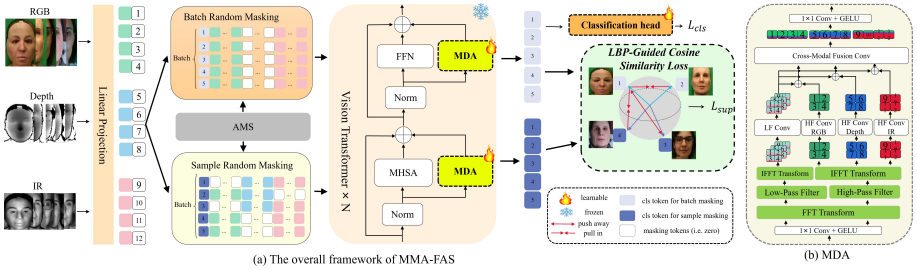
Recent FAS methods tend to use multi-modal inputs rather than unimodal inputs to achieve better FAS performance through the fusion of complementary information. Early approaches leverage CNNs to extract specific features from each modality and then fuse them in the latter stage. CMFL designs cross-modal focal loss to adjust the confidence of each channel in the extracted features [4]. CDCN [30] applies CDC into multi-modal FAS and proves the gradient information in features is important for discrimination capacity. ViTs are commonly used later due to the ability to achieve finer-grained cross-modal fusion than CNNs. MFAST [22] applies two transformers to extract features in two modalities. MFViT [12] proposes a multi-feature and multi-rank fusion strategy to learn multi-scale features. AMA [29] is the first to jointly consider the local descriptor, adapters, and pretraining framework in FAS. However, all these methods only leverage modality-invariant features while ignoring the fine-grained modality-specific knowledge, which is also beneficial for the FAS task.

2.2 Missing Modalities in Multi-modal Learning

The missing modality problem has received a lot of attention. Existing works usually randomly drop some modalities in the training stage to mimic the missing modality scenarios in the test stage. [28] leverage the autoencoder to reconstruct the missing modalities. [33] uses four specific encoders for unimodal input and then applies ViT to achieve cross-modal feature fusion. MMANet [27] leverages a deployment network with missing modalities to distill knowledge from the teacher network, which is trained with complete modalities. ShaSpec [24] leverages several CNN encoders to capture modality-specific and modality-invariant features for fusion. However, these methods lack of local fine-grained spoofing cues extraction and separation of spoofing features in different attack types and cannot be directly applied to ViT in multi-modal FAS.

3 Methods

Given N multi-modal FAS inputs $\{X_i, Y_i\}_{i=1}^{N}$, where $X_i = \{X_i^{RGB}, X_i^{Depth}, X_i^{IR}\}$ are input images with three modalities and Y_i are live/spoofing labels. In this paper, multi-modal vision transformer [2] is adopted as the baseline. Since finetuning the whole ViT for FAS is inefficient [29], we freeze the backbone and only tune MDAs and the classification head. As mentioned in Sec 1, training these multi-modal ViTs in missing modality scenarios heavily relies on modality-invariant



(a) The overall framework of MMA-FAS (b) MDA

Fig. 3: (a) The overall architecture of our MMA-FAS. We first apply MDA to extract modality-specific features for FAS. Then, we combine batch-level random masking strategy and sample-level random masking strategy to mimic the missing modality scenarios. With the LBP-guided contrastive loss, modality-specific features are extracted. (b) The details of MDA, where each modality corresponds to a convolution branch. These convolutions extract local features in each modality to benefit the FAS task.

features. To extract both modality-invariant and modality-specific features during training, we propose the MMA-FAS framework. Section 3.2 overviews the whole framework of MMA-FAS and modality-disentangle adapter. Section 3.3 develops two levels of masking strategy with LBP-guided contrastive loss and Section 3.4 presents the cosine function sampling strategy.

3.1 Difference from Existing Methods

To clarify the novelty of our MMA-FAS, we distinguish it from existing works and specify their shortcomings. In Section 1, we argue existing works cannot extract modality-invariant and modality-specific effectively. To be specific, MA-ViT [13] and FM-ViT [15] only focus on modality-invariant feature via attention mechanisms. Adaptive Multi-modal Adapter (AMA) [29] equally leverages modality-invariant features, since the same adapter is applied in different modal combinations, as shown in Fig. 1(a). On the contrary, Missing-Aware Prompts [11] utilizes specific prompts for each modal combinations for modality-specific features but ignores modality-invariant features. MMANet [27] distills modality-specific features from teacher network to the deployment network, as illustrated in Fig. 1(b). ShaSpec [24] adopts the shared encoder and specific encoders for modality-invariant and -specific features extraction. However, the specific encoders cannot extract fine-grained spoofing features for FAS by considering only the modality discrepancy of modality-specific features to distinguish different modalities rather than fine-grained features (e.g. spoofing cues for FAS), as demonstrated in Fig. 1(c). Different from these works, our MMA-FAS simultaneously considers modality-invariant and -specific features in each layer via frequency decomposition for fine-grained spoofing features extraction.

3.2 Missing Modality Adapter for FAS

Fig. 3 (a) depicts the proposed framework of MMA-FAS. Three modality inputs are first converted into the patch embeddings z^{RGB}, z^{Depth}, z^{IR} via linear

projection. To mimic the incomplete modalities in deployment, we introduce a Bernoulli mask $\Delta = \{\delta_1, \cdots, \delta_m\}$, where δ_i is set as 0 to mimic the corresponding missing modality. Thus, there are 7 possible patterns for Δ with $m = 3$ in total.[3] We apply two random masking strategies in our MMA-FAS and here we firstly introduce the sample-level random masking strategy. The batch-level random masking strategy will be introduced in the next section. The sample-level random masking strategy randomly selects Δ to mask z^{RGB}, z^{Depth} and z^{IR} for each sample for each mini-batch. The masking operation involves element-wise multiplying the Δ with the corresponding modality patch embeddings $\{z^{rgb}, z^{depth}, z^{ir}\} = \{z^{RGB}, z^{Depth}, z^{IR}\} \otimes \Delta$. After masking, these patch embeddings are concatenated as $z = \text{concat}(z^{cls}, z^{rgb}, z^{depth}, z^{ir})$, where z^{cls} is the class token and z^j ($j \in \{rgb, depth, ir\}$) denotes patch embeddings of modality j after masking.

Consequently, z is fed into the multi-modal ViT backbone of N_b transformer blocks for cross-modal fusion, which consists of LayerNorm, Multi Head Self Attention (MHSA) and Feed Forward Network (FFN) module. To enhance modality-specific features during training, we propose modality disentangle adapters (MDA) to extract both modality-invariant and modality-specific features by frequency decomposition.

As illustrated in Fig. 3 (b), MDA consists of five modules, including downsampling, frequency decomposition, feature extraction, feature fusion, and upsampling. The N_t input patch embeddings $z \in \mathbb{R}^{N_t \times D}$ are first downsampled to K dimension ($K \ll D$) with a downsampling module of a 1×1 convolution and a GELU activation. Specifically, the 1D patch tokens are reshaped into 2D for the next operation. Based on observations in Section 1, we then utilize a low-pass filter and a high-pass filter to separate modality-invariant and modality-specific features. The concatenated 2D feature is converted to the frequency map via FFT transform. Two filter masks with a $f \times f$ kernel are overridden on the frequency map for extracting low-frequency components and high-frequency components separately. The filter masks are two 0/1 binary matrices with the same size as the frequency map. For low-frequency components, the value is 1 inside the kernel and 0 outside, while for high-frequency components, the value is 0 inside the kernel and 1 outside. After filtering, two masked frequency maps are converted to low-frequency features and high-frequency features using the IFFT transform. Then, in order to extract features corresponding to specific frequencies, modality-invariant features are sent to 5×5 Low-Frequency convolutions (LF Conv), while modality-specific features are sent to 3×3 High-Frequency convolutions (HF Conv), since a small kernel (3x3) tends to extract high-frequency details, while a large kernel (5x5) is more sensitive to low-frequency features. Moreover, we use convolution instead of linear transformation due to the local information benefits FAS tasks greatly. The low-frequency features are added to their corresponding high-frequency features of each modality to ensure the

[3] The 7 possible Δ are $\{1,0,0\}$, $\{0,1,0\}$, $\{0,0,1\}$, $\{1,1,0\}$, $\{1,0,1\}$, $\{0,1,1\}$ and $\{1,1,1\}$, corresponding to 7 modal combinations (RGB, Depth, IR, RGB+Depth, RGB+IR, Depth+IR and RGB+Depth+IR).

richness of information in feature of every modalities. We utilize a cross-modal fusion convolution layer to interactively fuse features across multiple modalities. Finally, the extracted modality-invariant and modality-specific features are added and the 1D flatten sequences are upsampled with a 1×1 convolution and a GELU activation.

After feature extraction and cross-modal fusion, the class token is sent to the classification head for the final prediction $\hat{Y}_i$. The classification loss is $L_{cls} = \sum_{i=1}^{N} \mathrm{CE}(\hat{Y}_i, Y_i)$, where CE is binary cross entropy loss and Y_i is the label.

3.3 LBP-guided Missing Modality Contrastive Learning

Although MDA effectively extracts modality-invariant and modality-specific features, there lack of regulation of these two features. To further enhance the modality-invariant and modality-specific features, we apply contrastive loss to cluster images of the same modal combinations while separate images of different modal combinations. The ensures features in the same modal combination contain similar information, which is modality-invariant, while features in different modal combination contain their own unique information, which is modality-specific. However, the vanilla contrastive loss cannot be directly applied here since the modal combinations are different in each batch under the sample-level masking strategy, which means there may be no positive pairs in one batch.

To solve this problem, in MMA-FAS, we combine sample-level and batch-level random masking strategies to obtain sufficient positive and negative samples for computing the contrastive loss. Different from sample-level random masking strategy, the batch-level random masking strategy randomly selects one Δ in a mini-batch so that the samples within the same batch belong to the same modality combination. Therefore, we leverage the batch-level masking strategy to generate positive sample pairs with the same modal combinations and the sample-level masking strategy for negative sample pairs with different modal combinations. As shown in Fig. 3(a), the batch-level random masking strategy adopts one Δ for all the samples in one batch such that all the samples in the batch have the same modal combination, while the sample-level random masking strategy randomly selects different Δ for each sample in one batch. Let $\Delta^b \in \mathbb{R}^{B\times 3}$ and $\Delta^s \in \mathbb{R}^{B\times 3}$ denote the batch-level and sample-level masks for patch embeddings z^{RGB}, z^{Depth}, z^{IR} in a batch of size B, respectively. The two CLS tokens are concatenated as

$$\begin{aligned} z_s &= \mathrm{concat}(z_s^{cls}, z_s^{rgb}, z_s^{depth}, z_s^{ir}) \\ z_b &= \mathrm{concat}(z_b^{cls}, z_b^{rgb}, z_b^{depth}, z_b^{ir}) \end{aligned} \tag{1}$$

where z_j^{cls}, $j \in \{s, b\}$ means the CLS token under sample-level masking and batch-level masking, z_v^j, $j \in \{rgb, depth, ir\}, v \in \{s, b\}$ means the patch embeddings of three modalities after sample-level masking and batch-level masking. z_s and z_b are separately fed into the N_b blocks to extract modality-invariant and modality-specific features. z_s^{cls} and z_b^{cls} contain global spoofing cues for classification.

Negative Sample Pairs. We set the $z_{b,i}^{cls}$ as the anchor, where i indicates the ith samples in z_b^{cls}. To push away samples of different modal combinations and labels, we leverage samples under the sample-level masking strategy as negative pairs. The negative pair set is denoted as

$$b_i^- = \begin{cases} z_{b,n}^{cls}, & \text{if } Y_i \neq Y_n \\ z_{s,n}^{cls}, & \text{if } \Delta_i^b \neq \Delta_n^s \end{cases}. \tag{2}$$

Positive Sample Pairs. For the purpose of pulling close samples of the same modal combination and labels, we leverage samples under the batch-level masking strategy as positive pairs. To this end, we design the positive pair set as

$$b_i^+ = \{z_{b,p}^{cls}\}, \quad \text{if } Y_i = Y_p. \tag{3}$$

LBP-Guided Weighting. Different from other tasks, different attack types have different spoofing cues in FAS tasks [7], which means features of different attack types should be separated. Since attack types could be missing in some datasets, we introduce Local Binary Pattern (LBP) [1] features as the metric to distinguish different attack types. If two spoofing features have similar LBP features, these two features can be viewed as the same attack types. On the contrary, if two spoofing features have different LBP features, these two features belong to different attack types, which are separated. Therefore, to separate different attack types in positive pairs, we leverage LBP feature similarity to weigh the positive pairs. Given two spoofing features, an indicator determines whether they are the same attack types.

$$s(a, b) = \text{sim}(f_{LBP}(a), f_{LBP}(b)), \tag{4}$$

where $\text{sim}(\cdot, \cdot)$ denotes a similarity function and f_{LBP} is the LBP feature extractor. We use this indicator to weigh the positive pairs and formulate the contrastive loss [8].

$$L_{con} = \sum_{i=1}^{N} -\log \frac{\sum_{j=1}^{|b_i^+|} \exp(s_{i,j}^+/\tau) \cdot s_{i,j}^{LBP}}{\sum_{j=1}^{|b_i^+|} \exp(s_{i,j}^+/\tau) + \sum_{k=1}^{|b_i^-|} \exp(s_{i,k}^-/\tau)}, \tag{5}$$

where $s_{i,j}^+ = \text{sim}(z_{b,i}^{cls}, b_{i,j}^+)$, $s_{i,k}^- = \text{sim}(z_{b,i}^{cls}, b_{i,k}^-)$, and $s_{i,j}^{LBP} = s(z_{b,i}^{cls}, b_{i,j}^+)$. Note that $s_{i,j}^{LBP}$ is only computed for spoofing images with $Y_i = 0$.

Besides, we apply mutual information to show the effectiveness of the proposed loss.

Proposition 1 *The LBP-Guided Contrastive Loss promotes to learn modality-specific features by increasing the mutual information $I(X; X^+)$ but decreasing $I(X; X^-)$, where X, X^+ and X^- are random variables of anchor, positive samples and negative samples.*

The proof can be found in the supplementary materials. As a result, the overall loss can be written as

$$L = L_{cls} + \lambda \cdot L_{con}, \tag{6}$$

where λ is the trade-off parameter.

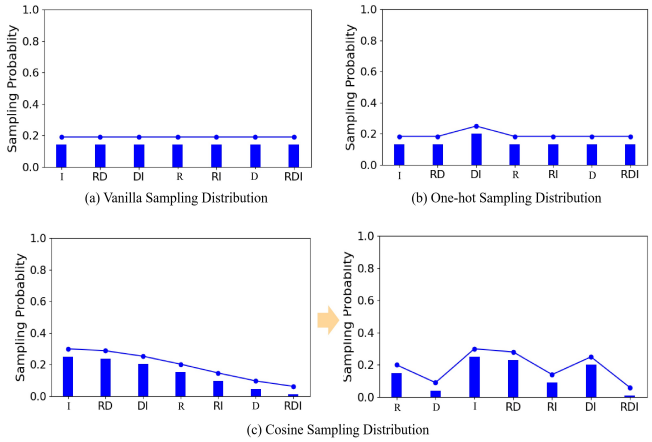


Fig. 4: Comparison of different sampling strategies. (a) The vanilla sampling strategy is uniformly distributed. (b) One-hot sampling aims to enhance one modal combination. (c) our AMS leverages the cosine function to dynamically adjust the probability.

Table 1: Comparisons with state-of-the-art approaches for intra-dataset protocol in WMCA 'seen' protocol and CASIA SURF. The bold font represents the best performance and the underlining font represents the second best performance.

Methods	Modalities							
	R	D	I	RD	RI	DI	RDI	mean±std
WMCA								
mmformer [33]	6.77	14.59	19.57	12.92	5.09	6.05	5.98	10.14±5.58
MMANet [27]	29.50	25.43	6.67	26.99	4.09	12.02	10.47	16.45±10.53
Vanilla ViT	12.84	14.08	12.30	12.68	13.67	13.54	7.68	12.39±2.17
FM-ViT [15]	**2.45**	**2.84**	3.10	2.58	2.83	2.86	**1.17**	2.54±0.64
MA-ViT [13]	5.64	3.75	2.15	2.46	2.94	2.57	1.51	3.00±1.35
AMA [29]	10.03	10.76	7.90	8.27	10.82	8.06	3.19	8.43±2.63
MAP [11]	11.14	7.43	5.24	8.79	3.87	5.03	4.43	6.56±2.66
MMA-FAS	4.42	3.63	**1.46**	**1.80**	**1.38**	**1.88**	1.25	**2.26±1.24**
CASIA-SURF								
mmformer [33]	22.82	5.03	31.88	5.10	25.40	7.14	7.31	14.95±11.34
MMANet [27]	24.06	3.45	24.90	**3.23**	19.56	4.38	4.00	11.94±10.33
Vanilla ViT	34.81	35.16	32.04	33.92	34.34	34.33	7.19	30.25±10.22
FM-ViT [15]	13.24	4.05	**3.54**	6.51	8.47	3.90	3.52	6.17±3.62
MA-ViT [13]	**12.67**	**2.66**	4.39	5.47	8.78	4.62	**2.12**	5.81±3.71
AMA [29]	28.57	26.58	28.29	27.15	28.33	25.89	11.55	25.19±6.09
MAP [11]	32.45	28.53	8.53	27.04	29.73	22.25	2.38	21.55±11.56
MMA-FAS	12.99	3.25	4.26	4.86	**6.58**	**3.87**	3.17	**5.56±3.47**

3.4 Adaptively Modal Combination Sampling

Random masking training in missing modality scenarios often results in imbalance training, i.e. the model converges faster in some modal combinations and slower in other modal combinations (weak combinations). To balance the training process of weak combinations, MMANet [27] applies an auxiliary classifier for the weak combination to encourage the model to focus more on the weak combination. However, as shown in Fig. 6 (b), by simply applying the auxiliary classifiers, MMANet only focuses on one weak modal combination, which may

cause performance degradation over other modal combinations. To balance the training process of all the modal combinations, we propose the AMS strategy to simultaneously balance the training process of all the modal combinations.

In batch-level and sample-level random masking strategies, the sampling probability distribution for every modal combinations is uniform, as shown in Fig. 4 (a). If we increase the probability of one modal combination, the model will face more samples of this modal combination, which is more focused on this modal combination, as shown in Fig. 4 (b). To balance all the modal combinations, we assign different sampling probabilities to each modal combination. To be specific, for each modal combination, MMANet [27] calculates its KL divergence of prediction distribution with the complete modalities as prediction discrepancy and we leverage these prediction discrepancy as imbalance score for all the modal combinations. We rank the imbalance scores from the largest to the smallest and calculate the cosine function of the imbalance scores as the sampling probability, as shown in Fig. 4(c). Finally, we reorder the the sampling probability as in Fig. 4 (d). It is worthy noting that we assign sampling probability for every modal combinations to encourage the model to synergistically focus on all the modal combinations.

4 Experiments

4.1 Dataset and Performance Metrics

Following [29], we experiment on WMCA [6], CASIA-SURF [32] and CASIA-SURF CeFA (CeFA) [14]. We use all three datasets to experiment in the missing modality scenario. As for evaluation metrics, we leverage Average Classification Error Rate (ACER) for intra-dataset experiments. The ACER on the testing set is determined by the Equal Error Rate (EER) threshold on dev sets for CASIA-SURF and CeFA, and the BPCER=1% threshold for WMCA [29].

4.2 Implementation Details

We adopt MultiViT-B [2] as our backbone with 3 tokenizers and 12 ViT blocks. The number and dimension of patch embeddings are $D = 768$ and $N_t = 589$. The kernel size of the filtering mask in MDA f is 2. The trade-off parameter λ is set to 0.1. We use Adam optimizer and the learning rate is set to 7e-4 with a cosine decline schedule. We train 100 epochs with batch size 32 and weight decay 5e-3. The code will be made available upon acceptance.

4.3 Experimental Results

To validate the ability to address missing modality problems in FAS, we evaluate MMA-FAS on various combinations of modalities under varying settings and compare with i) general methods directly employed to solve missing modality

Table 2: Comparisons of cross-dataset protocol. The models are trained on WMCA 'seen' protocol and tested on CASIA-SURF.

Methods	Modalities							
	R	D	I	RD	RI	DI	RDI	mean±std
mmformer [33]	47.93	49.92	49.49	49.79	45.54	49.82	48.76	48.75±1.58
MMANet [27]	55.48	40.27	53.46	62.06	56.14	50.01	62.26	54.24±7.57
Vanilla ViT	28.19	29.56	22.42	29.46	28.52	24.35	25.51	26.85±2.78
FM-ViT [15]	28.01	26.32	22.50	25.38	24.66	21.07	24.98	24.70±2.31
MA-ViT [13]	28.32	25.52	23.67	23.94	24.48	22.57	23.78	24.61±1.86
AMA [29]	26.97	**21.85**	22.92	22.37	23.89	26.57	**21.63**	23.74±2.19
MAP [11]	41.64	35.61	44.41	48.76	31.47	25.71	31.31	36.98±8.23
MMA-FAS	**24.23**	24.39	**22.20**	**17.56**	**21.45**	**20.97**	22.62	**21.91±2.59**

Table 3: Comparisons of cross attack protocol. We leverage CeFA (protocol4@3) and WMCA ('flexiblemask') in this protocol.

Methods	Modalities							
	R	D	I	RD	RI	DI	RDI	mean±std
Protocol4@3								
mmformer [33]	19.11	43.14	40.11	48.04	23.01	32.45	32.32	34.02±10.53
MMANet [27]	57.92	36.31	**5.18**	34.42	**10.24**	16.51	13.09	24.81±18.81
Vanilla ViT	26.84	25.29	26.18	26.96	25.76	26.23	13.59	24.40±4.80
FM-ViT [15]	30.53	34.20	28.53	29.23	26.47	20.19	18.43	26.79±5.64
MA-ViT [13]	26.88	35.51	25.14	26.99	24.55	22.16	19.81	25.86±4.96
AMA [29]	34.65	34.37	34.63	34.42	32.89	34.02	19.45	32.06±5.59
MAP [11]	52.34	42.86	40.73	37.66	52.30	34.83	23.47	40.59±10.12
MMA-FAS	**19.06**	**12.21**	14.34	**19.99**	20.91	**6.58**	**8.86**	**14.56±5.65**
WMCA Flexiblemask								
mmformer [33]	31.04	49.61	31.96	41.26	40.24	34.44	38.31	38.12±6.42
MMANet [27]	37.30	45.58	30.28	35.31	17.94	36.49	23.56	32.35±9.26
Vanilla ViT	21.03	53.51	42.52	16.73	19.06	32.49	21.18	29.50±13.95
FM-ViT [15]	22.34	46.86	38.59	20.28	18.47	31.11	25.78	29.06±10.46
MA-ViT [13]	22.68	45.14	36.61	21.15	16.48	35.52	21.17	28.39±10.62
AMA [29]	**20.73**	44.71	26.78	11.98	12.68	30.94	**9.15**	22.42±12.72
MAP [11]	28.40	60.56	49.77	35.29	33.72	40.14	12.09	37.13±15.47
MMA-FAS	21.10	**35.08**	**20.07**	**11.44**	**10.54**	**25.63**	9.65	**19.07±9.34**

problems in FAS (i.e., MMANet [27], mmformer [33], MAP (short for Missing-aware Prompts) [11]), and ii) recent methods in flexible modal FAS (i.e., FM-ViT [15] and MA-ViT [13]). We reproduce the results of FM-ViT [15] and MA-ViT [13] due to their lacking of codes. Besides, we directly apply AMA [29] in missing modality scenarios for comparison[4]. For simplicity, we abbreviate each modality combination in our experiments ({R, D, I, RD, RI, DI, RDI} are short for {RGB, Depth, IR, RGB+Depth, RGB+IR, Depth+IR, RGB+Depth+IR}) seperately.

Intra-dataset Protocol. Table 1 shows that MMA-FAS outperforms other methods on WMCA ('seen' protocol) and CASIA-SURF. This demonstrates the effectiveness of our missing modality adapters. Meanwhile, MMA-FAS has less variance than other methods, indicating our AMS mines the weak combinations

[4] AMA [29] is trained without modality masking but we train AMA with modality masking to mimic the missing modality. Besides, AMA are trained separately for each modal combination in [29] but we train AMA once for all modal combinations. Therefore, the results are different from those in [29]

Table 4: Ablation study on each component of the proposed MMA-FAS.

MultiViT	MDA	Contrastive	LBP-Guided	AMS	Mean ACER (%)
✓	✗	✗	✗	✗	12.39
✓	✓	✗	✗	✗	4.28
✓	✓	✓	✗	✗	3.59
✓	✓	✓	✓	✗	2.72
✓	✓	✓	✓	✓	2.26

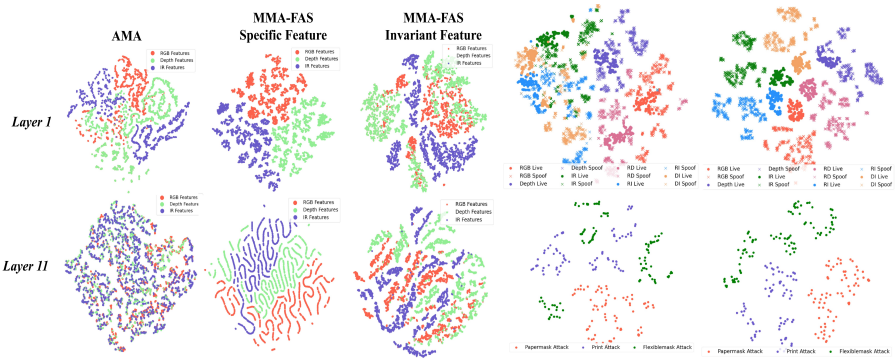


Fig. 5: TSNE visualization of (a) extracted features in the adapters from the low layer (layer 1) and high layer (layer 11). (b) Top two columns: extracted features in different modal combinations w/o and w contrastive loss. Bottom two columns: extracted features of different attack types w/o and w LBP-guided contrastive loss.

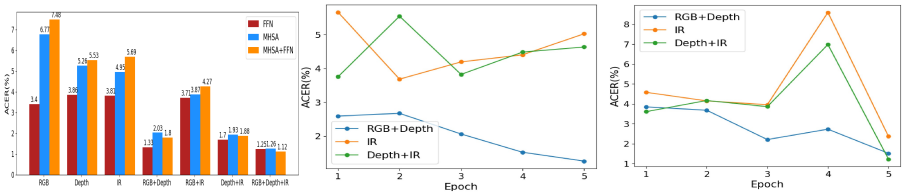
effectively. The great performance gap between MMA-FAS and other methods illustrates the essential role of modality-specific features. Besides, MMA-FAS achieves slightly higher performance with FM-ViT [15] and MA-ViT [13] while our MMA-FAS only contains 20× less training parameters than these methods (10.48M vs 271M).

Cross-dataset Protocol. To evaluate the cross-dataset generalization ability, we use WMCA for training and CASIA-SURF for testing. Table 2 shows that MMA-FAS achieves better and more stable performance, indicating the high-frequency components generalize well across datasets.

Cross-attack Protocol. We use 'protocol4@3' on CeFA and the 'flexiblemask' protocol on WMCA. The model is trained on fixed attack types and evaluated on unknown attack types. Table 3 shows that MMA-FAS also achieves the best performance in the more difficult settings.

4.4 Visualizations.

Fig. 2(b)(c)(d) visualize the frequency map of extracted features in AMA [29], MAP [11] and our MMA-FAS. It is obvious that AMA [29] and MAP [11] mainly focus on low-frequency and high-frequency regions, while our MMA-FAS takes both low-frequency and high-frequency information into account. Meanwhile, this visualization confirms that our motivation is correct.



(a) **(b)**

Fig. 6: (a) Ablation study for position of MMA-FAS. (b) The learning process of weak combinations without AMS (left) and with AMS (right).

Table 5: Dimension K for MDA.

K	Mean ACER (%)
16	3.03
32	3.75
64	2.72
78	4.45
128	3.18

Table 6: Length of frequency mask f for MDA.

f	Mean ACER (%)
2	2.26
3	3.98
4	5.84
5	8.27
6	5.67

Table 7: Results with various feature descriptors (FD).

FD	Mean ACER (%)
LBP [1]	2.26
HoG [10]	3.99
DoG [34]	3.18
SIFT [20]	2.82

Fig. 5(a) provides the TSNE visualizations on feature distribution in adapters in WMCA dataset. Different colors refer to features of different modalities. It is intuitive that the AMA captures modality-invariant features in both low and high layers, while MMA-FAS captures both modality-invariant and modality-specific features simultaneously. The modality-invariant features overlap in different modalities, while modality-specific features classify different modalities properly.

Fig. 5(b) visualize the distribution of the CLS token extracted by MMA-FAS in different modal combinations in WMCA dataset with and without LBP-guided supervised contrastive loss. It is obvious that only applying MDA without LBP-guided contrastive loss distinguishes different modal combinations preliminarily, which contain some modality-specific features. The distribution of live and spoofing samples is even not aligned. However, leveraging LBP-guided contrastive loss efficiently clusters samples according to modal combinations and categories, and live-spoofing sample separation is well aligned in different modal combinations. Meanwhile, Fig. 5(b) visualize the distribution of features of different attack types. Under the guidance of the LBP feature, the different attack types are successfully separated and the MultiViT extracts characteristic features specific to the attack types.

4.5 Ablation Study

We further validate the effectiveness of each component of MMA-FAS on WMCA. Table 4 shows MMA-FAS suffers worse ACER when removing any component.

Position of Adapters. As shown in Fig. 6(a), we yield the best performance by only applying adapters to FFN, since FFN contains more useful knowledge but MHSA is just used to calculate similarity.

Dimension K of Adapters. Table 5 shows that the best performance is obtained when the dimension K is 64. This is consistent with the result in [3].

Kernel Size f of Filtering in MDA. We evaluate different kernel sizes f of filtering masks in MDA. The smaller f is, the more high-frequency information is used for modality-specific feature extraction. Table 6 shows that when $f = 2$, the performance of MMA-FAS achieves the best. When the kernel size f increases, the performance of MMA-FAS decreases. This is because a larger kernel size results in less modality-specific feature extraction. On the contrary, small kernel size f ensures sufficient modality-specific features extracted by MDA.

Feature Descriptors in Contrastive Loss. We evaluate several common used feature descriptors (LBP [1], HoG [10], DoG [34] and SIFT [20]) in contrastive loss for comparison. As shown in Table 7, the LBP descriptors achieves the best performance, which is consistent with the results in [21].

Balanced Learning for Every Modal Combinations. To illustrate the effectiveness of AMS, we train the MMA-FAS model with sampling probability adjustment for 5 epochs. As shown in Fig. 6(b), given the weak combinations ['RGB+Depth', 'IR', 'Depth+IR'] with imbalanced scores from the largest to the smallest, MMANet [27] results in the performance increase of this modal combination, while performance decrease of other two modal combinations during training, which only increases the probability of 'RGB+Depth'. However, after adjusting sampling probabilities using our AMS, the performance of all three modal combinations increases, while the most imbalanced combination 'RGB+Depth' increases greatly. This demonstrates that our AMS encourages the model to balance all the modal combinations simultaneously.

5 Conclusion

In this paper, we propose a comprehensive framework delicately designed for FAS to tackle the missing modality problem. We propose modality-disentangle adapters, which extract and enhance modality-invariant and -specific features simultaneously from the view of frequency decomposition. We also combine batch-level and sample-level masking strategies to generate positive pairs and negative pairs for LBP-guided contrastive loss to further enhance modality-specific features. As for weak combinations, we propose a adaptively modal combination sampling strategy to dynamically adjust the probability of each combination. Extensive experiments demonstrate the effectiveness of MMA-FAS for missing modality scenarios.

Acknowledgment. This work was supported in part by the National Natural Science Foundation of China under Grant 62125109, Grant 61931023, Grant 61932022, Grant 62371288, Grant 62320106003, Grant 62301299, Grant T2122024, Grant 62120106007.

References

1. Ahonen, T., Hadid, A., Pietikäinen, M.: Face recognition with local binary patterns. In: Computer Vision-ECCV 2004: 8th European Conference on Computer Vision, Prague, Czech Republic, May 11-14, 2004. Proceedings, Part I 8. pp. 469–481. Springer (2004)

2. Bachmann, R., Mizrahi, D., Atanov, A., Zamir, A.: Multimae: Multi-modal multi-task masked autoencoders. In: European Conference on Computer Vision. pp. 348–367. Springer (2022)
3. Chen, S., Ge, C., Tong, Z., Wang, J., Song, Y., Wang, J., Luo, P.: Adaptformer: Adapting vision transformers for scalable visual recognition. Advances in Neural Information Processing Systems **35**, 16664–16678 (2022)
4. George, A., Marcel, S.: Cross modal focal loss for rgbd face anti-spoofing. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 7882–7891 (2021)
5. George, A., Marcel, S.: On the effectiveness of vision transformers for zero-shot face anti-spoofing. In: 2021 IEEE International Joint Conference on Biometrics (IJCB). pp. 1–8. IEEE (2021)
6. George, A., Mostaani, Z., Geissenbuhler, D., Nikisins, O., Anjos, A., Marcel, S.: Biometric face presentation attack detection with multi-channel convolutional neural network. IEEE transactions on information forensics and security **15**, 42–55 (2019)
7. Jia, Y., Zhang, J., Shan, S., Chen, X.: Single-side domain generalization for face anti-spoofing. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 8484–8493 (2020)
8. Khosla, P., Teterwak, P., Wang, C., Sarna, A., Tian, Y., Isola, P., Maschinot, A., Liu, C., Krishnan, D.: Supervised contrastive learning. Advances in neural information processing systems **33**, 18661–18673 (2020)
9. Kim, T., Kim, Y.: Suppressing spoof-irrelevant factors for domain-agnostic face anti-spoofing. IEEE Access **9**, 86966–86974 (2021)
10. Komulainen, J., Hadid, A., Pietikäinen, M.: Context based face anti-spoofing. In: 2013 IEEE Sixth International Conference on Biometrics: Theory, Applications and Systems (BTAS). pp. 1–8. IEEE (2013)
11. Lee, Y.L., Tsai, Y.H., Chiu, W.C., Lee, C.Y.: Multimodal prompting with missing modalities for visual recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 14943–14952 (2023)
12. Li, Z., Cui, Y., Wang, F., Liu, W., Yang, Y., Yu, Z., Jiang, B., Chen, H.: A multimodal face antispoofing method based on multifeature vision transformer and multirank fusion. Concurrency and Computation: Practice and Experience **35**(23), e7824 (2023)
13. Liu, A., Liang, Y.: Ma-vit: Modality-agnostic vision transformers for face anti-spoofing. arXiv preprint arXiv:2304.07549 (2023)
14. Liu, A., Tan, Z., Wan, J., Escalera, S., Guo, G., Li, S.Z.: Casia-surf cefa: A benchmark for multi-modal cross-ethnicity face anti-spoofing. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 1179–1187 (2021)
15. Liu, A., Tan, Z., Yu, Z., Zhao, C., Wan, J., Liang, Y., Lei, Z., Zhang, D., Li, S.Z., Guo, G.: Fm-vit: Flexible modal vision transformers for face anti-spoofing. arXiv preprint arXiv:2305.03277 (2023)
16. Liu, Y., Chen, Y., Dai, W., Gou, M., Huang, C.T., Xiong, H.: Source-free domain adaptation with contrastive domain alignment and self-supervised exploration for face anti-spoofing. In: European Conference on Computer Vision. pp. 511–528. Springer (2022)
17. Liu, Y., Chen, Y., Dai, W., Gou, M., Huang, C.T., Xiong, H.: Source-free domain adaptation with domain generalized pretraining for face anti-spoofing. IEEE Transactions on Pattern Analysis and Machine Intelligence (2024)

18. Liu, Y., Chen, Y., Dai, W., Li, C., Zou, J., Xiong, H.: Causal intervention for generalizable face anti-spoofing. In: 2022 IEEE International Conference on Multimedia and Expo (ICME). pp. 01–06. IEEE (2022)
19. Liu, Y., Chen, Y., Gou, M., Huang, C.T., Wang, Y., Dai, W., Xiong, H.: Towards unsupervised domain generalization for face anti-spoofing. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 20654–20664 (2023)
20. Lowe, D.G.: Object recognition from local scale-invariant features. In: Proceedings of the seventh IEEE international conference on computer vision. vol. 2, pp. 1150–1157. Ieee (1999)
21. Patel, K., Han, H., Jain, A.K.: Secure face unlock: Spoof detection on smartphones. IEEE Transactions on Information Forensics and Security **11**(10), 2268–2283 (2016)
22. Samar, A.R., Farooq, M.U., Tariq, T., Khan, B., Beg, M.O., Mumtaz, A.: Multimodal face anti-spoofing transformer (mfast). In: 2022 19th International Bhurban Conference on Applied Sciences and Technology (IBCAST). pp. 494–501. IEEE (2022)
23. Shao, R., Lan, X., Li, J., Yuen, P.C.: 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 (2019)
24. Wang, H., Chen, Y., Ma, C., Avery, J., Hull, L., Carneiro, G.: Multi-modal learning with missing modality via shared-specific feature modelling. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 15878–15887 (2023)
25. Wang, X., Zhang, K.Y., Yao, T., Zhou, Q., Ding, S., Dai, P., Ji, R.: Tf-fas: Twofold-element fine-grained semantic guidance for generalizable face anti-spoofing. In: European Conference on Computer Vision (ECCV). Springer (2024)
26. Wang, Z., Wang, Z., Yu, Z., Deng, W., Li, J., Gao, T., Wang, Z.: 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 (2022)
27. Wei, S., Luo, C., Luo, Y.: Mmanet: Margin-aware distillation and modality-aware regularization for incomplete multimodal learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 20039–20049 (2023)
28. Woo, S., Lee, S., Park, Y., Nugroho, M.A., Kim, C.: Towards good practices for missing modality robust action recognition. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 37, pp. 2776–2784 (2023)
29. Yu, Z., Cai, R., Cui, Y., Liu, X., Hu, Y., Kot, A.: Rethinking vision transformer and masked autoencoder in multimodal face anti-spoofing. arXiv preprint arXiv:2302.05744 (2023)
30. Yu, Z., Qin, Y., Li, X., Wang, Z., Zhao, C., Lei, Z., Zhao, G.: Multi-modal face anti-spoofing based on central difference networks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. pp. 650–651 (2020)
31. Zhang, K.Y., Yao, T., Zhang, J., Tai, Y., Ding, S., Li, J., Huang, F., Song, H., Ma, L.: Face anti-spoofing via disentangled representation learning. In: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XIX 16. pp. 641–657. Springer (2020)

32. Zhang, S., Liu, A., Wan, J., Liang, Y., Guo, G., Escalera, S., Escalante, H.J., Li, S.Z.: Casia-surf: A large-scale multi-modal benchmark for face anti-spoofing. IEEE Transactions on Biometrics, Behavior, and Identity Science **2**(2), 182–193 (2020)
33. Zhang, Y., He, N., Yang, J., Li, Y., Wei, D., Huang, Y., Zhang, Y., He, Z., Zheng, Y.: mmformer: Multimodal medical transformer for incomplete multimodal learning of brain tumor segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 107–117. Springer (2022)
34. Zhang, Z., Yan, J., Liu, S., Lei, Z., Yi, D., Li, S.Z.: A face antispoofing database with diverse attacks. In: 2012 5th IAPR international conference on Biometrics (ICB). pp. 26–31. IEEE (2012)
35. Zheng, G., Liu, Y., Dai, W., Li, C., Zou, J., Xiong, H.: Learning causal representations for generalizable face anti spoofing. In: ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 1–5. IEEE (2023)
36. Zheng, T.: Enfomax: Domain entropy and mutual information maximization for domain generalized face anti-spoofing (2023), `https://arxiv.org/abs/2302.08674`
37. Zheng, T., Li, B., Wu, S., Wan, B., Mu, G., Liu, S., Ding, S., Wang, J.: Mfae: Masked frequency autoencoders for domain generalization face anti-spoofing. IEEE Transactions on Information Forensics and Security (2024)
38. Zheng, T., Yu, Q., Chen, Z., Wang, J.: Famim: A novel frequency-domain augmentation masked image model framework for domain generalizable face anti-spoofing. In: ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 4470–4474. IEEE (2024)
39. Zhou, Q., Zhang, K.Y., Yao, T., Lu, X., Ding, S., Ma, L.: Test-time domain generalization for face anti-spoofing. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2024)
40. Zhou, Q., Zhang, K.Y., Yao, T., Lu, X., Yi, R., Ding, S., Ma, L.: Instance-aware domain generalization for face anti-spoofing. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 20453–20463 (2023)
41. Zhou, Q., Zhang, K.Y., Yao, T., Yi, R., Ding, S., Ma, L.: Adaptive mixture of experts learning for generalizable face anti-spoofing. In: Proceedings of the 30th ACM International Conference on Multimedia (ACM MM). pp. 6009–6018 (2022)
42. Zhou, Q., Zhang, K.Y., Yao, T., Yi, R., Sheng, K., Ding, S., Ma, L.: Generative domain adaptation for face anti-spoofing. In: European Conference on Computer Vision (ECCV). pp. 335–356. Springer (2022)