MA-ViT: Modality-Agnostic Vision Transformers for Face Anti-Spoofing

Ajian Liu[1,2] , **Yanyan Liang**[1*]

[1]School of Computer Science and Engineering, Faculty of Innovation Engineering, Macau University of Science and Technology, Macau
[2]CBSR&NLPR, Institute of Automation, Chinese Academy of Sciences, Beijing, China
ajianliu92@gmail.com, yyliang@must.edu.mo

Abstract

The existing multi-modal face anti-spoofing (FAS) frameworks are designed based on two strategies: halfway and late fusion. However, the former requires test modalities consistent with the training input, which seriously limits its deployment scenarios. And the latter is built on multiple branches to process different modalities independently, which limits their use in applications with low memory or fast execution requirements. In this work, we present a single branch based Transformer framework, namely Modality-Agnostic Vision Transformer (MA-ViT), which aims to improve the performance of arbitrary modal attacks with the help of multi-modal data. Specifically, MA-ViT adopts the early fusion to aggregate all the available training modalities' data and enables flexible testing of any given modal samples. Further, we develop the Modality-Agnostic Transformer Block (MATB) in MA-ViT, which consists of two stacked attentions named Modal-Disentangle Attention (MDA) and Cross-Modal Attention (CMA), to eliminate modality-related information for each modal sequences and supplement modality-agnostic liveness features from another modal sequences, respectively. Experiments demonstrate that the single model trained based on MA-ViT can not only flexibly evaluate different modal samples, but also outperforms existing single-modal frameworks by a large margin, and approaches the multi-modal frameworks introduced with smaller FLOPs and model parameters.

1 Introduction

Face Anti-Spoofing (FAS) aims to strengthen the face recognition system from a variety of presentation attacks (PAs) [Boulkenafet et al., 2017; Liu et al., 2018; Liu et al., 2016]. It has become an increasingly concern [Zhang et al., 2020a; Liu et al., 2020; Liu et al., 2021c; Liu et al., 2021e; Liu et al., 2021d] due to its wide applications in face payment, phone unlocking, and self-security inspection.
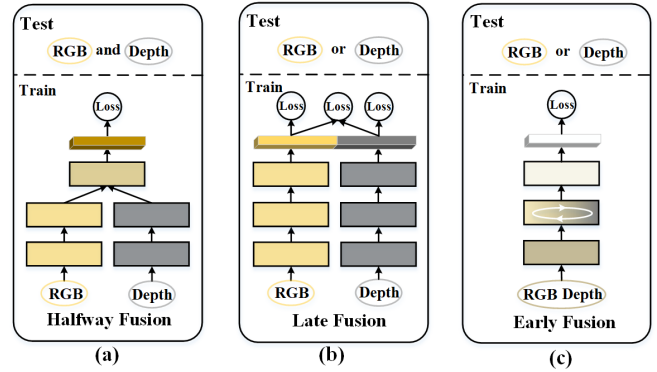
---

*Corresponding author

Figure 1: Comparison with existing multi-modal fusion strategies. (a) Halfway Fusion: it fuses multi-modal information by merging independent modal branches and requires the test modalities to be consistent with the training modalities. (b) Late Fusion: it fuses the multi-modal information at the decision level and requires complete modal branch for each input modality. (c) Early Fusion: it aggregates all available multi-modal data as input to realize flexible modal testing in a single branch structure.

With the increasingly advanced presentation attack instruments (PAIs) and acquisition sensors, face presentation attack detection (PAD) algorithms are also expanded from RGB spectrum [Liu et al., 2018; Shao et al., 2019; George and Marcel, 2019; Liu et al., 2019; Yu et al., 2020] to multi-spectrum to explore more reliable spoofing traces, which can be divided into halfway fusion framework [Zhang et al., 2020b; Parkin and Grinchuk, 2019; George et al., 2019] and late fusion framework [George and Marcel, 2021]. As shown in Fig. 1 (a), the halfway fusion is one of the commonly used fusion strategy, which combines the sub-networks of different modalities at a later stage via the feature concatenation. Although the halfway fusion can improve the robustness of PAD systems with the help of multi-modal information, their design concept is not intelligent enough, which requires the consistent training and testing modalities. If any modality disappears during testing, these methods would fail to distinguish live *vs.*fake faces and result in poor performance. In order to freely test any modal samples, as shown in Fig. 1 (b), the late fusion strategy retains a specific branch for each modality to capture different modal information independently, and fuses

the multi-modal information at the decision level (prediction embeddings or scores) to retain the function of any modal sample can be tested. However, these models may be slow to execute or large to store, limiting their use in applications or platforms with low memory or fast execution requirements, *e.g.*, mobile phones. To sum up, it is a significant work to design a framework can combine relevant information from the different modalities without introducing additional parameters (or a small number of acceptable parameters) to improve the performance over using only one modality.

Whether it is halfway or late fusion, the common is that different modal inputs need to be mapped to a higher semantic layer in advance for knowledge fusion. Inspired by the success of Transformer architecture in natural language tasks, several recent works have also applied transformer-based fusion between image-text, image-audio, and image-text-audio with self-attention. In this work, as shown in Fig. 1 (c), we adopt early fusion strategy by projecting all inputs from different modalities into a joint embedding space at input level and capture intra- and inter-modality interactions homogeneously within a pure transformer architecture. Compared with ConvNet-based framework, Transformer can not only capture local spoofing traces and establish their dependence on the patch tokens for a long time, but also can fuse multimodal information directly without aligning different modal features in advance. Therefore, abandoning the CNNs to mine spoofing traces, we adopt a single branch transformer framework to aggregate all the available training modalities' data and introduce the Modality-Agnostic Transformer Block (MATB) which consists of two cascaded attentions named Modal-Disentangle Attention (MDA) and Cross-Modal Attention (CMA) to improve the performance of any single-modal attacks at the inner- and inter modal levels, respectively. Specifically, the former eliminates modal-related information in the classification token under the guidance of a introduced modality token, and the latter supplements the modality-agnostic liveness information from other modalities. To sum up, the main contributions of this paper are summarized as follows:

- We present a novel Modality-Agnostic Vision Transformer (MA-ViT) framework to improve the performance of any single modal FAS system with the help of available multi-modal training data. Compared with the previous multi-modal fusion frameworks, our approach has the advantages of flexible testing of different modal samples and less introduction of additional model parameters.

- We develop the Modality-Agnostic Transformer Block (MATB) in MA-ViT with two effective attentions, namely Modal-Disentangle Attention (MDA) and Cross-Modal Attention (CMA), to achieve the modality-agnostic and complementary liveness features learning, respectively.

- Extensive experiments are conducted and demonstrate that the proposed MA-ViT can improve the performance of a single-modal system with the help of multi-modal data, with only acceptable FLOPs and model parameters are increased.

2 Modality-Agnostic ViT (MA-ViT)

2.1 Overall Framework

As shown in Fig. 2, our MA-ViT is built on the standard ViT [Dosovitskiy *et al.*, 2020] which consists of tokenization module, transformer encoder, and classification head. In addition, a learnable modality token is introduced to summarizes the modal information from patch tokens. A proposed MATB is inserted after Standard Transformer Block (STB) to disentangle the modal information and supplement the liveness features for the classification token, respectively. Let $\mathbf{I}^i$ represents the input in a training batch, where i can be r and d for the RGB and Depth samples, and also can be easily extended to other modalities.

Multi-Modal Tokenization Module. MA-ViT first splits any modal input $\mathbf{I}$ into a sequence of non-overlapping patches by a patch splitting module and then linearly projecting patches into tokens $\mathbf{x}_{pat} \in \mathbb{R}^{n \times D}$, where n is the number of resulting patches, and D is vector size through all of MA-ViT layers. Similar to BERT and DeiT [Touvron *et al.*, 2021], a learnable class token (CLS) $\mathbf{x}_{cls} = \mathbf{z}_{0,cls} \in \mathbb{R}^{1 \times D}$ and a modality token (MOD) $\mathbf{x}_{mod} = \mathbf{z}_{0,mod} \in \mathbb{R}^{1 \times D}$ are concatenated to the patch tokens, who serves as the image and modality agents for classification. Due to different modal images are collected by fixed position multi-spectral sensor, they share location information and exclusive spectrum space. Therefore, position embeddings $\mathbf{x}_{pos} \in \mathbb{R}^{(n+2) \times D}$ and spectrum embeddings $\mathbf{x}_{spe} \in \mathbb{R}^{(n+2) \times D}$ are added to each token to retain positional and modal informations. The tokenization process is expressed as follows:

$$\mathbf{z}_0^i = [\mathbf{x}_{cls}||\mathbf{x}_{mod}^i||\mathbf{x}_{pat}^i] + (\mathbf{x}_{pos} + \mathbf{x}_{spe}^i), \mathbf{z}_0^i \in \mathbb{R}^{N \times D} \quad (1)$$

where $||$ means token concatenation. $i \in \{r, d\}$ and $N = n+2$ represent the modal type and the total number of tokens. respectively. The resulting sequence $\mathbf{z}_0$ serves as input to the following transformer encoder.

Transformer Encoder with MATBs. After tokenization module, the sequences $\mathbf{z}_0^i$ ($i \in \{r, d\}$) from all modalities are mixed through STB and MATB, which constitute one transformer block of our MA-ViT. The complete transformer encoder consists of K transformer blocks by repeatedly stacking. As illustrated in Fig. 2, a MATB with MDA receives the output of the previous STB as input, and with CMA receives the current intra-modal sequences ($\mathbf{z}^i$) and the previous cross-modal sequence ($\mathbf{z}^{rd}$) as input. Similar to Multi-headed Self-Attention (MSA) in STB, a Layer normalization (LN) is applied before each module, and a residual shortcut is applied after each module. Specifically, the processing of k-th (where $k = 1, ..., K$) Transformer block can be expressed as (omit MLP operation):

$$\mathbf{z'}_k^i = \text{MSA}(\text{LN}(\mathbf{z}_{k-1}^i)) + \mathbf{z}_{k-1}^i$$
$$\mathbf{z}_k^{rd} = \text{MCMA}(\text{LN}(\mathbf{z}_{k-1}^r), \text{LN}(\mathbf{z}_{k-1}^d)) + \mathbf{z}_{k-1}^{rd} \quad (2)$$
$$\mathbf{z}_k^i = \text{MMDA}(\text{LN}(\mathbf{z'}_k^i)) + \mathbf{z'}_k^i$$

where MSA and MCMA (Multi-headed CMA) and the following MLP weights are shared. $\mathbf{z}_k^i$ and $\mathbf{z}_k^{rd}$ (cross-modal sequence) are the output of MATB (MDA) and MATB (CMA).
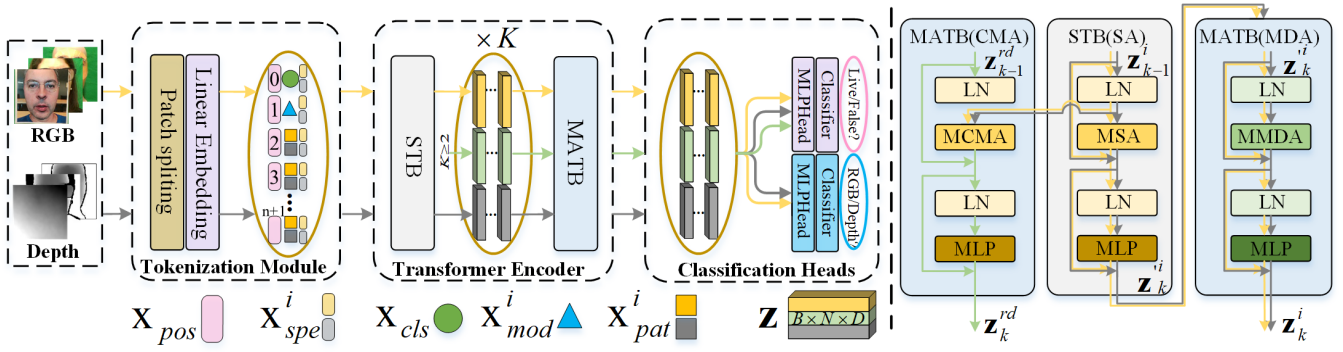
Figure 2: The architecture of Modality-Agnostic Vision Transformer (MA-ViT). It is built on a single ViT branch and consists of Multi-Modal Tokenization Module (Eq. 1), Transformer Encoder, and Modal Shared Classification Heads. (Eq. 3). A completed transformer encoder contains K transformer block, which consists of Standard Transformer Block (STB) and Modality-Agnostic Transformer Block (MATB). Specifically, the STB processes all modal sequences (yellow and black gray arrows) independently and outputs them into MATB (MDA). And the MATB (CMA) takes all modal sequences as inputs and outputs the cross-modal sequences (green arrow).

Modal Shared Classification Heads. See from Fig. 2, we provide two modal shared classification heads for all sequences to meet the requirements of any modal sample can be tested, which are supervised and jointly optimized by binary cross-entropy (BCE). The output sequences at the K-th block in Transformer encoder $\mathbf{z}_{K,cls}$ and $\mathbf{z}_{K,mod}$ are served as the agents for liveness and modality classification, respectively. The total loss L^{total} to minimize is given as:

$$L_{cls}^i = \text{BCE}(\text{MLP}^{cls}(\text{LN}(\mathbf{z}_{K,cls}^i)), y_{cls}),$$
$$L_{mod}^i = \text{BCE}(\text{MLP}^{mod}(\text{LN}(\mathbf{z}_{K,mod}^i)), y_{mod}), \quad (3)$$
$$L_{total}^i = L_{cls}^i + L_{cls}^{rd} + L_{mod}^i, i \in \{r, d\}$$

where L_{cls}^{rd} is the loss for cross-modal sequence $\mathbf{z}_K^{rd}$. The classification head is implemented by MLP with a single linear layer. y is the ground truth for sample $\mathbf{I}$. In detail, $y_{cls} = 0/1$ for attack/bonafide and $y_{mod} = 0/1$ for RGB/Depth, respectively.

2.2 Modality-Agnostic Transformer Block.

Multi-headed Modal-Disentangle Attention (MMDA). How to determine the modality-irrelated patch tokens is the primary task in MMDA. By analyzing the attention matrix, which is essentially a relevance maps, whose each row corresponds to a link for each token given the other tokens. In which the relevance map that corresponds to the CLS/MOD token links each of the tokens to the CLS/MOD token, and the strength of this link can be intuitively considered as an indicator of the contribution of each token to the classification.

See from the MATB (MDA) in Fig. 2, for any modal sequence $\mathbf{z}^i$, as shown in Eq. 4 and Fig. 3, we first compute $\mathbf{q}_{mod}$ of MOD token and $\mathbf{k}_{pat}^{mod}$ of patch tokens through two learnable parameters $\mathbf{W}_q^{mod}$ and $\mathbf{W}_k^{mod}$, respectively. Then, we compute the dot products of the $\mathbf{q}_{mod}$ with $\mathbf{k}_{pat}^{mod}$, divide each by $\sqrt{D/h}$, and apply a softmax function to obtain the modal relevance map $\mathbf{map}_{mod}$. Then, we identify the modality-irrelated (or weakly related) patch tokens by excluding strong links in $\mathbf{map}_{mod}$, which is completed by a threshold function $\Gamma_\lambda(\cdot)$. Specifically, $\Gamma_\lambda(\cdot)$ aims to find the

modality-related patch tokens by thresholding the modal relevance map to keep $\lambda(\in [0,1])$ proportional mass. It outputs a mask matrix $\mathbf{M}$ with values of 1 and 0, where 1/0 means to modality-related/-irrelated the patch tokens of the corresponding position.

In order to disentangle the modal information for classification, we take the mask matrix $\mathbf{M}$ as indicators to discard the modality-related patch tokens. As shown in Eq. 4, we first compute $\mathbf{q}_{cls}$ of CLS token, $\mathbf{k}_{pat}^{cls}$ and $\mathbf{v}_{pat}^{cls}$ of patch tokens through three learnable parameters $\mathbf{W}_q^{cls}$, $\mathbf{W}_k^{cls}$ and $\mathbf{W}_v^{cls}$, respectively. Then, we obtain the attention map $\mathbf{map}_{cls}$ in a similar way by $\mathbf{q}_{cls}$ and $\mathbf{k}_{pat}^{cls}$, while disconnecting the links with the modality-related patch tokens under the guidance of $\mathbf{M}$. See from the Eq. 4, this process can be completed with a selection function $\Gamma'_\mathbf{M}(\cdot)$ based on the index position of a given matrix $\mathbf{M}^i$. Finally, we successively obtain the weights by a softmax function to $\mathbf{map}_{cls}$ and the output of MDA by the weighted sum over all values of patch tokens $\mathbf{v}_{pat}^{cls}$. This process is denoted as:

$$[\mathbf{q}_{cls}, \mathbf{k}_{pat}^{cls}, \mathbf{v}_{pat}^{cls}] = [\mathbf{z}_{cls}\mathbf{W}_q^{cls}, \mathbf{z}_{pat}\mathbf{W}_k^{cls}, \mathbf{z}_{pat}\mathbf{W}_v^{cls}],$$
$$[\mathbf{q}_{mod}, \mathbf{k}_{pat}^{mod}] = [\mathbf{z}_{mod}\mathbf{W}_q^{mod}, \mathbf{z}_{pat}\mathbf{W}_k^{mod}],$$
$$\mathbf{map}_{mod} = \mathbf{q}_{mod}(\mathbf{k}_{pat}^{mod})^T/\sqrt{D/h},$$
$$\mathbf{map}_{cls} = \mathbf{q}_{cls}(\mathbf{k}_{pat}^{cls})^T/\sqrt{D/h}, \quad (4)$$
$$\mathbf{M} = \Gamma_\lambda(\mathbf{map}_{mod}),$$
$$\text{MDA}(\mathbf{z}^i) = \text{softmax}[\Gamma'_\mathbf{M}(\mathbf{map}_{cls})] \cdot \mathbf{v}_{pat}^{cls}$$

where D and h are the embedding dimension and number of heads, respectively. $\mathbf{W}_q$, $\mathbf{W}_k$ and $\mathbf{W}_v \in \mathbb{R}^{D \times (D/h)}$. $\Gamma'_\mathbf{M}(\cdot)$ is a selection function defined as:

$$\Gamma'_\mathbf{M}(\mathbf{A}) = \begin{cases} -\infty, & \mathbf{M}_{a,b} > 0 \\ \mathbf{A}_{a,b}, & \mathbf{M}_{a,b} = 0 \end{cases} \quad (5)$$

Similar to [Wang *et al.*, 2021], we discard the modality-related tokens by setting the attention values in $\mathbf{map}_{cls}$ to small enough constant. MMDA is an extension of MDA in which we run h mutual-attention operations in parallel. As
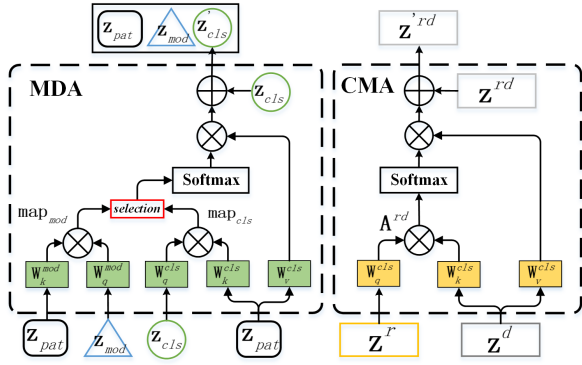
Figure 3: Implementation details of Modal-Disentangle Attention (MDA) and and Cross-Modal Attention (CMA).

shown in Fig. 2 and Eq. 2, we apply a residual shortcuts after the MDA and concatenate with patch tokens $\mathbf{z}_{pat}$ to obtain a new sequences $\mathbf{z}'$:

$$\mathbf{z}'^i_{cls} = \text{MMDA}(\mathbf{z}^i) + \mathbf{z}^i_{cls}, \mathbf{z}'^i = [\mathbf{z}'^i_{cls}||\mathbf{z}^i_{mod}||\mathbf{z}^i_{pat}] \quad (6)$$

where $||$ means token concatenation, the output sequences $\mathbf{z}'^i$ will be used as the input sequences of the next stage.

Multi-headed Cross-Modal Attention (MCMA). How to effectively produce a stronger modality-agnostic representation with the help of multi-modal data is the primary task in MCMA. Inspired by CDTrans [Xu *et al.*,], we summarize the common information of different modal sequences to construct cross-modal sequence output by MCMA, which is supervised by the same label. In other words, the MCMA further guides MSA to pay attention to modality-agnostic liveness features by sharing with its parameters.

See from the MATB (CMA) module in Fig. 2, it takes all modal sequences $\mathbf{z}^r$ and $\mathbf{z}^d$ as inputs and outputs cross-modal sequence $\mathbf{z}^{rd}$. Similar to MSA [Dosovitskiy *et al.*, 2020], as shown in Fig. 3, we first compute queries $\mathbf{q}^r$ of modal sequence $\mathbf{z}^r$, and keys $\mathbf{k}^d$, values $\mathbf{v}^d$ from another modal sequence $\mathbf{z}^d$, respectively. Then, the attention function $\mathbf{A}^{rd}$ is computed on the set of queries $\mathbf{q}^r$ simultaneously with all keys $\mathbf{k}^d$. Finally, the outputs of CMA is a weighted sum over all values $\mathbf{v}^d$, denoted as:

$$[\mathbf{q}^r, \mathbf{k}^d, \mathbf{v}^d] = [\mathbf{z}^r\mathbf{W}^{cls}_q, \mathbf{z}^d\mathbf{W}^{cls}_k, \mathbf{z}^d\mathbf{W}^{cls}_v],$$
$$\mathbf{A}^{rd} = \text{softmax}(\mathbf{q}^r(\mathbf{k}^d)^T/\sqrt{D/h}), \quad (7)$$
$$\text{CMA}(\mathbf{z}^i) = [\mathbf{A}^{rd} \cdot \mathbf{v}^d||\mathbf{A}^{dr} \cdot \mathbf{v}^r], i \in \{r, d\}$$

where the calculation process of $\mathbf{A}^{dr}$ and $\mathbf{v}^r$ is similar to that of $\mathbf{A}^{rd}$ and $\mathbf{v}^d$ by simply swapping the index r and d. $\mathbf{W}_q$, $\mathbf{W}_k$ and $\mathbf{W}_v$ are shared with MSA. $||$ means concatenating cross-modal sequences along the batch dimension. Similar to MDA in Eq. 6, we also apply a residual shortcuts after the CMA, and obtain the new cross-modal sequence $\mathbf{z}'^{rd} = \text{CMA}(\mathbf{z}^i) + \mathbf{z}^{rd}$ (The operation is not performed in the 1-th layer).

3 Experiments

3.1 Experimental Setup

Datasets & Protocols. We use three commonly used multi-modal and a single-modal FAS datasets for experiments, including CASIA-SURF (MmFA) [Zhang *et al.*, 2019], CASIA-SURF CeFA (CeFA) [Liu *et al.*, 2021b], WMCA [George *et al.*, 2019] and OULU-NUPU [Boulkenafet *et al.*, 2017] (OULU). MmFA consists of $1,000$ subjects with $21,000$ videos and each sample has 3 modalities, and provides a intra-testing protocol. CeFA covers 3 ethnicities, 3 modalities, $1,607$ subjects, and provides five protocols to measure the affect under varied conditions. We select the Protocol 1, 2, and 4 for experiments. WMCA contains a wide variety of 2D and 3D presentation attacks, which introduces 2 protocols: grandtest protocol which emulates the "seen" attack scenario and the "unseen" attack protocol that evaluates the generalization on an unseen attack.

Test Scenario Settings. We consider two test scenarios. The first is the commonly used setting that the test modalities need to be consistent with the training stage. In this scenario, we evaluate the intra-testing performances of each dataset based on the provided protocols, and evaluate the robustness of our approach through cross-testing experiments between these datasets. The second is the flexible modal test scenario, which means the tester can provide any single-modal or any multi-modal paired samples.

Evaluation Metrics. In intra-testing experiments, Attack Presentation Classification Error Rate (APCER), Bonafide Presentation Classification Error Rate (BPCER), and ACER are used for the metrics. The ACER on testing set is determined by the Equal Error Rate (EER) threshold on dev sets for MmFA, CeFA, OULU, and the BPCER=1% threshold for WMCA. TPR(@FPR=10^{-4}) is also provided for MmFA. For cross-testing experiments, Half Total Error Rate (HTER) is adopted as the metric, which computes the average of False Rejection Rate (FRR) and the False Acceptance Rate (FAR), and the threshold computed in dev set using EER criteria.

Implementation Details. Our models can be freely built on any version of ViT [Dosovitskiy *et al.*, 2020], which means that our MATB can be compatible with other improvements to ViT, such as DeiT [Touvron *et al.*, 2021] and Swin [Liu *et al.*, 2021f]. In our experiments, we adopt ViT-S/16 as the backbone through comparative experiments, which means the "Small" variant with $K = 12$.

We train our model with the ViT-S/16 backbone, which is initialized with weights provided by [Touvron *et al.*, 2021], and other newly added layers are randomly initialized. We resize all modal images to 224×224 and train all models with 50 epochs via Adam solver. All models are trained with a batch seize of 8 and an initial learning rate of 0.0001 for all epochs. We set $\lambda = 0.8$ in MDA according to comparative experiments. In the testing stage, for single-mdoal samples, we only activate the input interface of the corresponding modality (mainly the corresponding position and spectrum embeddings). As can be seen more clearly from Fig 2, the green line only outputs the results of multi-modal samples, while the yellow and gray lines output the results of corresponding single-modal samples respectively.

Method	seen	unseen							
		Flexiblemask	Replay	Fakehead	Prints	Glasses	Papermask	Rigidmask	mean±std
MC-PixBiS	1.80	49.70	3.70	0.70	0.10	16.00	0.20	3.40	10.50±16.70
MCCNN-OCCL-GMM	3.30	22.80	31.40	1.90	30.00	50.00	4.80	18.30	22.74±15.30
CMFL	1.70	12.40	1.00	2.50	0.70	33.50	1.80	1.70	7.60±11.20
ViT	3.79	14.56	1.58	4.14	0.83	36.00	2.82	2.19	7.93±13.80
MA-ViT	**1.45**	**9.76**	**0.93**	**0.55**	**0.00**	**14.00**	**0.00**	**1.46**	**3.81±5.67**

Table 1: Comparison of ACER (%) values on Protocol "seen" and "unseen" for the WMCA. Best results are bolded.

Method	APCER	BPCER	ACER	TPR(@FPR=10^{-4})
SEF	3.80	1.00	2.40	56.80
MS-SEF	2.30	**0.30**	1.30	81.40
ViT	3.52	9.50	6.51	44.33
MA-ViT	**0.78**	0.83	**0.80**	**82.83**

Table 2: The results on MmFA. A large TPR(%) and a lower ACER (%) indicate better performance. Best results are bolded.

Pro.	Method	APCER(%)	BPCER(%)	ACER(%)
1	PSMM	2.40±0.60	4.60±2.30	3.50±1.30
	ViT	4.55±7.29	1.83±1.59	3.19±4.07
	MA-ViT	**1.45±1.75**	**0.75±0.43**	**1.10±1.09**
2	PSMM	7.70±9.00	3.10±1.60	5.40±5.30
	ViT	3.48±1.81	0.92±0.12	2.20±0.85
	MA-ViT	**0.12±0.08**	**0.09±0.12**	**0.10±0.01**
4	PSMM	7.80±2.90	5.50±3.00	6.70±2.20
	Hulking	3.25±1.98	1.16±1.12	2.21±1.26
	Super	**0.62±0.43**	2.75±1.50	1.68±0.54
	BOBO	1.05±0.62	**1.00±0.66**	**1.02±0.59**
	ViT	20.44±17.12	7.42±3.96	13.92±10.45
	MA-ViT	2.10±1.47	1.17±0.38	1.64±0.89

Table 3: Evaluation results (%) on the Protocol 1, 2, and 4 of CeFA dataset. Note that a lower ACER value indicates better performance.

Method	Train(MmFA),Test(WMCA)	Train(WMCA),Test(MmFA)
Aux.(Depth)	24.54	12.35
ViT	31.87	22.79
CMA-ViT	27.97	16.97
MDA-ViT	24.84	13.30
MA-ViT	**20.63**	**10.41**

Table 4: The HTER (%) values from the cross-testing between MmFA and WMCA (Protocol "seen") datasets.

In order to perform a fair comparison with prior methods, only RGB and Depth data in WMCA [George *et al.*, 2019] are used for intra-testing experiments. Tab. 1 presents the comparisons of ACER to the SOTA ConvNet-based methods, including MC-PixBiS [George and Marcel, 2019], MCCNN-OCCL-GMM [George and Marcel, 2020], MC-ResNetDLAS [Parkin and Grinchuk, 2019] and CMFL [George and Marcel, 2021]. Compared with the previous best results, our MA-ViT achieves significantly better performance with a large margin in "seen" and "unseen" protocols: -0.25% over CMFL (1.70%), and -3.79% over CMFL (7.60%), respectively. It is worth noting that MA-ViT noticeably surpass these methods on two challenging sub-protocols of "unseen" protocol: when "Flexiblemask" and "Glasses" are not seen in the training stage. We analyze that the commonality of the two attacks is 3D facial structure, realistic color-texture, and only local regions contain spoofing traces, which are easy to be ignored by the ConvNet-based methods and play a long-term indicative role in ViT.

Cross-Testing Results. To evaluate the robustness, we conduct cross-testing experiments between models trained on MmFA and WMCA with Protocol "seen". We also introduce the previous SOTA ConvNet-based method Aux.(Depth) [Liu *et al.*, 2018] as a baseline. Tab. 4 lists the HTER of all methods trained on one dataset and tested on another dataset. From these results, MA-ViT outperforms Aux.(Depth) whether tested on WMCA (24.54% vs. 20.63%) or MmFA (12.35% vs. 10.41%). When comparing the results of MDA-ViT and CMA-ViT, we can conclude that MDA plays a more important role in improving the robustness. See Tab. 4 for details, when replacing the element of ViT from CMA to MDA, the results are further reduced from 27.97% to 24.84% when tested on WMCA and from 16.97% to 13.30% when tested on MmFA, respectively. Due to the mismatch of sensors, resolutions, attack types, and settings between different datasets, it is more effective to mine potential features by removing the interferences of modal information.

3.2 Fixed Modal Scenario Evaluations

On the three multi-modal datasets, we compare with the baseline method ViT that removed the MATB from MA-ViT, and the previous state-of-the-art (SOTA) methods.

Intra-Testing Results. For MmFA, we compare with the benchmark SEF [Zhang *et al.*, 2019] and the multi-scale version MS-SEF [Zhang *et al.*, 2020b], respectively. From the Tab. 2, we can observe that the performance of ViT is worse than SEF due to the lack of modal fusion ability. However, when equipped with MATB, MA-ViT improves the TPR from 44.33% to 82.83%, reduces ACER from 6.51% to 0.80%, and outperforms MS-SEF by a large margin, which uses ImageNet pretrain.

For CeFA, we compare our approach with the benchmark PSMM [Liu *et al.*, 2021b] and the top three methods introduced in the challenge [Liu *et al.*, 2021a] with the Protocol 4, *i.e.*, BOBO, Super and Hulking. It can be seen from Tab. 2 that the performance of ViT is superior to PSMM on Protocol 1 and Protocol 2 due to the pre-training model. Our MA-ViT further reduces the ACER on Protocol 1 and Protocol 2 to 1.10% and 0.10%, and outperforms the Hulking and Super methods with ACER gains of -0.04% (1.68% vs. 1.64%) and -0.57% (2.21% vs. 1.64%), respectively.

Method	Train	Test	MmFA			WMCA		
			APCER	BPCER	ACER	APCER	BPCER	ACER
Fixed modal testing								
SOTA	R	R	40.30	1.60	21.00	65.65	0.00	32.82
	D	D	6.00	1.20	3.60	11.77	0.31	6.04
	I	I	38.60	0.40	19.40	5.03	0.00	2.51
ViT	R	R	22.67	30.67	26.67	2.49	12.17	7.33
	D	D	3.02	7.50	5.26	4.75	2.61	3.68
	I	I	3.52	9.50	6.51	4.98	4.35	4.66
Flexible modal testing								
MA-ViT	R&D&I	R	8.88	14.00	11.43	7.69	3.48	5.59
	R&D&I	D	0.78	0.83	0.80	6.56	0.00	3.28
	R&D&I	I	0.67	2.33	1.50	3.39	0.87	2.13

Table 5: Comparison of flexible modal results (%) based on multi-modal datasets. The 'SOTA' means the method with public results on the corresponding dataset. R&D&I indicates the method receives RGB (R), Depth (D) and IR (I) paired samples as input.

3.3 Flexible Modal Scenario Evaluations

In order to explore the ability of our approach for modality-agnostic features learning, we conducted experiments on dataset MmFA, and Protocol "seen" of WMCA. Before reporting the results in flexible modal scenarios, we first list the results of the SOTA methods on RGB (R), Depth (D) and IR (I) modalities on each dataset, *i.e.*, MS-SEF for MmFA, and MC-CNN [George *et al.*, 2019] for WMCA.

By comparing the results in Tab. 5, we can draw the following two conclusions: (1) Our MA-ViT achieves the best results in each modality by training only one model. Such as compared with MS-SEF on MmFA, MA-ViT reduces the ACER of RGB, Depth and IR results by 9.57% (11.43% vs. 21.00%), 2.80% (0.80% vs. 3.60%) and 17.90% (1.50% vs. 19.40%) points where they are trained in an independent model. Experimental results show that our model can flexibly test samples with any modalities without retraining the model again. (2) Our MATB can improve the any single modal performance with the help of multi-modal data. This clearly demonstrates that the proposed MATB enables any modal sequences to mine modality-agnostic liveness features for flexible modal testing. There is a similar conclusion on WMCA when compared with MC-CNN.

3.4 Ablation Study

Effect of the Frameworks. In this experiment, we compare MA-ViT(S) with a CNN method Aux.(Depth) [Liu *et al.*, 2018], with a variety of ViT family architectures, *i.e.*, baseline ViT(S), "tiny" (T) and "base" (B) variants. In addition to the metric of ACER, FLOPs and parameters measured for all the models and shown as Tab. 6. Compared with ViT(S), our approach only introduces 0.35 (G) FLOPs and 1.06 (M) parameters, but has significant performance improvement on three modalities. When compared with MA-ViT(T) and MA-ViT(B), we find that the "small" variant is more suitable for face anti-spoofing tasks, which may be related to the dataset size and task category. Finally, it can be seen that MDA-ViT and CMA-ViT outperforms ViT(s) by a big margin, which improve the any single modal performance by eliminating modality-related information and fusing liveness features from another modal sequences.

Effect of the λ in MDA. In MDA, the λ determines the number of modality-irrelated patch tokens. The experimental results are shown in Tab. 6. When $\lambda = 0.8$ in MDA-ViT, all

Model	MmFA			WMCA			FLOPs (G)	Params (M)
	R	D	I	R	D	I		
ViT(S)	26.67	5.26	6.51	7.33	3.68	4.66	1.61	2.30
MDA-ViT($\lambda=0.7$)	14.08	3.17	1.75	6.95	5.49	6.60	1.92	3.36
MDA-ViT($\lambda=0.8$)	13.78	2.01	1.75	6.25	2.61	4.43	1.92	3.36
MDA-ViT($\lambda=0.9$)	16.74	2.06	2.22	6.60	5.49	4.76	1.92	3.36
CMA-ViT	**9.71**	3.25	1.53	8.97	**2.02**	2.13	1.64	2.30
MA-ViT(S)	11.43	**0.80**	**1.50**	**5.59**	3.28	**2.13**	1.96	3.36
MA-ViT(T)	12.89	1.67	3.64	5.77	4.38	2.92	0.49	0.86
MA-ViT(B)	18.00	1.61	5.31	5.64	3.94	2.70	7.79	13.26

Table 6: Ablation study with different architecture on MmFA and WMCA with Protocol "seen". ACER values are reported.
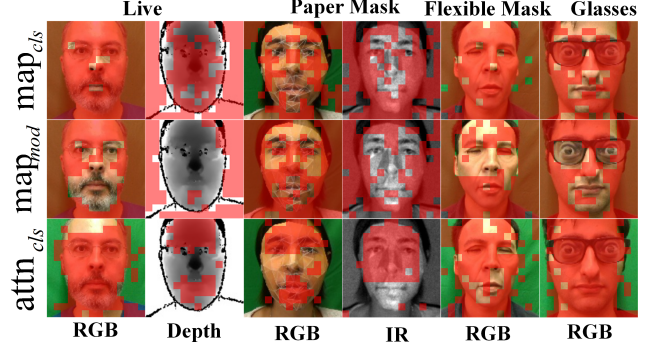

Figure 4: The mining process of informative patch tokens in each modal sequence which are covered by the red mask and obtained by thresholding the maps to keep 80% ($\lambda = 0.8$) of the mass. The sample is from the WMCA dataset, where the sample category is marked at the top.

experiments achieve the best results.

In Fig. 4, we visualize the patch tokens mined by the MDA in the last layer. For each modal sequence (each column), we visualize the areas of modality-irrelated patch tokens (the first row), classification-irrelated patch tokens (the second row), and the final informative patch tokens (the third row). We can see that without MDA, the informative patch tokens are distributed in global regions, including modality-irrelated region and background region. While, the distribution is adjusted after MDA, which refers to the patch tokens in facial region.

4 Conclusion

In this work, we present a pure transformer-based framework named MA-ViT, which aims to guide any modal sequence to learn modality-agnostic liveness features, and fuse the liveness complementary information from other modalities through the developed MMDA and MCMA respectively. Experiments show that our approach MA-ViT only introduces 0.35 (G) FLOPs and 1.06 (M) parameters, but gains significant improvements compared with baseline method ViT(S).

Acknowledgements

References

[Boulkenafet *et al.*, 2017] Zinelabinde Boulkenafet, Jukka Komulainen, Lei Li, Xiaoyi Feng, and Abdenour Hadid. Oulu-npu: A mobile face presentation attack database with real-world variations. In *FGR*, pages 612–618, 2017.

[Dosovitskiy *et al.*, 2020] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[George and Marcel, 2019] Anjith George and Sébastien Marcel. Deep pixel-wise binary supervision for face presentation attack detection. In *ICB*, pages 1–8, 2019.

[George and Marcel, 2020] Anjith George and Sébastien Marcel. Learning one class representations for face presentation attack detection using multi-channel convolutional neural networks. *IEEE TIFS*, 16:361–375, 2020.

[George and Marcel, 2021] Anjith George and Sébastien Marcel. Cross modal focal loss for rgbd face anti-spoofing. In *CVPR*, pages 7882–7891, 2021.

[George *et al.*, 2019] Anjith George, Zohreh Mostaani, David Geissenbuhler, Olegs Nikisins, André Anjos, and Sébastien Marcel. Biometric face presentation attack detection with multi-channel convolutional neural network. *IEEE TIFS*, 15:42–55, 2019.

[Liu *et al.*, 2016] Siqi Liu, Baoyao Yang, Pong C Yuen, and Guoying Zhao. A 3d mask face anti-spoofing database with real world variations. In *CVPRW*, pages 100–106, 2016.

[Liu *et al.*, 2018] Yaojie Liu, Amin Jourabloo, and Xiaoming Liu. Learning deep models for face anti-spoofing: Binary or auxiliary supervision. In *CVPR*, pages 389–398, 2018.

[Liu *et al.*, 2019] Ajian Liu, Jun Wan, Sergio Escalera, Hugo Jair Escalante, Zichang Tan, Qi Yuan, Kai Wang, Chi Lin, Guodong Guo, Isabelle Guyon, et al. Multi-modal face anti-spoofing attack detection challenge at cvpr2019. In *CVPRW*, pages 0–0, 2019.

[Liu *et al.*, 2020] Yaojie Liu, Joel Stehouwer, and Xiaoming Liu. On disentangling spoof trace for generic face anti-spoofing. In *ECCV*, pages 406–422, 2020.

[Liu *et al.*, 2021a] Ajian Liu, Xuan Li, Jun Wan, Yanyan Liang, Sergio Escalera, Hugo Jair Escalante, Meysam Madadi, Yi Jin, Zhuoyuan Wu, Xiaogang Yu, et al. Cross-ethnicity face anti-spoofing recognition challenge: A review. *IET Biometrics*, 10(1):24–43, 2021.

[Liu *et al.*, 2021b] Ajian Liu, Zichang Tan, Jun Wan, Sergio Escalera, Guodong Guo, and Stan Z Li. Casia-surf cefa: A benchmark for multi-modal cross-ethnicity face anti-spoofing. In *WACV*, pages 1179–1187, 2021.

[Liu *et al.*, 2021c] Ajian Liu, Zichang Tan, Jun Wan, Yanyan Liang, Zhen Lei, Guodong Guo, and Stan Z Li. Face anti-spoofing via adversarial cross-modality translation. *IEEE TIFS*, 16:2759–2772, 2021.

[Liu *et al.*, 2021d] Ajian Liu, Chenxu Zhao, Zitong Yu, Anyang Su, Xing Liu, Zijian Kong, Jun Wan, Sergio Escalera, Hugo Jair Escalante, Zhen Lei, et al. 3d high-fidelity mask face presentation attack detection challenge. In *ICCVW*, pages 814–823, 2021.

[Liu *et al.*, 2021e] Ajian Liu, Chenxu Zhao, Zitong Yu, Jun Wan, Anyang Su, Xing Liu, Zichang Tan, Sergio Escalera, Junliang Xing, Yanyan Liang, et al. Contrastive context-aware learning for 3d high-fidelity mask face presentation attack detection. *arXiv preprint arXiv:2104.06148*, 2021.

[Liu *et al.*, 2021f] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *CVPR*, pages 10012–10022, 2021.

[Parkin and Grinchuk, 2019] Aleksandr Parkin and Oleg Grinchuk. Recognizing multi-modal face spoofing with face recognition networks. In *CVPRW*, pages 0–0, 2019.

[Shao *et al.*, 2019] Rui Shao, Xiangyuan Lan, Jiawei Li, and Pong C Yuen. Multi-adversarial discriminative deep domain generalization for face presentation attack detection. In *CVPR*, pages 10023–10031, 2019.

[Touvron *et al.*, 2021] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablay-rolles, and Hervé Jégou. Training data-efficient image transformers & distillation through attention. In *ICML*, pages 10347–10357, 2021.

[Wang *et al.*, 2021] Pichao Wang, Xue Wang, Fan Wang, Ming Lin, Shuning Chang, Wen Xie, Hao Li, and Rong Jin. Kvt: k-nn attention for boosting vision transformers. *arXiv preprint arXiv:2106.00515*, 2021.

[Xu *et al.*,] Tongkun Xu, Weihua Chen, Pichao Wang, Fan Wang, Hao Li, and Rong Jin. Cdtrans: Cross-domain transformer for unsupervised domain adaptation. *arXiv preprint arXiv:2109.06165*.

[Yu *et al.*, 2020] Zitong Yu, Chenxu Zhao, Zezheng Wang, Yunxiao Qin, Zhuo Su, Xiaobai Li, Feng Zhou, and Guoying Zhao. Searching central difference convolutional networks for face anti-spoofing. In *CVPR*, pages 5295–5305, 2020.

[Zhang *et al.*, 2019] Shifeng Zhang, Xiaobo Wang, Ajian Liu, Chenxu Zhao, Jun Wan, Sergio Escalera, Hailin Shi, Zezheng Wang, and Stan Z Li. A dataset and benchmark for large-scale multi-modal face anti-spoofing. In *CVPR*, pages 919–928, 2019.

[Zhang *et al.*, 2020a] Ke-Yue Zhang, Taiping Yao, Jian Zhang, Ying Tai, Shouhong Ding, Jilin Li, Feiyue Huang, Haichuan Song, and Lizhuang Ma. Face anti-spoofing via disentangled representation learning. In *ECCV*, pages 641–657, 2020.

[Zhang *et al.*, 2020b] Shifeng Zhang, Ajian Liu, Jun Wan, Yanyan Liang, Guodong Guo, Sergio Escalera, Hugo Jair Escalante, and Stan Z Li. Casia-surf: A large-scale multi-modal benchmark for face anti-spoofing. *IEEE TBIOM*, 2(2):182–193, 2020.