Suppress and Rebalance: Towards Generalized Multi-Modal Face Anti-Spoofing

Xun Lin[1], Shuai Wang[1], Rizhao Cai[2], Yizhong Liu[1], Ying Fu[3], Wenzhong Tang[1], Zitong Yu[4*], Alex Kot[2]
[1]Beihang University [2]Nanyang Technological University [3]Beijing Institute of Technology
[4]Great Bay University
{linxun, wangshuai}@buaa.edu.cn yuzitong@gbu.edu.cn

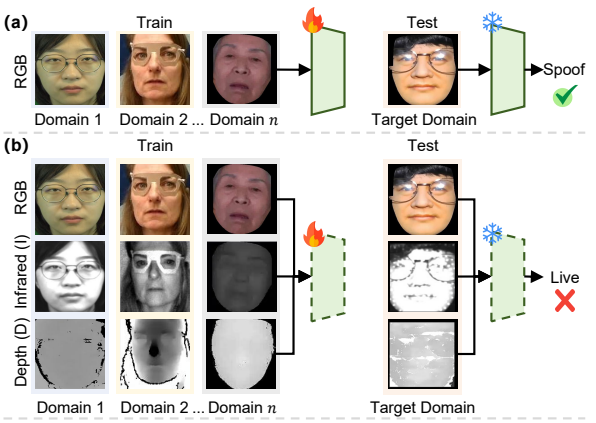
2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) | 979-8-3503-5300-6/24/$31.00 ©2024 IEEE | DOI: 10.1109/CVPR52733.2024.00028

Abstract

*Face Anti-Spoofing (FAS) is crucial for securing face recognition systems against presentation attacks. With advancements in sensor manufacture and multi-modal learning techniques, many multi-modal FAS approaches have emerged. However, they face challenges in generalizing to unseen attacks and deployment conditions. These challenges arise from (1) modality unreliability, where some modality sensors like depth and infrared undergo significant domain shifts in varying environments, leading to the spread of unreliable information during cross-modal feature fusion, and (2) modality imbalance, where training overly relies on a dominant modality hinders the convergence of others, reducing effectiveness against attack types that are indistinguishable by sorely using the dominant modality. To address modality unreliability, we propose the **U**ncertainty-Guided Cross-**Adapter** (**U-Adapter**) to recognize unreliably detected regions within each modality and suppress the impact of unreliable regions on other modalities. For modality imbalance, we propose a **Re**balanced Modality **Grad**ient Modulation (**ReGrad**) strategy to rebalance the convergence speed of all modalities by adaptively adjusting their gradients. Besides, we provide the first large-scale benchmark for evaluating multi-modal FAS performance under domain generalization scenarios. Extensive experiments demonstrate that our method outperforms state-of-the-art methods. Source codes and protocols are released on* `https://github.com/OMGGGGG/mmdg`.

1. Introduction

The remarkable success of deep neural networks has also brought concerns regarding their security vulnerabilities [56, 57, 64]. Especially face recognition (FR) systems, widely used in applications like surveillance and mobile payment [63], are susceptible to various face presentation attacks, including printed photos, video replays, and 3D wearable masks [3]. These vulnerabilities challenge the security of FR systems, limiting their broader application. As a result, face anti-spoofing (FAS) methods are developed to enhance the security of FR systems.



(c)

Method	Modality	HTER (%) ↓	AUC (%) ↑
SSDG	(a): RGB	**24.07**	**83.69**
SSDG*	(b): RGB+D+I	26.09 (+2.02)	82.03 (-1.66)

Figure 1. Illustration of DG scenarios in the context of (a) uni-modal and (b) multi-modal. (c) DG performance of SSDG [20] on our Protocol 1 (see Sec. 4.1). Though fed with more modalities, SSDG performs worse in multi-modal scenarios compared to uni-modal ones. * denotes our re-implemented multi-modal version.

Recently, deep learning has been widely applied in FAS, particularly in unimodal scenarios based on the RGB spectrum. While these models perform well in intra-dataset evaluations, their generalizability to unseen attacks and deployment environments remains limited [2]. Despite the improvement of robustness via domain generalization (DG) techniques [20, 28, 49, 50, 71], these approaches still tend to over-fit the attacks presented in source domains.

With the recent progress of multi-modal learning [31, 52, 54, 68] and the development of sensor manufacture [62], some large-scale multi-modal FAS datasets [15, 34, 45, 67] are proposed, which also drive the advancement in multi-modal FAS [13, 14, 33, 36, 58, 60, 61]. By expanding the input from RGB to multi-modal, *e.g.*, RGB, depth, and infrared, richer spoofing traces can be captured for detecting

*Corresponding author.

2575-7075/24/$31.00 ©2024 IEEE
DOI 10.1109/CVPR52733.2024.00028

various presentation attacks [63]. However, existing multi-modal methods fail to perform well under unseen deployment conditions. As illustrated in Fig. 1, though the current methods generalize well in unimodal scenarios, their performance abnormally declines when extended and applied to multi-modal FAS. *Why using more modalities is not better?* We argue that there are two main reasons: (1) **Modality unreliability.** Deploying in unseen environments or with various kinds of models of imaging sensors within the same modality, especially the depth and infrared shown in Fig. 1 (b), might introduce significant domain shifts [66]. Such shifts can result in unreliable multi-modal live/spoof feature extraction [51], misleading mutual modalities through cross-modal fusion. (2) **Modality imbalance.** In some multi-modal learning scenarios, models tend to overly rely on a dominant modality with the highest convergence speed (*i.e*. fast modality), which prevents itself from fully exploiting the other dominated modalities with relatively slower convergence speeds (*i.e*. slow modalities) [43]. In FAS, modality imbalance can cause significant negative impacts, especially when the fast modality is unsuitable or unavailable for detecting attacks in the target domain, relying on slow modalities that have not adequately converged cannot achieve satisfactory performance. This study addresses the cross-domain limitations of multi-modal FAS approaches due to modality unreliability and imbalance. As illustrated in Fig. 2, we introduce a Multi-Modal Domain Generalized (MMDG) framework for FAS. Within MMDG, we develop trainable Uncertainty-guided cross-Adapters (U-Adapter) to fine-tune Vision Transformers (ViT) [9] The U-Adapter address modality unreliability leveraging the uncertainty of each modality to suppress the focus on unreliable tokens during the cross-modal fusion. This prevents the excessive propagation of unreliable spoofing traces which tends to decrease the model's discriminative capability. Meanwhile, to solve modality imbalance, we design **Re**balanced modality **Grad**ient modulation strategy (**ReGrad**). Our ReGrad rebalance the convergence speed of all modalities by dynamically adjusting the backward propagated gradients of all trainable parameters in U-Adapters based on their conflict degree and convergence speed. This ensures that all modalities can be fully leveraged to resist various unseen attacks in the target domain. Our contributions are as follows.

- We propose the MMDG framework to enhance the domain generalizability by addressing the modality unreliability and imbalance. This is the first investigation for multi-modal FAS under multi-source DG scenarios.
- Within MMDG, we propose a **U**ncertainty-Guided Cross-**Adapter** (**U-Adapter**) for ViT-based multi-modal FAS, which can complementarily fuse cross-modal features and prevents each modality from being influenced by unreliable information caused by domain shifts.
- We design the **Re**balanced Modality **Grad**ient Modulation (**ReGrad**) for MMDG, a strategy that can monitor and rebalance the convergence speed of each modality by adaptively modulating gradients of U-Adapters.
- We build the first large-scale benchmark to evaluate the domain generalizability of multi-modal FAS approaches. Extensive experiments on this benchmark demonstrate the superiority and generalizability of MMDG.

2. Related Works

2.1. Domain Generalization for FAS

In FAS, domain generalization focuses on training a model using multiple source domains, with the intention that this model can work effectively on unseen target domains [18, 46]. Previous works verify the effectiveness of adversarial training [22, 37, 65], asymmetric triplet loss [20, 30] and controversial learning [38, 49] in developing a shared feature space across domains. Besides, some studies [50, 70, 71] investigate the style information of different domains, aiming to learn domain-invariant features by minimizing distinct style characteristics. Furthermore, meta-learning-based methods [4, 10, 21, 44, 69] are also discussed to simulate domain shifts during training, fostering the learning of a robust representative feature space to such variations. More recent works focus on parameter-efficient transfer learning [5, 6, 18, 48], enabling a pretrained ViT to efficiently adapt to unseen domains with a lower risk of over-fitting. However, these approaches are proposed for unimodal (*i.e*. RGB) FAS. Due to the ignorance of modality unreliability, there is a gap in adapting them to multi-modal scenarios. Simple multi-modal extensions of these approaches, through early fusion, late fusion, or even the substitution with multi-modal backbones, still exhibit limited performance under multi-modal DG scenarios.

2.2. Multi-Modal FAS

Multi-modal FAS refers to the use of multiple spectra, *e.g*., RGB, depth, and infrared, to reveal live and spoofing traces. Due to the complementary information among different modalities, traces that are undetectable in one modality can be easily captured in others [17, 36]. To fully exploit information from each modality, early methods employ channel-wise concatenation of inputted modalities [14, 42], or develop multiple separate branches for modality feature extraction followed by late fusion [25, 26, 47, 58]. Recent works introduce attention-based feature fusion techniques [8, 27] and adaptive cross-modal loss functions [13] to encourage the extraction of complementary information among modalities. Meanwhile, cross-modality translation is discussed [29, 35] to mitigate the semantic gap between different modalities. To further enhance the practical usability of FAS, Yu et al. [60, 61, 62] introduce flexible-modal

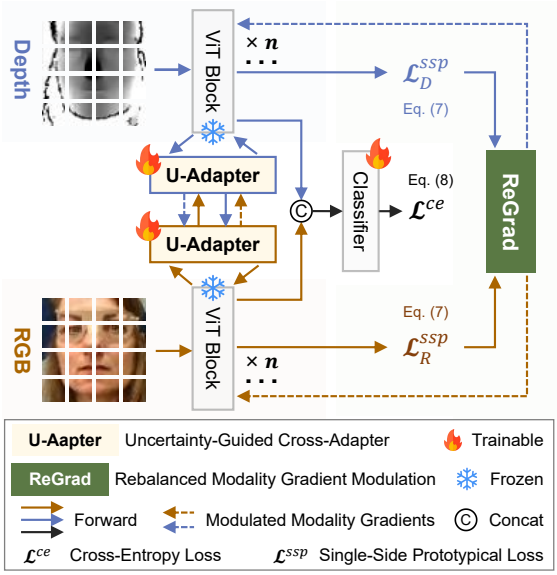


Figure 2. Overall structure of our MMDG. It consists of ViT-based backbones fine-tuned by the proposed U-Adapters with the modality rebalancing strategy called ReGrad. Each modality is designed with a branch for feature extraction and enables feature interaction and mutual complementarity with other modalities. For simplicity, we illustrate the two-modality scenario.

benchmarks that force the use of incomplete modalities during training or testing. In the context of flexible-modal FAS, cross-modality attention [33, 36] or multi-modal adapters [60, 61] are proposed for pretrained ViT to learn modality-agnostic live/spoofing features. However, the aforementioned methods fail to perform well in DG scenarios. The lack of sensitivity to modality unreliability and imbalance leads to an insufficient ability to resist domain shifts. Although previous works rebalance modalities by modulating gradients [43] or adjusting modality losses [11], these methods ignore domain shifts and are not suitable for FAS frameworks with complex cross-modal fusion modules.

3. Proposed MMDG

As shown in Fig. 2, our MMDG consists of ViT-based backbones fine-tuned by the proposed U-Adapters and ReGrad strategy, which are respectively designed for solving modality unreliability and imbalance.

3.1. Fine-tuning ViT with U-Adapters

In multi-modal FAS, DG scenarios involving significant domain shifts [32], including unseen face presentations attacks [51], different kinds of imaging sensors, and noise from low-quality sensors [66], often lead to the extraction of unreliable live/spoofing features from certain modal inputs. These unreliable features can propagate during cross-modal fusion, adversely affecting other modalities. To this end, we aim to identify unreliable local features within each modality and suppress the interaction of these features with other modalities, ensuring that each modality utilizes reliable information from other modalities as complementary traces. Specifically, we adopt uncertainty estimated via Monte Carlo sampling [12, 23, 40] to evaluate the feature unreliability in each modality, and based on the uncertainty, we reduce their weight in cross-modal attention-based feature fusion, minimizing their negative impact. Inspired by recent success in parameter-efficient transfer learning (PETL) by introducing adapters into pretrained ViT [5, 6, 18, 39, 60, 61], but distinctively, we propose the U-Adapter for fine-tuning, which enables cross-modal feature fusion and recognizes as well as suppresses the propagation of unreliable information from uncertainly detected regions caused by domain shifts. In Fig. 3, we present the process of fine-tuning ViT with the U-Adapters, including adopting an Uncertainty Estimation Module (UEM) to recognize unreliable tokens and employing the U-Adapter for uncertainty-aware cross-modal fusion. The ViT [9], comprising n transformer blocks each with Layer Normalization (LN), Multi-Head Self Attention (MHSA), and Multi-Layer Perceptron (MLP), is fine-tuned at the MLP's output using our U-Adapter. By representing our U-Adapter as $\mathcal{A}$ and taking RGB (R), depth (D), and infrared (I) as the inputted modalities, the fine-tuning process is formulated as follows.

$$\mathbf{x}_R^{out} = \mathcal{A}(\mathbf{x}_D^U, \mathbf{x}_D^3, \mathbf{x}_R^3) + \mathcal{A}(\mathbf{x}_I^U, \mathbf{x}_I^3, \mathbf{x}_R^3) + \mathbf{x}_R^3 + \mathbf{x}_R^4, \quad (1)$$

$$\mathbf{x}_D^{out} = \mathcal{A}(\mathbf{x}_R^U, \mathbf{x}_R^3, \mathbf{x}_D^3) + \mathbf{x}_D^3 + \mathbf{x}_D^4, \quad (2)$$

$$\mathbf{x}_I^{out} = \mathcal{A}(\mathbf{x}_R^U, \mathbf{x}_R^3, \mathbf{x}_I^3) + \mathbf{x}_I^3 + \mathbf{x}_I^4, \quad (3)$$

where for $m \in \{R, D, I\}$, $\mathbf{x}_m^U \in \mathbb{R}^{B\times L\times 1}$ are the estimated token uncertainty in the corresponding modality, and $\mathbf{x}_m^3, \mathbf{x}_m^4 \in \mathbb{R}^{B\times L\times C}$ are the outputs of the last LN and MLP, respectively. The meaning of other vectors is illustrated in Fig. 3. To prevent the modalities that experienced significant domain shifts, *i.e.* depth and infrared, from misleading each other, we prohibit their direct interaction.

Uncertainty Estimation Module. As discussed above, it is essential to recognize unreliable modal information for FAS. In deep learning, uncertainty reflects the level of confidence a model has in its predictions. When the model is less certain about a prediction, we consider that prediction to be unreliable [12]. When a model takes images with substantial noise caused by significant domain shifts as inputs, uncertainty estimation techniques [12] tend to assign elevated uncertainty values to regions more likely to be misclassified [32]. One well-known approach is the Bayesian Neural Network (BNN), widely applied in various vision tasks [16, 19]. BNN focuses on learning the posterior distribution of the model's weights rather than output a single value. Since the posterior distribution is always not tractable [12], approximate inference methods, such as Monte Carlo dropout (MCD) [7, 40, 53], are employed for estimation. In this work, we also use MCD to construct the UEM as a probabilistic representational model, which imposes a Bernoulli distribution over the weights of a model

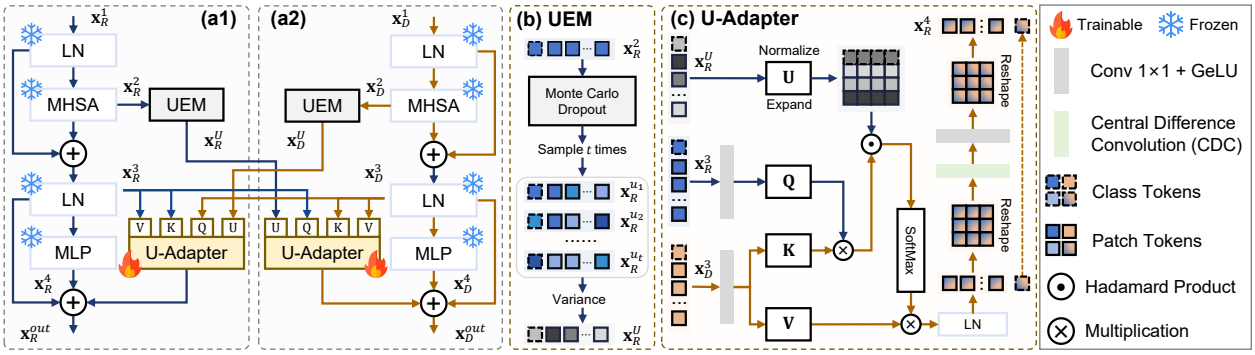


Figure 3. (a1)-(a2) Illustration of fine-tuning ViT with proposed U-Adapters, showcasing the interaction between the RGB (R) and Depth (D) modalities. Note that only parameters of U-Adapters are trainable. (b) Uncertainty Estimation Module (UEM) used for recognizing unreliable tokens. (c) Detailed structure of U-Adapter, which adopts cross-modal fusion and suppresses the interference of unreliable tokens on other modalities. After fusion, discriminative central difference information is integrated for fine-grained spoof representation.

without introducing extra parameters. As presented in Fig. 3 (b), we perform MCD instead of the vanilla dropout after each MHSA, simulating the effect of sampling from the posterior distribution by randomly dropping out neurons in MHSA. According to Bayesian probability theory, we consider $\mathcal{T} = \{\mathbf{x}_R^{u_1}, \mathbf{x}_R^{u_2}, \cdots, \mathbf{x}_R^{u_t}\}$ as empirical samples from the approximate distribution, and use their token-wise variance $\mathbf{x}_R^U = \mathrm{Var}(\mathcal{T}) \in \mathbb{R}^{B\times L\times 1}$ to reflect their unreliability.

Uncertainty-Guided Cross-Adapter. Different from existing cross-modal fusion [33, 36], our U-Adapter, guided by the estimated uncertainty, suppresses the interaction with unreliable tokens from other modalities to ensure the reliability of the fused features. As shown in Fig. 3 (c), U-Adapter interactively fuses features from different modalities and suppresses unreliable information during fusion. For example, in the depth branch (see Fig. 3 (a2)), U-Adapter takes the output of the second LN from another modality $\mathbf{x}_R^3$ as the *query* for attention-based feature fusion. The output of the second LN from its own modality $\mathbf{x}_D^3$ is used as *key* and *value*. Moreover, U-Adapter modulates the *query* tokens from the other modality based on their uncertainty. Since malicious presentation attacks result in subtle traces, we adopt central difference convolution (CDC) [59] to integrate the local details for ViT and explore fine-grained difference information from neighbor tokens, thus making the fused features more discriminative. By ignoring GeLU and vanilla convolution layers, the U-Adapter is formulated as follows.

$$\mathcal{A}(\mathbf{x}_R^U,\mathbf{x}_R^3,\mathbf{x}_D^3)=\mathcal{C}\left(\mathcal{S}\left(\frac{\mathrm{Q}(\mathbf{x}_R^3)\mathrm{K}(\mathbf{x}_D^3)^\intercal \odot \mathrm{U}(\mathbf{x}_R^U)}{\sqrt{n_k}}\right)\mathrm{V}(\mathbf{x}_D^3)\right), \quad (4)$$

where $\mathcal{C}(\cdot)$ denotes the CDC layer, $\mathcal{S}(\cdot)$ represents the softmax, and n_k is the number of token channels. $\mathrm{Q}(\cdot)$, $\mathrm{K}(\cdot)$, and $\mathrm{V}(\cdot)$ are the linear projections corresponding to the *query*, *key*, and *value*, respectively. $\odot$ is the Hadamard product. $\mathrm{U}(\mathbf{x}_R^U) = \exp(-r_e \cdot \mathbf{x}_R^U \times \mathbf{I})$ normalizes the inputted uncertainty using an exponential function, where r_e is a penalty intensity parameter, and $\mathbf{I} = [1, 1, \ldots, 1]_{1\times L}$ is adopted to expend the dimensions of x_R^U to match the size of $L \times L$. We show the meaning of other vectors in Figs. 3 (a1-a2). Note that Eq. (4) only provides the scenario where RGB tokens are fused into depth in Fig. 3 (c). The pairwise fusion among other modalities follows similarly.

3.2. Rebalanced Modality Gradient Modulation

Different modalities are suitable for distinct types of attacks and deployment conditions, *e.g.*, the depth modality being less sensitive to light and more apt for detecting replay attacks, and the infrared modality excelling in identifying wearable mask attacks and performing well in low-light conditions. To cope with various unknown attacks and deployment conditions, we aim to ensure that each modality can sufficiently converge, thereby fully leveraging their discriminative capability, rather than over-relying on a specific modality (*i.e.* fast modality) that shows superior performance in the source domain.

Current modality balancing methods [11, 43], designed for simple late fusion models, are limited in their applicability to FAS. Late fusion networks have been shown to underperform in FAS compared to hybrid fusion networks [33, 36] that allow the fusion both before the final classifier and during the feature extraction. Additionally, the trainable parameters of a late fusion network are only affected by the gradient of its own modality, whereas those of a hybrid fusion network are influenced by the gradients of all modalities. Moreover, gradients from different modalities may even conflict with each other, making the modulation more difficult. To this end, our ReGrad adaptively modulates the gradients of all modalities that simultaneously act on each trainable layer. Moreover, existing modality balancing methods fail to generalize well while balancing each modality due to the ignorance of domain shifts. Drawing inspiration from SSDG [20] and PMR [11], we introduce a single-side prototypical loss to supervise modality branches and monitor their convergence speed. This loss enables each modality to cluster according to domain information through internal impetus, which further mitigates the domi-

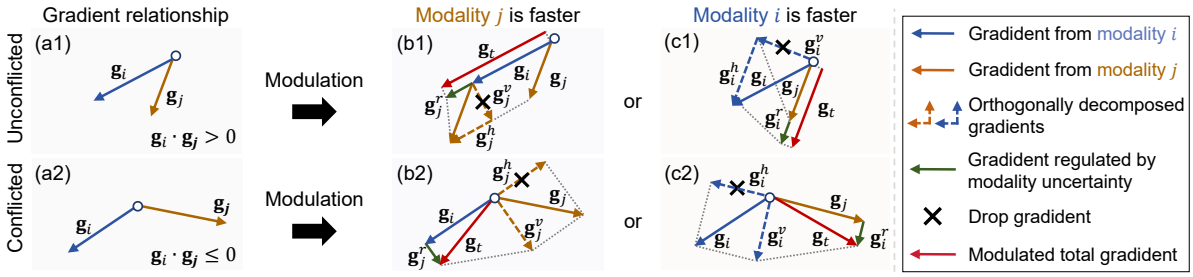


Figure 4. Illustration of gradient modulation via the proposed ReGrad in different scenarios: (Row 1) Non-conflicting (a1) and faster modality j (b1) or i (c1). (Row 2) Conflicting (a2) and faster modality j (b2) or i (c2).

nance of faster modalities.

Modulating Gradients for Each Modality. When the optimization direction is dominated by the gradients of the fast modality, the convergence degree of slow modalities is suppressed. The discriminative ability of slow modalities may even be weakened when there is a conflict in the gradient directions between modalities [55]. Inspired by the studies on solving domain conflicts [41] and multi-task conflicts [55] via gradient modulation, our ReGrad adaptively performs gradient modulation based on the conflict and convergence degree of each modality, making it applicable to complex hybrid fusion networks. The criteria of modality convergence speed will be discussed in the following section.

As shown in Fig. 4 (b1), if modality i converges slower than j and their gradients $\mathbf{g}_i$ and $\mathbf{g}_j$ are not conflicting, we suppress $\mathbf{g}_j$ and boost the learning of modality i. This involves decomposing $\mathbf{g}_j$ into orthogonal $\mathbf{g}_j^v$ and non-orthogonal $\mathbf{g}_j^h$ components relative to $\mathbf{g}_i$. We drop $\mathbf{g}_j^v$ to slow down modality j and use $\mathbf{g}_j^h$ to accelerate modality i. Furthermore, to prevent unreliable modalities that converge rapidly from misleading other modalities, we employ the estimated uncertainty of modality j to further constrain g_j. Conversely, if $\mathbf{g}_i$ conflicts with $\mathbf{g}_j$ and modality i remains a slow modal, it is unreasonable to use any part of $\mathbf{g}_j$ to boost the learning of $\mathbf{g}_i$. To prevent $\mathbf{g}_j$ from seriously slowing down the convergence speed of modality i, as shown in Fig. 4 (b2), we remove the non-orthogonal part of $\mathbf{g}_j^h$ that is opposite to $\mathbf{g}_i$. Meanwhile, we retain its orthogonal part $\mathbf{g}_j^v$ to allow modality j to learn at a slower speed and reduce $\mathbf{g}_j$'s interference with $\mathbf{g}_i$. Similarly, the uncertainty-based suppression is adopted. Cases where the modality i is faster are presented in Figs. 4 (c1)-(c2) and Eq. (5).

$$\mathrm{ReGrad}_2(\mathbf{g}_i,\mathbf{g}_j)=\begin{cases}\mathbf{g}_i+\dfrac{\mathbf{g}_i\cdot\mathbf{g}_j}{\|\mathbf{g}_i\|^2}\mathbf{g}_i\cdot\mathrm{U}(u_j) & \text{, in Fig. 4 (b1),}\\ \mathbf{g}_i+(\mathbf{g}_j-\dfrac{\mathbf{g}_i\cdot\mathbf{g}_j}{\|\mathbf{g}_i\|^2}\mathbf{g}_i)\cdot\mathrm{U}(u_j) & \text{, in Fig. 4 (b2),}\\ \dfrac{\mathbf{g}_i\cdot\mathbf{g}_j}{\|\mathbf{g}_j\|^2}\mathbf{g}_j\cdot\mathrm{U}(u_i)+\mathbf{g}_j & \text{, in Fig. 4 (c1),}\\ (\mathbf{g}_i-\dfrac{\mathbf{g}_i\cdot\mathbf{g}_j}{\|\mathbf{g}_j\|^2}\mathbf{g}_j)\cdot\mathrm{U}(u_i)+\mathbf{g}_j & \text{, in Fig. 4 (c2),}\end{cases} \tag{5}$$

where u_i and u_j are respectively the batch-wise mean uncertainty of the class tokens outputted by the last ViT block of the corresponding modality, denoting the modality uncertainty.

Single-Side Prototypical Loss. Prototype learning has been successfully applied in FAS and modality rebalancing. PMR [11] indicates that introducing prototypes to each modality in multi-modal learning to guide the learning process can mitigate modality imbalance. However, vanilla prototype learning is suboptimal under DG scenarios as it neglects the differences in samples across various domains. Inspired by SSDG [20], we propose the Single-Side Prototypical (SSP) loss with the concept of asymmetry representation and prototype learning. Specifically, we introduce prototypes for the live faces of all domains and the fake faces from each domain. Formulate the domain set as $\mathcal{D} = \{d_\ell, d_s^1, d_s^2, \ldots, d_s^{N_s}\}$, where N_s is the number of source domains, d_ℓ denotes the domain consists of all live faces, and d_s^i represents domain of spoof faces from the i-th dataset. For modality m, the prototype of domain $d \in \mathcal{D}$ is the center of all samples belonging to d in the feature space:

$$p_m^d = \frac{1}{N_d}\sum_{c_m \in d} c_m, \tag{6}$$

where c_m is a sample of d in modality m, and N_d denotes the total number of samples in d. Based on the prototypes of each domain, we aim to reduce the distance between samples and the prototype of their corresponding domain and increase the distance from other prototypes, thus preventing conflicts when samples from different domains are trained together. Therefore, following the form of cross-entropy, the SSP loss $\mathcal{L}_m^{ssp}$ of modality m is formulated as follows.

$$\mathcal{L}_m^{ssp}(c_m, p_m^{d_{gt}}) = -\log\frac{\exp(-\mathrm{ED}(c_m, p_m^{d_{gt}}))}{\sum_{d\in\mathcal{D}}\exp(-\mathrm{ED}(c_m, p_m^{d}))}, \tag{7}$$

where sample c_m belongs to domain d_{gt}. The function $\mathrm{ED}(\cdot)$ denotes the Euclidean distance of the input sample and its prototype. Our SSP loss encourages the features of each modality of the samples to be pulled to the corresponding domain's prototype. This optimization of each modality branch is driven by the intrinsic force within each modality. We mark the modality with lower $\mathcal{L}_m^{ssp}$ as a faster modality. Our final loss is expressed as follows.

$$\mathcal{L}^{final} = \mathcal{L}^{ce} + \sum_{m\in\{R,D,I\}} \lambda \cdot \mathcal{L}_m^{ssp}, \tag{8}$$

Table 1. Cross-dataset testing results under the fixed-modal scenarios (**Protocol 1**) among CASIA-CeFA (**C**), PADISI (**P**), CASIA-SURF (**S**), and WMCA (**W**). DG, MM, and FM are short for domain-generalized, multi-modal, and flexible-modal, respectively.

Method	Type	CPS → W		CPW → S		CSW → P		PSW → C		Average	
		HTER (%) ↓	AUC (%) ↑	HTER (%) ↓	AUC (%) ↑	HTER (%) ↓	AUC (%) ↑	HTER (%) ↓	AUC (%) ↑	HTER (%) ↓	AUC (%) ↑
SSDG [20]	DG	26.09	82.03	28.50	75.91	41.82	60.56	40.48	62.31	34.22	70.20
SSAN [50]	DG	17.73	91.69	27.94	79.04	34.49	68.85	36.43	69.29	29.15	77.22
IADG [71]	DG	27.02	86.50	23.04	83.11	32.06	73.83	39.24	63.68	30.34	76.78
ViTAF [18]	DG	20.58	85.82	29.16	77.80	30.75	73.03	39.75	63.44	30.06	75.02
MM-CDCN [58]	MM	38.92	65.39	42.93	59.79	41.38	61.51	48.14	53.71	42.84	60.10
CMFL [13]	MM	18.22	88.82	31.20	75.66	26.68	80.85	36.93	66.82	28.26	78.04
ViT + AMA [61]	FM	17.56	88.74	27.50	80.00	21.18	85.51	47.48	55.56	28.43	77.45
VP-FAS [60]	FM	16.26	91.22	24.42	81.07	21.76	85.46	39.35	66.55	25.45	81.08
ViT [9]	Baseline	20.88	84.77	44.05	57.94	33.58	71.80	42.15	56.45	35.16	67.74
MMDG	**Ours**	**12.79**	**93.83**	**15.32**	**92.86**	**18.95**	**88.64**	**29.93**	**76.52**	**19.25**	**87.96**

where L^{ce} denotes the cross-entropy loss, and λ is used to balance the trade-off between $\mathcal{L}_m^{ssp}$ and $\mathcal{L}^{ce}$. Here, we utilize $\mathcal{L}^{ce}$ to ensure the model's performance, while $\mathcal{L}_m^{ssp}$ is employed to balance all modalities and enhance generalizability. Since $\mathcal{L}_m^{ssp}$ is not modality-specific, it does not undergo gradient modulation, which we perform to $\mathcal{L}^{ce}$ only.

4. Experimental Evaluation

4.1. Multi-Modal FAS Benchmark in DG Scenarios

Datasets, Protocols, and Performance Metrics. With the increasing popularity of large-scale foundation models [24], the current trend is towards fully leveraging multiple and diverse source datasets/domains [1]. While training with multiple datasets has been widely studied for unimodal FAS [20, 50, 71], the multi-modal FAS community remains in the phase of single-dataset training [13, 33, 36]. Therefore, we propose the first large-scale multi-source benchmark to evaluate the DG performance of multi-modal FAS. This benchmark incorporates four commonly used multi-modal datasets: CASIA-CeFA (**C**) [34], PADISI-Face (**P**) [45], CASIA-SURF (**S**) [67], and WMCA (**W**) [15], each featuring a variety of attack types. For each dataset, we use RGB, depth (D), and infrared (I) modalities as inputs. Following previous works [20], we use Half Total Error Rate (HTER) and Area Under the Receiver Operating Characteristic Curve (AUC) for evaluation. We design three protocols to evaluate the cross-dataset performance under different unseen deployment conditions, *i.e*., fixed modalities, missing modalities, and limited source domains. For **Protocol 1**, inspired by the leave-one-out (LOO) protocols designed in unimodal DG scenarios [20], we propose four multi-modal LOO sub-protocols among **C**, **P**, **S**, and **W**. For example, the sub-protocol **CPS** $\rightarrow$ **W** represents we take **C**, **P**, and **S** as training sets, while **W** is testing set. In **Protocol 2**, for each LOO sub-protocol of **Protocol 1**, we design three test-time missing-modal scenarios, *i.e*., D is missing, I is missing, and both D and I are missing. In **Protocol 3**, we limit the number of source domains by proposing two sub-protocols, namely **CW** $\rightarrow$ **PS** and **PS** $\rightarrow$ **CW**. **Implementation Details.** We resize all RGB, depth, and infrared images to $224 \times 224 \times 3$. Each modality's input is divided into 14×14 patch tokens and a class token is attached, collectively serving as the input to the ViT. The hidden size of each token is default set to 768. We employ the Adam optimizer with a learning rate of 5×10^{-5} and a weight decay of 1×10^{-3} to train 70 epochs. The batch size is set to 32. Our classifier is a single-layer fully-connected network that reduces the dimensionality of the class token outputted by the final ViT block from 768 to 2. The number n of ViT blocks in MMDG is set to 12. ViT-b pretrained on ImageNet is adopted as the baseline.

4.2. Cross-dataset Testing

We compare our method on **Protocols 1-3** with three categories of FAS methods: (1) Multi-modal methods, including CMFL [13] and MM-CDCN [58]. (2) Flexible-modal FAS, such as ViT + AMA [61] and VP-FAS [60]. (3) Domain generalized FAS *i.e*., SSDG [20], SSAN [50], IADG [71], and ViTAF [18]. For methods not designed for unimodal FAS, we re-implement them either by employing a concatenation of multi-modal inputs or by replacing the backbone with our ViT+U-Adapter or vanilla ViT. All these methods are trained on the proposed protocols. We automatically use their hyper-parameters as described in the corresponding reference papers or optimally assign better ones.

Protocol 1: Fixed Modalities. We evaluated the cross-dataset performance of unimodal DG and multi-modal methods in the fixed modalities scenarios. Since unimodal DG methods cannot be directly applied to this protocol, we concatenate the images of the three modalities along the channel dimension. We then reduce the channel number to 3 using a trainable 1×1 convolution layer before feeding them to these methods. As shown in the Table 1, our MMDG achieves the best results. Compared to the second-ranked method, *i.e*. VP-FAS [60], there is an improvement of over 6% in both HTER and AUC. A higher performance boost occurs when compared to domain-generalized methods. This indicates MMDG's effectiveness in addressing modality unreliability and imbalance.

Protocol 2: Missing Modalities. As shown in Table 2, our MMDG outperforms other algorithms in all three scenarios of modality missing during testing. It's worth noting that ViT+AMA and VP-FAS are specifically designed

Table 2. Cross-dataset testing results under the test-time missing-modal scenarios (**Protocol 2**) among CASIA-CeFA (C), PADISI (P), CASIA-SURF (S), and WMCA (W). We report the average HTER (%) ↓ and AUC (%) ↑ on four sub-protocols, *i.e.*, **CPS** → **W**, **CPW** → **S**, **CSW** → **P**, and **PSW** → **C**. DG, MM, and FM are short for domain-generalized, multi-modal, and flexible-modal, respectively.

Method	Type	Missing D		Missing I		Missing D & I		Average	
		HTER (%) ↓	AUC (%) ↑	HTER (%) ↓	AUC (%) ↑	HTER (%) ↓	AUC (%) ↑	HTER (%) ↓	AUC (%) ↑
SSDG [20]	DG	38.92	65.45	37.64	66.57	39.18	65.22	38.58	65.75
SSAN [50]	DG	36.77	69.21	41.20	61.92	33.52	73.38	37.16	68.17
IADG [71]	DG	40.72	58.72	42.17	61.83	37.50	66.90	40.13	62.49
ViTAF [18]	DG	34.99	73.22	35.88	69.40	35.89	69.61	35.59	70.64
MM-CDCN [58]	MM	44.90	55.35	43.60	58.38	44.54	55.08	44.35	56.27
CMFL [13]	MM	31.37	74.62	30.55	75.42	31.89	74.29	31.27	74.78
ViT + AMA [61]	FM	29.25	77.70	32.30	74.06	31.48	75.82	31.01	75.86
VP-FAS [60]	FM	29.13	78.27	29.63	77.51	30.47	76.31	29.74	77.36
ViT	Baseline	40.04	64.69	36.77	68.19	36.20	69.02	37.67	67.30
MMDG	**Ours**	**24.89**	**82.39**	**23.39**	**83.82**	**25.26**	**81.86**	**24.51**	**82.69**

Table 3. Cross-dataset testing results under the limited source domain scenarios (**Protocol 3**) among CeFA-CeFA (**C**), PADISI-USC (**P**), CASIA-SURF (**S**), and WMCA (**W**).

Method	Type	CW → PS		PS→ CW	
		HTER (%)	AUC (%)	HTER (%)	AUC (%)
SSDG [20]	DG	25.34	80.17	46.98	54.29
SSAN [50]	DG	26.55	80.06	39.10	67.19
IADG [71]	DG	22.82	83.85	39.70	63.46
ViTAF [18]	DG	29.64	77.36	39.93	61.31
MM-CDCN [58]	MM	29.28	76.88	47.00	51.94
CMFL [13]	MM	31.86	72.75	39.43	63.17
ViT + AMA [61]	FM	29.25	76.89	38.06	67.64
VP-FAS [60]	FM	25.90	81.79	44.37	60.83
ViT [9]	Baseline	42.66	57.80	42.75	60.41
MMDG	**Ours**	**20.12**	**88.24**	**36.60**	**70.35**

for scenarios of modality missing and have achieved better results compared to other methods. However, under the DG premise, they did not consider the modality imbalance and unreliability issues brought by domain shift, resulting in a difference of over 5% in average HTER and AUC compared to our method. Another interesting observation is that SSAN, IADG, ViT, and ViT + AMA perform better in the **Missing D & I** scenario than in **Missing D** and **Missing I**. These methods lose more modalities but achieve performance improvement, indicating that unreliable information within the modalities leads to erroneous detection. This reflects the effectiveness of our approach in suppressing the interaction of unreliable information between modalities.

Protocol 3: Limited Source Domains. Compared to protocol 1, the limitation of source domains results in a more significant domain shift. We observe that our proposed MMDG achieves optimal results in both sub-protocols (see Table 3), even in the presence of substantial domain shift in **PS** → **CW**. This further underscores the effectiveness of our U-Adapter and ReGrad in enhancing domain generalizability in unseen multi-modal deployment environments.

4.3. Ablation Study and Discussion

In this section, we perform ablation analysis on U-Adapter and ReGrad to verify their individual contribution.

Effectiveness of U-Adapter. To validate the effectiveness of the proposed U-Adapter, we compare the network performance before and after removing the adapter. We also replace the backbone of existing DG methods (*i.e.*, SSDG, SSAN, and IADG) and multi-modal methods (CMFL) to

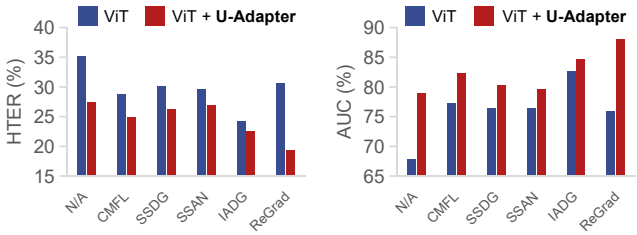


Figure 5. Ablation results on our U-Adapter. We report average HTER ↓ and AUC ↑ on four sub-protocols in **Protocol 1**.

Table 4. Ablation results on the proposed ReGrad. We report the average HTER and AUC on four sub-protocols in **Protocol 1**.

Backbone	Method	HTER (%)	AUC (%)
ViT [9]	-	31.14	74.81
ViT [9]	**ReGrad**	**30.59**	**75.94**
ViT + U-Adapter	SSDG [20]	26.13	80.31
ViT + U-Adapter	SSAN [50]	26.94	79.52
ViT + U-Adapter	IADG [71]	24.23	82.54
ViT + U-Adapter	PMR [11]	24.54	83.14
ViT + U-Adapter	OGM-GE [43]	29.32	76.54
ViT + U-Adapter	-	27.38	78.88
ViT + U-Adapter	**ReGrad** (Conflicted-only)	21.91	85.65
ViT + U-Adapter	**ReGrad** (Unconflicted-only)	21.23	85.51
ViT + U-Adapter	**ReGrad**	**19.48**	**87.48**

observe the performance differences when using ViT+U-Adapter as the backbone versus using ViT alone. As shown in Fig. 5, compared to using ViT as the backbone, these methods experience a performance boost when additional U-Adapters are employed. The performance improvement is especially pronounced when no additional method is used, and when our proposed ReGrad is integrated. This demonstrates the effectiveness of U-Adapter and its universality for existing DG methods and multi-modal strategies.

Effectiveness of ReGrad. Here, we try to answer two questions: (1) *Whether the idea to address domain shift through modality rebalancing, is more suitable compared to existing DG methods?* (2) *Is our ReGrad more capable of enhancing the performance of complex cross-modal fusion frameworks like ViT+U-Adapter compared to existing modality balancing methods?* From Table 4, we can observe that when ViT+U-Adapter is used as the backbone, our ReGrad surpasses existing DG methods and other modality rebalancing methods. Even for vanilla ViT, our ReGrad can enhance its performance. This indicates the effectiveness of addressing domain shifts in unseen scenarios by balancing

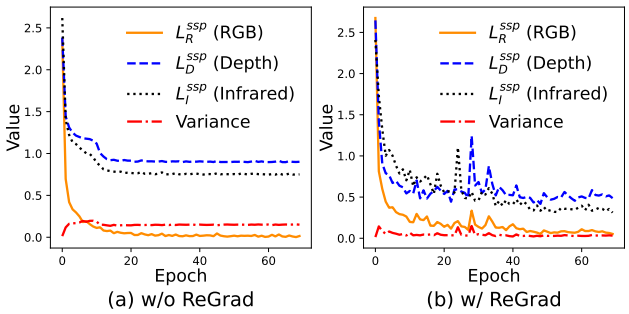


Figure 6. Modality balance degree during training. (a) MMDG w/o ReGrad. (b) MMDG w/ ReGrad. The red dashed line in each sub-figure represents the variance of $\mathcal{L}_R^{ssp}$, $\mathcal{L}_D^{ssp}$, and $\mathcal{L}_I^{ssp}$.

Table 5. Impact of estimated uncertainty in U-Adapters and ReGrad. We report the average HTER (%) ↓ and AUC (%) ↑ on four sub-protocols in **Protocol 1**.

Using Estimated Uncertainty		Metric Value	
U-Adapter: Eq. (4)	ReGrad: Eq. (5)	HTER (%) ↓	AUC (%) ↑
×	×	28.40	78.65
×	✓	26.13	81.60
✓	×	22.75	84.84
✓	✓	**19.48**	**87.48**

various modalities. The results in Table 4 also indicate that when compared to the SoTA modality balancing methods (*i.e.*, OGM-GE and PMR), our ReGrad is more suitable for ViT + U-Adapter, which involves hybrid feature fusion and complex cross-modal interactions.

Meanwhile, as illustrated in Fig. 6, the lower variance (right) of $\mathcal{L}_R^{ssp}$, $\mathcal{L}_D^{ssp}$, and $\mathcal{L}_I^{ssp}$ demonstrates that ReGrad obviously improves the balance degree of MMDG. We also notice that ReGrad causes slight fluctuations in the loss function values. This is due to the gradient modulation allowing the model to escape from local optima, seeking a direction with a higher degree of modality balance. Despite the fluctuations, MMDG can still converge after about 50 epochs. Additionally, we compare ReGrad with its two variants, which perform modulation only when modality gradients are conflicted and unconflicted, respectively. Although these two variants improve performance compared to not modulating gradients at all, their performance is inferior to modulating gradients in all situations. This validates the necessity of adaptive modulation under various scenarios.

Impact of the Uncertainty in U-Adapter and ReGrad. To testify whether the estimated uncertainty can enhance detection performance, we made a comparison by removing the parts that use uncertainty for constraints in U-Adapter and ReGrad. From Table 5, it can be observed that adding uncertainty to constrain the propagation of unreliable information in U-Adapter and the gradient of fast modalities in ReGrad both leads to performance improvement. When uncertainty is added to both modules, HTER and AUC can be further improved by over 3%. As shown in Fig. 7, MMDG can identify unreliable regions in depth (left) and infrared (right) modalities. Although the U-Adapters of these modalities focus on regions with uncertain predictions, due to our idea of suppressing uncertain regions, these regions fail to mislead mutual modalities, enabling the capture of more reliable spoof traces against domain shifts.

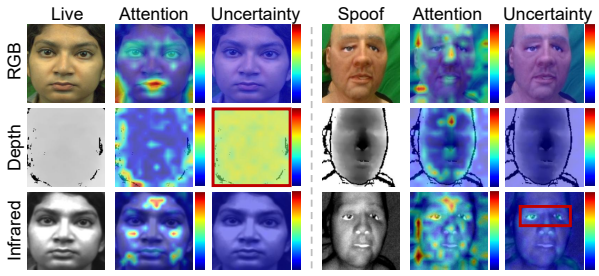


Figure 7. Visual attention and uncertainty maps of our MMDG. Uncertainly predicted regions are highlighted with red boxes. The redder color indicates the higher uncertainty or more attention.

Discussion. The results in Tables 1-5 and Figs. 5-6 validate the MMDG's generalizability by addressing the problems of modality unreliability and imbalance. This work marks a stride towards practical deployments and can benefit the FAS community. However, our MMDG's performance can still be improved. Continued efforts on this benchmark are necessary for further progress. Meanwhile, our method also has space for improvement, *e.g.*, further enhancing the degree of modality balance and adopting more efficient uncertainty estimation techniques.

5. Conclusion

We propose the first multi-modal DG framework for FAS, *i.e.* MMDG, to enhance generalizability by addressing modality unreliability and imbalance. To mitigate the interference of unreliable modalities, we propose U-Adapter to suppress uncertainly predicted regions during cross-modal feature fusion. Besides, we develop ReGrad to adaptively modulate the gradients of each modality branch to achieve modality balance, thus preventing an over-reliance on a dominant modality. We also establish the first large-scale benchmark for multi-modal FAS under DG scenarios. Extensive experiments demonstrate MMDG's effectiveness.

Acknowledgement. This research was done at the Rapid-Rich Object Search (ROSE) Lab, Nanyang Technological University, and supported by the National Key Research and Development Program of China (Grant No.2022YFB3207700), the National Natural Science Foundation of China (Grants No.62272022, 62306061, and 62331006), the NTU-PKU Joint Research Institute (sponsored by the Ng Teng Fong Charitable Foundation), the Science and Technology Foundation of Guangzhou Huangpu Development District (Grant No.2022GH15), and the Guangdong Provincial Regional Joint Fund - Regional Cultivation Project (Grant No.2023A1515140037)

References

[1] Sebastian Borgeaud, Arthur Mensch, Jordan Hoffmann, Trevor Cai, Eliza Rutherford, Katie Millican, George van den Driessche, Jean-Baptiste Lespiau, Bogdan Damoc, Aidan Clark, Diego de Las Casas, Aurelia Guy, Jacob Menick, Roman Ring, Tom Hennigan, Saffron Huang, Loren Maggiore, Chris Jones, Albin Cassirer, Andy Brock, Michela Paganini, Geoffrey Irving, Oriol Vinyals, Simon Osindero, Karen Simonyan, Jack W. Rae, Erich Elsen, and Laurent Sifre. Improving language models by retrieving from trillions of tokens. In *Proceedings of the International Conference on Machine Learning*, pages 2206–2240, 2022.

[2] Rizhao Cai and Changsheng Chen. Learning deep forest with multi-scale local binary pattern features for face anti-spoofing. *arXiv*, abs/1910.03850, 2019.

[3] Rizhao Cai, Haoliang Li, Shiqi Wang, Changsheng Chen, and Alex C. Kot. Drl-fas: A novel framework based on deep reinforcement learning for face anti-spoofing. *IEEE Transactions on Information Forensics and Security*, 16:937–951, 2021.

[4] Rizhao Cai, Zhi Li, Renjie Wan, Haoliang Li, Yongjian Hu, and Alex C. Kot. Learning meta pattern for face anti-spoofing. *IEEE Transactions on Information Forensics and Security*, 17:1201–1213, 2022.

[5] Rizhao Cai, Yawen Cui, Zhi Li, Zitong Yu, Haoliang Li, Yongjian Hu, and Alex C. Kot. Rehearsal-free domain continual face anti-spoofing: Generalize more and forget less. In *Proceedings of the IEEE International Conference on Computer Vision*, 2023.

[6] Rizhao Cai, Zitong Yu, Chenqi Kong, Haoliang Li, Changsheng Chen, Yongjian Hu, and Alex C. Kot. S-adapter: Generalizing vision transformer for face anti-spoofing with statistical tokens. *arXiv*, abs/2309.04038, 2023.

[7] Pamela Carreno, Dana Kulic, and Michael Burke. Adapting neural models with sequential monte carlo dropout. In *Conference on Robot Learning*, pages 1542–1552, 2022.

[8] Pengchao Deng, Chenyang Ge, Xin Qiao, Hao Wei, and Yuan Sun. Attention-aware dual-stream network for multimodal face anti-spoofing. *IEEE Transactions on Information Forensics and Security*, 18:4258–4271, 2023.

[9] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *Proceedings of the International Conference on Learning Representations*, 2021.

[10] Zhekai Du, Jingjing Li, Lin Zuo, Lei Zhu, and Ke Lu. Energy-based domain generalization for face anti-spoofing. In *Proceedings of the ACM International Conference on Multimedia*, pages 1749–1757, 2022.

[11] Yunfeng Fan, Wenchao Xu, Haozhao Wang, Junxiao Wang, and Song Guo. PMR: prototypical modal rebalance for multimodal learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 20029–20038, 2023.

[12] Yarin Gal and Zoubin Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *Proceedings of the International Conference on Machine Learning*, pages 1050–1059, 2016.

[13] Anjith George and Sébastien Marcel. Cross modal focal loss for RGBD face anti-spoofing. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7882–7891, 2021.

[14] Anjith George and Sébastien Marcel. Learning one class representations for face presentation attack detection using multi-channel convolutional neural networks. *IEEE Transactions on Information Forensics and Security*, 16:361–375, 2021.

[15] 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 Transactions on Information Forensics and Security*, 15:42–55, 2020.

[16] Hongji Guo, Hanjing Wang, and Qiang Ji. Uncertainty-guided probabilistic transformer for complex action recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 20020–20029, 2022.

[17] Shuangpeng Han, Rizhao Cai, Yawen Cui, Zitong Yu, Yongjian Hu, and Alex C. Kot. Hyperbolic face anti-spoofing. *arXiv*, abs/2308.09107, 2023.

[18] Hsin-Ping Huang, Deqing Sun, Yaojie Liu, Wen-Sheng Chu, Taihong Xiao, Jinwei Yuan, Hartwig Adam, and Ming-Hsuan Yang. Adaptive transformers for robust few-shot cross-domain face anti-spoofing. In *Proceedings of the European Conference on Computer Vision*, pages 37–54, 2022.

[19] Kaixiang Ji, Feng Chen, Xin Guo, Yadong Xu, Jian Wang, and Jingdong Chen. Uncertainty-guided learning for improving image manipulation detection. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 22456–22465, 2023.

[20] Yunpei Jia, Jie Zhang, Shiguang Shan, and Xilin Chen. Single-side domain generalization for face anti-spoofing. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8481–8490, 2020.

[21] Yunpei Jia, Jie Zhang, and Shiguang Shan. Dual-branch meta-learning network with distribution alignment for face anti-spoofing. *IEEE Transactions on Information Forensics and Security*, 17:138–151, 2022.

[22] Fangling Jiang, Qi Li, Pengcheng Liu, Xiang-Dong Zhou, and Zhenan Sun. Adversarial learning domain-invariant conditional features for robust face anti-spoofing. *International Journal of Computer Vision*, 131(7):1680–1703, 2023.

[23] Alex Kendall, Vijay Badrinarayanan, and Roberto Cipolla. Bayesian segnet: Model uncertainty in deep convolutional encoder-decoder architectures for scene understanding. In *Proceedings of the British Machine Vision Conference*, 2017.

[24] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloé Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross B. Girshick. Segment anything. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4015–4026, 2023.

[25] Chenqi Kong, Kexin Zheng, Shiqi Wang, Anderson Rocha, and Haoliang Li. Beyond the pixel world: a novel acoustic-based face anti-spoofing system for smartphones. *IEEE Transactions on Information Forensics and Security*, 17: 3238–3253, 2022.

[26] Chenqi Kong, Kexin Zheng, Yibing Liu, Shiqi Wang, Anderson Rocha, and Haoliang Li. M3fas: An accurate and robust multimodal mobile face anti-spoofing system. *arXiv preprint arXiv:2301.12831*, 2023.

[27] Kaicheng Li, Hongyu Yang, Binghui Chen, Pengyu Li, Biao Wang, and Di Huang. Learning polysemantic spoof trace: A multi-modal disentanglement network for face anti-spoofing. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 1351–1359, 2023.

[28] Zhi Li, Rizhao Cai, Haoliang Li, Kwok-Yan Lam, Yongjian Hu, and Alex C. Kot. One-class knowledge distillation for face presentation attack detection. *IEEE Transactions on Information Forensics and Security*, 17:2137–2150, 2022.

[29] Zhi Li, Haoliang Li, Xin Luo, Yongjian Hu, Kwok-Yan Lam, and Alex C. Kot. Asymmetric modality translation for face presentation attack detection. *IEEE Transactions on Multimedia*, 25:62–76, 2023.

[30] Chen-Hao Liao, Wen-Cheng Chen, Hsuan-Tung Liu, Yi-Ren Yeh, Min-Chun Hu, and Chu-Song Chen. Domain invariant vision transformer learning for face anti-spoofing. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 6087–6096, 2023.

[31] Xun Lin, Shuai Wang, Jiahao Deng, Ying Fu, Xiao Bai, Xinlei Chen, Xiaolei Qu, and Wenzhong Tang. Image manipulation detection by multiple tampering traces and edge artifact enhancement. *Pattern Recognition*, 133:109026, 2023.

[32] Mattia Litrico, Alessio Del Bue, and Pietro Morerio. Guiding pseudo-labels with uncertainty estimation for source-free unsupervised domain adaptation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7640–7650, 2023.

[33] Ajian Liu and Yanyan Liang. Ma-vit: Modality-agnostic vision transformers for face anti-spoofing. In *Proceedings of the International Joint Conference on Artificial Intelligence*, pages 1180–1186, 2022.

[34] 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 *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1178–1186, 2021.

[35] 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 Transactions on Information Forensics and Security*, 16:2759–2772, 2021.

[36] Ajian Liu, Zichang Tan, Zitong Yu, Chenxu Zhao, Jun Wan, Yanyan Liang, Zhen Lei, Du Zhang, Stan Z. Li, and Guodong Guo. FM-ViT: Flexible modal vision transformers for face anti-spoofing. *IEEE Transactions on Information Forensics and Security*, 18:4775–4786, 2023.

[37] Yaojie Liu and Xiaoming Liu. Spoof trace disentanglement for generic face anti-spoofing. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(3):3813–3830, 2023.

[38] Yuchen Liu, Yabo Chen, Mengran Gou, Chun-Ting Huang, Yaoming Wang, Wenrui Dai, and Hongkai Xiong. Towards unsupervised domain generalization for face anti-spoofing. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 20654–20664, 2023.

[39] Anwei Luo, Rizhao Cai, Chenqi Kong, Xiangui Kang, Jiwu Huang, and Alex C. Kot. Forgery-aware adaptive vision transformer for face forgery detection. *arXiv*, abs/2309.11092, 2023.

[40] Yuki Mae, Wataru Kumagai, and Takafumi Kanamori. Uncertainty propagation for dropout-based bayesian neural networks. *Neural Networks*, 144:394–406, 2021.

[41] Lucas Mansilla, Rodrigo Echeveste, Diego H. Milone, and Enzo Ferrante. Domain generalization via gradient surgery. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 6610–6618, 2021.

[42] Olegs Nikisins, Anjith George, and Sébastien Marcel. Domain adaptation in multi-channel autoencoder based features for robust face anti-spoofing. In *Proceedings of the International Conference on Biometrics*, pages 1–8, 2019.

[43] Xiaokang Peng, Yake Wei, Andong Deng, Dong Wang, and Di Hu. Balanced multimodal learning via on-the-fly gradient modulation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8228–8237, 2022.

[44] Yunxiao Qin, Zitong Yu, Longbin Yan, Zezheng Wang, Chenxu Zhao, and Zhen Lei. Meta-teacher for face anti-spoofing. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(10):6311–6326, 2022.

[45] Mohammad Rostami, Leonidas Spinoulas, Mohamed E. Hussein, Joe Mathai, and Wael Abd-Almageed. Detection and continual learning of novel face presentation attacks. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 14831–14840, 2021.

[46] Rui Shao, Xiangyuan Lan, Jiawei Li, and Pong C. Yuen. Multi-adversarial discriminative deep domain generalization for face presentation attack detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 10023–10031, 2019.

[47] Tao Shen, Yuyu Huang, and Zhijun Tong. Facebagnet: Bag-of-local-features model for multi-modal face anti-spoofing. In *IEEE Conference on Computer Vision and Pattern Recognition Workshop*, pages 1611–1616, 2019.

[48] Koushik Srivatsan, Muzammal Naseer, and Karthik Nandakumar. Flip: Cross-domain face anti-spoofing with language guidance. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 19685–19696, 2023.

[49] Yiyou Sun, Yaojie Liu, Xiaoming Liu, Yixuan Li, and Wen-Sheng Chu. Rethinking domain generalization for face anti-spoofing: Separability and alignment. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 24563–24574, 2023.

[50] Zhuo Wang, Zezheng Wang, Zitong Yu, Weihong Deng, Jiahong Li, Tingting Gao, and Zhongyuan Wang. Domain generalization via shuffled style assembly for face anti-spoofing. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4113–4123, 2022.

[51] Junru Wu, Xiang Yu, Buyu Liu, Zhangyang Wang, and Manmohan Chandraker. Uncertainty-aware physically-guided proxy tasks for unseen domain face anti-spoofing. *arXiv*, abs/2011.14054, 2020.

[52] Wuyuan Xie, Kaimin Wang, Yakun Ju, and Miaohui Wang. pmbqa: Projection-based blind point cloud quality assessment via multimodal learning. In *Proceedings of the ACM International Conference on Multimedia*, page 3250–3258, 2023.

[53] Lan Xixi, Changchun Zou, Cheng Peng, and Caowei Wu. Uncertainty quantification in intelligent-based electrical resistivity tomography image reconstruction with monte carlo dropout strategy. *IEEE Transactions on Geoscience and Remote Sensing*, 61:1–16, 2023.

[54] Peng Xu, Xiatian Zhu, and David A Clifton. Multimodal learning with transformers: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023.

[55] Tianhe Yu, Saurabh Kumar, Abhishek Gupta, Sergey Levine, Karol Hausman, and Chelsea Finn. Gradient surgery for multi-task learning. In *Proceedings of the Neural Information Processing Systems*, 2020.

[56] Yi Yu, Wenhan Yang, Yap-Peng Tan, and Alex C Kot. Towards robust rain removal against adversarial attacks: A comprehensive benchmark analysis and beyond. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6013–6022, 2022.

[57] Yi Yu, Yufei Wang, Wenhan Yang, Shijian Lu, Yap-Peng Tan, and Alex C Kot. Backdoor attacks against deep image compression via adaptive frequency trigger. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12250–12259, 2023.

[58] Zitong Yu, Yunxiao Qin, Xiaobai Li, Zezheng Wang, Chenxu Zhao, Zhen Lei, and Guoying Zhao. Multi-modal face anti-spoofing based on central difference networks. In *IEEE Conference on Computer Vision and Pattern Recognition Workshop*, pages 2766–2774, 2020.

[59] 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 *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5294–5304, 2020.

[60] Zitong Yu, Rizhao Cai, Yawen Cui, Ajian Liu, and Changsheng Chen. Visual prompt flexible-modal face anti-spoofing. *arXiv*, abs/2307.13958, 2023.

[61] Zitong Yu, Rizhao Cai, Yawen Cui, Xin Liu, Yongjian Hu, and Alex C. Kot. Rethinking vision transformer and masked autoencoder in multimodal face anti-spoofing. *arXiv*, abs/2302.05744, 2023.

[62] Zitong Yu, Ajian Liu, Chenxu Zhao, Kevin H. M. Cheng, Xu Cheng, and Guoying Zhao. Flexible-modal face anti-spoofing: A benchmark. In *IEEE Conference on Computer Vision and Pattern Recognition Workshop*, pages 6346–6351, 2023.

[63] Zitong Yu, Yunxiao Qin, Xiaobai Li, Chenxu Zhao, Zhen Lei, and Guoying Zhao. Deep learning for face anti-spoofing: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(5):5609–5631, 2023.

[64] Zitong Yu, Rizhao Cai, Zhi Li, Wenhan Yang, Jingang Shi, and Alex C. Kot. Benchmarking joint face spoofing and forgery detection with visual and physiological cues. *IEEE Transactions on Dependable and Secure Computing*, pages 1–15, 2024.

[65] Haixiao Yue, Keyao Wang, Guosheng Zhang, Haocheng Feng, Junyu Han, Errui Ding, and Jingdong Wang. Cyclically disentangled feature translation for face anti-spoofing. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 3358–3366, 2023.

[66] Qingyang Zhang, Haitao Wu, Changqing Zhang, Qinghua Hu, Huazhu Fu, Joey Tianyi Zhou, and Xi Peng. Provable dynamic fusion for low-quality multimodal data. In *Proceedings of the International Conference on Machine Learning*, pages 41753–41769, 2023.

[67] 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 transactions on biometrics, behavior, and identity science*, 2(2):182–193, 2020.

[68] Taoying Zhang, Hesong Li, Qiankun Liu, Xiaoyong Wang, and Ying Fu. Mgt: Modality-guided transformer for infrared and visible image fusion. In *Chinese Conference on Pattern Recognition and Computer Vision*, pages 321–332, 2023.

[69] Qianyu Zhou, Ke-Yue Zhang, Taiping Yao, Ran Yi, Shouhong Ding, and Lizhuang Ma. Adaptive mixture of experts learning for generalizable face anti-spoofing. In *Proceedings of the ACM International Conference on Multimedia*, pages 6009–6018, 2022.

[70] Qianyu Zhou, Ke-Yue Zhang, Taiping Yao, Ran Yi, Kekai Sheng, Shouhong Ding, and Lizhuang Ma. Generative domain adaptation for face anti-spoofing. In *Proceedings of the European Conference on Computer Vision*, pages 335–356, 2022.

[71] Qianyu Zhou, Ke-Yue Zhang, Taiping Yao, Xuequan Lu, Ran Yi, Shouhong Ding, and Lizhuang Ma. Instance-aware domain generalization for face anti-spoofing. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 20453–20463, 2023.