Preprint version; final version will be available at https://openaccess.thecvf.com
The IEEE / CVF International Conference on Computer Vision (2025)
Published by: IEEE & CVF

Group-wise Scaling and Orthogonal Decomposition for Domain-Invariant Feature Extraction in Face Anti-Spoofing

Seungjin Jung[1] Kanghee Lee[2] Yonghyun Jeong[3]
Haeun Noh[1] Jungmin Lee[2] Jongwon Choi[1,2*]

[1]Department of Artificial Intelligence, Chung-Ang University, Seoul, Korea
[2]Department of Advanced Imaging, Chung-Ang University, Seoul, Korea
[3]Naver Cloud, Seongnam, Korea

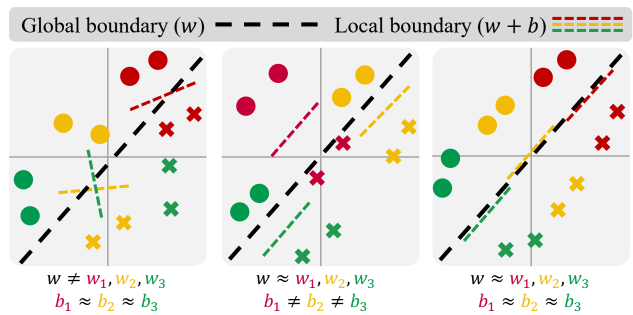
Abstract

Domain Generalizable Face Anti-Spoofing (DGFAS) methods effectively capture domain-invariant features by aligning the directions (weights) of local decision boundaries across domains. However, the bias terms associated with these boundaries remain misaligned, leading to inconsistent classification thresholds and degraded performance on unseen target domains. To address this issue, we propose a novel DGFAS framework that jointly aligns weights and biases through Feature Orthogonal Decomposition (FOD) and Group-wise Scaling Risk Minimization (GS-RM). Specifically, GS-RM facilitates bias alignment by balancing group-wise losses across multiple domains. FOD employs the Gram-Schmidt orthogonalization process to decompose the feature space explicitly into domain-invariant and domain-specific subspaces. By enforcing orthogonality between domain-specific and domain-invariant features during training using domain labels, FOD ensures effective weight alignment across domains without negatively impacting bias alignment. Additionally, we introduce Expected Calibration Error (ECE) as a novel evaluation metric for quantitatively assessing the effectiveness of our method in aligning bias terms across domains. Extensive experiments on benchmark datasets demonstrate that our approach achieves state-of-the-art performance, consistently improving accuracy, reducing bias misalignment, and enhancing generalization stability on unseen target domains. Code: https://github.com/SeungjinJung/GD-FAS

1. Introduction

With the rapid increase in the use of face recognition (FR) [18, 37] technology in critical applications such as payment systems and identity verification, ensuring the security of these systems has become more important than ever. FR systems are now widely deployed not only at airport entry points but also in personal devices, making it crucial for them to effectively defend against presentation attacks, in which adversaries employ various techniques to impersonate legitimate users. In this context, face anti-spoofing (FAS) technology has emerged as an essential defense mechanism for safeguarding FR systems.

Recent FAS studies have focused on Domain Generalizable FAS (DGFAS) to overcome the cross-domain perfor-



(a) Contrastive Loss (b) SA-FAS (c) Ours

Figure 1. **Bias term and weight term across domains**: Decision boundaries are defined by the classifier's weights and biases. Domain invariance maintains consistent distinguishability across domains, typically by sharing a common weight term. In contrast, domain specificity represents domain gaps arising from variations in bias terms. (a) Domain-wise contrastive loss enhances domain-specific separability with bias alignment but struggles to extract domain-invariant features. (b) SAFAS aligns local decision boundaries for improved domain invariance, yet residual domain gaps persist, particularly due to bias misalignment. (c) Our approach addresses this by jointly aligning weights and biases using Group-wise Scaling (GS) and Feature Orthogonal Decomposition (FOD), establishing a unified and consistent decision boundary.

*Corresponding author. Email: choijw@cau.ac.kr

mance drop across various domains, especially the adaptability to new environments, camera sensors, and attack types unseen during training [20, 21, 27, 28, 30, 35]. Such DGFAS research primarily focuses on representation learning to extract domain-invariant features by training models to perform effectively with a single classifier. One approach enhances representations by generating two distinct augmented patch images from the same image and aligning them in the feature space [30]. Another approach segregates domain-specific features and aligns the directional consistency between spoofness and liveness [28]. With the rise of Large Language Models (LLMs), recent studies have leveraged text embedding features extracted from LLMs as guidance to improve the extraction of domain-invariant features [19, 22, 27, 34].

Despite advancements in capturing domain-invariant features, DGFAS methods still encounter significant challenges arising from misaligned bias terms associated with local decision boundaries. As illustrated in Fig. 1-(a), contrastive loss enhances domain-specific separability and aligns bias terms across domains but fails to effectively extract domain-invariant features. In contrast, SAFAS [28] aligns classifier weights across domains yet remains limited due to persistent bias misalignment, as depicted in Fig. 1-(b). Addressing this challenge requires explicitly aligning the bias terms of local decision boundaries. A detailed discussion addressing this bias misalignment issue in Sec. 3.3.

To address this challenge, we propose a novel DG-FAS framework called GD-FAS, comprising Group-wise Scaling Risk Minimization (GS-RM) and Feature Orthogonal Decomposition (FOD). GS-RM aligns the bias terms of classifiers by adaptively balancing group-wise losses across domains through a scaling factor applied at each training iteration. This approach ensures consistent separation of real and spoofed samples across multiple domains. FOD decomposes feature representations into orthogonal domain-invariant and domain-specific subspaces using the Gram-Schmidt orthogonalization process. By constraining domain-specific features to their designated subspace, FOD aligns weight terms across domains without interfering with bias alignment, as shown in Fig. 1-(c). Consequently, our method effectively aligns both bias and weight across domains, thereby enhancing domain invariance and improving cross-domain generalization performance.

Furthermore, we introduce Expected Calibration Error (ECE) as an evaluation metric to quantitatively assess the reliability and effectiveness of our method in aligning bias terms across different domains. By explicitly decomposing feature representations into orthogonal domain-invariant and domain-specific subspaces, our method achieves state-of-the-art performance, significantly enhancing accuracy, reducing bias misalignment, and improving generalization stability on unseen target domains.

This work makes the following key contributions:

- We propose a novel face anti-spoofing approach that ensures robust and consistent performance across diverse domains by decomposing into domain-invariant and domain-specific components orthogonally.
- To enhance domain invariance, we introduce Group-wise Scaling Risk Minimization (GS-RM), which employs a loss-adaptive scaling factor to align classifier bias terms by balancing learning across domains. This ensures consistent and reliable classification of real and spoof samples in multiple domain scenarios.
- To mitigate the domain gap, our approach introduces the Feature Orthogonal Decomposition (FOD), which employs a contrastive loss guided by domain indices to construct domain-specific feature spaces. By enforcing orthogonality between domain-specific and domain-invariant features, this method effectively aligns the weight terms across domains while leaving the bias term.
- We introduce Expected Calibration Error (ECE) to quantitatively validate the effectiveness and reliability of bias term alignment across domains. Comprehensive experiments on standard benchmarks demonstrate that our method achieves state-of-the-art performance, consistently ensuring strong adaptability and stability in unseen target domains.

2. Related Work

Face Anti-Spoofing (FAS) research aims to enhance the security of biometric authentication in facial recognition systems. Initially, FAS was studied using single-domain datasets [23, 38, 39], but this approach struggles with generalizability when encountering unseen domains. This limitation led to the development of cross-domain FAS approaches, categorized as follows: Domain Adaptation FAS (DAFAS) [31, 40, 43] and Domain Generalizable FAS (DG-FAS) [20, 21, 27, 28, 30, 35]. DAFAS requires access to target datasets, whereas DGFAS does not. Therefore, DGFAS is more suitable for real-world applications.

2.1. Domain Generalizable Face Anti-Spoofing.

DGFAS is to learn common features for domain-agnostic use by training only on labeled source domain datasets. One study proposed adversarial loss and a triplet loss-based approach [13]. The patch-based method [30] makes two different augmented patches from the same image similar for learning good representation. The CLIP-based method [27] aligns an ensemble of the class description with image features for improving generality. On the other hand, the instance-level approach [44] focuses on extracting general representations by reducing the sensitivity of features to varying styles.

With the growing prominence of Large Language Models (LLMs), recent research has utilized text embedding

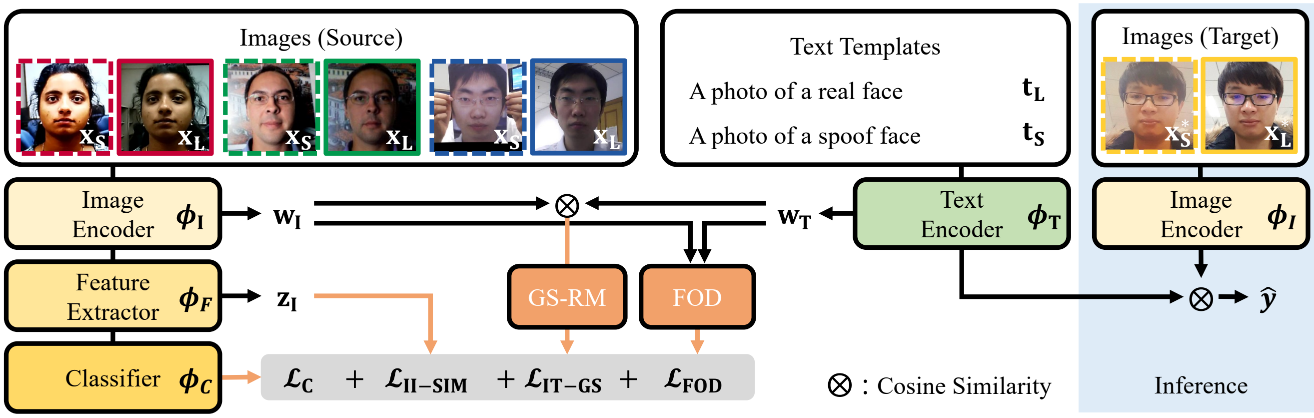


Figure 2. **Framework overview**: Our framework, based on CLIP [27], achieves domain generality by decomposing representations into invariant and specific components within the Text-Image embedding space. (1) GS-RM balances group-wise loss across domains to enhance domain invariance. (2) FOD leverages the Gram-Schmidt process to decompose domain-invariant and domain-specific components, effectively mitigating major spurious correlations. (3) The feature extractor and classifier enhance the embedding of images within the space, ensuring robust feature representation.

features derived from LLMs to guide and enhance the extraction of domain-invariant features [19, 22, 27, 34]. Since these methods focus only on domain-invariant features, they struggle to account for domain-specific components embedded within invariant features, such as spurious correlations.

2.2. Invariant Learning

Machine learning models learn complex networks by minimizing training loss through Empirical Risk Minimization (ERM) [29]. However, ERM has a clear limitation for generalization, as models tend to learn biases from training data. To counter this, Invariant Risk Minimization (IRM) [1, 2] was proposed, encouraging networks to avoid dataset biases by enforcing an optimal classifier over the representation space that is consistent across multiple environments. Despite its promise, IRM still remains challenging due to the complexity of the optimization problem [16, 25]. To handle this, EQRM [6] proposed a new probabilistic aspect of the optimization problem, assuming target and test domains are drawn underlying (meta) distribution over multiple domains. Invariant learning has been adapted to diverse applications such as FAS [26], semantic segmentation [10], and person Re-ID [14, 33].

Recently, SAFAS [28] proposed training domain-specific classifiers and aligning their decision directions, demonstrating that Invariant Risk Minimization (IRM) can enhances domain invariance in DGFAS. However, SAFAS still suffers from unavoidable domain gaps caused by misaligned domain-specific components, particularly the bias terms of the classifiers. In contrast, our approach explicitly decomposes feature representations into domain-invariant and domain-specific subspaces, effectively aligning both weights and biases to mitigate these domain gaps.

3. Methodology

3.1. Problem Setup

The goal of DGFAS is to detect spoofing or presentation attacks on datasets except those used for training. We denote the image space as $\mathcal{X} = \mathbb{R}^{H\times W\times 3}$. Here, H and W represent the height and width of the image, respectively, with three channels corresponding to the RGB color model to encode color information for each pixel. Each image input is labeled as $\mathcal{Y} = \{\mathit{Live}(0),\ \mathit{Spoof}(1)\}$, and the dataset's domain type is denoted as $\mathcal{E} = \{1, \cdots, e\}$, with a total of e domains. We further derive the dataset $\mathcal{D} = \{(\mathbf{x}, y, e) | \mathbf{x} \in \mathcal{X}, y \in \mathcal{Y}, e \in \mathcal{E}\}$. We assume domain $\mathcal{D}$ is given as training data. Then, our goal is to train model Φ:

$$\Phi : \mathbf{x}^* \rightarrow \{\mathit{Live}\,(0),\ \mathit{Spoof}\,(1)\}, \tag{1}$$

where Φ consists of feature extractor ϕ_f and classifier ϕ_c. Φ is trained with training data and distinguishes well whether new samples $\mathbf{x}^*$ are lives or not. New samples $\mathbf{x}^*$ are defined on an unseen domain $\mathcal{E}^*$.

Fig. 2 illustrates the proposed framework, comprising the text embedding network $\phi_\mathbf{T}(\mathbf{t}) : \mathbf{x} \rightarrow \mathbf{w_T}$, the image embedding network $\phi_\mathbf{I}(\mathbf{x}) : \mathbf{x} \rightarrow \mathbf{w_I}$, the feature extractor $\phi_\mathbf{F}(\mathbf{w_I}) : \mathbf{w_I} \rightarrow \mathbf{z}$, and the classifier $\phi_\mathbf{C}(\mathbf{z}) : \mathbf{z} \rightarrow \hat{y}$, with Group-wise Scaling Risk Minimization (GS-RM) and Feature Orthogonal Decomposition (FOD) modules. During the inference phase, the framework predicts spoofness by calculating the cosine similarity between the text embedding feature $\mathbf{w_T}$ and the image embedding features $\mathbf{w_I}$ embedded by the text and image encoders. We explain the details regarding the learning losses of the proposed framework in the following sections.

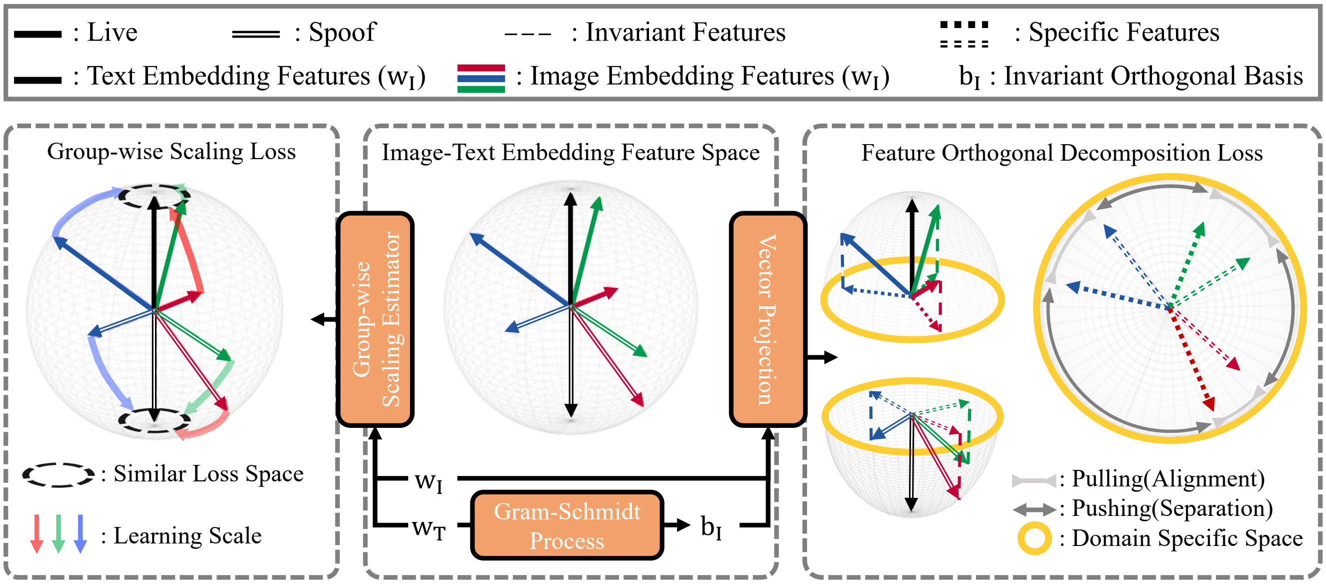


Figure 3. **Insight for GS-RM and FOD**: The center panel illustrates text features and the features of each domain represented within a hypersphere. The left panel demonstrates GS, which penalizes low-loss (high-similarity) groups and boosts high-loss (low-similarity) groups to achieve balanced learning quantities across groups. The right panel utilizes orthogonal bases and orthogonal projections to decompose domain-specific components from image embedding vectors. Subsequently, FOD aligns the domain-specific components orthogonally to the invariant components by leveraging domain labels during training.

3.2. Preliminary

ERM optimizes the network without accounting for domain environments, which can result in the network learning more from certain domains and less from others, finally leading to biased information. Invariant Risk Minimization (IRM) [2] is proposed to solve the problem of Empirical Risk Minimization. IRM aims to optimize networks across all domains, and the objective formula is defined as follows:

$$\mathcal{L}_{\text{IRM}} := \min_{\Phi} \frac{1}{|\mathcal{E}|} \sum_{e \in \mathcal{E}} \mathcal{R}^e(\Phi) + ||\nabla_{w|w=1} \mathcal{R}^e(w \cdot \Phi)||^2, \quad s.t.\ \forall e \in \mathcal{E},\ \Phi \in \arg\min_{\Phi*} \mathcal{R}^e(\Phi*), \tag{2}$$

where $\mathcal{R}^e(\Phi) := \mathbb{E}_{(\mathbf{x},y)\sim\mathcal{D}_e}[\ell_{\text{ce}}(\Phi(\mathbf{x}), y)]$ and $\mathcal{D}_e := \{(\mathbf{x}, y) | \mathbf{x} \in \mathcal{X}, y \in \mathcal{Y}, e = e\}$. Eq. 2 forces the network to learn domain invariant features. However, IRM has difficulty in solving the optimization problem. [16, 25].

To deal with the issue, SAFAS proposed a Projected Gradient (PG-IRM) [28], as follows:

$$\begin{aligned} \mathcal{L}_{\text{PG-IRM}} &:= \min_{\phi_f, \phi_{c(1)}, \cdots, \phi_{c(e)}} \frac{1}{|\mathcal{E}|} \sum_{e \in \mathcal{E}} \mathcal{R}^e(\phi_f, \phi_c), \\ s.t.\ &\exists \phi_{c(e)} \in \arg\min_{\phi_{c(e)}} \mathcal{R}^e(\phi_f, \phi_{c(e)}), \\ &\forall e \in \mathcal{E},\ \phi_{c(e)} \in \gamma_\alpha(\phi_{c(e)}), \end{aligned} \tag{3}$$

where γ_α means α-adjacency set, which is the space where the classifier is Euclidean projected. PG-IRM trains classifiers $\phi_{c(e)}$ corresponding to each domain e and makes each classifier similar by projecting interpolation. Specifically, the classifier aligns the direction of translation between spoofness and liveness. However, each classifier becomes strongly entangled with domain-specific components due to overfitting on individual domain datasets, which manifests as a bias term. Consequently, these bias terms result in gaps that degrade generalization performance across domains.

3.3. Geometrical Motivation

Our geometrical motivation aims to improve generality by orthogonally decomposing the feature space into domain-invariant and domain-specific feature spaces, defined as $\mathbf{b_I}$ and $\mathbf{b_S}$, respectively. Generality is established as follows: We assume that a feature $\mathbf{f}$ is always combination of both domain-invariant $\mathbf{f_I}$ and domain-specific $\mathbf{f_S}$ components, i.e., $\mathbf{f} = \mathbf{f_I} + \mathbf{f_S}$. When applying a linear transformation (classifier) computed as $\mathbf{f} \cdot w^T$, domain bias arises from $\mathbf{f_S} \cdot w^T$ and is compensated by adding a bias term b. However, since $\mathbf{f_S}$ induces various local boundaries due to differences across domains, as shown in Fig. 1, aligning bias across all domains is challenging. To ensure universal applicability across all domains, two conditions must be satisfied: First, $\mathbf{f_I}$ must exist within the orthogonal basis space $\mathbf{b_I}$ derived from optimal w^*. Second, $\mathbf{f_S}$ must exist in $\mathbf{b_S}$ orthogonal to $\mathbf{b_I}$, called domain specific feature space. Under these conditions, $\mathbf{f_S} \cdot w^T$ stays zero due to the orthogonality of the inner product, ensuring that b is consistently aligned as zero across domains.

3.4. Group-wise Scaling Risk Minimization

We introduce Group-wise Scaling Risk Minimization (GS-RM) to enhance domain-invariance, drawing inspiration from [15]. The GS-RM mechanism is designed to achieve alignment in the bias term by ensuring consistent classifier performance across all domains, thereby mitigating overfitting to any specific domain dataset. In contrast, PG-IRM [27] enforces alignment among classifier weights that are individually overfitted to their respective domains, consequently leading to misalignment of bias terms across domains.

To induce GS-RM, we first define group-wise loss for each group dataset. We define group dataset as $\mathcal{D}_g = \{(\mathbf{x}, y)|\mathbf{x} \in \mathcal{X}, (y, e) = g\}$, where $g \in \mathcal{G} = \mathcal{Y} \times \mathcal{E}$. Then, we formulate group-wise losses as follows:

$$\mathcal{L}_g(\ell_{\text{ce}}, \Phi) = \mathbb{E}_{(\mathbf{x},y)\sim\mathcal{D}_g}[\ell_{\text{ce}}(f(\mathbf{x}; \Phi), y)], \tag{4}$$

where ℓ_{ce} represents the cross-entropy loss, and $f(\cdot;\cdot)$ denotes the network, of which the input is to the left of the semicolon and the network module to the right.

Second, we normalize group-wise loss to establish clear distinctions in the current iteration. Normalized loss is calculated by subtracting the mean and dividing it by the standard deviation for group-wise losses, defined as follows:

$$\bar{\mathcal{L}}_g = \frac{\mathcal{L}_g - \mathbb{E}_g[\mathcal{L}_g]}{\sqrt{\mathbb{E}_g[\mathcal{L}_g^2] - \mathbb{E}_g^2[\mathcal{L}_g]}} \tag{5}$$

Then, we introduce a loss-adaptive scaler, called the Group-wise Scaling Estimator (Fig. 3), which assigns a higher scaling factor to high-loss scores and a lower scaling factor to low-loss scores.

$$\sigma(\bar{\mathcal{L}}_g; \alpha, \beta) := \frac{\beta}{1 + exp(-\bar{\mathcal{L}}_g/\alpha)} - \frac{\beta}{2} + 1, \tag{6}$$

where α and β are $\log(|g|)/2$ and 1.5. Each parameter was empirically determined through experimentation.

Finally, we use the group-wise loss (Eq. 4), group-wise normalized loss (Eq. 5) and GS Estimator (Eq. 6) to define GS-RM as follows:

$$\mathcal{L}_{\text{GS-RM}} := \min_{\Phi} \frac{1}{|\mathcal{G}|} \sum_{g\in\mathcal{G}} \sigma(\bar{\mathcal{L}}_g(\Phi))\mathcal{L}_g(\Phi) \tag{7}$$

Image-Text Similarity Loss We define image embedding vector $\mathbf{w_I} := f(\mathbf{x}; \phi_\mathbf{I})$ and text embedding vector $\mathbf{w_T} := f(\mathbf{t}; \phi_\mathbf{T})$. Then, we formulate the group-wise image-text similarity loss, defined as follows:

$$\mathcal{L}_{\text{IT}_g} = \mathbb{E}_{(\mathbf{x},y)\sim\mathcal{D}_g}[\ell_{\text{ce}}(\mathbf{w_I} \otimes \mathbf{w_T}, y)], \tag{8}$$

where $\otimes$ means cosine similarity. We further apply GS-RM for the loss to achieve consistent separation of real and spoofed samples across multiple domains. Then, the image-text similarity loss with GS-RM is defined as follows:

$$\mathcal{L}_{\text{IT-GS}} = \frac{1}{|\mathcal{G}|} \sum_{g\in\mathcal{G}} \sigma(\bar{\mathcal{L}}_{\text{IT}_g})\mathcal{L}_{\text{IT}_g}, \tag{9}$$

where σ denotes GS Estimator (Eq. 6).

The text embedding vector serves as effective guidance for extracting domain-invariant features [19, 22, 27, 34]. Therefore, we utilize this vector as classifier weights, ensuring that it satisfies the first condition mentioned in Sec. 3.3 by balancing performance across domains.

3.5. Feature Orthogonal Decomposition

SAFAS [27] and FOD both aim to align the weight terms across different domains. However, while SAFAS misaligns the bias term across domains, FOD effectively preserves it. Before describing Feature Orthogonal Decomposition (FOD), we introduce the vector projection to decompose representation. The vector projection is the orthogonal projection of $\mathbf{v}$ onto the line spanned by $\mathbf{u}$ and is defined as:

$$\text{proj}_\mathbf{u}(\mathbf{v}) = \frac{<\mathbf{v}, \mathbf{u}>}{<\mathbf{u}, \mathbf{u}>}\mathbf{u}, \tag{10}$$

where $<\mathbf{v}, \mathbf{u}>$ denotes the inner product of $\mathbf{v}$ and $\mathbf{u}$. The Gram-Schmidt process [4] is a way of finding a set of two or more vectors that are perpendicular to each other using Eq. 10. It helps to obtain an orthogonal basis $\mathbf{b_I}$ from text embedding vectors, treating them as invariant vectors. We now formulate invariant features as follows:

$$\mathbf{f_I} = \text{proj}_{\mathbf{b_I}}(\mathbf{w_I}) \tag{11}$$

Then, we derive the domain-specific feature $\mathbf{f_S}$ by removing the domain-invariant feature $\mathbf{f_I}$ from the image embedding features $\mathbf{w_I}$. As a result, the domain-invariant $\mathbf{f_I}$ and domain-specific features $\mathbf{f_S}$ become fully orthogonalized, ensuring their inner product is always zero.

Next, we train domain-specific components using these domain-specific features, explicitly aligning the classifier's bias terms across different domains.

Feature Orthogonal Decomposition Loss We define the FOD loss within the domain-specific space using domain labels, as follows:

$$\mathcal{L}_{\text{FOD}} = \sum_{i=1}^{N} \frac{-1}{|P(i)|} \sum_{p\in P(i)} \log \frac{\exp(\mathbf{f_S}^i \cdot \mathbf{f_S}^p/\tau)}{\sum_{j,j\neq i}^{N} \exp(\mathbf{f_S}^i \cdot \mathbf{f_S}^j/\tau)}, \tag{12}$$

where $P(i)$ denotes the positive sample set defined as $P(i) := \{\mathbf{f_S}^j | \exists e^j = e^i\}$, τ represents temperature parameter, and N and superscript indicate the number of batch samples and the index of batch samples, respectively.

Method	conference	OCI → M		OMI→C		OCM→I		ICM→O		Average	
		HTER	AUC	HTER	AUC	HTER	AUC	HTER	AUC	HTER	AUC
MMD-AAE [17]	CVPR 18	27.08	83.19	44.59	58.29	31.58	75.18	40.98	63.08	36.06	69.94
MADDG [26]	CVPR 19	17.69	88.06	24.50	84.51	22.19	84.99	27.98	80.02	23.09	84.40
SDA [32]	AAAI 21	15.40	91.80	24.50	84.40	15.60	90.10	23.10	84.30	19.65	87.65
SSDG-R [13]	CVPR 20	7.38	97.17	10.44	95.94	11.71	96.59	15.61	91.54	11.29	95.29
SSAN-R [35]	CVPR 22	6.67	98.75	10.00	96.67	8.88	96.79	13.72	93.63	9.82	96.46
PatchNet [30]	CVPR 22	7.10	98.46	11.33	94.58	13.40	95.67	11.82	95.07	10.90	95.95
SAFAS [28]	CVPR 23	5.95	96.55	8.78	95.37	6.58	97.54	10.00	96.23	3.60	99.00
FLIP-MCL* [27]	ICCV 23	7.50	98.19	1.30	99.19	5.33	97.92	5.19	98.48	4.83	98.45
CFPL [19]	CVPR 24	3.09	99.45	2.56	99.10	5.43	98.41	3.33	99.05	3.60	99.00
BUDoPT [22]	ECCV 24	<u>0.95</u>	<u>99.70</u>	2.85	98.03	4.40	98.54	**2.26**	98.78	2.62	98.76
TF-FAS [34]	ECCV 24	3.44	99.42	**0.81**	<u>99.92</u>	**2.24**	**99.67**	**2.26**	**99.48**	<u>2.19</u>	**99.96**
GD-FAS	Ours	**0.42**	**99.88**	<u>0.93</u>	**99.99**	<u>3.33</u>	<u>98.61</u>	2.64	<u>99.34</u>	**1.83**	<u>99.46</u>
FLIP-MCL† [27]	ICCV 23	4.95	98.11	0.54	99.98	4.25	99.07	2.31	<u>99.63</u>	3.01	99.19
CFPL† [19]	CVPR 24	1.43	99.28	2.56	99.10	5.43	98.41	2.50	99.42	2.98	99.05
BUDoPT† [22]	ECCV 24	<u>0.40</u>	<u>99.99</u>	<u>0.26</u>	<u>99.96</u>	**1.38**	99.69	<u>1.60</u>	99.51	**0.91**	99.78
TF-FAS† [34]	ECCV 24	1.49	99.80	0.58	**99.99**	<u>1.56</u>	**99.89**	**1.43**	**99.93**	1.27	**99.90**
GD-FAS†	Ours	**0.00**	**100.00**	**0.19**	99.95	1.67	<u>99.71</u>	1.92	99.50	<u>0.95</u>	<u>99.79</u>

Table 1. **Comparison results on the benchmark**: Best results considered across epochs. The bold font and the underline represent the best and the second, respectively. † indicates the use of an extra source dataset (CelebA-Spoof [36]]), and ∗ means reproducing.

3.6. Image-Image Similarity Loss and Class Loss

To mitigate the data scarcity in FAS, we adopt a contrastive learning strategy inspired by [30] in the image feature space $\mathcal{Z} \ni \mathbf{z}$. The image features $\mathbf{z}$ are mapped from the image embedding feature $\mathbf{w}_\mathbf{I}$ via a multiple layer perceptron [25].

Image-Image Similarity Loss we calculate Image-image similarity loss using the InfoNCE loss, where positive pairs are constructed from two different views of the same image, and negative pairs are sampled from other images.

$$\mathcal{L}_{\text{II-SIM}} = \sum_{i=1}^{2N} \frac{-1}{|P(i)|} \sum_{j \in P(i)} \log \frac{\exp(\mathbf{z}_i \cdot \mathbf{z}_j / \tau)}{\sum_{k=1,k\neq i}^{2N} \exp(\mathbf{z}_i \cdot \mathbf{z}_k / \tau)}, \tag{13}$$

where $\mathbf{z}$ means image feature vector defined as $\mathbf{z} = f(\mathbf{x}; \phi_\mathbf{I}, \phi_\mathbf{F})$ and $P(i)$ denotes the positive sample.

Class Loss We then compute a class loss to enhance spoof-live discrimination in the image feature space.

$$\mathcal{L}_\mathbf{C} = \mathbb{E}[\ell_{\text{ce}}(f(\mathbf{x}; \phi_\mathbf{I}, \phi_\mathbf{F}, \phi_\mathbf{C}), y)] \tag{14}$$

Total Loss Finally, for training the proposed framework, the total loss is as follows:

$$\mathcal{L}_{\text{Total}} = \mathcal{L}_{\text{IT-GS}} + \lambda_1 \mathcal{L}_{\text{FOD}} + \lambda_2 \mathcal{L}_{\text{II-SIM}} + \mathcal{L}_{\text{C}}, \tag{15}$$

where λ_1 and λ_2 are 0.8 and 0.1 as hyper-parameters, respectively. Our framework is trained in an end-to-end manner by minimizing the overall objective function

4. Experiments

4.1. Experimental Setting

Dataset We conduct experiments on four widely used FAS datasets: CASIA [42] (C), OULU-NPU [3] (O), Idiap Replay-Attack [5] (I), and MSU-MFSD [36] (M). Following the protocol in previous works and SAFAS [28], we treat each domain as a separate domain and use the leave-one-out test protocol to evaluate cross-domain generalization.

Specifically in Tab. 1, OCI →M implies that the model is trained on O, C and I, while it is tested on M. OMI →C, OCM →I and ICM →O are defined in a same manner.

Implement Details The experiments were conducted on the Nvidia RTX A6000. We use 224x224 image size, 16 batch size, Adam optimizer with learning rate 3-e06 and weight decay 1-e06, and StepLR scheduler with step size 10 and gamma 0.1. To crop the face region used for learning and inference, we adopted the face information via a Multi-Task Convolutional Neural Network (MTCNN) [41] with a selection of padding value 0.6.

Metrics For the evaluation, we use traditional FAS metrics [27, 28], called Halt Total Error Rate (HTER) and Area Under the Curve (AUC). In addition, we introduce another metric for model reliability, called Expected Calibration Error (ECE) [24]. To measure the reliability, calibration error is defined as the difference between the model's prediction probability and actual probability. Since calculating the ac-

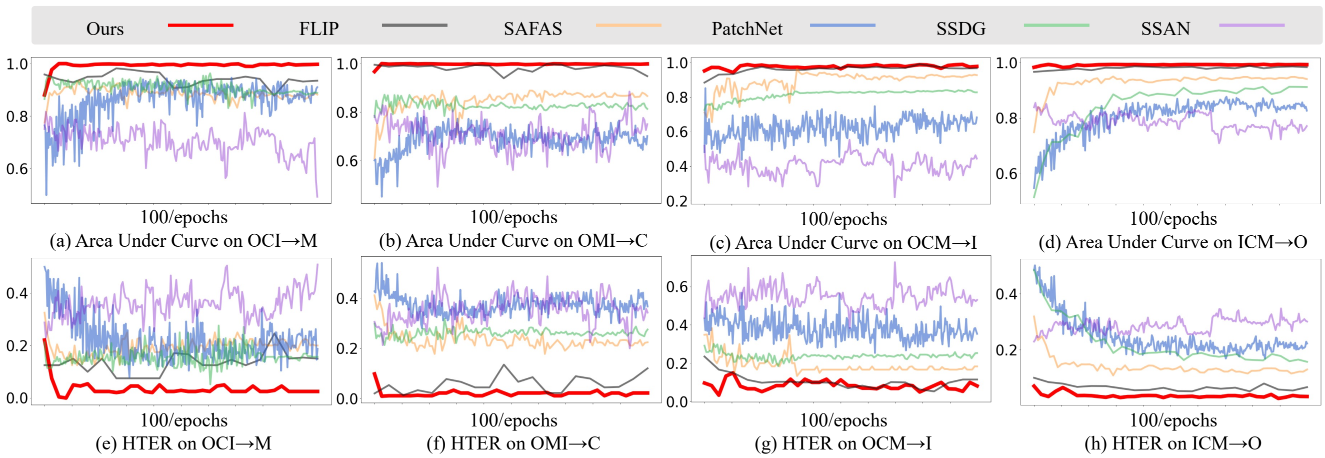


Figure 4. **Learning stability**: Changes in performance measurements according to epoch settings for each domain.

Method	OCI→M	OMI→C	OCM→I	ICM→O
SSDG-R [13]	8.40	11.09	8.94	9.47
SSAN-R [35]	9.58	5.92	12.03	11.10
PatchNet [30]	19.98	20.98	24.31	34.91
SAFAS [28]	18.12	17.77	7.77	13.85
FLIP-MCL [27]	3.94	**1.45**	2.62	5.45
GD-FAS(Ours)	**3.39**	1.82	**2.37**	**2.86**

Table 2. **Model reliability**: Comparative measurement of Expected Calibration Error (ECE) in various scenarios.

tual probability is not feasible, it is approximated using the accuracy of samples, defined as follows:

$$\mathbf{ECE} = \sum_{m=1}^{M} \frac{|B_m|}{n} |\mathbf{acc}(B_m) - \mathbf{conf}(B_m)|, \qquad (16)$$

where B_m is m-th data bin divided into M equal intervals about predictive probabilities, $\mathbf{acc}(B_m)$ means the accuracy of B_m, and $\mathbf{conf}(B_m)$ means the expectation of prediction probability of B_m. A lower ECE score indicates greater confidence in the model's predictions.

4.2. Comparison Results on the Benchmark

In this section, we provide quantitative comparison results across various approaches, covering different domain shifts to comprehensively evaluate the effectiveness of our methods in diverse scenarios. Following a traditional protocol, we select results based on two criteria: the highest AUC-HTER and the lowest HTER, determined by the corresponding threshold. Regarding the HTER metric, Tab. 1 demonstrates that our approach achieves state-of-the-art performance in scenarios like OCI→M while attaining the second-best performance in OMI→C and OCM→I. Overall, our method outperforms the second-best method by 0.36 points. When incorporating an extra source dataset, our approach achieves state-of-the-art performance in cases OCI→M and OMI→C. Notably, it produces perfect predictions in the OCI→M scenario.

4.3. Evaluate Reliability and Bias Term Alignment

The confidence score is directly influenced by misaligned bias terms across different domains, as they affect the classification threshold. Therefore, a lower Expected Calibration Error (ECE) score indicates better alignment, whereas a higher ECE score reflects poorer alignment. As shown in Tab. 2, our method outperforms other methods in most cases. Notably, SAFAS exhibits a higher (worse) ECE score, primarily due to misalignment in the classifier's bias.

4.4. Learning Stability

The convergence stability observed in our model, as shown in Fig. 4, underscores its ability to steadily improve and maintain performance throughout the training epochs. This stability demonstrates the robustness of our approach, thereby providing a reliable solution for real-world applications where domain shifts are frequent.

In contrast, other methods, such as SAFAS, PatchNet, SSDG, and SSAN, exhibit considerable fluctuations during training, indicating their instability in adapting to domain shifts. Specifically, FLIP, the most recent approach, achieves competitive performance only with early stopping applied, as its performance degrades in later epochs. This is particularly evident in Fig. 4, where FLIP's HTER performance worsens significantly without early stopping.

Overall, our model's superior stability and consistent convergence, as shown in both AUC and HTER, reflect its effectiveness in handling domain shifts and maintaining reliability over extended training epochs. This highlights the practical utility of our approach in the real world.

Module		OCI→M			OMI→C			OCM→I			ICM→O		
GS-RM	FOD	HTER	AUC	ECE	HTER	AUC	ECE	HTER	AUC	ECE	HTER	AUC	ECE
		2.92	99.35	5.60	2.22	99.59	2.02	5.17	98.19	3.55	3.98	99.09	3.21
✓		2.50	99.42	4.70	2.22	99.59	2.02	4.83	98.08	1.58	3.75	99.14	2.61
	✓	2.08	99.83	**1.60**	**0.93**	99.95	**1.00**	**3.33**	**99.17**	**1.48**	3.38	99.15	**2.38**
✓	✓	**0.42**	**99.88**	3.39	**0.93**	**99.99**	1.82	**3.33**	98.61	2.37	**2.64**	**99.34**	2.86

Table 3. **Ablation study on each component for GD-FAS**: Quantitative results.

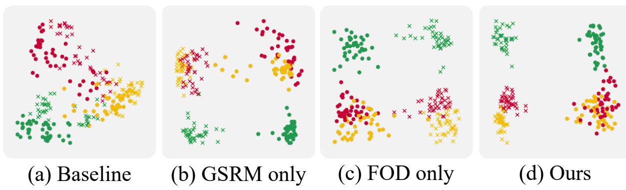


Figure 5. **Ablation study on each component for GD-FAS**: 2D PCA Visualization for OCI→M according to each module

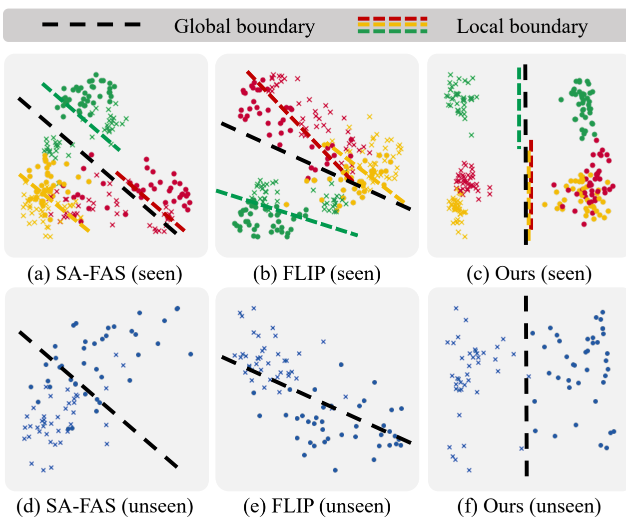


Figure 6. **Comparative Visual Analysis**: 2D PCA Visualization for OCI→M according to different models

4.5. Visual Analysis

We employed PCA to preserve the global structure of the data, capturing its overall patterns and directionality for visual analysis. As depicted in Fig. 6-(c), our method effectively isolates domain-specific information into distinct, separable spaces while maintaining learning consistency across domains. Consequently, as shown in Fig. 6-(f), our approach establishes a universal boundary that robustly separates previously unseen datasets. In contrast, Fig. 6-(d) and (e) illustrate that previous models have difficulty in establishing a universal boundary in the unseen domain, resulting in poor separation between spoof and liveness in feature space.

4.6. Ablation Study

We conducted an ablation study to evaluate the effectiveness of GSRM and FOD on domain generalization and model reliability, as summarized in Tab. 3. The results demonstrate that the combination of GSRM and FOD consistently achieves state-of-the-art performance in most scenarios. Notably, the full model achieves the best results in the OCI→M setting (HTER: 0.42%, AUC: 99.88%) and the OMI→C setting (HTER: 0.93%, AUC: 99.99%). Concerning ECE, there exists a trade-off between model reliability and other performance metrics. While FOD enhances both reliability and generalization, GS-RM primarily focuses on improving generalization. As a result, using only FOD can significantly reduce ECE, but our full model achieves a more balanced trade-off between FAS performance and calibration reliability.

In Fig. 5, panels (a) and (b) illustrate that GSRM achieves consistent separation of real and spoofed samples across multiple domains, while panels (c) and (d) demonstrate that FOD effectively suppresses domain-specific information within the domain-separable space.

5. Conclusion

In this paper, we introduce GD-FAS, a novel Domain Generalizable Face Anti-Spoofing (DGFAS) framework designed to align classifier biases and weights across different domains through Group-wise Scaling Risk Minimization (GS-RM) and Feature Orthogonal Decomposition (FOD). GS-RM facilitates bias alignment by employing a loss-adaptive scaling factor to balance group-wise losses across multiple domains, ensuring consistent separation of real and spoof samples. FOD leverages the Gram-Schmidt orthogonalization process to explicitly decompose feature representations into domain-invariant and domain-specific subspaces. By enforcing orthogonality between these feature spaces using domain labels, FOD ensures effective alignment of weights across domains without negatively impacting bias alignment. Extensive experimental evaluations on standard benchmarks demonstrate that GD-FAS achieves state-of-the-art performance, exhibiting superior reliability, reduced bias misalignment, and enhanced generalization stability and adaptability in unseen domain scenarios.

Acknowledgement

This work was partly supported by the Institute of Information & Communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) [IITP-2023(2024)-RS-2024-00418847, Graduate School of Metaverse Convergence support program; RS-2021-II211341, Artificial Intelligence Graduate School Program (Chung-Ang University)].

References

[1] Kartik Ahuja, Ethan Caballero, Dinghuai Zhang, Jean-Christophe Gagnon-Audet, Yoshua Bengio, Ioannis Mitliagkas, and Irina Rish. Invariance principle meets information bottleneck for out-of-distribution generalization. *Advances in Neural Information Processing Systems*, 34: 3438–3450, 2021. 3

[2] Martin Arjovsky, Léon Bottou, Ishaan Gulrajani, and David Lopez-Paz. Invariant risk minimization. *arXiv preprint arXiv:1907.02893*, 2019. 3, 4

[3] Z. Boulkenafet, J. Komulainen, Lei. Li, X. Feng, and A. Hadid. OULU-NPU: A mobile face presentation attack database with real-world variations. 2017. 6

[4] Ward Cheney and David Kincaid. Linear algebra: Theory and applications. *The Australian Mathematical Society*, 110: 544–550, 2009. 5, 11

[5] Ivana Chingovska, André Anjos, and Sébastien Marcel. On the effectiveness of local binary patterns in face anti-spoofing. 2012. 6

[6] Cian Eastwood, Alexander Robey, Shashank Singh, Julius Von Kügelgen, Hamed Hassani, George J Pappas, and Bernhard Schölkopf. Probable domain generalization via quantile risk minimization. *Advances in Neural Information Processing Systems*, 2022. 3

[7] George et al. Biometric face presentation attack detection with multi-channel convolutional neural network. *TIFS*, 2019. 15

[8] Liu et al. Casia-surf cefa: A benchmark for multi-modal cross-ethnicity face anti-spoofing. In *WACV*, 2021. 15

[9] Zhang et al. Casia-surf: A large-scale multi-modal benchmark for face anti-spoofing. *TBBIS*, 2020. 15

[10] Rui Gong, Wen Li, Yuhua Chen, and Luc Van Gool. Dlow: Domain flow for adaptation and generalization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2477–2486, 2019. 3

[11] Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. On calibration of modern neural networks. In *International conference on machine learning*, pages 1321–1330. PMLR, 2017. 14

[12] Zong-Wei Hong, Yu-Chen Lin, Hsuan-Tung Liu, Yi-Ren Yeh, and Chu-Song Chen. Domain-generalized face anti-spoofing with unknown attacks. In *2023 IEEE International Conference on Image Processing (ICIP)*, pages 820–824. IEEE, 2023. 15

[13] Yunpei Jia, Jie Zhang, Shiguang Shan, and Xilin Chen. Single-side domain generalization for face anti-spoofing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8484–8493, 2020. 2, 6, 7, 15

[14] Xin Jin, Cuiling Lan, Wenjun Zeng, and Zhibo Chen. Feature alignment and restoration for domain generalization and adaptation. *arXiv preprint arXiv:2006.12009*, 2020. 3

[15] Seungjin Jung, Seungmo Seo, Yonghyun Jeong, and Jongwon Choi. Scaling of class-wise training losses for post-hoc calibration. In *Proceedings of the 40th International Conference on Machine Learning*, pages 15421–15434. PMLR, 2023. 5

[16] Pritish Kamath, Akilesh Tangella, Danica Sutherland, and Nathan Srebro. Does invariant risk minimization capture invariance? In *International Conference on Artificial Intelligence and Statistics*, pages 4069–4077. PMLR, 2021. 3, 4

[17] Haoliang Li, Sinno Jialin Pan, Shiqi Wang, and Alex C Kot. Domain generalization with adversarial feature learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018. 6

[18] Jingzhi Li, Zidong Guo, Hui Li, Seungju Han, Ji-won Baek, Min Yang, Ran Yang, and Sungjoo Suh. Rethinking feature-based knowledge distillation for face recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 20156–20165, 2023. 1

[19] Ajian Liu, Shuai Xue, Jianwen Gan, Jun Wan, Yanyan Liang, Jiankang Deng, Sergio Escalera, and Zhen Lei. Cfpl-fas: Class free prompt learning for generalizable face anti-spoofing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 222–232, 2024. 2, 3, 5, 6, 15

[20] Shubao Liu, Ke-Yue Zhang, Taiping Yao, Mingwei Bi, Shouhong Ding, Jilin Li, Feiyue Huang, and Lizhuang Ma. Adaptive normalized representation learning for generalizable face anti-spoofing. In *Proceedings of the 29th ACM international conference on multimedia*, pages 1469–1477, 2021. 2

[21] Shubao Liu, Ke-Yue Zhang, Taiping Yao, Kekai Sheng, Shouhong Ding, Ying Tai, Jilin Li, Yuan Xie, and Lizhuang Ma. Dual reweighting domain generalization for face presentation attack detection. *arXiv preprint arXiv:2106.16128*, 2021. 2

[22] Si-Qi Liu, Qirui Wang, and Pong C Yuen. Bottom-up domain prompt tuning for generalized face anti-spoofing. In *European Conference on Computer Vision*. Springer, 2025. 2, 3, 5, 6, 14, 15

[23] Yaojie Liu, Amin Jourabloo, and Xiaoming Liu. Learning deep models for face anti-spoofing: Binary or auxiliary supervision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 389–398, 2018. 2

[24] Mahdi Pakdaman Naeini, Gregory Cooper, and Milos Hauskrecht. Obtaining well calibrated probabilities using bayesian binning. In *Twenty-Ninth AAAI Conference on Artificial Intelligence*, 2015. 6

[25] Elan Rosenfeld, Pradeep Ravikumar, and Andrej Risteski. The risks of invariant risk minimization. In *International Conference on Learning Representations*, 2021. 3, 4, 6

[26] 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/CVF conference on computer vision and pattern recognition*, pages 10023–10031, 2019. 3, 6

[27] Koushik Srivatsan, Muzammal Naseer, and Karthik Nandakumar. Flip: Cross-domain face anti-spoofing with language guidance. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 19685–19696, 2023. 2, 3, 5, 6, 7, 14, 15

[28] 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/CVF Conference on Computer Vision and Pattern Recognition*, pages 24563–24574, 2023. 2, 3, 4, 6, 7

[29] V. Vapnik. Principles of risk minimization for learning theory. In *Advances in Neural Information Processing Systems*. Morgan-Kaufmann, 1991. 3

[30] Chien-Yi Wang, Yu-Ding Lu, Shang-Ta Yang, and Shang-Hong Lai. Patchnet: A simple face anti-spoofing framework via fine-grained patch recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20281–20290, 2022. 2, 6, 7

[31] Guoqing Wang, Hu Han, Shiguang Shan, and Xilin Chen. Cross-domain face presentation attack detection via multi-domain disentangled representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6678–6687, 2020. 2

[32] Jingjing Wang, Jingyi Zhang, Ying Bian, Youyi Cai, Chunmao Wang, and Shiliang Pu. Self-domain adaptation for face anti-spoofing. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2021. 6

[33] Wenhao Wang, Shengcai Liao, Fang Zhao, Cuicui Kang, and Ling Shao. Domainmix: Learning generalizable person re-identification without human annotations. *arXiv preprint arXiv:2011.11953*, 2020. 3

[34] Xudong Wang, Ke-Yue Zhang, Taiping Yao, Qianyu Zhou, Shouhong Ding, Pingyang Dai, and Rongrong Ji. Tf-fas: twofold-element fine-grained semantic guidance for generalizable face anti-spoofing. In *European Conference on Computer Vision*, pages 148–168. Springer, 2025. 2, 3, 5, 6, 14

[35] 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/CVF Conference on Computer Vision and Pattern Recognition*, pages 4123–4133, 2022. 2, 6, 7, 15

[36] Di Wen, Anil. K Jain, and Hu Han. Face Spoof Detection with Image Distortion Analysis. *IEEE Trans. Information Forensic and Security*, 2015 (To Appear). 6, 14

[37] Xiao Yang, Chang Liu, Longlong Xu, Yikai Wang, Yinpeng Dong, Ning Chen, Hang Su, and Jun Zhu. Towards effective adversarial textured 3d meshes on physical face recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4119–4128, 2023. 1

[38] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 650–651, 2020. 2

[39] 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/CVF Conference on Computer Vision and Pattern Recognition*, pages 5295–5305, 2020. 2

[40] 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. 2

[41] Kaipeng Zhang, Zhanpeng Zhang, Zhifeng Li, and Yu Qiao. Joint face detection and alignment using multitask cascaded convolutional networks. *IEEE signal processing letters*, 23 (10):1499–1503, 2016. 6

[42] Zhiwei Zhang, Junjie Yan, Sifei Liu, Zhen Lei, Dong Yi, and Stan Z Li. A face antispoofing database with diverse attacks. In *2012 5th IAPR international conference on Biometrics (ICB)*, pages 26–31. IEEE, 2012. 6

[43] Qianyu Zhou, Ke-Yue Zhang, Taiping Yao, Ran Yi, Kekai Sheng, Shouhong Ding, and Lizhuang Ma. Generative domain adaptation for face anti-spoofing. In *European Conference on Computer Vision*, pages 335–356. Springer, 2022. 2

[44] 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/CVF Conference on Computer Vision and Pattern Recognition*, pages 20453–20463, 2023. 2

Group-wise Scaling and Orthogonal Decomposition for Domain-Invariant Feature Extraction in Face Anti-Spoofing

Supplementary Material

A. Additional Visual Analysis

In this section, we describe additional visual analyses that were not covered in detail in Section 4.6 of the main paper. Specifically, we provide experimental validation of our method using PCA-based visualizations.

A.1. Feature Orthogonal Decomposition

For PCA analysis and visualization, we calculate the 2D principal components based on the source dataset (training data) which indicates the (seen) domains. The target dataset (test data) which indicates the (unseen) domains is then projected onto these 2D principal components. Finally, we visualize the features of the source and target datasets projected onto the derived principal components.

As shown in Fig. A-(a), (b), (c), and (d), all cases demonstrate strong learning consistency. However, there are some mismatches between the domain-separable feature space and the decision boundary, indicating that they are not parallel to each other. Specifically, Fig. A-(b) exhibits little skewed decision boundary in the Idaip domain, (c) in the CASIA domain, and (d) in the MSU domain. Referring to Tab 1 in the main paper, the alignment quality correlates with the HTER scores. The most aligned case, (a), achieves the lowest HTER of 0.42, followed by (b) with 0.93, (d) with 2.64, and finally, (c), which has the highest HTER of 3.33.

Moreover, Fig. A-(e), (f), (g), and (h) further confirm that better alignment quality enhances the alignment of the domain-specific separable space in the target dataset.

This analysis validates that aligning the domain-specific feature space with the decision boundary significantly contributes to improving domain generality.

A.2. Decomposition into Invariant and Specific Features

For PCA analysis and visualization, we conduct the 2D principal components analysis for invariant features which are projected onto the invariant basis features. The invariant basis features are driven by Gram-Schmidt process [4]. Specific features are then obtained by subtracting the invariant features from the original features. Finally, we visualize the invariant features and specific features separately to analyze their behavior.

The domain-invariant features should demonstrate that they are challenging to distinguish across different domains. In Fig. B-(a), (b), (c), and (d), this is evident as the liveness features are difficult to associate with specific domains. In contrast, spoofness features show some level of domain separability. This might be from the fact that liveness typically represents a single characteristic, while spoofness encompasses a variety of attack types, such as print attacks, video attacks, and partial attacks. However, they still prove their invariance to domains.

Specific features, on the other hand, should neither clearly distinguish spoofness from liveness nor differentiate between domains, as they are unrelated to domain invariance. Moreover, since they are projected onto the principal components of domain-invariant features, domain separability should also be minimal. As expected, Fig. B-(e), (f), (g), and (h) depict indistinct patterns, where neither spoofness nor domain information can be identified.

The overall visualizations confirm that our method effectively and explicitly decomposes domain-invariant components from domain-specific components, achieving the intended decomposition.

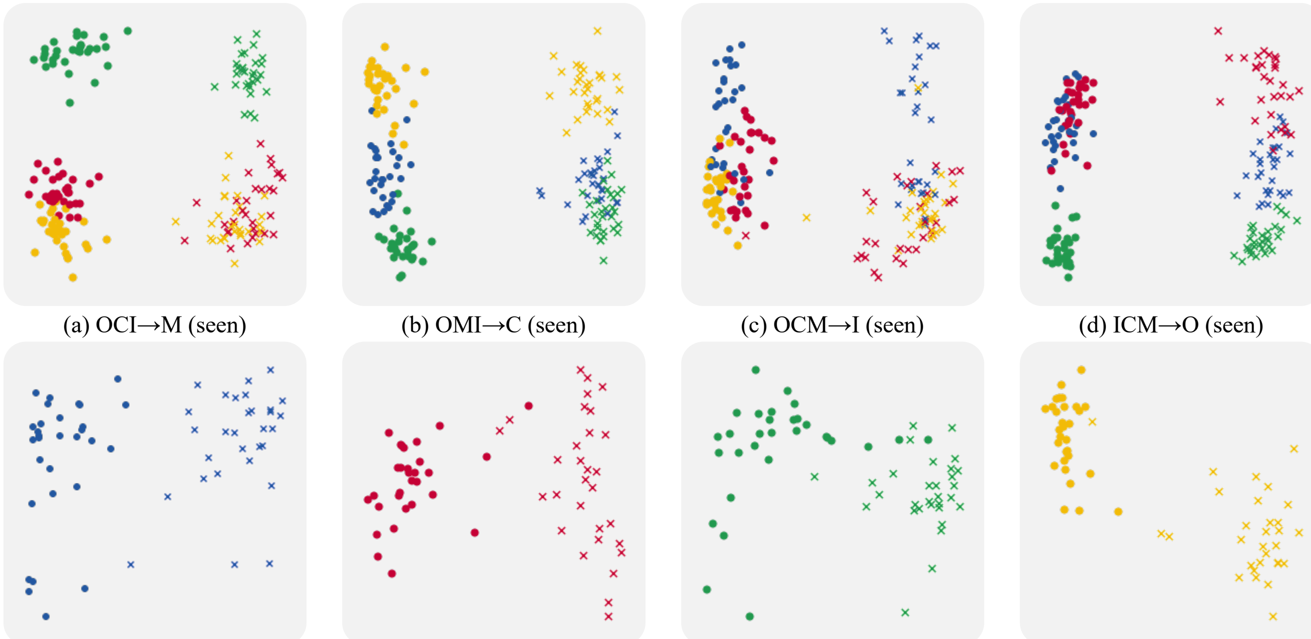
A.3. 3D visualization

We aim to intuitively examine the domain-invariant and domain-specific components. To achieve this, we further conduct 3D principal components analysis and project the features onto them for visualization, as shown in Fig. C. In the figure, the x-axis refers to the horizontal axis when the sphere is viewed from the front. Accordingly, at 0º on the x-axis, the domain-invariant components are highlighted; at 90º, the domain-specific components are emphasized; and at 40º, both components are visible. Additionally, the seen visualization represents data from the training set, while the unseen visualization includes both target and training datasets.

As proposed in Sec. A.1, Fig. C (seen, 0º) demonstrates that learning consistency is well-aligned. Furthermore, as discussed in Sec A.2, Fig. C (seen, 90º) shows that the domain-specific components can distinguish domains but cannot differentiate between liveness and spoofness. These results confirm that the domain-invariant and domain-specific components are effectively and explicitly separated. We also provide 3D visualized GIF images, including ours, FLIP, and SAFAS.

B. Model Reliability

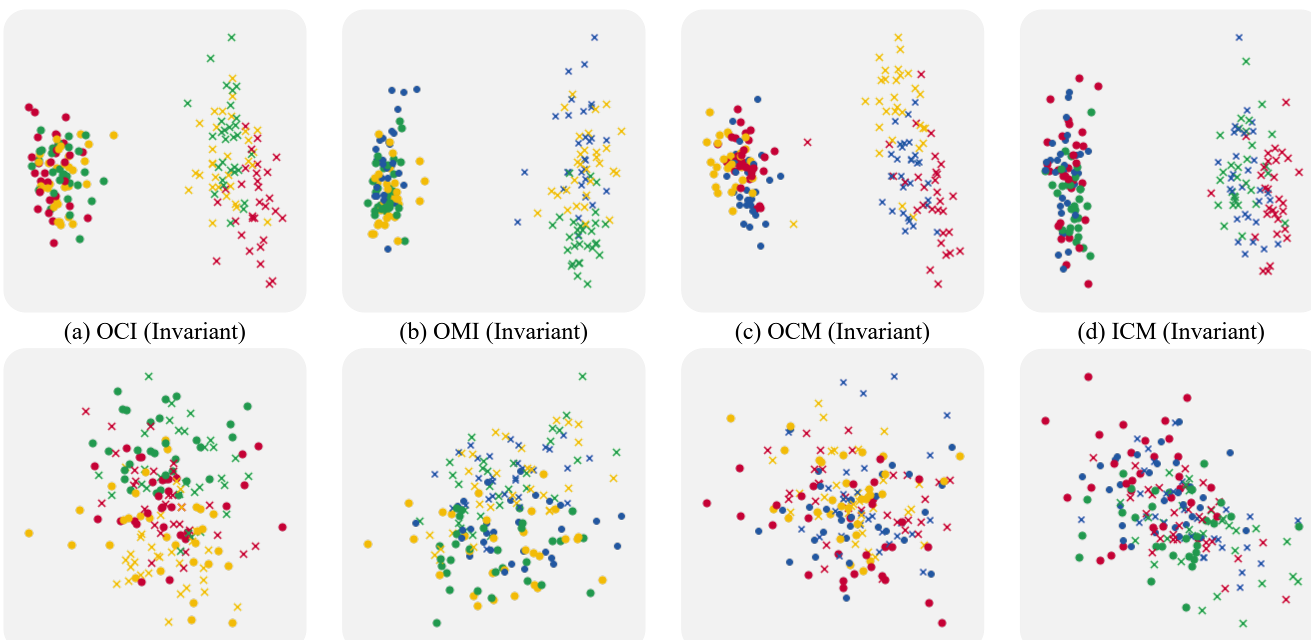
Regardless of how advanced AI models become, the implementation of fundamental safety mechanisms remains essential for their deployment in real-world scenarios, particularly within security systems. This necessity arises be-



(a) OCI→M (seen) (b) OMI→C (seen) (c) OCM→I (seen) (d) ICM→O (seen)

(e) OCI→M (unseen) (f) OMI→C (unseen) (g) OCM→I (unseen) (h) ICM→O (unseen)

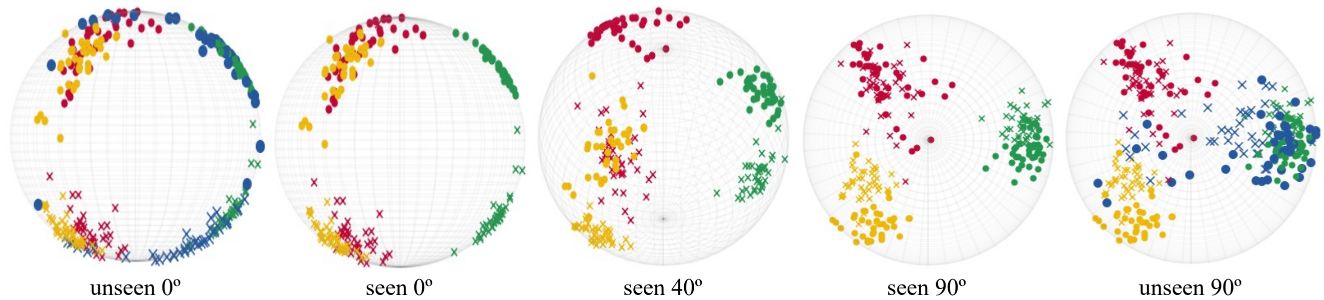
Figure A. **Visual analysis**: 2D PCA Visualization across domains. The term (seen) refers to the training dataset, while (unseen) refers to the target dataset that was not encountered during training. ○ and × represent liveness and spoofness, respectively. The datasets are represented as follows: ○, ×: CASIA, ○, ×: Idaip Replay-Attack, ○, ×: MSU-MFSD, ○, ×: OULU-NPU.



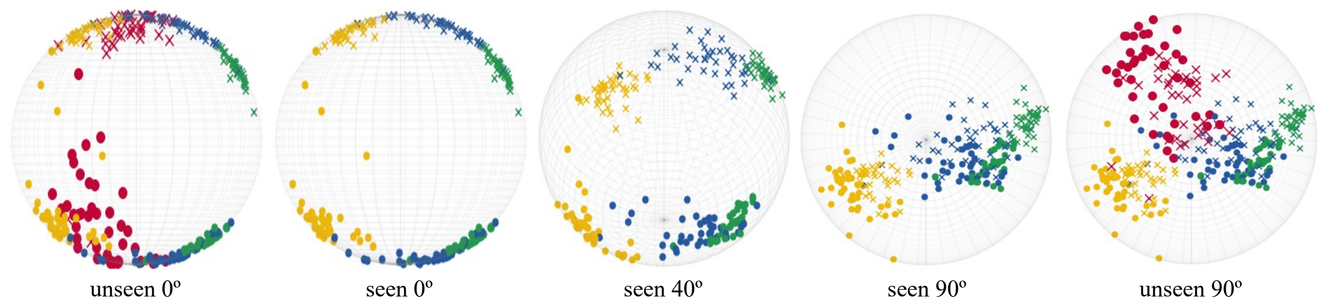
(a) OCI (Invariant) (b) OMI (Invariant) (c) OCM (Invariant) (d) ICM (Invariant)

(e) OCI (Specific) (f) OMI (Specific) (g) OCM (Specific) (h) ICM (Specific)

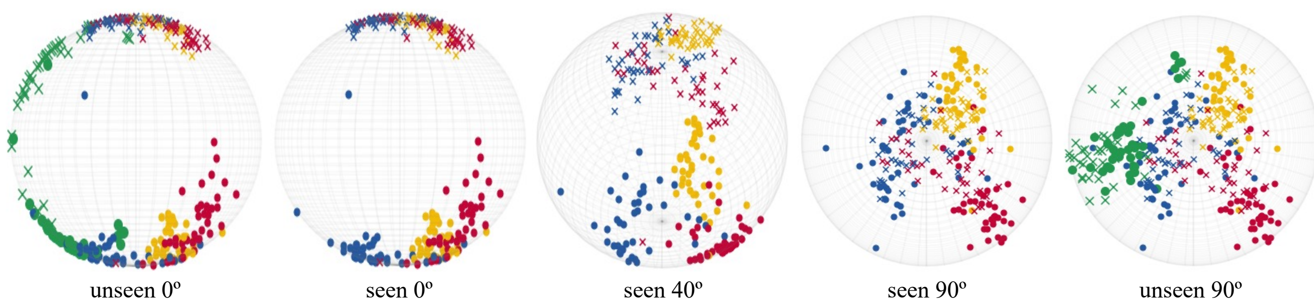
Figure B. **Decomposition into invariance and specificity for domain**: 2D PCA Visualization across domains. The term (Invariant) refers to features projected onto the invariant space, while (specific) refers to features projected onto the domain-specific separable space. ○ and × represent liveness and spoofness, respectively. The datasets are represented as follows: ○, ×: CASIA, ○, ×: Idaip Replay-Attack, ○, ×: MSU-MFSD, ○, ×: OULU-NPU.



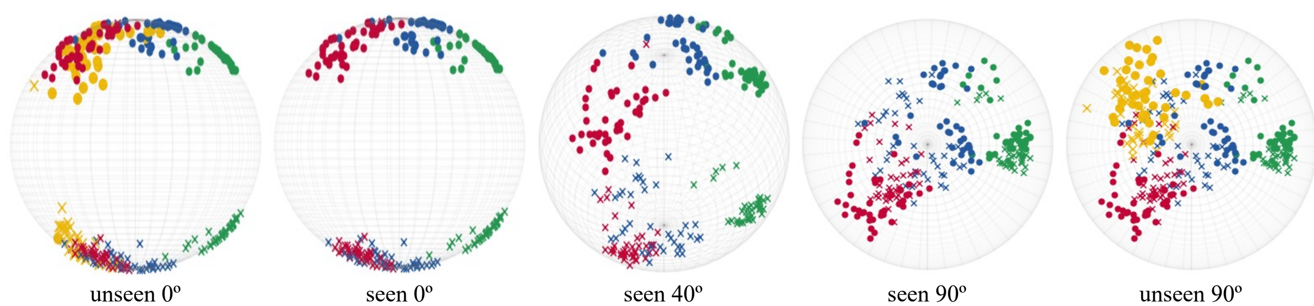
(a) OCI→M



(b) OMI→C

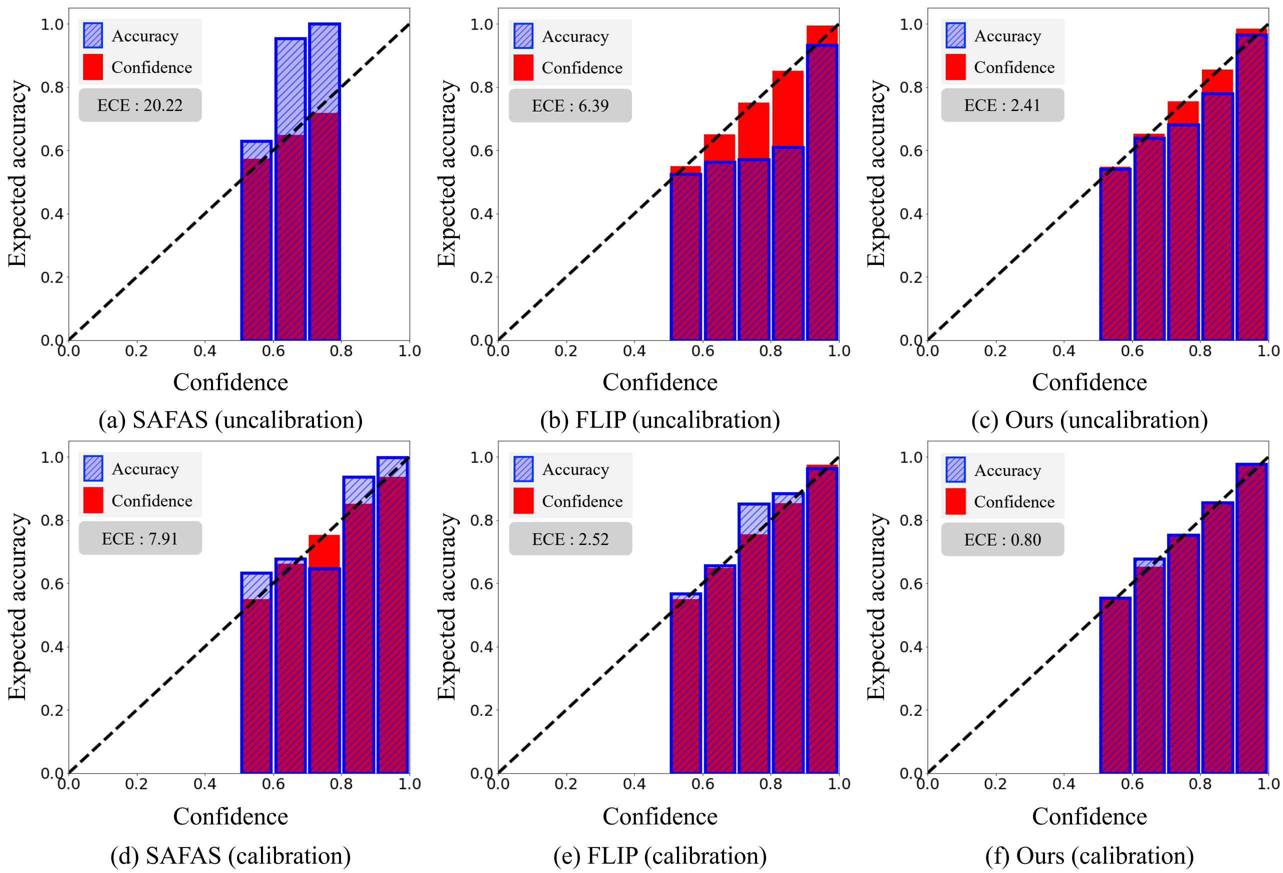


(c) OCM→I



(d) ICM→O

Figure C. **3D visualization projected on sphere**: PCA 3D Visualization across domains. The terms (0º), (40º), and (90º) represent the angles of rotation about the vertical axis. ○ and × represent liveness and spoofness, respectively. The datasets are represented as follows: ○, ×: CASIA, ○, ×: Idaip Replay-Attack, ○, ×: MSU-MFSD, ○, ×: OULU-NPU.



(a) SAFAS (uncalibration) (b) FLIP (uncalibration) (c) Ours (uncalibration)

(d) SAFAS (calibration) (e) FLIP (calibration) (f) Ours (calibration)

Figure D. **Reliable Diagram**: Comparison with Previous Methods. The terms (uncalibration) and (calibration) are distinguished based on whether the uncertainty has been calibrated.

Methods	C→I	C→M	C→O	I→C	I→M	I→O	M→C	M→I	M→O	O→C	O→I	O→M
FLIP-MCL† [27]	10.57	7.15	3.91	0.68	7.22	4.22	0.19	5.88	3.95	0.19	5.69	8.40
BUDoPT† [22]	4.33	2.62	4.98	0.48	1.83	4.14	**0.00**	2.45	1.87	0.44	2.53	1.43
TF-FAS† [34]	3.06	**1.59**	3.78	0.69	**1.34**	**2.50**	0.11	2.31	**1.40**	1.5	4.02	1.59
GD-FAS†	**2.98**	3.92	**3.36**	**0.19**	2.50	3.36	**0.00**	**2.23**	2.59	**0.00**	**1.83**	**0.42**

Table A. **Leave-three-domains-out protocol: † indicates the use of an extra source dataset (CelebA-Spoof [36]).**

cause model reliability is often expressed through the probability values associated with predictions; however, recent AI models have demonstrated a tendency toward overconfidence [11]. Unfortunately, this critical aspect has been largely overlooked in prior studies.

In this section, we aim to underscore the importance of model reliability and delineate how our approach differs from previous methods by discussing the Expected Calibration Error (ECE) and the reliability diagrams introduced in Sec 4.1 of this paper.

A reliability diagram is an analytical tool that visually displays the discrepancy between predicted probability values and the actual probabilities. We conduct a visual analysis of the dataset both with and without uncertainty calibration, utilizing a post-hoc uncertainty calibration method that employs the temperature scaling technique. For this purpose, we split the test dataset into a validation set and a test set with a 2:8 ratio. The validation set is used to calibrate the reliability, while the test set is used to evaluate.

As shown in Fig. D, our method maintains a slightly overconfident state, whereas FLIP is significantly overconfident and, conversely, SAFAS is underconfident. In the calibrated diagrams below, the slightly under and overconfident regions are notably reduced. Here, we observe that our method before calibration outperforms other methods even after calibration.

Methods		SSAN-R [35]	SSDG-R [13]	DGUA-FAS [12]	BUDoPT [22]	GD-FAS
MI→C	HTER	25.56	19.86	19.22	5.33	**2.22**
	AUC	83.89	86.46	86.81	98.92	**99.15**
MI→O	HTER	24.44	27.92	20.05	5.94	**3.75**
	AUC	82.86	78.72	88.75	98.37	**98.75**

Table B. **Leave-two-domains-out protocol**

Methods	CS→W		SW→C		CW→S		Average	
	HTER	AUC	HTER	AUC	HTER	AUC	HTER	AUC
FLIP-MCL† [27]	4.46	**99.16**	9.66	96.69	11.71	95.21	8.61	97.02
CFPL† [19]	**4.40**	99.11	8.13	96.70	8.50	97.00	7.01	97.60
GD-FAS†	7.88	97.60	**4.01**	**98.59**	**5.87**	**98.04**	**5.92**	**98.08**

Table C. **Leave-one-domain-out protocol with different datasets**: † indicates the use of an extra source dataset (CelebA Spoof)

C. Additional Quantitative Experiments

We conducted experiments under more challenging settings, including the leave-two-domain-out (Sec. C.1) and leave-three-domain-out (Sec. C.2) protocols. Additionally, we evaluated our method on a different DGFAS dataset to assess its generalizability beyond the original benchmark(Sec. C.3). We evaluate all experiments using two standard metrics: Half Total Error Rate (HTER) and Area Under the ROC Curve (AUC).

C.1. Leave-three-domains-out protocol

Tab. A presents the results under the leave-three-domains-out protocol, where the model is trained on a single domain and evaluated on the remaining unseen domains. Our method, GD-FAS, consistently outperforms existing state-of-the-art approaches across most domain shifts. Notably, it achieves the best performance in particularly challenging scenarios such as O→C and M→C(HTER: 0.00%) and I→C (HTER: 0.19%). These results highlight the strong generalizability of our approach across highly diverse and difficult domain pairs.

C.2. Leave-two-domains-out protocol

Tab. B presents the performance under the leave-two-domains-out protocol, where the model is trained on two domains and evaluated on the remaining unseen domains—consistent with the experimental setup used in BUDoPT [22]. In both MI→C scenarios, GD-FAS achieves the lowest HTER (2.22% and 3.75%) and the highest AUC (99.15% and 98.75%), significantly outperforming prior methods such as BUDoPT [22], DGUA-FAS [12], and SSDG-R [13]. These results demonstrate the effectiveness of our group-wise scaling and orthogonal decomposition techniques in capturing domain-invariant features, even under limited training conditions and complex inter-domain shifts.

C.3. Different dataset

Our experiments were conducted on three diverse datasets—Surf (S) [9], CeFA (C) [8], and WMCA (W) [7]—under the leave-one-domain-out protocol to evaluate cross-dataset generalization.

As shown in Tab. C, while FLIP-MCL and CFPL achieve high AUC scores, GD-FAS demonstrates the most balanced performance, achieving the best average results across all domains (HTER: 5.92%, AUC: 98.08%). Notably, in the SW→C setting, GD-FAS outperforms all baselines with a significantly lower HTER (4.01%) and higher AUC (98.59%), highlighting its robustness against challenging cross-domain shifts.

D. Discussion of Computational Cost

In this section, we discuss the computational cost of GD-FAS.

Training Efficiency The additional computational overhead introduced by the GS-RM and FOD losses is approximately 10M FLOPs, which is negligible compared to the overall training complexity of FLIP-MCL (88.6G FLOPs). GD-FAS also adopts the same backbone architecture as FLIP-MCL and simply replaces the original MSE loss with the proposed GS-RM and FOD losses. As a result, GD-FAS maintains a comparable training cost to FLIP-MCL. In contrast, prior methods often rely on significantly more resource-intensive components, such as large language models (LLMs) or two-stage training pipelines.

Inference Efficiency GD-FAS achieves an inference time of 0.01 seconds per frame, matching the computational cost of FLIP-MCL due to their shared inference pipeline. In comparison, other baseline methods typically incur slightly higher computational overhead and latency.

Prediction	OCI→M HTER / AUC / ECE	OMI→C HTER / AUC / ECE	OCM→I HTER / AUC / ECE	ICM→O HTER / AUC / ECE
$\mathbf{w_I} \otimes \mathbf{w_T}$	**0.42 / 99.88 / 3.39**	**0.93 / 99.99 / 1.82**	**3.33 / 98.61 / 2.37**	**2.64 / 99.34 / 2.86**
Φ_C	2.50 / 99.44 / 5.02	2.22 / 99.71 / 5.15	4.83 / 98.75 / 3.51	3.33 / 99.28 / 2.41

Table D. **Embedding Features ($\mathbf{w_I} \otimes \mathbf{w_T}$) vs. Classifier ($\phi_C$)**

batch size	OCI→M HTER / AUC / ECE	OMI→C HTER / AUC / ECE	OCM→I HTER / AUC / ECE	ICM→O HTER / AUC / ECE
3	5.00 / 98.77 / 6.38	3.15 / 99.24 / 15.81	6.67 / 98.72 / 2.24	2.96 / 99.24 / 9.41
8	2.50 / 99.85 / 3.60	2.22 / 98.84 / 8.67	5.00 / 98.94 / 2.78	3.01 / 99.20 / 5.41
16	**0.42** / 99.88 / 3.39	**0.93** / **99.99** / 1.82	**3.33** / 98.61 / 2.37	**2.64 / 99.34 / 2.86**
24	**0.42** / **99.90** / 2.79	**0.93** / 99.86 / 0.80	**3.33** / **99.25** / 2.79	2.96 / **99.34** / 5.77
32	**0.42** / 99.73 / **2.53**	1.11 / 99.76 / **0.72**	3.50 / 99.24 / **1.84**	3.01 / 99.07 / 6.72

Table E. **Influence of Batch-size**

Method	Target	OCI→M	OMI→C	OCM→I	ICM→O
Gaussian Noise	Image	7.5	3.52	3.5	4.77
Random Vector	Text	7.08	4.44	11.83	7.36
a Pohoto of {dog or cat}	Text	3.3	12.04	5.60	15.09
a Photo of {live of fake} face	Text	2.5	1.11	5.17	4.07
Baseline (FLIP-MCL)	None	**0.42**	**0.93**	**3.33**	**2.64**

Table F. **Robustness across Image and Text Quality**

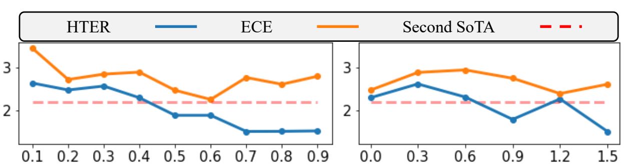
E. Classifier vs Embedding Features

In this section, we explain why GD-FAS detects spoofing attacks using embedding features rather than relying on the classifier output, as illustrated in Fig. 2. As shown in Tab. D, predictions based on embedding features outperform those based on the classifier. This performance gap arises because the embedding space explicitly separates domain-invariant and domain-specific information through orthogonal decomposition. In contrast, the classifier operates on features prior to this decomposition, failing to fully exploit the decomposed representations.

F. Analysis of Batch-Size Influence

We analyze the influence of batch size on model performance. CLIP-based methods typically use a batch size of 3 per domain due to the large model size, resulting in an effective batch size equal to the per-domain batch size multiplied by the number of domains.

As shown in Tab. E, a batch size of 16 is sufficient for stable and effective training. Increasing the batch size beyond 16 yields marginal gains while requiring GPU memory in excess of 48GB, which may not be practical for most setups.



(a) Hyperparameter λ_1 (b) Hyperparameter β

Figure E. **Hyperparameter Sensitivity**: Default λ_1=0.8, β=1.5

G. Robustness across Image and Text Quality

We analyze the robustness of GD-FAS to variations in both image and text quality. GD-FAS utilizes the same text templates as FLIP-MCL, comprising six distinct templates for each class—spoof and live—to guide representation learning. To evaluate robustness, we conduct experiments under four distinct conditions: one related to image quality and three to text quality. For the image case, we apply Gaussian noise to the input images to simulate visual degradation. For the text cases: (1) we replace text embeddings with random vectors, (2) we use only a single incorrect-class template, and (3) we use only a single correct-class template.

As descriebed in Tab. F, while both image and text degradations lead to moderate performance drops, degradation in text quality—particularly the use of random vectors or incorrect-class templates—results in a more substantial decline. These findings suggest that text embeddings play a crucial role in enabling GD-FAS to extract domain-invariant features effectively.

H. Sensitivity of Hyperparameter

We performed a series of analytical experiments to examine how FOD and GS-RM contribute to overall performance, as shown in Fig. E. The hyperparameters λ_1 and β correspond to FOD and GS-RM, respectively. While our method exhibits slight sensitivity to hyperparameter variations, the combined use of FOD and GS-RM consistently improves performance as their contributions are strengthened. We selected the final hyperparameter values based on HTER, as it directly reflects the model's domain generalization ability.