Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Domain Generalization for Face Anti-Spoofing with Joint Frequency-Domain Phase and Spatial-Domain Modulation

weihua liu
Quanzhou University of Information Engineering
tongming wan
state Grid Jiangxi Information Telecommunication Company
shu liu
sliu35@csu.edu.cn
Central South University

Research Article

Keywords: Face anti-spoofing, domain-invariant features, domain generalization, spectrum domain, frequency filter module

Posted Date: November 19th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11777889/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Domain Generalization for Face Anti-Spoofing with Joint Frequency-Domain Phase and Spatial-Domain Modulation

Weihua Liu, *Member, IEEE*, Tongming Wan, Shu Liu, *Member, IEEE*

***Abstract*—Most existing domain generalization methods for face anti-spoofing rely only on spatial-domain features, making them vulnerable to performance drops when facing different environments, attack materials, or camera devices. To solve this problem, we propose a new Spatial and Spectrum Dual-Domain Generalization (SSDDG) approach that effectively combines both spatial and frequency information. Our key insight reveals that phase components in the frequency domain inherently provide stable, domain-invariant features. Building upon this discovery, we develop the SSDDG framework comprising three core components: (1) A frequency branch with an adaptive filter to extract and adjust important phase features, (2) A spatial branch using multi-scale transformers to identify live face characteristics, and (3) A multi-scale interpolation alignment module operates on both spatial and spectral domains to achieve precise feature alignment across multiple scales. We also design two auxiliary supervisions - an interaction loss for cross-domain feature fusion and an adversarial loss for compact live feature distribution learning. Extensive experiments show that our SSDDG approach is effective, competitive with recent domain generalization methods on four public datasets, and particularly effective in limited-source transfer without target-domain access. The code will be made publicly available upon acceptance of the paper.**

***Index Terms*—Face anti-spoofing, domain-invariant features, domain generalization, spectrum domain, frequency filter module**

I. Introduction

FACE recognition technology [1] has become an indispensable component of computer vision, offering substantial convenience and security enhancements for commercial applications while driving economic growth and innovation. However, the widespread adoption of this technology has also introduced significant security vulnerabilities. Specifically, the continuous evolution of face spoofing techniques (high-quality prints, video replays, 3D masks) poses escalating threats to recognition systems. To address these challenges, Face Anti-Spoofing (FAS) has emerged as a critical research direction, thereby sparking renewed investigation into robust countermeasures against presentation attacks.

FAS research can be broadly categorized into traditional handcrafted methods and deep learning-based approaches. In the early stages, researchers primarily relied on traditional feature descriptors such as Local Binary Patterns (LBP) [2], [3] and histogram of oriented gradients [4], [5]. These methods exploit the textural differences between live faces and spoofing attacks; however, their effectiveness is limited by their dependence on expert knowledge and inherent unreliability. With the advancement of FAS techniques and the availability of large-scale datasets, deep learning-based methods have gained significant attention [6]–[15]. Furthermore, multimodal approaches [16] incorporating auxiliary features such as depth maps [17] and near-infrared (NIR) spectrum [18] have been developed to enhance generalization performance and robustness. The underlying principle of depth-based methods lies in the distinct depth variations between live and fake faces. For instance, in a frontal view, the nose region of a live face typically exhibits greater proximity to the camera compared to other facial regions, whereas printed photos maintain uniform depth. NIR-based methods effectively utilize reflectance differences between live and spoof faces, although their imaging quality degrades at longer distances [11].

Deep learning-based FAS faces two primary challenges. First, models tend to overfit training datasets, leading to poor generalization in real-world deployment. This occurs because they learn dataset-specific discriminative features that may not transfer well to unseen data [19]. Second, while existing methods achieve strong performance in intra-dataset evaluations, they often fail to maintain this accuracy in practical applications. This performance gap stems from significant domain shifts between source (training) and target (testing) environments, caused by variations in imaging devices, lighting conditions, and spoofing materials.

To address these challenges, researchers have employed both Domain Generalization (DG) and Domain Adaptation (DA) strategies in FAS. DG methods [20], [21] focus on learning domain-invariant features without target domain access, utilizing techniques like domain adversarial learning [13] and invariant risk minimization [22] to extract cross-domain representations, along with data augmentation [23], meta-learning [21], and semi-supervised approaches to enhance

This work was supported by the National Natural Science Foundation of China under Grant U24A20333, Natural Science Foundation of Fujian Province under Grant 2025J011608, Hunan Provincial Natural Science Foundation of China under Grant 2023JJ30700, and the Fundamental Research Funds for the Central Universities of Central South University. We are grateful for resources from the High Performance Computing Center of Central South University. (*Corresponding authors: Shu Liu*)

Weihua Liu is with the School of Electronics and Communication Engineering, Quanzhou University of Information Engineering, Quanzhou 362000, China. (E-mail: lwh86117@163.com)

Tongming Wan is with State Grid Jiangxi Information Telecommunication Company, Nanchang 330000, China, and also with the School of Computer Science and Engineering, Central South University, Changsha 410083, China. (E-mail: wtm114@csu.edu.cn)

Shu Liu are with the School of Computer Science and Engineering, Central South University, Changsha 410083, China (email: dwq_csu@csu.edu.cn; sliu35@csu.edu.cn)

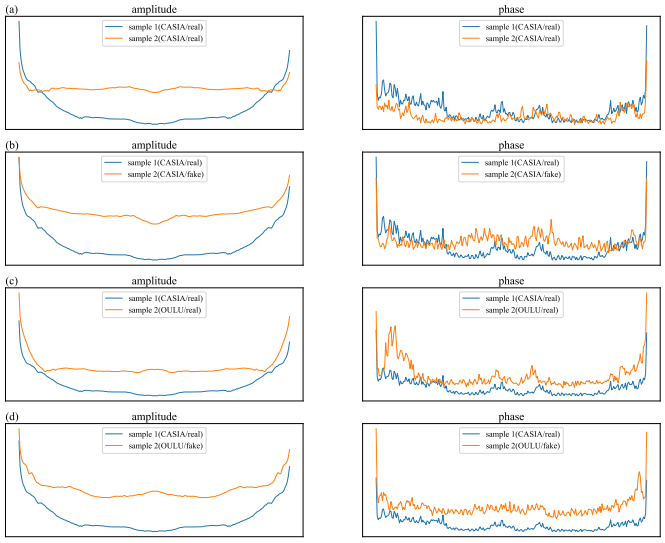


Fig. 1: Frequency spectrum information diagram on live and spoofing samples. The notation (e.g., "sample 1 (CASIA/real)") indicates sample type (live/fake) and dataset origin. It shows that the phase components exhibit remarkable consistency across domains for both live and spoofed faces, while amplitude information shows significant domain-specific variations.

generalization. In contrast, DA methods [24], [25] leverage available target domain data during training to specifically adapt models to new environments, offering more precise but less flexible solutions compared to DG approaches.

Most FAS methods based on DG or DA extract representative features from RGB images. DIFEX [26], known for image classification, demonstrated that the phase information obtained through Fourier transform can effectively represent the domain-invariant features, thereby enhancing the generalization ability across unseen domains. Based on this, we visualize the frequency spectrum to further verify if the phase information is truly effective in extracting domain-invariant features in the face anti-spoofing task. Four pairs of images from the CASIA [27] and OULU [28] datasets are randomly selected for comparison experiments in this paper.

As demonstrated in Fig. 1, our spectral analysis reveals a critical finding: while amplitude characteristics show significant domain-specific variations in both live and spoofed faces (panels a-d), the phase components maintain remarkable cross-domain consistency. This key observation establishes Fourier phase features as particularly valuable for face anti-spoofing, as they inherently preserve domain-invariant liveness characteristics that remain robust against domain shifts. The figure illustrates this through four comparative scenarios: (a) intra-dataset live samples, (b) intra-dataset live vs. spoofed samples, (c) cross-dataset live samples, and (d) cross-dataset live vs. spoofed samples - all consistently showing phase stability versus amplitude variability.

Motivated by this discovery, we propose a Spatial-Spectral Dual-Domain Generalization (SSDDG) framework that strategically leverages this phase consistency property. Our approach systematically: (1) decomposes facial images into complementary spatial and spectral representations, (2) employs an adaptive frequency filtering mechanism to amplify domain-invariant phase channels containing essential liveness cues, and (3) intelligently fuses these enhanced spectral features with spatial features through our novel alignment modules. To enhance cross-domain feature alignment and generalization, we introduce two complementary mechanisms: (1) an approximation loss that explicitly minimizes the feature distance between spatial and spectral branches to enforce domain-invariant representation learning, and (2) domain adversarial learning applied to the fused features to further improve generalization by enforcing compact feature distributions that are indistinguishable across domains. This dual-optimization strategy effectively bridges the gap between spatial and frequency domains while maintaining discriminative power for live/spoof classification. The main contributions of this work are summarized as follow:

- We propose a novel SSDDG framework that effectively integrates spatial and frequency domain information by leveraging their complementary strengths, motivated by the observation that phase components contain intrinsic domain-invariant liveness cues for face anti-spoofing.
- For spectral processing, we design an Adaptive Frequency Modulation (AFM) module that dynamically filters phase information while implementing domain-aware feature recalibration, enhancing the model's ability to generalize across domains through weighted fusion of source-domain invariant features.
- We introduce a Multi-scale Interpolation Alignment (MIA) module that establishes cross-domain feature correspondence through dual-branch multi-scale alignment, reinforced by two complementary supervision mechanisms: approximation loss for cross-domain feature consistency, and adversarial learning for compact live feature distribution modeling.

II. Related Work

A. Face Anti-spoofing Methods

FAS methods include traditional handcrafted feature based and deep learning based. According to face feature attributes, handcrafted features mainly focus on skin texture, optical flow motion, physiological signals, etc. The texture-based approach lies in the differences in texture details between live and fake face images, due to multiple device flippings. Boulkenafet et al. [29] extracted color LBP features from faces and combined LBP histograms of each color channel for classification. In addition, the input of continuous frames can also capture optical flow motion or physiological signals. The physiological signal-based approach is a new biomedical technology named remote photo2plethysmo graphy (rPPG) [30]. Dense optical flow [31] was utilized to extract motion-based liveness features. Yu et al. [32] constructed spatio-temporal rPPG maps from face videos, and applied the transformer to fully explore the global relationship for liveness representation.

Yang et al. [6] first used a deep convolutional neural network to extract features, and then used a support vector machine for

binary classification to solve the problem. Xu et al. [7] used temporal features among multiple frames and local and dense features in convolutional operations with a novel LSTM-CNN network to discriminate between live and fake faces. Cai et al. [8] combined deep reinforcement learning to learn local features from image patches and proposed a novel CNN-RNN network to respectively extract local and global features for classification. PatchNet [33] recast FAS as a fine-grained patch recognition problem, with the corresponding fine-grained attack classes based on prosthetic materials and capture devices. It supervised the model to learn a more compact distribution in live samples, while ensuring the spoof samples more dispersed in the feature space and similarity of the two view features. PDLE [34] treated the FAS problem as a regression task. It designed a new labeling technique for generating discretized labels that show the amount of information associated with the live image, and then predicted the liveness score using the regression network. Liu et al. [17] presented a unified framework for face anti-spoofing, including both attack and defense systems. The attack system combined live faces to generate attack samples with known patterns, and then forced the defense system ineffective via adversarial training.

Some literatures have been reported to use Fourier transforms to deal with the FAS task. Roy et al. [35] hypothesized that, compared to a live sample, the 2D Fourier spectra of an attack sample have fewer high-frequency components; thus an auxiliary branch of 2D Fourier spectra was added. The results indicated that generating Fourier features is less computationally expensive than depth features, facilitating faster computation during the training stage. Jourabloo et al. [36] proposed a novel idea that a spoof face can be decomposed into a spoof noise and a live face. They performed Fast Fourier Transform (FFT) on the spoof noise and computed the maximum value in the high-frequency band, so as to detect the presence of attack information. GDA [25] employed a data augmentation method called spectrum mixup in intra-domain to further expand target data distributions, ensuring better generalization and smaller intra-domain gap. The above methods leverage Fourier features either as an auxiliary supervision or as a data augmentation method to enhance the model generalization performance. However, the phase information in the Fourier transform has not been specifically investigated, which carries high-level semantics. Instead, we extract the Fourier phase as domain-invariant features in this work.

B. Multi-modal based Face Anti-spoofing Methods

To facilitate face anti-spoofing research, despite the limitations of subjects and modalities, Zhang et al. [37] introduced a large-scale multi-modal dataset, namely CASIA-SURF, and achieved a comprehensive representation by combining RGB, depth and infrared (IR) modal features in a three-branch architecture. George and Marcel [9] introduced a DenseNet architecture with separate RGB and Depth channels to address face anti-spoofing, and proposed a cross-modal focal loss function to supervise the individual channels. Yu et al. [38] extended Central Difference convolution to a multi-modal version aimed at capturing inherent deception patterns between different modalities. Meanwhile, combined with the ViT-based backbones, MMDG [39] proposed the uncertainty-guided crossadapter and the rebalanced modal gradient modulation strategies to solve the modal unreliability problem and the modal imbalance problem.

C. Domain Generalization/Domain Adaptation

Although substantial progress has been achieved in intra-domain FAS, challenges remain in real applications. Due to the unpredictable nature of photographic equipment, lighting conditions or attack materials, the domain shift between the source and target domains often exists, so that a trained model fails to generalize well to the unseen domain.

Recent DG-FAS studies have further explored robust optimization, frequency shortcuts, and semantic consistency. GAC-FAS [40] aligns domain-wise gradient updates at ascending points with empirical-risk-minimization updates to encourage convergence to a flatter minimum. FSV-FAS [41] analyzes frequency-shortcut learning in FAS and introduces frequency-aware reconstruction, dynamic frequency masking, and style-inhibited modulation to prevent over-reliance on narrow frequency bands. DCRN [42] employs dense semantic correspondence and dual consistency regularization for fine-grained feature alignment, while DSCM-FAS [43] models both inter-sample semantic relations through a graph and intra-sample consistency through patch-level regularization. These methods establish optimization geometry, frequency-bias mitigation, and semantic consistency as important directions for DG-FAS. They do not, however, explicitly exploit the cross-domain stability of Fourier phase for joint multi-scale spatial-spectral feature alignment. The proposed SSDDG framework is therefore complementary to these studies: rather than only suppressing frequency shortcuts or enforcing semantic consistency, it preserves phase-based structural cues and aligns them with multi-scale spatial representations.

DG [10] posits the existence of a universal feature space that encompasses both the observed multiple source domains and the unobserved yet associated target domain. Within this space, the model learned from the observed source domains is capable of effectively generalizing to the unobserved domain [11]. DA [25] utilizes a small portion of the available data in the target domain to reduce the distribution differences between the source and target domains. Most DG/DA methods align the feature space between source and target domains to deal with domain shift, which have achieved some results on FAS in cross-domain scenarios.

Jia et al. [13] introduced an innovative end-to-end framework for single-side domain generalization, termed SSDG, aimed at augmenting the generalization capabilities of the model. Within this framework, they devised a single-side adversarial learning mechanism and an asymmetric triplet loss function, meticulously crafted to disperse the features of fake faces within the feature space while compacting those of live faces. This strategic design facilitates the achievement of a more discernible class boundary, thereby enhancing the model's efficacy in generalizing to the target domain. Kim et al. [14] proposed a test-time adaptive model and designed a

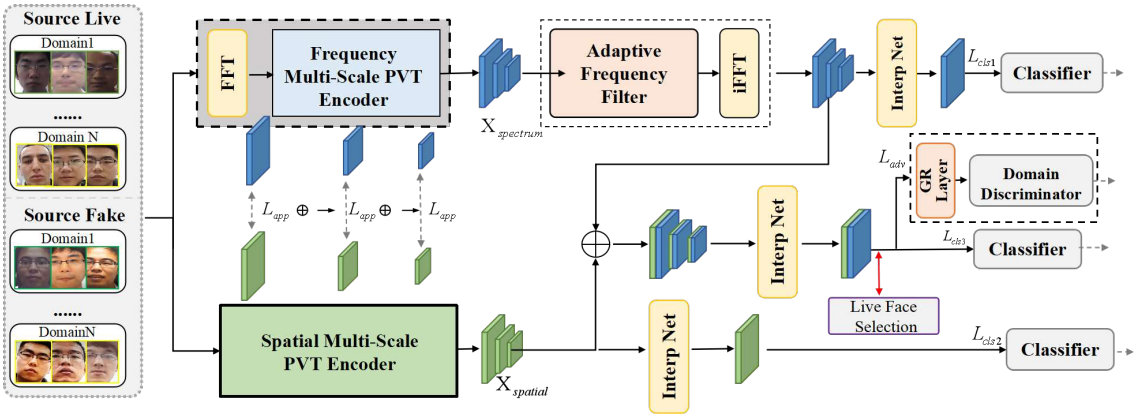


Fig. 2: An overview of the proposed SSDDG method. The input faces adorned with distinct color borders serve as exemplars from diverse domains. Different background colors represent different branches.

style selective normalization to identify the style of individual domains. SDA [24] explored a domain adaptor based on meta-learning, which operates solely within the target domain. It leverages self-supervised learning to generate pseudo-labels and then refines them iteratively through a domain-specific network. Liu et al. [44] first obtained a generalized feature space with clustered live faces and dispersed fake faces, then trained a domain agnostic classifier by a low-rank decomposition. Drawing inspiration from the theoretical error bounds in domain generalization, Wang et al. [23] proposed negative data augmentation to synthesize fake faces using live faces for training, without requiring real-world attack samples. They also designed a global loss and a local loss for learning the generalized feature space, respectively. In this paper, we intend to transform the original RGB image to spectrum domain and then derive the domain-invariant representations, which are general and transferable to different domains, thus contributing to the cross-domain FAS task.

III. Proposed Method

Due to the diversity of devices, lighting conditions, and attack materials, there is a huge domain shift between the source and target domains. Consequently, discovering features that remain consistent across domains becomes crucial in addressing this issue. We believe that the phase information, which encompasses high-level semantic details, can serve as a domain-invariant feature to enhance the model's proficiency in distinguishing real faces from counterfeit ones, ultimately improving its overall performance.

A. Overview

To minimize domain shift and enhance the extraction of domain-invariant features for face anti-spoofing, we propose a dual-branch domain generalization framework that simultaneously captures live cues in both spatial and spectral domains (see Fig. 2).

Our model architecture consists of two parallel branches designed to separately process spatial and spectral information, which are subsequently fused to achieve robust domain-invariant representations. For comprehensive face representation across both domains, we employ a Pyramid Vision Transformer (PVT) as our multi-scale backbone network, as shown in Fig. 3. The PVT encoder extracts hierarchical features at three scales: $F_1 : \frac{H}{4} \times \frac{W}{4}(C_1)$, $F_2 : \frac{H}{8} \times \frac{W}{8}(C_2)$ and $F_3 : \frac{H}{16} \times \frac{W}{16}(C_3)$. The framework then performs hierarchical feature fusion by computing differential representations between corresponding spatial and spectral layers. This approach effectively reveals multi-scale liveness cues embedded in different frequency bands while maintaining spatial coherence.

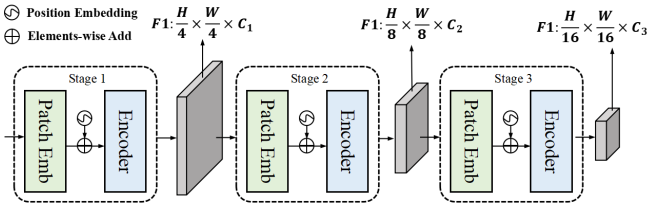


Fig. 3: Illustration of multi-scale PVT encoder.

As illustrated in Fig. 2, the spectrum branch extracts domain-invariant features through a frequency-based multi-scale PVT encoder and a frequency filter, which enhance intrinsic liveness cues by suppressing irrelevant facial features in the frequency domain. Specifically, we first apply 2D FFT to RGB face images to derive phase components from their frequency-domain representations. The PVT encoder then processes these components, extracting multi-scale features, denote as $X_{\text{freq}} = \text{Concat}(f^{\text{f}}_{\text{low}}, f^{\text{f}}_{\text{mid}}, f^{\text{f}}_{\text{high}})$ from three hierarchical stages. These features are further refined by the proposed frequency filtering module, which amplifies discriminative cues while suppressing noise, thereby improving generalization. Moreover, after reconstructing the features into the original latent space via 2D inverse FFT (iFFT), we further align the multi-scale representations using an interpolation network. Finally, a binary cross-entropy loss is applied for frequency-domain FAS classification.

In the spatial branch, facial liveness cues are extracted by a PVT encoder in the spatial domain, yielding multi-scale features denoted as $X_{\text{spatial}} = \text{Concat}(f^{\text{s}}_{\text{low}}, f^{\text{s}}_{\text{mid}}, f^{\text{s}}_{\text{high}})$. Similar to the frequency branch, an interpolation network is employed to align these multi-scale features, enabling effective

classification in the spatial domain.

To ensure robust generalization across both spatial and spectral domains, we introduce an approximation loss to enforce consistency between the learned representations of the two branches. The outputs of these branches, X_{spatial} and X_{freq}, are fused along the channel dimension through concatenation: $X_{\text{fused}} = \text{Concat}\,(X_{\text{spatial}}, X_{\text{freq}}) \in \mathbb{R}^{C\times H\times W}$, where $C = C_s + C_f$ denotes the total channel dimension, and H, W represent spatial dimensions. To further enhance domain-invariant feature learning, we employ an adversarial loss on separately selected live face features. This mitigates domain-specific variations and improves generalization to unseen domains.

During inference, the fused features from both spatial and spectral branches are used for classification. By integrating spectral features with spatial representations, our proposed SSDDG method enhances the learning of intrinsically domain-invariant features, thereby improving generalization to unseen domains.

B. Spectrum-Branch Feature learning

Given that Fourier phase components are known to preserve global semantic structures and exhibit strong robustness to domain shifts in visual data, we propose a novel frequency-domain learning framework for face anti-spoofing. Our spectrum branch consists of four key components: Frequency Domain Transformation, Phase Extraction, Adaptive Frequency Modulation, and Spatial Domain Reconstruction. The process begins by transforming the input RGB image into the frequency domain via FFT. For a given intermediate feature map $X = \mathbb{R}^{C\times H\times W}$, its frequency-domain representation $X_F = \mathbb{R}^{C\times H\times W}$ is computed using a channel-wise FFT operation:

$$\mathrm{X}_F(u,v) = \sum_{h=0}^{H-1}\sum_{w=0}^{W-1} \mathrm{X}(h,w) e^{-j2\pi(\frac{h}{H}u + \frac{w}{W}v)} \quad (1)$$

where u and v denote frequency domain coordinates and h and w represent spatial positions. The complex-valued X_F contains both real $\mathcal{R}(\mathbf{X})$ and imaginary $\mathcal{I}(\mathbf{X})$ components. To extract domain-invariant features, we focus on the Fourier phase spectrum, which is computed per frequency channel as:

$$\mathbf{X}_{freq}(u,v) = \mathcal{P}(\mathbf{X})(u,v) = \arctan\left[\frac{\mathcal{I}(\mathbf{X}_F)(u,v)}{\mathcal{R}(\mathbf{X}_F)(u,v)}\right]. \quad (2)$$

This phase representation captures global latent texture information while being inherently robust to appearance variations across domains. Subsequently, X_{freq} is processed by our Adaptive Frequency Modulation Module, which learns to emphasize frequency components that contain salient liveness cues while suppressing domain-specific artifacts. Finally, the processed spectrum is transformed back to the spatial domain via the Inverse Fast Fourier Transform (iFFT), yielding the reconstructed feature map:

$$\mathrm{X}(h,w) = \frac{1}{H\cdot W}\sum_{h=0}^{H-1}\sum_{w=0}^{W-1} \mathrm{X}_F(u,v) e^{j2\pi(\frac{h}{H}u + \frac{w}{W}v)} \quad (3)$$

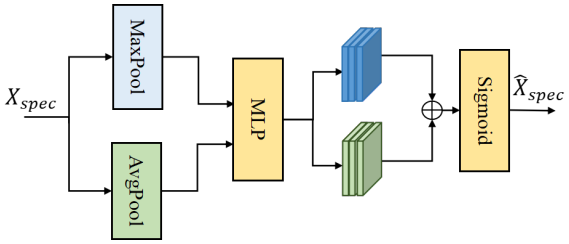


Fig. 4: Illustration of proposed adaptive frequency modulation.

The reconstructed spatial features are subsequently processed by the Interpolation Network to achieve cross-scale feature alignment. This enables the network to simultaneously learn both fused representations and independent live cues, thereby enhancing generalization capability and improving FAS performance on unseen domains.

Adaptive Frequency Modulation Module. The proposed AFM module functions as an adaptive spatial mask designed to preserve frequency components containing salient liveness cues through channel-wise weight recalibration, as illustrated in Fig. 4. Unlike conventional filtering approaches, our module incorporates a squeeze-and-excitation block [45] to achieve three key advantages: (1) effective aggregation of global information, (2) identification of channel-wise interdependencies, and (3) enhancement of representational capacity. The module's innovative design utilizes Fourier phase components as input to the channel attention mechanism, enabling simultaneous preservation of both high-level semantic features and low-level texture/noise characteristics through intelligent weight redistribution in the Fourier phase domain. Furthermore, the integrated channel attention mechanism significantly improves the network's ability to extract and utilize liveness information across diverse source domains, thereby enhancing both feature utilization efficiency and discriminative power.

Multi-scale Interpolation Alignment Module. The proposed MIA module, as shown in Fig. 5 is an architecture-agnostic component designed to be seamlessly integrated into the SSDDG pipeline. It aims to unify and spatially align multi-scale features F_1, F_2, F_3 output by the preceding AFM module. Specifically, MIA first normalizes the channel dimensions of these features using 1×1 convolutions. To align their spatial resolutions, a hybrid depthwise-separable upsampling strategy—combining learnable transposed convolutions with bilinear interpolation—is applied to F_2, F_3. The upsampled features are then element-wise summed. Finally, the fused feature is refined by a spatial alignment module, as shown in Fig. 6, consisting of grouped spatial attention followed by deformable convolution, which enhances spatial coherence and adaptability to geometric variations.

C. Spatial-Spectrum Feature Fusion

The spatial branch adopts a straightforward design. In this independent branch, both live and spoofing face images are directly processed by a PVT for multi-scale feature encoding. These hierarchical features are then aligned and fused through

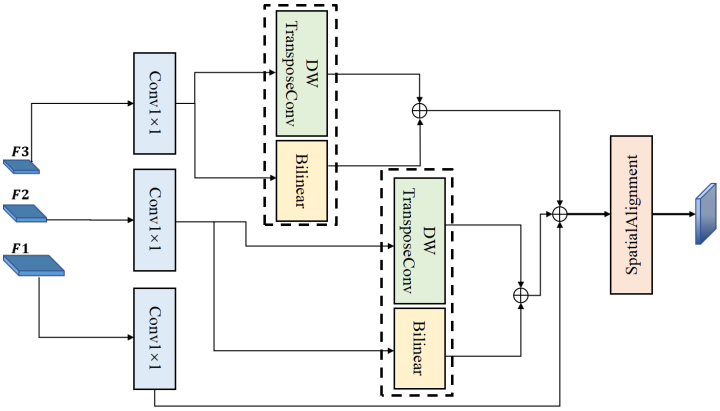


Fig. 5: Illustration of multi-scale interpolation alignment module.

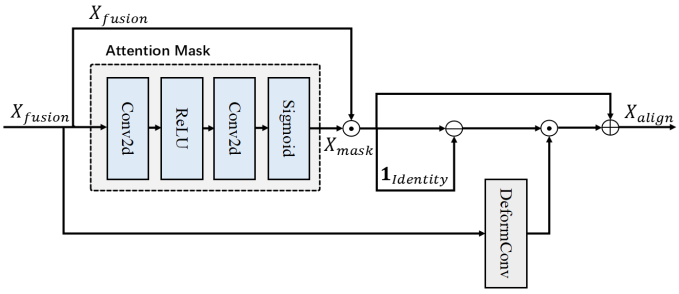


Fig. 6: Illustration of spatial alignment module.

a MIA module, generating a unified representation for subsequent cross-entropy loss supervision.

To enhance the network's generalization capability and mitigate representation bias between spatial and spectral domains, we propose a cross-domain fusion strategy. Specifically, features from both spatial and spectral branches are concatenated at corresponding scales. After alignment via the MIA module, we selectively optimize live features through single-side adversarial learning (inspired by SSDG [13] but with modifications). Our adversarial loss function is formulated as:

$$\min_{D}\max_{G} L_{adv}(G,D) = -E_{x,y\sim X_r,Y_D}\sum_{n=1}^{N}\log D(G(x)) \tag{4}$$

where Y_D represents the source domain labels of the live faces, G is the feature generator, and D the domain discriminator. To jointly optimize G and D, we insert a gradient reversal layer [46] G. During backpropagation, the adversarial loss is scaled by $-\lambda$ to mitigate noise interference in early training phases.

D. Loss Function

The proposed SSDDG framework employs a comprehensive loss function system comprising three key components to supervise its dual-branch architecture. First, binary cross-entropy losses ($L_{cls1}, L_{cls2}, L_{cls3}$) are applied to both the spatial and spectral branches as well as their fusion output, ensuring effective live/spoof classification within each domain. Second, a domain approximation loss (L_{app}) is introduced to minimize the L2 distance between spatial features $\mathrm{X}_{spatial}$ and the domain-invariant features $\mathrm{X}_{Spectrum}$, calculated as the mean squared difference across batch samples to align cross-domain representations while preserving domain-specific characteristics. This loss function is formulated as:

$$L_{app} = \frac{1}{N}\sum_{j=1}^{N}\left(\mathrm{X}_{(j,spectrum)} - \mathrm{X}_{(j,spatial)}\right)^2 \tag{5}$$

where N is a mini-batch size, and $\mathrm{X}_{(j,spatial)}$ and $\mathrm{X}_{(j,spectrum)}$ are the j^{th} image feature and domain-invariant feature in the mini-batch. We use the L2 norm to computer the distance between $\mathrm{X}_{(j,spectrum)}$ and $\mathrm{X}_{(j,spatial)}$ for our approximation loss function.

Finally, these components are integrated with the adversarial loss (L_{adv}) through a weighted combination, where balancing parameters λ_1 and λ_2 coordinate the trade-off between branch-specific classification, cross-domain feature alignment, and adversarial generalization. This unified loss formulation enables the framework to simultaneously optimize discriminative capability, domain invariance, and generalization performance.

$$L_{SSDDG} = L_{cls1} + L_{cls2} + +L_{cls3} + \lambda_1 L_{app} + \lambda_2 L_{adv} \tag{6}$$

where λ_1 and λ_2 are the balanced parameters.

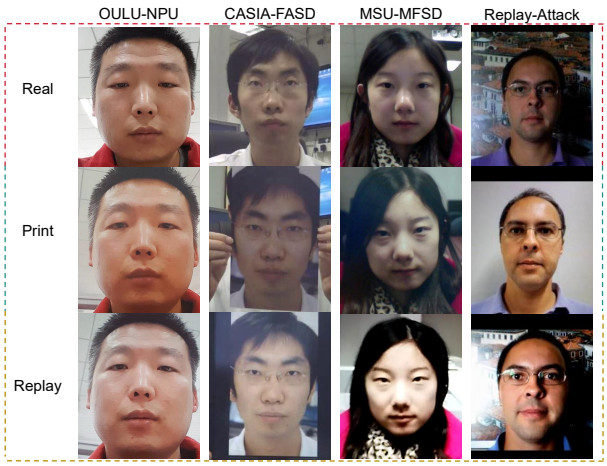


Fig. 7: Samples of the live and fake faces from the OULU-NPU [28], CASIA-FASD [27], MSU-MFSD [47] and Idiap Replay-Attack [2].

IV. Experiment

A. Datasets and Metrics

In order to analyze the generalization ability of the proposed method in this paper, we implement our experiments on four commonly public datasets including OULU-NPU [28] (O), CASIA-FASD [27] (C), Idiap Replay-Attack [2] (I), and MSU-MFSD [47] (M). Specific descriptions of the datasets are given below. The samples from the four datasets, which include live faces, print attacks and replay attacks, are shown in Fig. 7.

1) Datasets: Below is a detailed introduction to the datasets that have been used.

OULU-NPU [28] provides a composition of 5,940 live and attack videos from 55 objects. These videos are collected under three different lighting and backgrounds and recorded

with six different mobile phones. Two printers and two video players are used to simulate the diversity of devices an intruder would use.

CASIA-FASD (CASIA face anti-spoofing dataset) [27] contains 600 video compositions of 50 subjects. For each subject, three cameras with different video qualities are used for acquisition in natural scenarios, and then the three attack types are distorted photo attacks, cropped photo attacks and replay attacks. Thus, there are 12 sequences (3 live and 9 fake) for each subject.

Idiap Replay-Attack [2] provides 1,300 video compositions from 50 objects. For each object, four live video sequences are collected in two different environments, a light-lit environment with a uniform background and a daylight-lit environment with a complex scene. Similar to CASIA-FASD [46], three types of spoofing in stationary and handheld conditions include print attacks, digital photo attacks, and video attacks.

MSU-MFSD (MSU mobile face spoofing dataset) [47] provides 280 video compositions from 35 subjects. These data are captured with two types of cameras on a laptop and a mobile phone.

2) Metrics: Drawing inspiration from the pioneering work of [13], we adopt the Half Total Error Rate (HTER) and the Area Under the Curve (AUC) as the key evaluation metrics for assessing the effectiveness of our face anti-spoofing approach. HTER is obtained from FAR and FRR together. False Acceptance Rate (FAR) refers to the rate at which the model judges a fake face as a live face, and False Rejection Rate (FRR) refers to the rate at which the model judges a live face as a fake face, and the mean of FRR and FAR is the Half Total Error Rate, which is defined as follows:

$$\mathrm{HTER} = \frac{\mathrm{FAR} + \mathrm{FRR}}{2} \tag{7}$$

B. Implementation Details

The data in the above four datasets are in the form of videos, so we extract images at certain interval and use MTCNN algorithm [48] for face detection and face alignment. The preprocessed data is resized to 256×256×3 as the input of the network. The same setup rules as in [13], we randomly extract one frame per video for training and two frames per video for testing. In order to prevent the model from overfitting the training data, in addition to the augmentation methods in [13], we also use mixup [49] for data augmentation. To train the FAS task, we use ResNet18 [50] as the feature generator with the AdamW optimizer under an initial learning rate and weight decay of 5e-4 and 5e-2, respectively. We train the model with maximum 400 epochs and batch size of 60, whereas decreasing the learning rate through CosineAnnealingLR in most set-ups. Due to the small number of test sets for O&C&I to M, this experiment uses an initial value of 0.001, which is decayed by 10 at epoch 150 and 250.

C. Experiment in Leave-One-Out Setting

Above four datasets are used to evaluate the performance of the proposed SSDDG method. We randomly select three datasets as training set and the remaining one as testing set, i.e., Leave-One-Out (LOO) strategy. Thus, there are four sets of experimental results, which are O&C&I to M, O&M&I to C, O&C&M to I and I&C&M to O. We compare SSDDG with representative recent DG methods reported under the same dataset splits, including GAC-FAS, FSV-FAS, and DCRN, as shown in Table I. SSDDG obtains 3.78%/99.35%, 7.18%/97.79%, 5.67%/98.42%, and 7.56%/98.31% in terms of HTER/AUC for the four protocols, respectively. It achieves the best HTER and AUC in O&C&I to M and O&M&I to C, and the best AUC in I&C&M to O. GAC-FAS remains stronger in O&C&M to I and achieves a lower HTER in I&C&M to O. Overall, SSDDG outperforms DCRN on all four protocols and FSV-FAS on three protocols, while the comparison with GAC-FAS is split across the four settings. These results indicate that spatial-spectral phase fusion is complementary to gradient alignment and frequency-shortcut mitigation.

Several recent DG-FAS methods also avoid target-domain access during training, so target-free operation alone does not establish superiority. The main distinction of SSDDG is its explicit use of the Fourier phase as a stable representation and its multi-scale alignment with the spatial branch. FSV-FAS instead suppresses shortcut bands, whereas DCRN emphasizes feature-consistency learning. The comparison therefore clarifies both the value and the current limitations of the proposed phase-oriented formulation, and identifies combining phase modulation with shortcut mitigation and semantic regularization as a promising direction for future work.

D. Experiment in Limited Source Domains Setting

To evaluate the robustness of our method in practical scenarios where source domains are typically scarce and image quality varies substantially, we designed a constrained experimental setup. We selected two low-resolution datasets as source domains (Idiap [2] and MSU [47]) and two high-resolution datasets as target domains (OULU [28] and CASIA [27]), thereby creating a challenging cross-domain evaluation framework. Table II additionally includes the recent GAC-FAS, FSV-FAS, and DCRN results under the same M&I source-domain setting. SSDDG achieves 13.41% HTER and 92.14% AUC for M&I to C, as well as 18.36% HTER and 92.52% AUC for M&I to O, without accessing target-domain data during training. It obtains the best HTER and AUC on M&I to C, and the best AUC on M&I to O, where FSV-FAS achieves a lower HTER. These results indicate strong limited-source generalization while showing that model ranking depends on the target domain and evaluation metric.

The results suggest that preserving phase-based structural cues is particularly beneficial when only two low-resolution source domains are available. This advantage complements recent shortcut-mitigation and semantic-consistency methods, which perform strongly in the standard leave-one-out setting but do not explicitly model Fourier phase stability.

E. Ablation Study

We conduct ablation studies to evaluate each component of our method, decomposing the SSDDG network into two branches: the F branch (spectral domain via FFT) and the

TABLE I: Comparison results between the proposed method and state-of-the-art methods for domain generalization.

Method	O&C&I → M		O&M&I → C		O&C&M → I		I&C&M → O	
	HTER(%)	AUC(%)	HTER(%)	AUC(%)	HTER(%)	AUC(%)	HTER(%)	AUC(%)
MADDG [20]	17.69	88.06	24.50	84.51	22.19	84.99	27.98	80.02
MDRL [10]	17.02	90.01	19.68	87.43	20.87	86.72	25.02	81.47
SSDG-M [13]	16.67	90.47	23.11	85.45	18.21	94.61	25.17	81.83
RFM [21]	13.89	93.98	20.27	88.16	17.30	90.48	16.45	91.16
SDA [24]	15.40	91.80	24.50	84.40	15.60	90.10	23.10	84.30
DRDG [51]	12.43	95.81	19.05	88.79	15.56	91.79	15.63	91.75
MetaSSN [14]	10.80	95.30	20.50	87.40	11.30	95.00	16.40	90.90
GDA [25]	9.20	98.00	12.20	93.00	10.00	96.00	14.40	92.60
Cross-ADD [52]	11.64	95.27	17.51	89.98	15.08	91.92	14.27	93.04
EPCR [53]	10.00	95.50	17.80	89.80	13.50	96.00	14.40	93.30
UDG-FAS [54]	7.14	97.31	11.44	95.59	6.28	98.61	12.18	94.36
GAC-FAS [40]	8.60	97.16	8.20	95.16	**4.29**	**98.87**	**5.00**	97.56
FSV-FAS [41]	5.95	98.52	7.33	97.86	5.45	98.77	10.88	96.29
DCRN [42]	4.05	99.12	7.38	97.57	6.17	98.22	8.33	98.17
Ours (SSDDG)	**3.78**	**99.35**	**7.18**	**97.79**	5.67	98.42	7.56	**98.31**

Note: GAC-FAS and FSV-FAS values are taken from their original papers; DCRN values are reported in DSCM-FAS under the unified protocol. DCRN additionally uses pixel/mask supervision, so all comparisons are reported for completeness rather than as identical supervision settings.

TABLE II: Comparison results on limited source domains for domain generalization.

Method	M&I → C		M&I → O	
	HTER(%)	AUC(%)	HTER(%)	AUC(%)
IDA [47]	45.16	58.80	54.52	42.17
MADDG [20]	41.02	64.33	39.35	65.10
SSDG-M [13]	31.89	71.29	36.01	66.88
DRDG [51]	31.28	71.50	33.35	69.14
MetaSSN [14]	31.44	72.08	28.12	76.77
Cross-ADD [52]	28.75	75.78	26.28	80.46
EBDG [55]	27.97	75.84	25.94	78.28
SSAN [56]	25.56	83.89	24.44	82.56
NDA-based [23]	21.78	83.01	21.41	86.35
DiVT-M [57]	20.11	86.71	23.61	85.73
GAC-FAS [40]	16.91	88.12	17.88	89.67
FSV-FAS [41]	23.33	84.99	**16.16**	91.72
DCRN [42]	15.00	91.81	19.31	89.43
Ours (SSDDG)	**13.41**	**92.14**	18.36	**92.52**

Note: GAC-FAS reports these M&I results directly; FSV-FAS and DCRN values follow the unified protocol reported in the DSCM-FAS study.

S branch (spatial domain), both supervised by binary cross-entropy loss, the comparison results are shown in Table III. The complete model achieves an arithmetic mean of 6.05% HTER and 98.47% AUC over the four protocols, improving consistently over both the spectral and spatial branches. Removing AFM increases the mean HTER from 6.05% to 14.51% and decreases AUC from 98.47% to 93.78%, showing that adaptive phase modulation is essential for retaining discriminative liveness cues. Removing MIA raises the mean HTER to 10.21% and lowers AUC to 95.71%, demonstrating the importance of multi-scale alignment. The approximation loss is likewise important: without it, the mean HTER and AUC become 13.23% and 93.56%, respectively. Finally, removing adversarial learning increases the mean HTER to 8.70% and reduces AUC to 97.12%. These larger errors also appear consistently across all four protocols, confirming that the two auxiliary supervisions jointly improve cross-domain robustness.

F. Visualizations

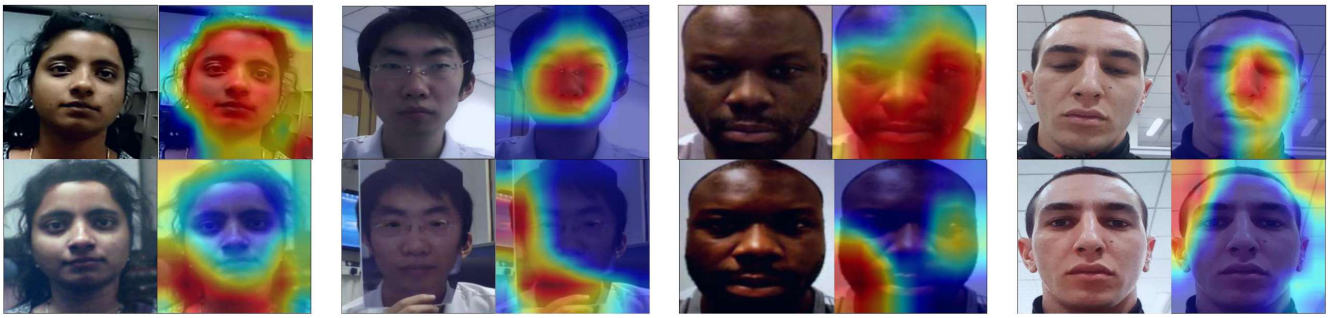
To better understand our model's decision-making process, we employ Grad-CAM [58] to visualize class activation mappings (CAMs) for both live and spoofing samples, as shown in fig8Fig. 8. The visualizations reveal two key observations: (1) For live faces, our method consistently focuses on the inner facial regions while effectively suppressing background interference, demonstrating its ability to extract domain-invariant liveness features. (2) For spoofing attacks, the model dynamically attends to attack-specific discriminative regions—highlighting edges in printed photographs or boundary artifacts between facial regions and backgrounds for different attack types. These adaptive attention patterns validate our SSDDG framework's capability to distinguish diverse presentation attacks through domain-aware feature analysis.

G. Discussion

The experimental results demonstrate the effectiveness of our approach in the limited-source settings. Across the representative recent baselines, SSDDG obtains the best HTER and AUC in two LOO settings, the best AUC in a third, and consistently outperforms DCRN and FSV-FAS in most settings. These observations provide valuable insights for further enhancing the model's adaptability to diverse real-world conditions. Our work focuses on RGB-based face anti-spoofing, a widely adopted solution that offers an optimal balance between deployment feasibility and security requirements. This approach has shown particular promise in general authentication scenarios where hardware cost and compatibility are key considerations.

TABLE III: Evaluations of different components of the proposed method.

Module	O&C&I → M		O&M&I → C		O&C&M → I		I&C&M → O	
	HTER(%)	AUC(%)	HTER(%)	AUC(%)	HTER(%)	AUC(%)	HTER(%)	AUC(%)
F Branch w/ Cls1	12.45	93.75	18.13	92.14	15.32	92.78	19.35	86.35
S Branch w/ Cls2	10.40	95.07	17.55	88.99	14.32	92.43	17.97	89.26
Fusion w/o AFM	10.32	95.16	16.63	93.25	15.64	93.53	15.45	93.16
Fusion w/o MIA	8.23	97.42	11.53	94.52	8.67	96.23	12.41	94.67
Fusion w/o App	8.41	96.68	14.89	93.25	13.45	93.35	16.15	90.94
Fusion w/o Adv	5.48	98.74	10.33	96.42	9.28	96.74	9.70	96.56
Fusion w/ All	**3.78**	**99.35**	**7.18**	**97.79**	**5.67**	**98.42**	**7.56**	**98.31**



O&C&I to M O&M&I to C O&C&M to I I&C&M to O

Fig. 8: Grad-CAM [58] visualizations of our SSDDG method under LOO strategy. The first and second rows show the live and fake faces, respectively.

V. Conclusions

Our work presents a novel dual-branch framework that significantly enhances cross-domain generalization in face anti-spoofing. The key innovations include: (1) leveraging FFT-transformed phase information to capture domain-invariant features, (2) developing a adaptive frequency modulation module for feature redistribution, and (3) introducing a multi-scale interpolation alignment module coupled with an approximation loss to precisely bridge the gap between spatial and spectral domains. This integrated approach enables more effective feature fusion than using either domain in isolation, as demonstrated by our experimental results. Under the recent leave-one-out comparisons, SSDDG outperforms DCRN on all four protocols and FSV-FAS on three protocols, while remaining competitive with GAC-FAS. In the limited-source setting without target-domain access, it obtains the best HTER and AUC for M&I to C and the best AUC for M&I to O. These findings confirm the value of phase information for domain generalization in FAS and motivate combining phase-aware fusion with gradient alignment, dynamic frequency masking, and semantic consistency regularization in future work. Building on the current RGB-domain success, we will further extend this framework to multi-modality scenarios [59], [60], particularly focusing on near-infrared spectral characteristics for enhanced anti-spoofing performance.

References

[1] J. Guo, X. Zhu, C. Zhao, D. Cao, Z. Lei, and S. Z. Li, "Learning meta face recognition in unseen domains," in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 6162–6171.

[2] I. Chingovska, A. Anjos, and S. Marcel, "On the effectiveness of local binary patterns in face anti-spoofing," in *2012 BIOSIG-Proceedings of the International Conference of Biometrics Special Interest Group*, 2012, pp. 1–7.

[3] J. Määttä, A. Hadid, and M. Pietikäinen, "Face spoofing detection from single images using micro-texture analysis," in *2011 international joint conference on Biometrics (IJCB)*, 2011, pp. 1–7.

[4] J. Komulainen, A. Hadid, and M. Pietikäinen, "Context based face anti-spoofing," in *2013 IEEE Sixth International Conference on Biometrics: Theory, Applications and Systems (BTAS)*, 2013, pp. 1–8.

[5] J. Yang, Z. Lei, S. Liao, and S. Z. Li, "Face liveness detection with component dependent descriptor," in *2013 International Conference on Biometrics (ICB)*, 2013, pp. 1–6.

[6] J. Yang, Z. Lei, and S. Z. Li, "Learn convolutional neural network for face anti-spoofing," *arXiv preprint arXiv:408.5601*, 2014.

[7] Z. Xu, S. Li, and W. Deng, "Learning temporal features using lstm-cnn architecture for face anti-spoofing," in *2015 3rd IAPR asian conference on pattern recognition (ACPR)*. IEEE, 2015, pp. 141–145.

[8] R. Cai, H. Li, S. Wang, C. Chen, and A. C. Kot, "Drl-fas: A novel framework based on deep reinforcement learning for face anti-spoofing," *IEEE Transactions on Information Forensics and Security*, vol. 16, pp. 937–951, 2020.

[9] A. George and S. Marcel, "Cross modal focal loss for rgbd face anti-spoofing," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 7882–7891.

[10] G. Wang, H. Han, S. Shan, and X. 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*, 2020, pp. 6678–6687.

[11] Z. Yu, Y. Qin, X. Li, C. Zhao, Z. Lei, and G. Zhao, "Deep learning for face anti-spoofing: A survey," *IEEE Transactions on Pattern Analysis & Machine Intelligence*, vol. 45, no. 05, pp. 5609–5631, 2023.

[12] W. Liu, X. Wei, T. Lei, X. Wang, H. Meng, and A. K. Nandi, "Data-fusion-based two-stage cascade framework for multimodality face anti-spoofing," *IEEE Transactions on Cognitive and Developmental Systems*, vol. 14, no. 2, pp. 672–683, 2021.

[13] Y. Jia, J. Zhang, S. Shan, and X. Chen, "Single-side domain generaliza-

tion for face anti-spoofing," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 8484–8493.
[14] Y.-E. Kim, W.-J. Nam, K. Min, and S.-W. Lee, "Style selective normalization with meta learning for test-time adaptive face anti-spoofing," *Expert Systems with Applications*, vol. 214, p. 119106, 2023.
[15] Z. Li, Y. Cui, F. Wang, W. Liu, Y. Yang, Z. Yu, B. Jiang, and H. Chen, "Representation constraint-based dual-channel network for face antispoofing," *IET Computer Vision*, 2023.
[16] ——, "A multimodal face antispoofing method based on multifeature vision transformer and multirank fusion," *Concurrency and Computation: Practice and Experience*, vol. 35, no. 23, p. e7824, 2023.
[17] A. Liu, Z. Tan, Y. Liang, and J. Wan, "Attack-agnostic deep face anti-spoofing," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 6335–6344.
[18] G. Heusch, A. George, D. Geissbühler, Z. Mostaani, and S. Marcel, "Deep models and shortwave infrared information to detect face presentation attacks," *IEEE Transactions on Biometrics, Behavior, and Identity Science*, vol. 2, no. 4, pp. 399–409, 2020.
[19] A. Torralba and A. A. Efros, "Unbiased look at dataset bias," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2011, pp. 1521–1528.
[20] R. Shao, X. Lan, J. Li, and P. 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*, 2019, pp. 10 023–10 031.
[21] R. Shao, X. Lan, and P. C. Yuen, "Regularized fine-grained meta face anti-spoofing," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 07, 2020, pp. 11 974–11 981.
[22] Y. Sun, Y. Liu, X. Liu, Y. Li, and W.-S. Chu, "Rethinking domain generalization for face anti-spoofing: Separability and alignment," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 24 563–24 574.
[23] W. Wang, P. Liu, H. Zheng, R. Ying, and F. Wen, "Domain generalization for face anti-spoofing via negative data augmentation," *IEEE Transactions on Information Forensics and Security*, vol. 14, no. 8, pp. 2333–2344, 2023.
[24] J. Wang, J. Zhang, Y. Bian, Y. Cai, C. Wang, and S. Pu, "Self-domain adaptation for face anti-spoofing," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 35, no. 4, 2021, pp. 2746–2754.
[25] Q. Zhou, K.-Y. Zhang, T. Yao, R. Yi, K. Sheng, S. Ding, and L. Ma, "Generative domain adaptation for face anti-spoofing," in *European Conference on Computer Vision*, 2022, pp. 335–356.
[26] W. Lu, J. Wang, H. Li, Y. Chen, and X. Xie, "Domain-invariant feature exploration for domain generalization," *Transactions on Machine Learning Research*, 2022.
[27] Z. Zhang, J. Yan, S. Liu, Z. Lei, D. Yi, and S. Z. Li, "A face anti-spoofing database with diverse attacks," in *2012 5th IAPR international conference on Biometrics (ICB)*, 2012, pp. 26–31.
[28] Z. Boulkenafet, J. Komulainen, L. Li, X. Feng, and A. Hadid, "Oulu-npu: A mobile face presentation attack database with real-world variations," in *2017 12th IEEE international conference on automatic face & gesture recognition (FG 2017)*, 2017, pp. 612–618.
[29] Z. Boulkenafet, J. Komulainen, and A. Hadid, "Face anti-spoofing based on color texture analysis," in *2015 IEEE international conference on image processing (ICIP)*, 2015, pp. 2636–2640.
[30] G. Heusch and S. Marcel, "Remote blood pulse analysis for face presentation attack detection," *Handbook of Biometric Anti-Spoofing: Presentation Attack Detection*, pp. 267–289, 2019.
[31] L. Feng, L.-M. Po, Y. Li, X. Xu, F. Yuan, T. C.-H. Cheung, and K.-W. Cheung, "Integration of image quality and motion cues for face anti-spoofing: A neural network approach," *Journal of Visual Communication and Image Representation*, vol. 38, pp. 451–460, 2016.
[32] Z. Yu, X. Li, P. Wang, and G. Zhao, "Transrppg: Remote photoplethysmography transformer for 3d mask face presentation attack detection," *IEEE Signal Processing Letters*, vol. 28, pp. 1290–1294, 2021.
[33] C.-Y. Wang, Y.-D. Lu, S.-T. Yang, and S.-H. 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*, 2022, pp. 20 281–20 290.
[34] Y. Kwak, M. Jung, H. Yoo, J. Shin, and C. Kim, "Liveness score-based regression neural networks for face anti-spoofing," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2023, pp. 1–5.
[35] K. Roy, M. Hasan, L. Rupty, M. S. Hossain, S. Sengupta, S. N. Taus, and N. Mohammed, "Bi-fpnfas: Bi-directional feature pyramid network for pixel-wise face anti-spoofing by leveraging fourier spectra," *Sensors*, vol. 21, no. 8, p. 2799, 2021.
[36] A. Jourabloo, Y. Liu, and X. Liu, "Face de-spoofing: Anti-spoofing via noise modeling," in *Computer Vision–ECCV 2018: 15th European Conference, Munich, Germany, September 8-14, 2018, Proceedings, Part XIII 15*, 2018, pp. 297–315.
[37] S. Zhang, A. Liu, J. Wan, Y. Liang, G. Guo, S. Escalera, H. J. Escalante, and S. Z. Li, "Casia-surf: A large-scale multi-modal benchmark for face anti-spoofing," *IEEE Transactions on Biometrics, Behavior, and Identity Science*, vol. 2, no. 2, pp. 182–193, 2020.
[38] Z. Yu, Y. Qin, X. Li, Z. Wang, C. Zhao, Z. Lei, and G. 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*, 2020, pp. 650–651.
[39] X. Lin, S. Wang, R. Cai, Y. Liu, Y. Fu, W. Tang, Z. Yu, and A. Kot, " Suppress and Rebalance: Towards Generalized Multi-Modal Face Anti-Spoofing ," in *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2024, pp. 211–221. [Online]. Available: https://doi.ieeecomputersociety.org/10.1109/CVPR52733.2024.00028
[40] B. M. Le and S. S. Woo, "Gradient alignment for cross-domain face anti-spoofing," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 188–199.
[41] J. Cao and C. Ma, "Towards generalized face anti-spoofing from a frequency shortcut view," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2025, pp. 1005–1015.
[42] Y. Liu, Z. Li, and L. Wu, "Dual consistency regularization for generalized face anti-spoofing," *IEEE Transactions on Information Forensics and Security*, vol. 20, pp. 2171–2183, 2025.
[43] H. Xing, S. Y. Tan, F. Qamar, and Y. Jiao, "Dual semantic consistency module for domain generalizable face anti-spoofing," *Journal of King Saud University–Computer and Information Sciences*, vol. 38, p. 829, 2026.
[44] M. Liu, J. Mu, Z. Yu, K. Ruan, B. Shu, and J. Yang, "Adversarial learning and decomposition-based domain generalization for face anti-spoofing," *Pattern Recognition Letters*, vol. 155, pp. 171–177, 2022.
[45] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 7132–7141.
[46] Y. Ganin and V. Lempitsky, "Unsupervised domain adaptation by backpropagation," in *International conference on machine learning*, 2015, pp. 1180–1189.
[47] D. Wen, H. Han, and A. K. Jain, "Face spoof detection with image distortion analysis," *IEEE Transactions on Information Forensics and Security*, vol. 10, no. 4, pp. 746–761, 2015.
[48] K. Zhang, Z. Zhang, Z. Li, and Y. Qiao, "Joint face detection and alignment using multitask cascaded convolutional networks," *IEEE signal processing letters*, vol. 23, no. 10, pp. 1499–1503, 2016.
[49] H. Zhang, M. Cisse, Y. N. Dauphin, and D. Lopez-Paz, "mixup: Beyond empirical risk minimization," *arXiv preprint arXiv:1710.09412*, 2017.
[50] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.
[51] S. Liu, K.-Y. Zhang, T. Yao, K. Sheng, S. Ding, Y. Tai, J. Li, Y. Xie, and L. Ma, "Dual reweighting domain generalization for face presentation attack detection," in *IJCAI*, 2021, pp. 867–873.
[52] H. Huang, Y. Xiang, G. Yang, L. Lv, X. Li, Z. Weng, and Y. Fu, "Generalized face anti-spoofing via cross-adversarial disentanglement with mixing augmentation," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022, pp. 2939–2943.
[53] Z. Wang, Z. Yu, X. Wang, Y. Qin, J. Li, C. Zhao, X. Liu, and Z. Lei, "Consistency regularization for deep face anti-spoofing," *IEEE Transactions on Information Forensics and Security*, vol. 18, pp. 1127–1140, 2023.
[54] Y. Liu, Y. Chen, M. Gou, C.-T. Huang, Y. Wang, W. Dai, and H. Xiong, "Towards unsupervised domain generalization for face anti-spoofing," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 20 654–20 664.
[55] Z. Du, J. Li, L. Zuo, L. Zhu, and K. Lu, "Energy-based domain generalization for face anti-spoofing," in *Proceedings of the 30th ACM International Conference on Multimedia*, 2022, pp. 1749–1757.
[56] Z. Wang, Z. Wang, Z. Yu, W. Deng, J. Li, T. Gao, and Z. Wang, "Domain generalization via shuffled style assembly for face anti-spoofing," in *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022, pp. 4113–4123.
[57] C.-H. Liao, W.-C. Chen, H.-T. Liu, Y.-R. Yeh, M.-C. Hu, and C.-S. Chen, "Domain invariant vision transformer learning for face anti-spoofing," in

2023 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2023, pp. 6087–6096.

[58] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-cam: Visual explanations from deep networks via gradient-based localization," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 618–626.

[59] S. Liu, Z. Ul Abideen, T. Wan, I. Shahzad, A. Waseem, and Y. Pan, "Must-gan mfas: Multi-semantic spoof tracer gan with transformer layers for multi-modal face anti-spoofing," *The Computer Journal*, vol. 68, no. 8, pp. 891–907, 2025.

[60] Z. U. Abideen, S. Liu, and T. Wan, "Robust multi-modal face anti-spoofing: Techniques for missing modality handling and fusion," in *International Conference on Neural Information Processing*, 2025, pp. 77–91.