Computers, Materials & Continua

Tech Science Press

https://doi.org/10.32604/cmc.2026.085087

ARTICLE

Semantically Anchored Test-Time Domain Generalization for Face Anti-Spoofing

Xiaosong Chang, Liang Shi* and Ao Zhang

School of Computer Science, Jiangsu University of Science and Technology, Zhenjiang, China

*Corresponding Author: Liang Shi. Email: jsjxy_sl@just.edu.cn

Received: 09 May 2026; Accepted: 08 July 2026; Published: 13 August 2026

ABSTRACT: To ensure the reliability of biometric authentication, Face Anti-Spoofing (FAS) models must accurately detect presentation attacks. However, due to the highly complex distribution shifts caused by variations in style, cross-domain generalization remains a significant challenge. Test-Time Domain Generalization (TTDG) has recently surfaced as an innovative framework, facilitating the adaptation of unseen samples to source-domain characteristics through the strategic utilization of learned style bases. Nevertheless, existing TTDG methods optimize randomly initialized style bases solely through statistical objectives, leaving a critical research gap: the lack of explicit semantic constraints inevitably leads to hierarchical semantic inconsistency and weakens subtle spoofing cues. To fill this gap, we introduce the Semantic-Anchored Test-Time Domain Generalization (SA-TTDG) framework. The core novelty of our approach lies in introducing a Text-Anchored Style Projection (TASP), which utilizes rich linguistic priors from Vision-Language Models (VLMs) to initialize and strongly constrain learnable style bases. By anchoring these bases to explicit semantic concepts, TASP encourages semantic consistency across hierarchical feature representations. Furthermore, to fully exploit these semantically aligned style representations, we design a Semantic Prompt Modulation (SPM) module driven by a Style-Query Cross-Attention (SQ-CA) mechanism. Instead of using static queries, SPM dynamically retrieves multi-level style cues to generate domain-sensitive prompts, which effectively modulate the original visual features. This process helps enhance subtle spoofing-related cues while preserving the underlying content structure. Evaluations under standard leave-one-domain-out protocols demonstrate that the proposed framework consistently reduces cross-domain classification errors compared to existing statistical TTDG baselines.

KEYWORDS: Face anti-spoofing; domain generalization; TTDG; prompt modulation

1 Introduction

To secure facial recognition systems against diverse presentation attacks (PAs)—which span from flat-printed photographs [1] and looped video replays [2] to sophisticated 3D facial masks [3]—a broad spectrum of FAS techniques has been extensively investigated in recent literature [4–7]. Earlier deep learning approaches employed multi-channel convolutional neural networks to jointly exploit complementary facial cues for presentation attack detection, demonstrating the effectiveness of deep feature learning over handcrafted descriptors [8]. Notably, recent authoritative studies have introduced diverse strategies—ranging from deep pixel-wise binary supervisions [9] and neural architecture search with central difference convolutions [10,11] to pseudo-negative sample synthesis [12] and multimodal deep fusion [13]—to extract more nuanced intrinsic spoofing features. While these deep learning-based models achieve impressive

results on intra-dataset evaluations, they frequently exhibit severe performance degradation when exposed to unseen target domains.

Recent surveys on deep learning-based authentication systems have further highlighted the increasing reliance of smart devices on biometric authentication modalities, while emphasizing the security risks posed by presentation attacks and the necessity of robust anti-spoofing mechanisms [14]. Even with substantial advancements in modern imaging hardware and recognition algorithms, spoofing attempts can still bypass security measures if the deployment conditions—such as the attack medium, sensor characteristics, ambient illumination, or screen artifacts—differ from the training distribution. Accordingly, the primary focus of this study shifts away from detecting known attacks under constrained settings. Instead, we aim to enhance the cross-domain generalization of FAS models when confronted with novel appearance variations and style shifts in unseen environments.

To tackle this cross-domain vulnerability, researchers have widely introduced Domain Generalization (DG) paradigms to modern vision systems. In the broader context of visual recognition, advanced frameworks have leveraged style perturbation and consistency regularizations [15] to enforce domain-agnostic feature learning. When generalized to the specific Face Anti-Spoofing (FAS) task, early representative efforts predominantly utilized multi-domain feature alignment and distribution regularizations [16] to bridge domain discrepancies, alongside meta-learning strategies [17] designed to foster cross-domain model robustness.

Nevertheless, the majority of these techniques concentrate heavily on extracting domain-invariant features strictly during the training stage. Consequently, their effectiveness often drops precipitously when tested against novel domains that present distribution shifts far beyond the scope of the original source data.

Recently, Test-Time Adaptation (TTA) has emerged as a powerful paradigm to mitigate distribution shifts by dynamically adjusting a pre-trained model to target data on the fly during inference [18]. In the specific context of Face Anti-Spoofing, Source-Free Domain Adaptation (SFDA) has also been explored to adapt models to unseen target domains without requiring access to the original source data, often by leveraging domain generalized pre-training [19]. Furthermore, to handle continually drifting domains, continual test-time adaptation (CTTA) has been actively investigated, where domain consistency learning is often leveraged to prevent error accumulation [20]. Although effective, these paradigms typically require continuous parameter updates at test time, which is computationally expensive and susceptible to model collapse in resource-constrained deployment. To overcome these limitations, Test-Time Domain Generalization (TTDG) has surfaced as a more efficient and robust alternative, attracting increasing attention in the FAS community [21]. Different from conventional DG methods that only rely on source-domain optimization, TTDG calibration strategies actively exploit the instance-specific stylistic properties of arriving samples to bridge the representational divide between source and target environments. A representative strategy is to project the test-sample style representation onto a learnable basis space and use the projected style to calibrate visual features without updating model parameters.

Although this paradigm is effective, its style-basis construction is still largely driven by statistical learning. The style bases are usually initialized randomly and optimized implicitly according to source-domain objectives. As a result, the learned bases may capture domain-dependent variations, but their semantic meaning remains ambiguous. This becomes problematic for FAS because spoofing evidence is often subtle and distributed across different feature levels, such as texture degradation, display reflection, paper artifacts, illumination changes, and color distortion. If hierarchical style statistics are projected onto an unconstrained basis space, the correspondence between low-level visual patterns and high-level attack-related cues may be weakened, leading to semantic inconsistency during test-time style calibration.

Motivated by this observation, we propose to make the style projection process semantically grounded. Instead of treating style bases as purely statistical prototypes, we introduce a Text-Anchored Style Projection (TASP) module that uses CLIP-derived textual embeddings as semantic anchors. Specifically, we construct a set of FAS-related style descriptors, such as blur, noise, low illumination, screen reflection, and paper texture, and encode them with the CLIP text encoder. These textual embeddings provide explicit semantic references for initializing and constraining the learnable style bases, as conceptually illustrated in Fig. 1. The detailed vocabulary design and prompt construction are described in Section 3.2. In this way, the basis space is encouraged to preserve interpretable style semantics while still adapting to real visual statistics from source-domain facial images.

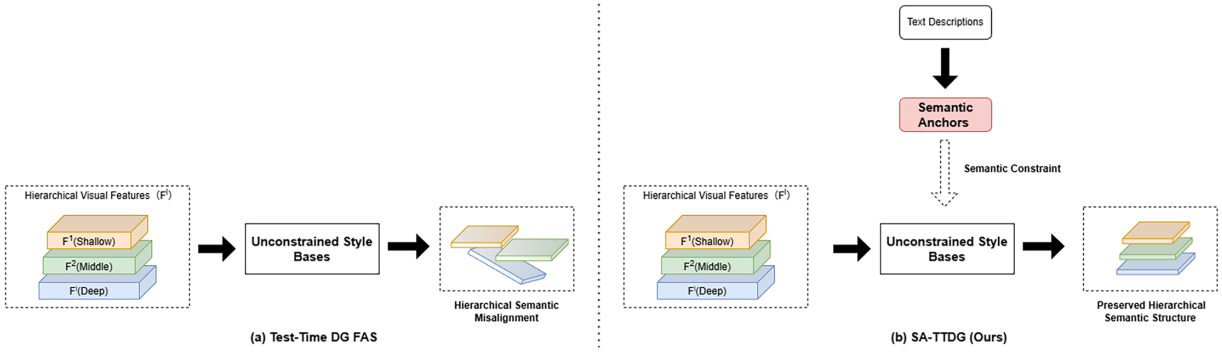

Figure 1: Differences between our method and existing test-time DG methods. (**a**) Without explicit semantic guidance, projecting hierarchical visual features onto unconstrained style bases may cause layer-wise distortion, resulting in hierarchical semantic misalignment. (**b**) SA-TTDG uses semantic anchors extracted from text descriptions to explicitly constrain the style bases, enabling the preservation of hierarchical semantic structure during test-time adaptation.

Furthermore, contemporary VLM-based FAS methods [22] mainly exploit linguistic priors through global prompts or learnable query tokens. Such prompts are useful for introducing high-level semantic guidance, but they are not specifically designed to interact with instance-dependent test-time style variations. To better exploit the semantically aligned style representation, we further introduce a Semantic Prompt Modulation (SPM) module. SPM uses Style-Query Cross-Attention (SQ-CA) to dynamically retrieve style cues from the projected style representation and generate domain-sensitive prompts for feature modulation. Therefore, SA-TTDG bridges semantic anchoring and test-time style calibration in a unified framework.

In summary, the main contributions of this paper are as follows:

- We identify the semantic ambiguity of unconstrained style bases as a key limitation in existing TTDG-based FAS methods. To address this issue, we propose Text-Anchored Style Projection (TASP), which introduces CLIP-derived textual anchors into the style-basis space and encourages each basis to maintain an explicit semantic meaning.
- We design Semantic Prompt Modulation (SPM) with Style-Query Cross-Attention (SQ-CA), which converts semantically aligned style representations into instance-conditioned prompts for adaptive feature modulation.
- Evaluations under four standard leave-one-domain-out cross-domain FAS protocols demonstrate that coupling semantic anchoring with dynamic style-query interaction consistently reduces cross-domain classification errors compared to existing statistical TTDG baselines.

2 Related Work

2.1 Domain Generalization for Face Anti-Spoofing

Separating authentic users from malicious presentation attacks (e.g., printed photographs, digital replays, and physical masks) remains the core task of FAS. Despite demonstrating impressive accuracy in closed-set, single-domain scenarios, most contemporary FAS architectures suffer from severe generalization degradation when evaluated on previously unseen domains. Therefore, Domain Generalization (DG) has been widely studied in FAS to improve the robustness of models trained on multiple source domains without accessing target-domain data, a fundamental challenge extensively reviewed in recent literature [23].

Existing DG-FAS methods mainly focus on learning domain-invariant or domain-robust representations from source domains. Early representative methods adopt adversarial learning [6,24] to reduce domain discrepancies. Specifically, MADDG [24] introduces a multi-adversarial learning strategy to encourage domain-invariant feature learning, while SSDG [16] learns a single-side domain generalization framework by compacting bona fide samples and separating spoof samples in the feature space. Beyond adversarial paradigms, other approaches leverage unsupervised domain adaptation [25], disentangled representation learning [26], and multi-domain feature alignment to comprehensively regularize cross-domain distributions. In addition to these global alignment efforts, recent literature has explored patch-oriented models [27] and style assembly mechanisms [28] to isolate localized and subtle presentation traces, such as regional anomalies and structural texture flaws.

Despite their effectiveness, most DG-FAS methods learn a fixed model from source domains and do not explicitly adjust the feature representation according to the style of each test sample. When the target domain contains unseen camera properties, illumination conditions, compression artifacts, or presentation media, the learned source-domain representation may still be insufficient for robust generalization. This motivates recent studies to explore test-time style calibration for FAS.

Test-time domain generalization for FAS is different from conventional test-time adaptation because it does not rely on target labels, iterative optimization, or model parameter updates during inference. Kim et al. [29] explored style-related normalization for adaptive FAS, while Park et al. [30] shifted the style statistics of a test sample toward source-domain styles. More recently, Zhou et al. [21] introduced a style projection mechanism that represents the test-sample style using learnable style bases, enabling feature calibration in a feed-forward manner. These methods demonstrate the importance of test-time style information, but their style bases are mainly constructed as statistical prototypes without explicit semantic supervision.

In contrast, our SA-TTDG focuses on the semantic reliability of the style-basis space. Rather than relying on randomly initialized bases, we initialize and regularize the style bases with text-derived semantic anchors. This design makes each basis associated with an interpretable style factor and helps reduce semantic ambiguity during test-time projection. Therefore, our method is not a redesign of the FAS backbone, but a semantic enhancement of the TTDG style projection framework.

2.2 Prompt Learning for Face Anti-Spoofing

Recently, Vision-Language Models (VLMs) have demonstrated unprecedented success in learning highly transferable representations through large-scale image-text pre-training. As extensively reviewed in recent literature [31,32], these large pre-trained foundation models significantly benefit a wide range of downstream visual tasks by bridging the semantic gaps between different modalities. Building upon these robust multimodal alignments, pioneering works like CLIP [33] leverage natural language supervision to construct domain-agnostic feature spaces. Inspired by this capability, prompt learning [22,34] has

been widely introduced to adapt VLMs to downstream scenarios. Instead of manually designing fixed prompts, these methods optimize learnable context tokens or query vectors to better capture task-specific semantics. Recently, several VLM-based methods have been explored for generalizable FAS to enhance facial representations and improve robustness across domains. For example, CFPL-FAS [35] learns content- and style-aware prompts. Collectively, these studies demonstrate that linguistic priors are highly effective for improving cross-domain FAS.

Different from these VLM-based methods, SA-TTDG uses textual semantics to anchor the style projection space itself. The textual descriptors are not used merely as classification prompts; instead, they serve as semantic references for constructing and constraining learnable style bases. Moreover, the proposed SQ-CA module dynamically converts the semantically aligned style representation into instance-conditioned prompts, which makes the prompt modulation process directly responsive to test-time style variations.

Overall, existing DG-FAS methods mainly focus on learning source-domain robust representations, while existing TTDG methods use test-time style statistics but usually rely on statistically learned style bases. Meanwhile, VLM-based FAS methods exploit linguistic priors mostly through image-level prompts or representation learning. The gap addressed in this work is how to use textual semantics to explicitly regularize the style projection space for test-time domain generalization. SA-TTDG addresses this gap by introducing semantic anchors for style-basis construction and style-conditioned prompt modulation for instance-specific adaptation.

3 Proposed Methodology

As illustrated in Fig. 2, we propose Semantic Anchored Test-Time Domain Generalization (SA-TTDG) for face anti-spoofing. Given a test image, an image encoder extracts visual features F and hierarchical style statistics s_t from multiple layers. SA-TTDG aims to adapt s_t in a semantically constrained projection space, thereby improving robustness to unseen domain shifts while preserving the hierarchical semantic structure of facial representations.

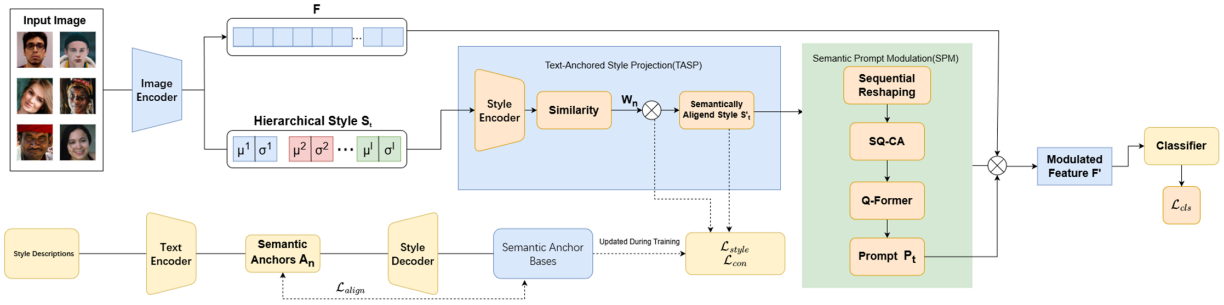

Figure 2: Our SA-TTDG is built on an image encoder and adapts face anti-spoofing models through semantic-anchored style projection and prompt modulation. The framework contains two key components: (1) text-anchored style projection (TASP), which uses text-derived semantic anchors to constrain style bases and obtain semantically aligned style representations; and (2) semantic prompt modulation (SPM), which transforms the aligned style representation into dynamic prompts via SQ-CA and Q-Former. The generated prompts modulate the visual features F to produce F' for robust classification. Auxiliary losses are applied during training to preserve semantic consistency, while no model update is required at inference.

Specifically, SA-TTDG consists of two key modules: Text-Anchored Style Projection (TASP) and Semantic Prompt Modulation (SPM). TASP leverages text-derived semantic anchors to constrain the style bases and generate a semantically aligned style representation s'_t. SPM converts s'_t into dynamic prompts

via Style-Query Cross-Attention (SQ-CA) and Q-Former, which further modulate the visual features F for robust classification.

3.1 Feature Extraction

Given an input image x_t, we employ a CLIP ViT-B/16 image encoder E_{img} as the visual backbone. Since CLIP is trained in a vision-language aligned representation space, its visual features provide a suitable foundation for the subsequent text-guided style projection. The encoder produces visual representations from multiple Transformer blocks, allowing us to capture both low-level appearance variations and high-level facial cues. Let $F_t^l \in \mathbb{R}^{H_l \times W_l \times C_l}$ denote the intermediate feature map extracted from the l-th block of E_{img}, where H_l, W_l, and C_l indicate the height, width, and channel dimension, respectively. We use the feature representations processed by Adaptive Instance Normalization (AdaIN) [36] as the original feature F for classification and feature modulation, while the intermediate features are used to construct hierarchical style statistics. Following the widely used feature-statistic representation in style modeling, we compute the channel-wise mean and standard deviation of each intermediate feature map. For the t-th sample, the statistics at the l-th layer are defined as:

$$\mu^l(x_t) = \frac{1}{H_l W_l} \sum_{h=1}^{H_l} \sum_{w=1}^{W_l} F_{t,h,w}^l, \tag{1}$$

$$\sigma^l(x_t) = \sqrt{\frac{1}{H_l W_l} \sum_{h=1}^{H_l} \sum_{w=1}^{W_l} \left(F_{t,h,w}^l - \mu^l(x_t)\right)^2}. \tag{2}$$

Here, $\mu^l(x_t), \sigma^l(x_t) \in \mathbb{R}^{C_l}$ describe the first-order and second-order style statistics of the l-th layer, respectively.

Finally, we concatenate the statistics from all selected layers to obtain the hierarchical style representation:

$$s_t = [\mu^1(x_t), \sigma^1(x_t), \mu^2(x_t), \sigma^2(x_t), \ldots, \mu^L(x_t), \sigma^L(x_t)]. \tag{3}$$

The resulting s_t summarizes multi-level style information of the test sample and serves as the input to TASP.

3.2 Text-Anchored Style Projection

Given the hierarchical style representation s_t extracted in Section 3.1, TASP projects it into a semantically constrained style-basis space. Instead of randomly initializing the style bases, we construct them from text-derived semantic anchors, so that each basis is associated with an explicit style concept from the beginning. This design provides a more interpretable projection space for test-time style projection.

Design Rationale of Semantic Anchor Vocabulary. The semantic anchor vocabulary is designed to provide human-interpretable references for the style-basis space. Instead of constructing anchors according to dataset identities or closed-set attack labels, we organize the vocabulary around transferable style factors that frequently appear in face anti-spoofing scenarios. Specifically, the descriptors are grouped into three categories: image quality and artifacts, illumination and environmental variations, and medium-specific spoofing cues. These groups correspond to common sources of cross-domain appearance variation in FAS, such as camera quality, video transmission, environmental lighting, replay screens, printed media, and geometric distortion. The vocabulary is not intended to exhaustively enumerate all possible attack labels.

Instead, it summarizes transferable visual style factors that may appear across different spoofing media and capture devices.

Table 1 lists representative descriptors used for constructing semantic anchors. Descriptors within the same factor group may be semantically close, such as "blurred", "out-of-focus", and "motion-blurred". We keep such descriptors because CLIP maps them to nearby but not identical semantic regions, providing intra-factor diversity while preserving group-level interpretability. In practice, the descriptor set is expanded to $N = 64$ anchors by adding semantically related variants within each category.

Table 1: Representative semantic anchor descriptors used in TASP. The descriptors are grouped according to FAS-related image quality degradation, illumination variation, and medium-specific spoofing cues.

Category	Factor	Style Descriptors
Quality & Artifacts	Blur	"blurred", "out-of-focus", "motion-blurred", "unsharp"
	Noise	"noisy", "grainy", "speckled", "pixelated"
	Compression	"jpeg-compressed", "blocky", "lossy-compressed"
Illumination & Env	Low-light	"dark", "under-exposed", "dim", "poorly-lit"
	Over-exposure	"bright", "over-exposed", "glaring", "washed-out"
	Color Shift	"warm-toned", "cool-toned", "yellowish", "bluish"
Medium Specific	Screen Replay	"screen-reflected", "moiré-patterned", "glowing"
	Print Attack	"paper-textured", "flat-looking", "color-faded"
	Distortion	"fisheye-distorted", "barrel-distorted", "warped"

Prompt Template and Anchor Encoding. Let $\mathcal{T} = \{T_n\}_{n=1}^{N}$ denote the semantic descriptor set, where each T_n corresponds to a style concept in the anchor vocabulary. For each descriptor, we use a unified prompt template to construct the textual input:

$$P_n = \text{"A photo of a } T_n \text{ face"}. \tag{4}$$

The prompt P_n is then encoded by the frozen CLIP text encoder E_{txt} to obtain the semantic anchor:

$$A_n = E_{txt}(P_n) \in \mathbb{R}^{D_{txt}}. \tag{5}$$

All textual anchors are normalized before being used to initialize or constrain the style bases. Using a unified prompt template avoids introducing additional prompt-specific variables and allows the effect of semantic vocabulary to be analyzed more directly.

Semantic-Visual Space Bridging. The textual semantic space and the visual style-statistic space have different dimensions and distributions. To bridge this gap, we introduce two projection adapters: a text-to-style decoder Ψ and a style-to-semantic encoder Φ. The text-to-style decoder $\Psi: \mathbb{R}^{D_{txt}} \to \mathbb{R}^{D_s}$ maps the text-derived semantic anchor A_n into the hierarchical visual style space, where $D_s = 2\sum_{l=1}^{L} C_l$ denotes the dimension of the concatenated mean and standard deviation statistics from all selected layers. The generated representation is used to initialize the corresponding style basis:

$$b_n^0 = \Psi(A_n). \tag{6}$$

In contrast, the style-to-semantic encoder $\Phi: \mathbb{R}^{D_s} \to \mathbb{R}^{D_{txt}}$ maps a visual style representation back to the CLIP textual embedding space. This allows us to measure the semantic correspondence between a visual

style vector and a textual anchor. It is used both for semantic regularization of the learnable style bases and for computing the anchor similarity of a test sample during inference. In this way, Ψ injects linguistic semantics into the style-basis space, while Φ projects visual style representations back to the semantic space for anchor-based comparison and regularization.

Semantic Basis Regularization. During training, the style bases $\mathcal{B} = \{b_n\}_{n=1}^{N}$ are learnable and updated by backpropagation. This allows them to absorb realistic style statistics from source-domain facial images. However, directly optimizing the bases may cause semantic drift. To maintain their correspondence with the original semantic anchors, we impose a semantic alignment constraint.

Specifically, we project each learnable basis back into the semantic space through a style encoder Φ and align it with its frozen anchor:

$$\mathcal{L}_{align} = \frac{1}{N}\sum_{n=1}^{N}\left[1 - \mathrm{CosSim}\left(\Phi(b_n), \mathrm{sg}(A_n)\right)\right], \tag{7}$$

where $\mathrm{sg}(\cdot)$ denotes the stop-gradient operation. This loss encourages each learnable basis to remain close to its corresponding semantic prototype while still allowing adaptation to empirical visual statistics.

In addition, different style bases should preserve sufficient diversity in their second-order statistics. We therefore impose an orthogonal regularization on the normalized variance bases $\hat{\Sigma}_b$:

$$\mathcal{L}_{orthstd} = \frac{1}{N(N-1)}\sum_{i\neq j}(G_\sigma)_{ij}^2, \quad G_\sigma = \hat{\Sigma}_b\hat{\Sigma}_b^T. \tag{8}$$

The overall style-basis regularization is defined as:

$$\mathcal{L}_{style} = \lambda_{align}\mathcal{L}_{align} + \mathcal{L}_{orthstd}. \tag{9}$$

Although semantic anchors constrain the style bases, excessive style projection may still affect content-related facial structures or spoofing cues. To reduce this risk, we introduce a content consistency loss that encourages the original feature and its style-augmented counterpart to preserve consistent content information.

Given an original feature F, we randomly select a style basis (μ_k^b, σ_k^b) and generate a style-augmented feature:

$$F_{aug} = \sigma_k^b \cdot \frac{F - \mu(F)}{\sigma(F)} + \mu_k^b. \tag{10}$$

Global average pooling is then applied to F and F_{aug} to obtain v_{orig} and v_{aug}, respectively. The content consistency loss is formulated as:

$$\mathcal{L}_{con} = -\log\frac{\exp(\mathrm{sim}(v_{orig}, v_{aug})/\tau)}{\sum_j \exp(\mathrm{sim}(v_{orig}, v_j)/\tau)}, \tag{11}$$

where $\mathrm{sim}(\cdot,\cdot)$ denotes cosine similarity and τ is the temperature coefficient. This loss encourages the representation to be less sensitive to style variations while preserving discriminative facial content.

Style Projection. At inference, the style bases are frozen and no parameter update is required. Given the hierarchical style representation s_t of a test sample, we first project it into the semantic space and compute its similarity with each semantic anchor:

$$d_n = \mathrm{CosSim}(\Phi(s_t), A_n). \tag{12}$$

The projection weight for each basis is obtained by:

$$w_n = \frac{\exp(d_n/\tau)}{\sum_{j=1}^{N} \exp(d_j/\tau)}. \tag{13}$$

Finally, the semantically aligned style representation is reconstructed as:

$$s'_t = \sum_{n=1}^{N} w_n b_n. \tag{14}$$

The resulting s'_t retains the domain-specific style information of the test sample while being guided by explicit semantic anchors, and is subsequently fed into SPM for prompt generation.

3.3 Semantic Prompt Modulation

After obtaining the semantically aligned style representation s'_t from TASP, the next step is to convert it into prompts that can guide feature modulation for face anti-spoofing. A straightforward strategy is to treat s'_t as a conditioning vector and fuse it with visual features through concatenation-based projection or feature-wise modulation. However, such a global fusion treats all style information uniformly and ignores the fact that different prompt queries may require different layer-wise style cues. For example, blur and noise are more related to shallow visual patterns, while replay or print artifacts may involve more structural information from deeper layers. To address this issue, we introduce Semantic Prompt Modulation (SPM). Furthermore, inspired by the remarkable success of transformer architectures in broader visual recognition tasks [37,38], as well as their recent pioneering applications in FAS [13,39], we design a Style-Query Cross-Attention (SQ-CA) mechanism and integrate a Q-Former within SPM to adapt learnable prompt queries according to the hierarchical style representation.

Hierarchical Style Sequence. Since s'_t is constructed from the style statistics of multiple layers, we preserve its hierarchical layer-level structure instead of flattening it into a single global vector. Specifically, the aligned style representation is reshaped into a hierarchical style sequence:

$$S'_{\mathrm{seq}} = [\psi(s'^1_t), \psi(s'^2_t), \cdots, \psi(s'^L_t)] \in \mathbb{R}^{L\times D}, \tag{15}$$

where s'^l_t denotes the aligned style statistics of the l-th layer, and $\psi(\cdot)$ is a linear projection that maps each layer-wise style vector into a D-dimensional embedding. In this way, each token in S'_{seq} corresponds to the style information from a specific feature level.

Style-Query Cross-Attention. The detailed architecture of the proposed SQ-CA module is illustrated in Fig. 3. We utilize standard multi-head attention (MHA) to construct the cross-attention SQ-CA module.

Let $Q_{\mathrm{learn}} \in \mathbb{R}^{N_q\times D}$ denote a set of learnable prompt queries. Instead of directly using these static queries, SQ-CA adapts them according to the style sequence S'_{seq}. Specifically, Q_{learn} is used as the query, while S'_{seq} serves as both the key and value:

$$\Delta Q = \mathrm{MHA}(Q_{\mathrm{learn}}, S'_{\mathrm{seq}}, S'_{\mathrm{seq}}), \tag{16}$$

where $\mathrm{MHA}(\cdot)$ denotes multi-head cross-attention. The retrieved style-aware information is then integrated into the original queries through a residual connection:

$$Q_{\mathrm{dyn}} = \mathrm{LayerNorm}(Q_{\mathrm{learn}} + \Delta Q). \tag{17}$$

Through this operation, the learnable queries are transformed into dynamic queries conditioned on the style distribution of the current test sample.

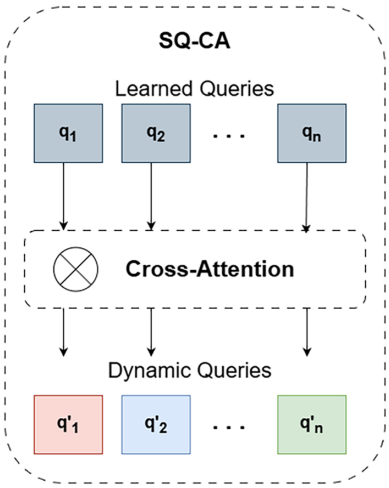

Figure 3: Detailed architecture of the SQ-CA module. This module takes the extracted visual features and semantic queries as inputs, and performs cross-modal interaction to generate the semantically rectified style representation.

Finally, the dynamic queries Q_{dyn} are fed into the Q-Former to interact with the visual feature F, producing the prompt representation P_t. The generated prompt is then used to modulate the original visual feature, resulting in the adapted feature F' for final classification. Compared with static prompt learning, SPM allows the prompt generation process to be conditioned on instance-specific style variations, making the model more responsive to unseen domain shifts.

3.4 Training and Inference

Training. During training, we optimize the learnable components of SA-TTDG using labeled source-domain samples. Given an input image x and its binary label y, where $y = 1$ denotes a bona fide face and $y = 0$ denotes a spoof attack, the classifier is supervised by the standard binary cross-entropy loss:

$$\mathcal{L}_{cls} = -[y \log \hat{y} + (1 - y) \log(1 - \hat{y})], \tag{18}$$

where $\hat{y}$ denotes the predicted liveness probability.

The overall training objective combines the classification loss, the style-basis regularization loss, and the content consistency loss:

$$\mathcal{L}_{total} = \mathcal{L}_{cls} + \mathcal{L}_{style} + \lambda_c \mathcal{L}_{con}. \tag{19}$$

Here, $\mathcal{L}_{style}$ constrains the learnable style bases to remain semantically aligned with the text-derived anchors while preserving sufficient statistical diversity, and $\mathcal{L}_{con}$ prevents the style projection process from corrupting the content-discriminative information. In this way, the model learns semantically grounded style bases and domain-sensitive prompts without relying on unconstrained random basis optimization.

Inference. During inference, all model parameters are fixed. For each unseen test sample, SA-TTDG first extracts the hierarchical style representation s_t and obtains the semantically aligned style representation s'_t through TASP according to Eqs. (11)–(13). Then, SPM transforms s'_t into dynamic prompts, which modulate the visual feature F for final liveness classification. The entire inference process is performed in

a single feed-forward manner, without gradient-based model updates, iterative optimization, or test-time parameter tuning.

3.5 Differences from Existing TTDG and VLM-FAS Methods

SA-TTDG differs from existing TTDG-based FAS methods mainly in how the style-basis space is constructed. Previous TTDG methods usually represent test-sample style statistics with randomly initialized or source-domain-derived bases, which act as implicit statistical prototypes. In contrast, SA-TTDG initializes and regularizes the style bases with CLIP-derived textual anchors, so that each basis is associated with an interpretable style factor such as blur, noise, illumination shift, screen reflection, or paper texture.

SA-TTDG also differs from VLM-FAS methods that mainly use textual prompts or vision-language representations for image-level classification or feature learning. Here, textual semantics are used to regularize the style projection process itself, while SQ-CA converts the semantically aligned style representation into instance-conditioned prompts. Thus, the contribution is not a redesign of the visual backbone or a direct combination of CLIP and attention, but a semantic anchoring mechanism for test-time style projection.

4 Experiments

4.1 Experimental Setup

Datasets. We evaluate SA-TTDG on four widely used public FAS benchmarks: OULU-NPU (O) [40], CASIA-FASD (C) [1], Idiap Replay-Attack (I) [2], and MSU-MFSD (M) [41]. These datasets cover complementary acquisition settings used in cross-domain FAS evaluation. OULU-NPU reflects mobile and smartphone-based authentication with different illumination and device conditions. CASIA-FASD contains print and replay attacks captured by low- and medium-quality cameras, representing early camera-based access-control settings. Idiap Replay-Attack focuses on controlled replay attacks captured with webcam-like devices, which is relevant to desktop or kiosk-style verification. MSU-MFSD includes attacks captured with multiple cameras and smartphones, providing additional device diversity. Although these benchmarks are standard in FAS research, they are mostly collected under relatively controlled conditions and cannot fully represent high-security deployments such as border control or financial authentication.

Evaluation Protocols. To assess cross-domain performance, we adopt the widely recognized leave-one-domain-out testing scheme. Specifically, our empirical evaluations are structured across four distinct adaptation scenarios: I&C&M → O, O&C&M → I, O&C&I → M, and O&M&I → C. During each phase, the model is optimized on a combination of three source datasets and subsequently evaluated on the single withheld dataset, which serves as the unfamiliar target environment.

Evaluation Metrics. We adopt Half Total Error Rate (HTER) and Area Under the Curve (AUC) as evaluation metrics. HTER is defined as the average of the false acceptance rate (FAR) and the false rejection rate (FRR):

$$\mathrm{HTER} = \frac{\mathrm{FAR} + \mathrm{FRR}}{2}. \tag{20}$$

A lower HTER indicates fewer classification errors, while a higher AUC reflects stronger overall discrimination between bona fide and spoof samples.

4.2 Implementation Details

All datasets used in this study consist of video data. We use FFmpeg [42] to extract frames from videos and Dlib [43] to detect faces in the extracted frames. For each video, we uniformly sample 12 frames. After

frame extraction, we verify whether each sampled frame contains a valid face. If fewer than 12 valid facial frames are obtained from a video, the existing valid frames are repeated for padding. All face images are resized and cropped to a resolution of 224 × 224 pixels. During training, we apply random rotation, color jittering, random erasing, and horizontal flipping for data augmentation. We implement the proposed SA-TTDG framework using PyTorch. All experiments are conducted on a single ASUS GeForce RTX 5070 Ti GPU.

We employ the CLIP ViT-B/16 architecture as our primary visual backbone, ensuring that both the image and text encoders remain in a frozen state throughout the entire training phase, and the CLIP text encoder is used to generate semantic anchors from the predefined style descriptors. For hierarchical style-statistic extraction, we collect patch-token features from the 3rd, 6th, 9th, and 12th Transformer blocks of CLIP ViT-B/16. For each selected block, the class token is removed, and the remaining 14 × 14 patch tokens are used to compute channel-wise mean and standard deviation. Thus, the number of selected feature levels is $L = 4$, and the channel dimension of each selected layer is $C_l = 768$. The statistics from the four selected layers are concatenated and projected into the 512-dimensional CLIP-aligned semantic space through style-to-semantic adapters. For AdaIN-style feature projection, the normalization statistics are computed from the final projected visual feature map, while the hierarchical style statistics are used to estimate the semantic-anchor projection weights. The number of semantic style bases is set to $N = 64$. The number of learnable prompt queries in SPM is set to $N_q = 16$, limiting the Q-Former depth to a single layer. The style-to-semantic adapters are implemented as Linear–LayerNorm projection modules, while other projection layers follow a compact MLP design when dimensional transformation is required. The architecture employs a hidden dimension size of 512, integrated with a dropout rate of 0.1 across the relevant network layers. Specifically, Ψ maps the CLIP textual embedding to the visual style-statistic space, while Φ maps the visual style representation back to the CLIP textual embedding space. This bottleneck design bridges the semantic and visual style spaces without introducing excessive additional capacity. The learnable style bases, projection adapters, SQ-CA module, Q-Former, and classification head are optimized using labeled source-domain data.

The model is trained for 300 epochs using the Adam optimizer with an initial learning rate of 1×10^{-5} and a batch size of 12. The loss weights are set to $\lambda_{align} = 10.0$ and $\lambda_c = 0.4$ unless otherwise specified. For fair comparison, all internal ablation variants use the same backbone, input resolution, preprocessing strategy, data augmentation, optimizer, training schedule, and evaluation protocol. For external baseline methods, we report the results from their original papers under the same leave-one-domain-out protocols when available, since many published FAS methods use different backbone architectures and training pipelines. Since the training code and random seeds of many external baselines are not consistently available, we do not report standard deviations for external methods. Instead, we evaluate the stability of our own method and key internal variants by repeating them with three random seeds, i.e., 24, 25, and 26. The semantic anchor vocabulary is kept fixed across different runs, while parameter initialization, data shuffling, and data augmentation randomness are changed.

4.3 Comparisons to the State-of-the-Art Methods

To provide a comprehensive evaluation, we divide the comparison into two groups: comparisons with conventional DG-based FAS methods and comparisons with recent TTDG-based methods. Tables 2 and 3 document the empirical results, with Figs. 4 and 5 offering visual evidence to support our comparative analysis.

Table 2: Comparison with state-of-the-art FAS methods in four standard cross-domain benchmarks.

Methods	I&C&M to O		O&C&M to I		O&C&I to M		O&M&I to C	
	HTER (%)	AUC (%)	HTER (%)	AUC (%)	HTER (%)	AUC (%)	HTER (%)	AUC (%)
MADDG [24]	27.98	80.02	22.19	84.99	17.69	88.06	24.50	84.51
SSDG [16]	25.17	81.83	18.21	94.61	16.67	90.47	23.11	85.45
RFM [44]	16.45	91.16	17.30	90.48	13.89	93.98	20.27	88.16
SDA [45]	23.10	84.30	15.60	90.10	15.40	91.80	24.50	84.40
DRDG [46]	15.63	91.75	15.56	91.79	12.43	95.81	19.05	88.79
ANRL [47]	15.67	91.90	16.03	91.04	10.83	96.75	17.85	89.26
SSAN [28]	19.51	88.17	14.00	94.58	10.42	94.76	16.47	90.81
EBDG [48]	15.66	92.02	18.69	92.28	9.56	97.17	18.34	90.01
IADG [49]	11.45	94.50	11.04	93.15	8.45	96.99	12.74	94.00
PatchNet [27]	11.82	95.07	13.40	95.67	7.10	98.46	11.33	94.58
TTDG [21]	10.00	95.70	6.50	97.98	7.91	96.83	8.14	96.49
SA-TTDG (Ours)	**8.37**	**97.94**	**3.87**	**99.41**	**5.00**	**98.56**	**4.20**	**99.48**

Note: Bold entries indicate the best performance.

Table 3: Comparison with test-time domain generalization methods.

Methods	I&C&M to O		O&C&M to I		O&C&I to M		O&M&I to C	
	HTER (%)	AUC (%)	HTER (%)	AUC (%)	HTER (%)	AUC (%)	HTER (%)	AUC (%)
TF-Cal [50]	13.29	93.71	19.75	90.35	12.08	95.58	14.26	92.10
DCN [51]	15.52	90.44	18.75	87.23	14.16	95.19	15.74	91.51
Sty.-Pro [52]	25.17	81.83	18.21	94.61	16.67	90.47	23.11	85.45
TTDG [21]	10.00	95.70	6.50	97.98	7.91	96.83	8.14	96.49
SA-TTDG (Ours)	**8.37**	**97.94**	**3.87**	**99.41**	**5.00**	**98.56**	**4.20**	**99.48**

Note: Bold entries indicate the best performance.

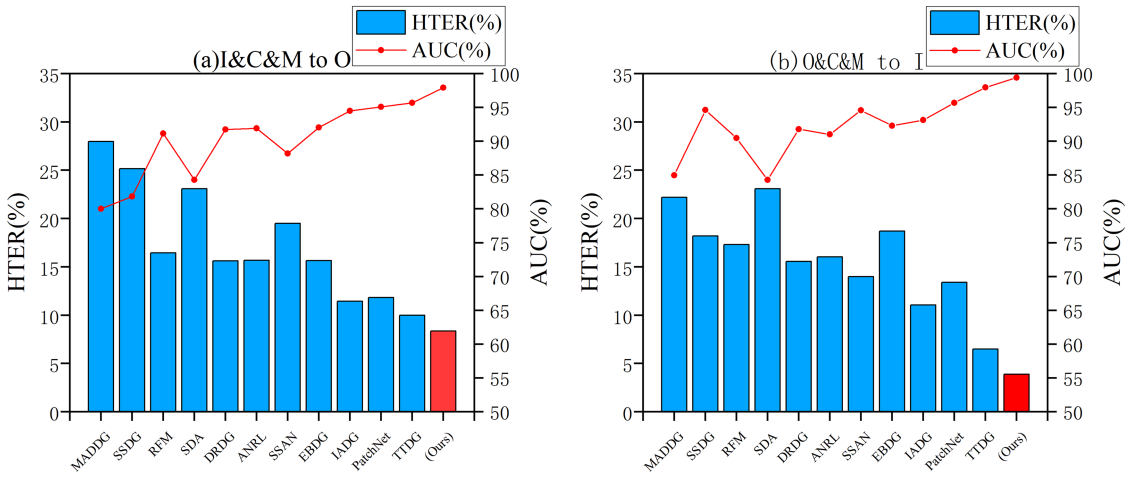

Figure 4: (Continued)

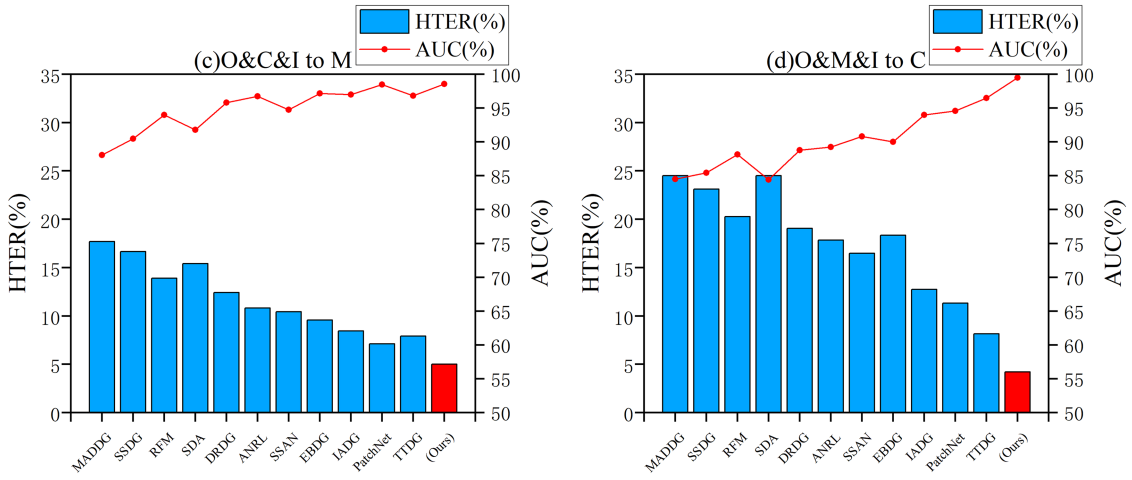

Figure 4: Comparison with the state-of-the-art FAS methods on four testing domains. HTER is represented by bars, and AUC by lines. (**a**) Performance evaluated on the OULU-NPU (O) target domain; (**b**) performance evaluated on the CASIA-FASD (C) target domain; (**c**) performance evaluated on the Idiap replay-attack (I) target domain; (**d**) performance evaluated on the MSU-MFSD (M) target domain.

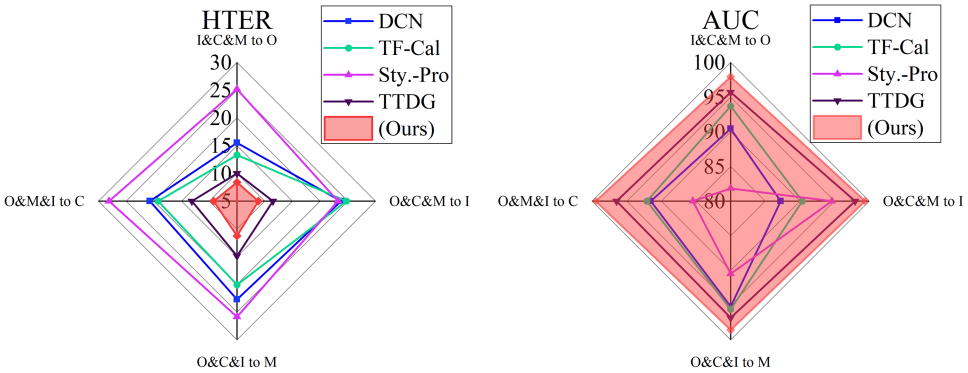

Figure 5: Comparison with test-time domain generalization methods. In terms of HTER, a diminished enclosed area on the chart directly reflects enhanced algorithmic performance. On the contrary, higher effectiveness is represented by an expanded area when evaluating the AUC metric.

Comparison with DG-based FAS methods. Table 2 and Fig. 4 compare –SA-TTDG with representative DG-based FAS methods, including adversarial learning, meta-learning, disentangled representation learning, attention-based methods, and patch-level modeling strategies. As shown in Table 2, SA-TTDG achieves the best overall performance across all four testing protocols. Specifically, our method obtains HTERs of 8.37%, 3.87%, 5.00%, and 4.20% on the I&C&M to O, O&C&M to I, O&C&I to M, and O&M&I to C protocols, respectively. Compared with strong DG-FAS baselines such as IADG and PatchNet, SA-TTDG consistently reduces the HTER while improving the AUC on most protocols.

Table 3 and Fig. 5 illustrate the comparison between our proposed method and existing test-time domain generalization approaches. In this context, SA-TTDG demonstrates highly competitive performance. The top-performing method is TTDG which introduces a test-time style projection mechanism. While TTDG generally outperforms most methods, its performance is still limited by the use of randomly initialized style bases, which leads to semantic ambiguity during adaptation and prevents effective utilization of high-level semantic information. In contrast, our SA-TTDG method overcomes this bottleneck by introducing

the TASP and SPM modules. These modules successfully integrate the robust semantic understanding capabilities of vision-language models with the dynamic style adaptation mechanisms of TTDG. Under the O&M&I to C protocol, SA-TTDG reduces the HTER from 8.14% to 4.20%. This result suggests that coupling semantic anchoring with style alignment can improve test-time style projection for cross-domain FAS.

4.4 Ablation Studies

To further verify the effectiveness of each key design in SA-TTDG, we conduct ablation studies under the same leave-one-out cross-domain protocols. The analysis focuses on four aspects: the style-query interaction mechanism in SPM, the initialization strategy of style bases in TASP, the contribution of different loss terms, and the stability of key variants under different random seeds.

Effect of Style-Query Interaction. Table 4 compares three interaction strategies in the SPM module: Static Query, MLP Adapter, and the proposed SQ-CA. Static Query directly uses the original learnable queries without incorporating test-time style information, while MLP Adapter injects the aligned style representation into the queries through a simple global mapping. Compared with these two alternatives, SQ-CA obtains the most favorable overall results among the compared interaction strategies. This indicates that merely adding style information is insufficient. Preserving the hierarchical structure of style features and allowing different queries to selectively attend to different style cues are both important for generating domain-sensitive prompts.

Table 4: Analysis of interaction mechanisms in the SPM module.

Methods	I&C&M to O		O&C&M to I		O&C&I to M		O&M&I to C	
	HTER (%)	AUC (%)	HTER (%)	AUC (%)	HTER (%)	AUC (%)	HTER (%)	AUC (%)
Static Query [35]	13.75	94.44	12.25	96.45	9.17	96.23	12.92	96.00
MLP	12.64	96.20	7.15	97.06	9.58	97.79	9.81	98.07
SQ-CA (Ours)	**8.37**	**97.94**	**3.87**	**99.41**	**5.00**	**98.56**	**4.20**	**99.48**

Note: Bold entries indicate the best performance.

Computational Overhead of SQ-CA. To evaluate the additional test-time cost introduced by SQ-CA, we compare the parameter number, FLOPs, and inference latency of different variants. As shown in Table 5, adding SQ-CA only slightly increases the computational cost. Compared with SA-TTDG without SQ-CA, the proposed SQ-CA module increases the latency from 11.35 to 11.72 ms/img, corresponding to a relative increase of 3.26%. Meanwhile, the FLOPs only increase from 17.59G to 17.61G. These results suggest that SQ-CA improves the trade-off between performance and test-time overhead.

Table 5: Test-time overhead introduced by SQ-CA.

Method	Params (M)	FLOPs (G)	Latency (ms/img)	Δ Latency (%)
TTDG	86.05	17.58	11.20	–
SA-TTDG w/o SQ-CA	86.12	17.59	11.35	0.00
SA-TTDG w/SQ-CA (Ours)	**88.22**	**17.61**	**11.72**	**+3.26**

Note: Bold entries indicate the best performance.

Effect of Style Basis Initialization. Table 6 evaluates different style basis initialization strategies, including FPS selection, TTSS [30], K-Means clustering, TTSP [21], and our semantic-anchor initialization.

The results show that source-domain-based strategies such as FPS and K-Means do not consistently improve generalization, since the selected or clustered bases may overfit the source-domain style distribution. While TTSS attempts to mitigate distribution shifts by aligning the style statistics of target samples toward the nearest source domain, this rigid mapping ignores the collaborative potential of other related source domains, thereby limiting the overall adaptability of the model. TTSP provides stronger flexibility, but they still lack explicit semantic constraints. In contrast, our SA-TTDG consistently achieves the lowest HTER and highest AUC on all testing domains. This demonstrates that semantic anchors provide a more stable and domain-agnostic basis space for test-time style projection.

Table 6: Impact of style basis initialization strategies.

Methods	I&C&M to O		O&C&M to I		O&C&I to M		O&M&I to C	
	HTER (%)	AUC (%)	HTER (%)	AUC (%)	HTER (%)	AUC (%)	HTER (%)	AUC (%)
TTSS [30]	15.10	93.30	12.50	92.91	9.58	95.88	13.14	93.65
K-Means	13.40	92.81	15.12	89.56	13.75	93.50	13.51	92.72
FPS Selection	11.35	95.14	10.12	95.42	10.00	94.64	12.40	94.55
TTSP [21]	10.00	95.70	6.50	97.98	7.91	96.83	8.14	96.49
SA-TTDG (Ours)	**8.37**	**97.94**	**3.87**	**99.41**	**5.00**	**98.56**	**4.20**	**99.48**

Note: Bold entries indicate the best performance.

Vocabulary Sensitivity of Semantic Anchors. To further examine whether the performance gain comes from meaningful semantic priors rather than merely additional learnable bases, we evaluate several variants of the anchor vocabulary. As shown in Table 7, the original FAS-oriented anchors achieve the best overall performance across all protocols. Replacing descriptors with close synonyms or using an alternative prompt template only causes mild performance changes, indicating that SA-TTDG is not overly sensitive to exact word choices. However, reducing the number of anchors or replacing them with unrelated textual concepts leads to clear performance degradation. This suggests that both sufficient anchor coverage and semantic relevance are important for constructing a reliable style-basis space. Compared with random bases without text initialization, the original semantic anchors consistently achieve lower HTER and higher AUC, supporting the usefulness of CLIP-derived linguistic priors.

Table 7: Vocabulary sensitivity analysis of semantic anchors.

Anchor Vocabulary	I&C&M to O		O&C&M to I		O&C&I to M		O&M&I to C	
	HTER (%)	AUC (%)	HTER (%)	AUC (%)	HTER (%)	AUC (%)	HTER (%)	AUC (%)
Original Anchors (Ours)	**8.37**	**97.94**	**3.87**	**99.41**	**5.00**	**98.56**	**4.20**	**99.48**
Synonym Anchors	8.54	96.74	4.05	99.12	5.46	98.31	4.46	99.21
Template Variant	8.42	97.44	5.21	98.33	5.29	98.46	4.25	99.27
Reduced Anchors (75%)	9.22	96.91	4.78	98.86	5.84	97.99	5.74	98.72
Reduced Anchors (50%)	11.45	95.50	7.04	97.15	8.45	97.35	6.37	98.08
Unrelated Text Anchors	13.82	94.87	8.24	95.67	9.72	96.36	9.62	95.58
Random Bases w/o Text	10.00	95.70	6.50	97.98	7.91	96.83	8.14	96.49

Note: Bold entries indicate the best performance.

Effect of Loss Components. Table 8 investigates the contribution of different loss terms in SA-TTDG, including the classification loss $\mathcal{L}_{cls}$, the content consistency loss $\mathcal{L}_{con}$, the semantic alignment loss $\mathcal{L}_{align}$, and the variance orthogonal regularization $\mathcal{L}_{orthstd}$. When only $\mathcal{L}_{cls}$ is used, the model obtains relatively high

HTERs on all four protocols, indicating that source-domain classification supervision alone is insufficient for learning robust style bases. Introducing $\mathcal{L}_{con}$ generally improves performance by encouraging the style-adapted features to preserve content-discriminative information. Adding $\mathcal{L}_{align}$ further reduces the HTER on most protocols, suggesting that constraining the learnable bases with text-derived semantic anchors helps alleviate semantic drift during basis optimization.

Table 8: Effectiveness of individual loss components in the SA-TTDG framework.

Loss Components				I&C&M to O		O&C&M to I		O&C&I to M		O&M&I to C	
$\mathcal{L}_{cls}$	$\mathcal{L}_{con}$	$\mathcal{L}_{align}$	$\mathcal{L}_{orthstd}$	HTER (%)	AUC (%)	HTER (%)	AUC (%)	HTER (%)	AUC (%)	HTER (%)	AUC (%)
✓				15.96	90.77	16.40	91.65	16.16	92.44	18.53	89.77
✓	✓			14.10	93.00	13.13	94.92	15.37	93.43	16.63	93.40
✓		✓		13.22	95.48	12.50	96.35	13.13	96.07	10.25	96.65
✓	✓	✓		12.08	96.42	9.17	97.52	9.92	97.82	8.92	97.88
✓		✓	✓	11.63	94.32	10.25	95.78	9.10	95.06	12.03	93.95
✓	✓	✓	✓	**8.37**	**97.94**	**3.87**	**99.41**	**5.00**	**98.56**	**4.20**	**99.48**

Note: Bold entries indicate the best performance.

The role of $\mathcal{L}_{orthstd}$ is slightly different. When used together with $\mathcal{L}_{cls}$ and $\mathcal{L}_{align}$, it improves the diversity of variance bases but does not always yield the best performance by itself, indicating that diversity regularization should be combined with content preservation. When all loss terms are jointly optimized, SA-TTDG achieves the best results on all protocols, with HTERs of 8.37%, 3.87%, 5.00%, and 4.20%, respectively. These results show that $\mathcal{L}_{con}$, $\mathcal{L}_{align}$, and $\mathcal{L}_{orthstd}$ play complementary roles: $\mathcal{L}_{con}$ preserves discriminative content, $\mathcal{L}_{align}$ maintains semantic consistency, and $\mathcal{L}_{orthstd}$ enhances style-basis diversity.

Stability under Different Random Seeds. To examine whether the reported improvements are sensitive to random initialization, we repeat SA-TTDG and two key internal variants under three random seeds, i.e., 24, 25, and 26. For each run, the semantic anchor vocabulary is kept fixed, while parameter initialization, data shuffling, and data augmentation randomness are changed. Table 9 reports the mean and standard deviation of HTER and AUC across the three runs. The results show that SA-TTDG consistently outperforms the two internal variants with small standard deviations, suggesting that the improvements brought by semantic anchoring and style-query interaction are stable under different random initializations.

Table 9: Stability analysis of key variants under three random seeds.

Methods	I&C&M to O		O&C&M to I		O&C&I to M		O&M&I to C	
	HTER (%)	AUC (%)	HTER (%)	AUC (%)	HTER (%)	AUC (%)	HTER (%)	AUC (%)
w/o SQ-CA	12.64 ± 0.42	96.20 ± 0.31	7.15 ± 0.28	97.06 ± 0.24	9.58 ± 0.36	97.79 ± 0.22	9.81 ± 0.39	98.07 ± 0.26
Random Bases w/o Text	10.00 ± 0.35	95.70 ± 0.27	6.50 ± 0.24	97.98 ± 0.19	7.91 ± 0.31	96.83 ± 0.25	8.14 ± 0.34	96.49 ± 0.28
SA-TTDG (Ours)	**8.37 ± 0.18**	**97.94 ± 0.12**	**3.87 ± 0.10**	**99.41 ± 0.06**	**5.00 ± 0.14**	**98.56 ± 0.09**	**4.20 ± 0.12**	**99.48 ± 0.05**

Note: Bold entries indicate the best performance.

4.5 Visualization and Analysis

T-SNE [53] visualization of the semantic anchor space. To further verify whether the proposed semantic anchors provide an interpretable and semantically meaningful projection space, we visualize both

the text-derived semantic anchors and the test samples in a joint 2D space using t-SNE. For each test image, we extract its hierarchical style statistics s_t and project it into the semantic space via $\Phi(s_t)$.

As shown in Fig. 6a, the predefined semantic anchors inherently form structured clusters corresponding to their major style categories. Crucially, Fig. 6b demonstrates that test samples from the four distinct benchmark datasets (OULU-NPU, CASIA-FASD, Idiap Replay-Attack, and MSU-MFSD) are well-aligned within this space, indicating that our framework effectively mitigates domain shift. Furthermore, as illustrated in Fig. 6c, test samples correctly gather around their semantically related anchors based on their intrinsic style factors. This compelling observation indicates that SA-TTDG does not merely learn unconstrained statistical bases, but successfully constructs a strictly interpretable projection space for robust test-time style calibration.

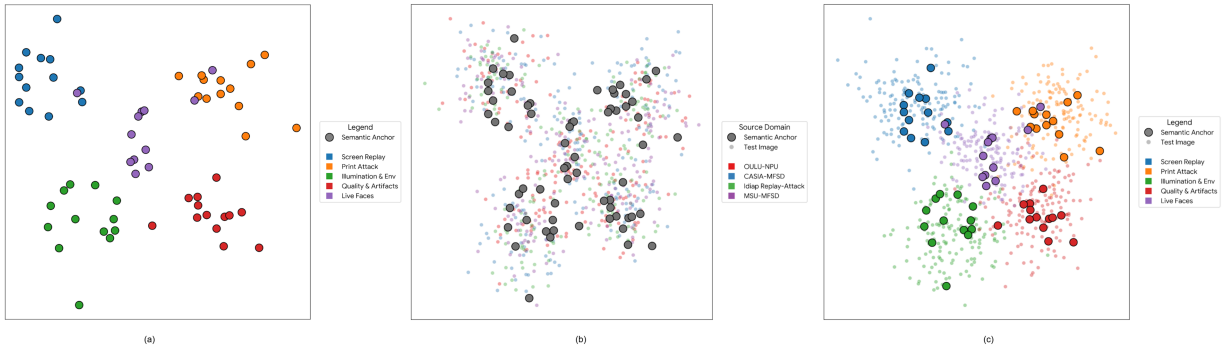

Figure 6: t-SNE visualization of the proposed semantic anchor space. (**a**) Semantic anchors form distinct structural clusters based on text. (**b**) Test samples from four domains align well with the anchors, demonstrating effective domain shift mitigation. (**c**) Test images correctly gather around semantically related anchors, showing that SA-TTDG constructs a strictly interpretable projection space.

Hyperparameter analysis. During the experimental process, the balance of weights among different loss components significantly impacts overall performance. Therefore, we investigate the influence of the hyperparameter λ_c on the SA-TTDG framework.

As illustrated in Fig. 7, under the I&C&M to O testing protocol, increasing λ_c up to 0.4 leads to a continuous decrease in HTER and a steady increase in AUC, indicating a progressive improvement in model performance. The model achieves optimal performance at $\lambda_c = 0.4$. However, further increasing this weight results in a slight performance degradation. Consequently, we observe that an overly small λ_c fails to adequately facilitate the training process and tends to disrupt or lose the inherent content and structural features of the face. Conversely, an excessively large λ_c leads to over-regularization. This over-regularization restricts the model's capacity for flexible adaptation to domain-specific features, ultimately causing a slight decline in performance.

Based on these experimental findings and empirical validations, setting $\lambda_c = 0.4$ across all experiments constitutes the most rational configuration. Similarly, we analyze the impact of the number of style bases, N, on the SA-TTDG framework under the identical I&C&M to O testing protocol. A smaller N results in an insufficient style representation capacity, whereas a larger N introduces excessive noise. Both extremes inevitably lead to suboptimal experimental outcomes. Therefore, we establish $N = 64$ as the default value for all experiments.

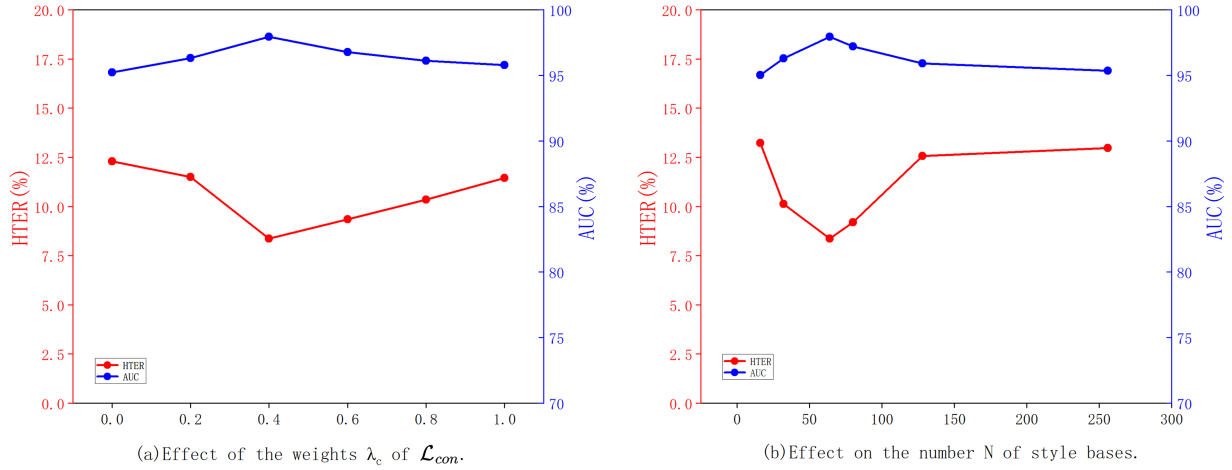

Figure 7: Hyperparameter analysis under the I&C&M to O test protocol. (**a**) Effect of the hyperparameter weight λ_c for the content consistency loss; (**b**) effect of the number of style bases N on the model performance.

5 Conclusion

In this work, we propose using language priors as explicit semantic anchors to constrain the style-basis space for domain-generalized FAS. To this end, we present the SA-TTDG framework, which utilizes two core modules: TASP, designed to enforce these semantic constraints, and by integrating a SQ-CA mechanism and a Q-Former within the SPM, developed to adaptively modulate instance-specific visual features, our approach effectively promotes cross-domain generalization.

Limitations and Future Work. Although SA-TTDG improves cross-domain performance by anchoring style bases with textual semantics, several limitations remain. First, the current experiments are conducted on standard public FAS benchmarks, which are mostly collected under relatively controlled conditions. Therefore, the results should be interpreted under these benchmark protocols and may not fully reflect high-security real-world deployments such as border control, financial authentication, or large-scale surveillance-assisted verification. Second, the current anchor vocabulary may not fully cover completely unprecedented style shifts or unseen attack materials. When the target domain contains attack artifacts that are far from all predefined anchors, such as new 3D mask materials or highly diverse acquisition conditions in datasets like SiW, SiW-M, and WMCA, the estimated anchor similarities may become less reliable. Third, although the current model extracts hierarchical style statistics from multiple CLIP Transformer blocks, we do not exhaustively investigate more fine-grained per-category sub-bases or flat vs. hierarchical anchor allocation strategies. In future work, we will explore automatic anchor expansion, open-vocabulary anchor retrieval, evaluation on more diverse real-world datasets, and deployment-oriented constraints such as sensor type, latency, and environmental variability.

Acknowledgement: The author, Xiaosong Chang, expresses sincere gratitude to his supervisors for their invaluable guidance, support, and encouragement throughout this research. His mentorship played a significant role in shaping the direction and quality of the study. The author is also deeply thankful to his family and friends for their constant support, understanding, and patience during this academic journey.

Funding Statement: The authors received no specific funding for this study.

Author Contributions: Xiaosong Chang was solely responsible for the conception, design, software development, data collection, analysis, and interpretation of results. He also prepared the original draft, carried out revisions, managed

visualizations, and handled the overall administration of the research. Liang Shi and Ao Zhang supervised the research, providing critical feedback and oversight. All authors reviewed and approved the final version of the manuscript.

Availability of Data and Materials: The data used in this study are available from the corresponding author upon reasonable request.

Ethics Approval: Not applicable.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Zhang Z, Yan J, Liu S, Lei Z, Yi D, Li SZ. A face antispoofing database with diverse attacks. In: 2012 5th IAPR International Conference on Biometrics (ICB); 2012 Mar 29–Apr 1; New Delhi, India. New York, NY, USA: IEEE; 2012. p. 26–31.
2. Chingovska I, Anjos A, Marcel S. On the effectiveness of local binary patterns in face anti-spoofing. In: 2012 BIOSIG-Proceedings of the International Conference of Biometrics Special Interest Group (BIOSIG); 2012 Sep 6–7; Darmstadt, Germany. New York, NY, USA: IEEE; 2012. p. 1–7.
3. Liu A, Zhao C, Yu Z, Wan J, Su A, Liu X, et al. Contrastive context-aware learning for 3D high-fidelity mask face presentation attack detection. IEEE Trans Inf Forensics Secur. 2022;17(11):2497–507. doi:10.1109/tifs.2022.3188149.
4. Xing H, Tan SY, Qamar F, Jiao Y. Face anti-spoofing based on deep learning: a comprehensive survey. Appl Sci. 2025;15(12):6891. doi:10.3390/app15126891.
5. Yu Z, Qin Y, Li X, Zhao C, Lei Z, Zhao G. Deep learning for face anti-spoofing: a survey. IEEE Trans Pattern Anal Mach Intell. 2022;45(5):5609–31. doi:10.1109/tpami.2022.3215850.
6. Liu A, Tan Z, Wan J, Liang Y, Lei Z, Guo G, et al. Face anti-spoofing via adversarial cross-modality translation. IEEE Trans Inf Forensics Secur. 2021;16:2759–72. doi:10.1109/tifs.2021.3065495.
7. Shao R, Perera P, Yuen PC, Patel VM. Federated generalized face presentation attack detection. IEEE Trans Neural Netw Learn Syst. 2022;35(1):103–16. doi:10.1109/tnnls.2022.3172316.
8. George A, Mostaani Z, Geissenbuhler D, Nikisins O, Marcel S. Biometric face presentation attack detection with multi-channel convolutional neural network. IEEE Trans Inf Forensics Secur. 2019;15:42–55. doi:10.1109/tifs.2019.2916652.
9. George A, Marcel S. Deep pixel-wise binary supervision for face anti-spoofing. IEEE Trans Biom Behav Identity Sci. 2020;2(2):182–93. doi:10.1109/tbiom.2021.3065526.
10. Yu Z, Zhao C, Wang Z, Qin Y, Su Z, Li X, et al. Searching central difference convolutional networks for face anti-spoofing. IEEE Trans Pattern Anal Mach Intell. 2021;43(9):3028–43. doi:10.1109/cvpr42600.2020.00534.
11. Yu Z, Wan J, Qin Y, Li X, Li SZ, Zhao G. NAS-FAS: static-dynamic central difference network search for face anti-spoofing. IEEE Trans Pattern Anal Mach Intell. 2020;43(9):3005–23.
12. Wang H, Shi Y, Feng J, Yu Z, Tao Z. PNSS: unknown face presentation attack detection with pseudo negative sample synthesis. Comput Mater Contin. 2025;83(2):3097–112. doi:10.32604/cmc.2025.061019.
13. Ameenulhakeem DWO, Uçan ON. A lightweight multimodal deep fusion network for face antis poofing with cross-axial attention and deep reinforcement learning technique. Comput Mater Contin. 2025;85(3):5671–702. doi:10.32604/cmc.2025.070422.
14. Shende SW, Tembhurne JV, Ansari NA. Deep learning based authentication schemes for smart devices in different modalities: progress, challenges, performance, datasets and future directions. Multimed Tools Appl. 2024;83(28):71451–93. doi:10.1007/s11042-024-18350-5.
15. Zhao Y, Zhong Z, Zhao N, Sebe N, Lee GH. Style-hallucinated dual consistency learning: a unified framework for visual domain generalization. Int J Comput Vis. 2024;132(3):837–53.
16. Jia Y, Zhang J, Shan S, Chen X. Single-side domain generalization for face anti-spoofing. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition; 2020 Jun 14–19; Online. Piscataway, NJ, USA: IEEE; 2020. p. 8484–93.

17. Qin Y, Zhao C, Zhu X, Wang Z, Yu Z, Fu T, et al. Meta-teacher for face anti-spoofing. IEEE Trans Pattern Anal Mach Intell. 2022;44(10):6306–20. doi:10.1109/TPAMI.2021.3091167.
18. Liang J, He R, Tan T. A comprehensive survey on test-time adaptation under distribution shifts. Int J Comput Vis. 2025;133(1):31–64. doi:10.1007/s11263-024-02181-w.
19. Liu Y, Chen Y, Dai W, Gou M, Huang CT, Xiong H. Source-free domain adaptation with domain generalized pretraining for face anti-spoofing. IEEE Trans Pattern Anal Mach Intell. 2024;46(8):5430–48. doi:10.1109/tpami.2024.3370721.
20. Ye Y, Wei W, Zhang L, Ding C, Zhang Y. Domain consistency learning for continual test-time adaptation in image semantic segmentation. Pattern Recognit. 2025;165:111585. doi:10.1016/j.patcog.2025.111585.
21. Zhou Q, Zhang KY, Yao T, Lu X, Ding S, Ma L. Test-time domain generalization for face anti-spoofing. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition; 2024 Jun 17–21; Seattle, WA, USA. Piscataway, NJ, USA: IEEE; 2024. p. 175–87.
22. Han T, Xu Z, Li W, Hu H, He X, He S, et al. Learning generic and specific prompts with contrastive constraints for multi-task visual scene understanding. Neurocomputing. 2025;657(7):131586. doi:10.1016/j.neucom.2025.131586.
23. Zhou K, Liu Z, Qiao Y, Xiang T, Loy CC. Domain generalization: a survey. IEEE Trans Pattern Anal Mach Intell. 2022;45(4):4396–415.
24. Shao R, Lan X, Li J, Yuen PC. 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 Jun 16–20; Long Beach, CA, USA. Piscataway, NJ, USA: IEEE; 2019. p. 10023–31.
25. Li H, Li W, Cao H, Wang S, Huang F, Kot AC. Unsupervised domain adaptation for face anti-spoofing. IEEE Trans Inf Forensics Secur. 2018;13(7):1794–809. doi:10.1109/tifs.2018.2801312.
26. Chen Z, Yao T, Sheng K, Ding S, Tai Y, Li J, et al. Generalizable representation learning for mixture domain face anti-spoofing. Pattern Recognit. 2021;114(2):107878. doi:10.1609/aaai.v35i2.16199.
27. Wang CY, Lu YD, Yang ST, Lai SH. 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 Jun 18–24; New Orleans, LA, USA. Piscataway, NJ, USA: IEEE; 2022. p. 20281–90.
28. Wang Z, Wang Z, Yu Z, Deng W, Li J, Gao T, et al. Domain generalization via shuffled style assembly for face anti-spoofing. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition; 2022 Jun 18–24; New Orleans, LA, USA. Piscataway, NJ, USA: IEEE; 2022. p. 4123–33.
29. Kim YE, Nam WJ, Min K, Lee SW. Style selective normalization with meta learning for test-time adaptive face anti-spoofing. Expert Syst Appl. 2023;214(2):119106. doi:10.1016/j.eswa.2022.119106.
30. Park J, Han DJ, Kim S, Moon J. Test-time style shifting: handling arbitrary styles in domain generalization. In: ICML'23: Proceedings of the 40th International Conference on Machine Learning; 2023 Jul 23–29; Honolulu, HI, USA. p. 27114–31.
31. Qi Y, Li H, Song Y, Wu X, Luo J. How vision-language tasks benefit from large pre-trained models: a survey. IEEE Trans Multimed. 2026;28(8):1188–210. doi:10.1109/tmm.2025.3632653.
32. Du Y, Liu Z, Li J, Zhao WX. A survey of vision-language pre-trained models. IEEE Trans Knowl Data Eng. 2022;35(7):6687–706. doi:10.24963/ijcai.2022/762.
33. Radford A, Kim JW, Hallacy C, Ramesh A, Goh G, Agarwal S, et al. Learning transferable visual models from natural language supervision. In: Proceedings of the 38 th International Conference on Machine Learning; 2021 Jul 18–24; Online. Cambridge, MA, USA: PMLR; 2021. p. 8748–63.
34. Zhou K, Yang J, Loy CC, Liu Z. Learning to prompt for vision-language models. IEEE Trans Pattern Anal Mach Intell. 2022;45(2):2139–52. doi:10.1109/cvpr52688.2022.01631.
35. Liu A, Xue S, Gan J, Wan J, Liang Y, Deng J, et al. CFPL-FAS: class free prompt learning for generalizable face anti-spoofing. In: Proceedings of the 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition; 2024 Jun 17–21; Seattle, WA, USA. Piscataway, NJ, USA: IEEE; 2024. p. 222–32.
36. Jing Y, Yang Y, Feng Z, Ye J, Yu Y, Song M. Neural style transfer: a review. IEEE Trans Vis Comput Graph. 2019;26(11):3365–85. doi:10.1109/tvcg.2019.2921336.

37. Han K, Wang Y, Chen H, Chen X, Guo J, Liu Z, et al. A survey on vision transformer. IEEE Trans Pattern Anal Mach Intell. 2022;45(1):87–110. doi:10.1109/tpami.2022.3152247.
38. Khan S, Naseer M, Hayat M, Zamir SW, Khan FS, Shah M. Transformers in vision: a survey. ACM Comput Surv. 2022;54(10s):1–41. doi:10.1145/3505244.
39. Liu A, Tan Z, Yu Z, Zhao C, Wan J, Liang Y, et al. FM-ViT: flexible modal vision transformers for face anti-spoofing. IEEE Trans Inf Forensics Secur. 2023;18:4775–86.
40. Boulkenafet Z, Komulainen J, Li L, Feng X, Hadid A. 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 May 30–Jun 3; Washington, DC, USA. Piscataway, NJ, USA: IEEE; 2017. p. 612–8.
41. Wen D, Han H, Jain AK. Face spoof detection with image distortion analysis. IEEE Trans Inf Forensics Secur. 2015;10(4):746–61. doi:10.1109/tifs.2015.2400395.
42. Tomar S. Converting video formats with FFmpeg. Linux J. 2006;2006(146):10.
43. King DE. Dlib-ml: a machine learning toolkit. J Mach Learn Res. 2009;10:1755–8.
44. Shao R, Lan X, Yuen PC. Regularized fine-grained meta face anti-spoofing. In: Proceedings of the 34th AAAI Conference on Artificial Intelligence; 2020 Feb 7–12; New York, NY, USA. Palo Alto, CA, USA: AAAI Press; 2020. p. 11974–81.
45. Wang J, Zhang J, Bian Y, Cai Y, Wang C, Pu S. Self-domain adaptation for face anti-spoofing. In: Proceedings of the 35th AAAI conference on artificial intelligence; 2021 Feb 2–9; Online. Palo Alto, CA, USA: AAAI Press; 2021. p. 2746–54.
46. Liu S, Zhang KY, Yao T, Sheng K, Ding S, Tai Y, et al. Dual reweighting domain generalization for face presentation attack detection. arXiv:2106.16128. 2021.
47. Liu S, Zhang KY, Yao T, Bi M, Ding S, Li J, et al. Adaptive normalized representation learning for generalizable face anti-spoofing. In: Proceedings of the 29th ACM International Conference on Multimedia; 2021 Oct 20–24; Online. New York, NY, USA: ACM; 2021. p. 1469–77.
48. Du Z, Li J, Zuo L, Zhu L, Lu K. Energy-based domain generalization for face anti-spoofing. In: Proceedings of the 30th ACM International Conference on Multimedia; 2022 Oct 10–14; Lisbon, Portugal. New York, NY, USA: ACM; 2022. p. 1749–57.
49. Zhou Q, Zhang KY, Yao T, Lu X, Yi R, Ding S, et al. Instance-aware domain generalization for face anti-spoofing. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition; 2023 Jun 17–24; Vancouver, BC, Canada. Piscataway, NJ, USA: IEEE; 2023. p. 20453–63.
50. Zhao X, Liu C, Sicilia A, Hwang SJ, Fu Y. Test-time Fourier style calibration for domain generalization. In: Proceedings of the 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition; 2023 Jun 17–24; Vancouver, BC, Canada. Piscataway, NJ, USA: IEEE; 2023. p. 17992–8001.
51. Jiang Y, Wang Y, Zhang R, Xu Q, Zhang Y, Chen X, et al. Domain-conditioned normalization for test-time domain generalization. In: Computer vision—ECCV 2022 Workshops (ECCV 2022). Cham, Switzerland: Springer; 2022. p. 291–307.
52. Huang W, Chen C, Li Y, Li J, Li C, Song F, et al. Style projected clustering for domain generalized semantic segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition; 2023 Jun 17–24; Vancouver, BC, Canada. Piscataway, NJ, USA; 2023. p. 3061–71.
53. Van der Maaten L, Hinton G. Visualizing data using t-SNE. J Mach Learn Res. 2008;9(86):2579–605.