Breaking the Trade-off: Orthogonal Semantic Decoupling for Generalizable and Fair Deepfake Detection

Zhongyu Shi[1,3], **Siyu Peng**[1], **Yimin Kang**[3], **Ruiyang Xia**[4], **Yizhi Fang**[1], **Weiping Wen**[1*], **Zeyu Gu**[2] and **Sai Cheng**[2]
[1]School of Software and Microelectronics, Peking University
[2]Xiaomi Corporation, Beijing, China
[3]School of Computer Science, Xiangtan University
[4]School of Electronic Engineering, Xidian University

Abstract

Deepfake detection faces dual challenges in real-world deployment: cross-domain generalization and demographic fairness. Existing approaches often struggle with a trade-off between these goals. Generalization-oriented detectors can over-rely on demographic shortcuts, while fairness constraints tend to steer optimization away from the most discriminative decision boundary. To address this, we propose Orthogonal Semantic Decoupling (OSD), a framework that decouples demographic semantics from forgery cues. Specifically, we perform Singular Value Decomposition on the pre-trained weights of a vision-language model, freezing the principal semantic subspace while learning parameter-efficient low-rank experts in the residual subspace. The experts comprise Demographic Semantic Experts, a set of experts specialized via group sampling and routed based on the similarities between image embeddings and text embeddings of predefined descriptions; and a Universal Forgery Expert, which captures forgery features transferable across domains and demographics. Extensive experiments across multiple benchmarks demonstrate that our approach outperforms state-of-the-art methods in both generalization and fairness, breaking the trade-off. The code is available at https://github.com/sonder-lin/osd-deepfake-detection.

1 Introduction

Recently, the rapid advances in generative AI techniques have significantly reduced the cost of synthesizing highly indistinguishable images [Rosberg *et al.*, 2023; Wu and De La Torre, 2023; Kim *et al.*, 2025]. Facial deepfakes, which focus on manipulating personal expressions, attributes, and identity, have severely threatened individual privacy and reputation as well as public safety [Wang *et al.*, 2024]. Consequently, developing reliable and robust face deepfake detection methods has become an urgent need.

*Corresponding author.

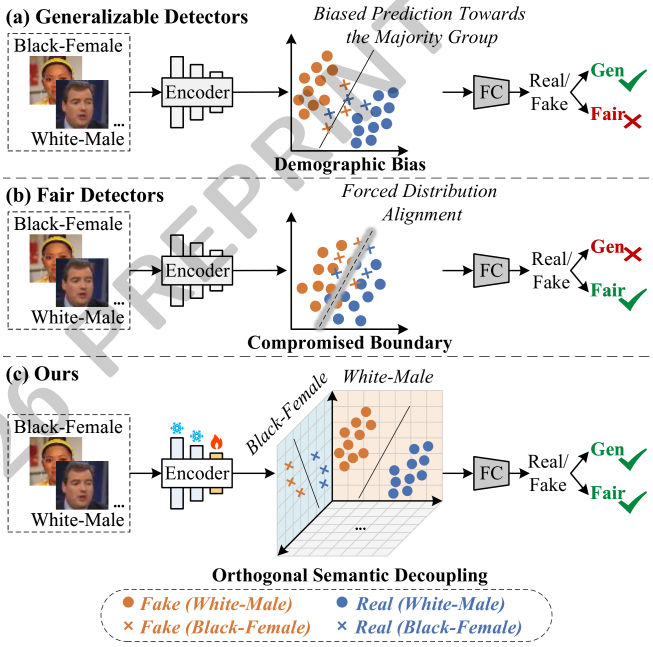

Figure 1: Comparison of (a) generic detectors with demographic bias, (b) fairness-aware training with a compromised decision boundary, and (c) our OSD that orthogonally decouples demographic semantics and forgery cues.

Early deepfake detectors primarily relied on manipulation-specific artifacts [Li and Lyu, 2019; Li *et al.*, 2020a; Qian *et al.*, 2020]. Such approaches suffer substantial performance drops on unseen manipulations [Luo *et al.*, 2021; Xia *et al.*, 2026]. To improve detector generalization, prior works have explored learning manipulation-invariant features [Liu *et al.*, 2021; Tan *et al.*, 2023], augmenting training data with generalized forgery traces [Shiohara and Yamasaki, 2022; Chen *et al.*, 2022], and leveraging pretrained visual foundation models [Radford *et al.*, 2021; Ojha *et al.*, 2023; Wu *et al.*, 2025; Yan *et al.*, 2025]. Besides, given that faces are demographically correlated, detector performance appears biased across different demographic groups, as shown empirically in [Lin *et al.*, 2025]. To improve detection fairness, recent studies have explored feature disentanglement [Lin *et al.*, 2024], risk-

based group-aware optimization [Ju *et al.*, 2024], and causal interventions [Cheng *et al.*, 2025b] to mitigate demographic bias and reduce performance disparities across subgroups.

However, prior works typically target either generalization or fairness in isolation. Generalization-oriented optimization is often dominated by the gradients of the majority-group population [Lin *et al.*, 2024], thereby biasing models toward majority-favored cues (Figure 1(a)). Fairness-oriented constraints can reduce subgroup disparities to improve detection performance under minority-group populations. Such enforced decision-boundary alignment, however, impacts discriminability and ultimately decreases cross-domain generalization [Cheng *et al.*, 2025b] (Figure 1(b)). As a result, it seems that there may be a trade-off between generalization and fairness, which impacts real-world detector deployments.

To improve both generalization and fairness, we revisit a structural perspective on where forgeries come from. Since facial deepfake images are always manipulated from real faces, there are high semantic similarities within the extracted feature space. The determination of authenticity can be restricted to a smaller, semantically consistent subspace.

Based on this insight, we propose Orthogonal Semantic Decoupling (OSD), a framework that constructs an orthogonal, low-rank mixture-of-experts in the residual subspace of a vision-language model (VLM). Specifically, OSD applies Singular Value Decomposition to the pretrained weights of the VLM to orthogonally construct the principal subspace and the residual subspace. Within the residual subspace, OSD instantiates a set of mutually orthogonal Demographic Semantic Experts (DSEs), which are selected through a text-guided and training-free router to adapt to demographic-specific distribution shifts. In parallel, a Universal Forgery Expert (UFE) is introduced to capture transferable face forgery cues shared across domains and demographics. By factorizing the decision space into mutually orthogonal demographic subgroup subspaces, OSD suppresses cross-domain interference and demographic-induced spurious correlations, while enabling a clear within-subgroup distinction in a compact subspace (Figure 1(c)), thereby effectively breaking the trade-off.

The contributions of this paper are summarized as follows:

- We propose Orthogonal Semantic Decoupling (OSD), a framework for generalizable and fair deepfake detection. OSD operates in the residual visual feature subspace and explicitly decouples Demographic Semantic Experts (DSEs) and a Universal Forgery Expert (UFE), which suppresses demographic–forgery correlations.
- We develop a training strategy for OSD that encourages the model to focus on demographic-specific and demographic-agnostic forgery cues, respectively. Based on the image–text alignment in VLM, DSEs are accurately selected with a text-guided and training-free router during inference.
- Extensive experiments on multiple face deepfake benchmarks demonstrate improvements in detection generalization as well as fairness compared with state-of-the-art methods, effectively breaking the generalization–fairness trade-off.

2 Related Work

2.1 Generalization in Deepfake Detection

Improving cross-domain generalization under unseen manipulations remains an active research focus. Existing generalization-oriented methods broadly fall into two categories: forgery-pattern learning and adaptation of pretrained visual foundation models.

Forgery pattern learning improves transferability by mining low-level artifacts or statistical inconsistencies shared across manipulation methods. F3-NET [Qian *et al.*, 2020] and SPSL [Liu *et al.*, 2021] model spectral anomalies in the frequency domain, while LGRAD [Tan *et al.*, 2023] learns more generalizable forgery representations via gradient-based responses. In the spatial domain, FACE X-RAY [Li *et al.*, 2020a] and IID [Huang *et al.*, 2023] exploit blending-boundary cues and identity-consistency constraints, respectively. In addition, reconstruction-based anomaly detection methods, such as RECCE [Cao *et al.*, 2022] and DIRE [Wang *et al.*, 2023], increase sensitivity to unseen generators using reconstruction errors or inversion-related features. To reduce reliance on a specific training distribution, SBI [Shiohara and Yamasaki, 2022] and SLADD [Chen *et al.*, 2022] further introduce self-supervised or self-synthesized strategies to increase the diversity of forged samples.

Adaptations of visual foundation models have emerged as an important trend, leveraging the generalizable representations provided by visual foundation models to improve generalization. UNIFD [Ojha *et al.*, 2023] performs classification in a fixed, general-purpose feature space, reducing overfitting to specific generator distributions. LASTED [Wu *et al.*, 2025] introduces semantic-level priors into detector training, for example via language-guided contrastive learning. EFFORT [Yan *et al.*, 2025] shows that naive fine-tuning may cause feature-space collapse and proposes an SVD-based orthogonal subspace decomposition: it freezes the principal components to preserve pretrained semantics and learns forgery patterns only in the residual subspace. Building on this design, OMNIAID [Guo *et al.*, 2025] introduces hybrid semantic and artifact experts, but targets universal AI-generated image detection rather than deepfake detection.

2.2 Fairness in Deepfake Detection

As deepfake detection continues to advance, performance disparities across demographic groups, such as race and gender, have received growing attention. Trinh et al. [Trinh and Liu, 2021] argue that imbalanced training distributions are a primary driver of group disparities and can encourage detectors to rely on spurious correlations. Building on this line of work, Xu et al. [Xu *et al.*, 2024] construct large-scale, fine-grained attribute annotations and a slice-based evaluation protocol, enabling more precise measurement of demographic bias and potential failure risks. To mitigate these issues, recent studies aim to dynamically reduce the influence of sensitive attributes during representation learning or decision making. Lin et al. [Lin *et al.*, 2024] propose the PFG framework, which uses decoupled learning to extract demographic-invariant forgery features. From a risk-optimization perspective, Ju et al. [Ju *et al.*, 2024] introduce a CVaR-based fairness objective to im-

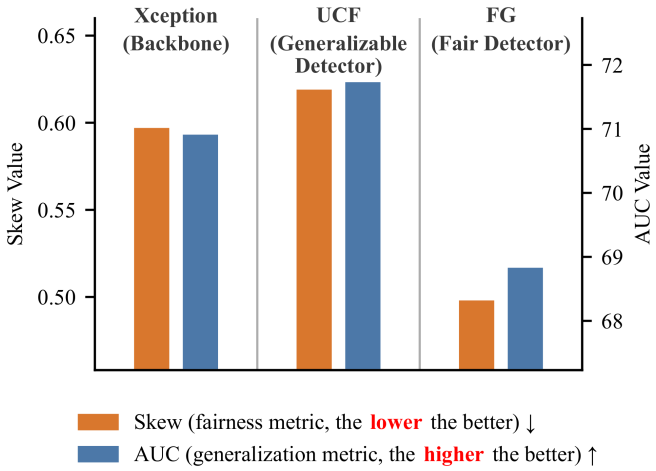

Figure 2: Illustration of the generalization–fairness trade-off in prior works. AUC and Skew [Geyik *et al.*, 2019] are reported on CDFv2, with all methods trained on FF++. UCF [Yan *et al.*, 2023] and FG [Lin *et al.*, 2024] adopt Xception [Chollet, 2017] as backbones.

prove robustness across groups by targeting the worst-case group or samples. Cheng et al. [Cheng *et al.*, 2025b] connect fairness to cross-domain generalization via causal analysis and mitigate demographic bias through demographic-aware rebalancing and demographic-agnostic feature aggregation.

3 Methodology

3.1 Problem Definition

Face deepfake detection is formulated as a binary classification problem: given a face image x, a detector outputs $\hat{y} \in \{0, 1\}$, where $\hat{y} = 0$ denotes a real face and $\hat{y} = 1$ denotes a synthesized (forged) face. In practical deployment, we consider two objectives for face deepfake detection: cross-domain generalization and demographic fairness. Let $a \in \mathcal{A}$ denote a sensitive attribute (e.g., race or gender). We quantify generalization by a metric $\mathcal{G}(\cdot)$ and fairness by a metric $\mathcal{F}(\cdot)$ that captures performance disparities across subpopulations. A desirable detector should perform well on both objectives, remaining robust under distribution shifts while behaving consistently across demographic groups. However, prior works typically optimize these objectives in isolation, i.e., improving either $\mathcal{G}$ or $\mathcal{F}$ while overlooking the other, as illustrated in Figure 2. Therefore, our goal is to develop a detector that simultaneously improves generalization and fairness, achieving strong performance on both $\mathcal{G}$ and $\mathcal{F}$.

3.2 Overview of the Proposed Framework

To improve both fairness and cross-domain generalization in face deepfake detection, we propose Orthogonal Semantic Decoupling (OSD). Our key observation is that demographic attributes, such as race and gender, define relatively stable high-level semantics: they shift the facial appearance distribution but are unrelated to whether an image is forged, making them a natural basis for forgery-detection subspace partitioning. Based on this, we perform deepfake detection within smaller, semantically consistent, and mutually independent subspaces. Figure 3 provides an overview of OSD.

OSD operationalizes this idea on top of a pretrained vision–language model (VLM) by freezing a shared principal subspace for general representations and learning only a set of low-rank residual experts. Concretely, we instantiate *Demographic Semantic Experts* indexed by race $\times$ gender, and route each input to the corresponding expert based on the similarity between image embeddings and text embeddings of predefined demographic descriptions; in parallel, a *Universal Forgery Expert* is always activated to extract demographic-invariant manipulation cues. The final prediction combines expert-specific forgery cues from the routed Demographic Semantic Expert with universal forgery cues from the Universal Forgery Expert. To encourage complementary and separable representations, we enforce pairwise orthogonality among all Demographic Semantic Experts as well as orthogonality between each expert and the frozen principal subspace. We adopt a two-stage training procedure that learns demographic-specific and demographic-agnostic forgery cues separately, improving both cross-group consistency and cross-domain generalization.

3.3 Demographic-Conditioned Residual Experts

Let $W \in \mathrm{R}^{d_o \times d_i}$ denote the pretrained weight of a self-attention projection in a vision–language model (VLM) (e.g., one of the $Q/K/V$ or output projection matrices). We compute the SVD $W = U\Sigma V^\top$ and keep the top r_{main} singular components (with corresponding U_r, Σ_r, V_r) to form a frozen principal subspace

$$W_{\text{main}} := U_r \Sigma_r V_r^\top. \tag{1}$$

We define the residual component as $W_{\text{res}} = W - W_{\text{main}}$. Both the Demographic Semantic Experts and the Universal Forgery Expert are initialized from this residual subspace by using the leading singular directions of W_{res} to initialize each expert's low-rank factors. On top of this, each expert e learns a low-rank residual update

$$\Delta W_e := U_e \,\mathrm{diag}(s_e) V_e^\top, \tag{2}$$

where $U_e \in \mathrm{R}^{d_o \times r_{\text{exp}}}$, $V_e \in \mathrm{R}^{d_i \times r_{\text{exp}}}$, $s_e \in \mathrm{R}^{r_{\text{exp}}}$, and $r_{\text{exp}} \ll \min(d_o, d_i)$. Given an input x, we obtain its image embedding $z_I = f_I(x)$ and select the expert index $e^*(x)$ using a text-guided router (see Sec. 3.4 for construction details). Given an input feature $h \in \mathrm{R}^{d_i}$, the layer output is

$$y = W_{\text{main}} h + \lambda_s \Delta W_{e^*(x)} h + \lambda_f \Delta W_f h, \tag{3}$$

Here, f denotes the Universal Forgery Expert, and λ_s and λ_f are constant fusion weights for the routed Demographic Semantic Expert and the Universal Forgery Expert, respectively.

To prevent different experts from updating the model along the same directions and interfering with each other, we impose an orthogonality regularization inspired by OMNI-AID [Guo *et al.*, 2025] on the subspace spanned by the low-rank residual updates. Concretely, we encourage each active expert's bases (U_e, V_e) to be orthogonal to the frozen principal subspace (U_r, V_r) and, for DSEs, to the previously trained demographic experts. We define the orthogonality loss as

$$L_{\text{orth}}^{(a)} = \frac{1}{|\mathcal{C}_a|} \sum_{j \in \mathcal{C}_a} \left(\|\bar{U}_a^\top \bar{U}_j\|_F + \|\bar{V}_a^\top \bar{V}_j\|_F \right), \tag{4}$$

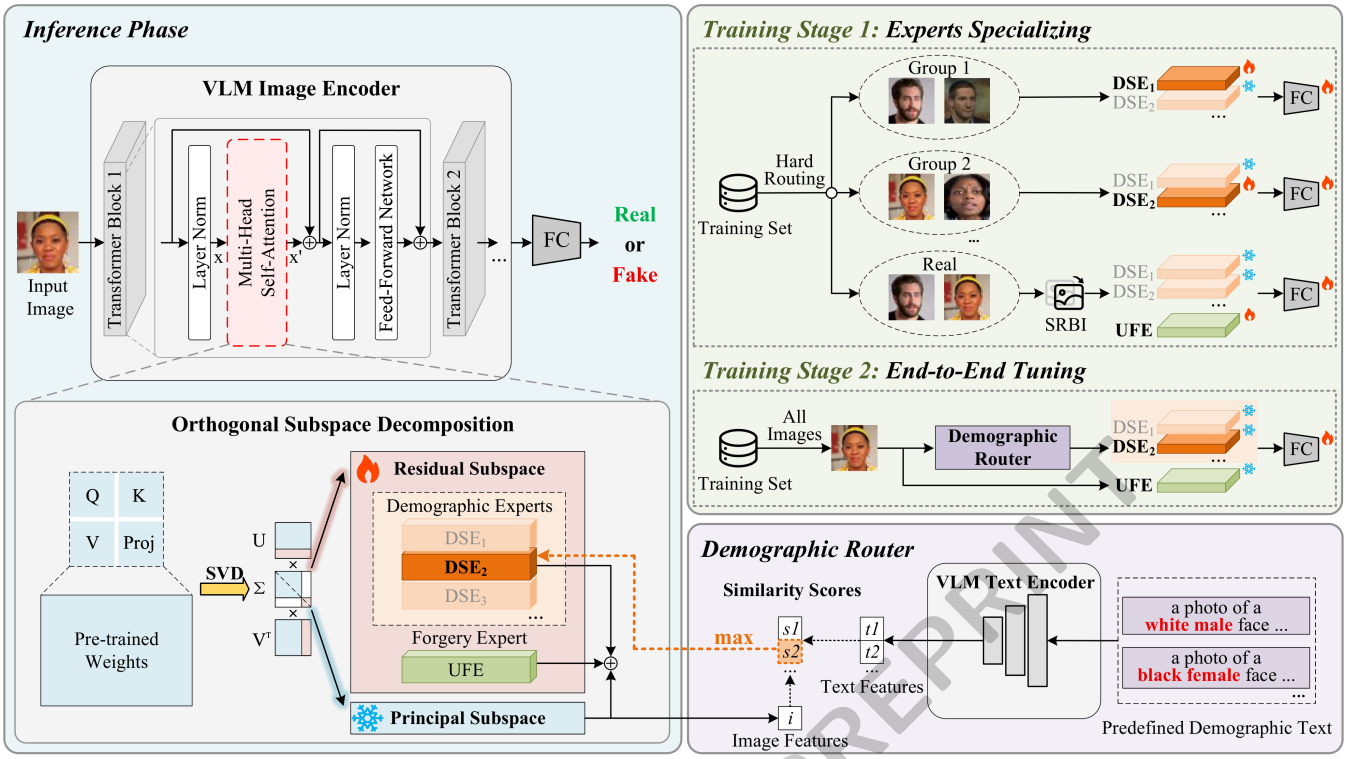

Figure 3: Overview of the proposed OSD framework. Given an input face image, a text-guided router selects the corresponding Demographic Semantic Expert (DSE) based on image–text similarity, while the Universal Forgery Expert (UFE) is always activated. Both experts operate as low-rank residual updates on top of a frozen principal subspace. Orthogonality constraints are enforced among all experts and the principal subspace to ensure decoupled representations. The final prediction combines outputs from the routed DSE and the UFE.

where $\bar{U}$ and $\bar{V}$ denote column-normalized bases, and M denotes the frozen principal subspace. We set $\mathcal{C}_{e_i} = \{M, e_1, \ldots, e_{i-1}\}$ for the i-th DSE and $\mathcal{C}_f = \{M\}$ for the UFE, and average the loss over all SVD-MoE projections.

3.4 Specialization to Balanced Integration

To effectively decouple demographic semantics from forgery artifacts and avoid expert entanglement under end-to-end joint optimization, OSD adopts a two-stage training strategy that progresses from specialization to integration. The core idea is to first strengthen the independent modeling capability of different types of experts, and then perform holistic collaboration under a fixed decoupled structure.

Expert Specialization. In the first stage, we train the Demographic Semantic Experts (DSEs) and the Universal Forgery Expert (UFE) separately to establish clear functional roles. For DSEs, we apply group-aware supervision with group sampling: for each sample, only the corresponding demographic expert is updated, allowing appearance and texture variations related to race and gender to be absorbed by expert specialization rather than interfering with forgery modeling. Additionally, we adopt class-balanced sampling to alleviate within-group real/fake imbalance during DSE training.

To encourage the UFE to learn transferable forgery cues rather than manipulation-specific artifacts, we train it on all real images in the training set, together with synthetic forged samples generated from these real faces via Self-Reconstruction-Blended Images (SRBI). Specifically, we extract facial boundaries and landmarks, obtain latent-space reconstructed frames using a pretrained autoencoder, and then randomly select the target from either the original frame or the reconstructed frame. Pixel-level blending with dynamic masks is applied to introduce more diverse statistical discontinuities and forgery patterns, thereby encouraging the UFE to learn more general artifact representations.

End-to-End Tuning. To promote collaborative inference between the DSEs and the UFE, OSD enters an integration stage to stabilize how expert outputs are combined for final decisions. Accordingly, we freeze the principal subspace as well as all experts' parameters, and fine-tune only the classification head with cross-entropy loss. To mitigate imbalance across demographic subgroups, we use demographic-weighted sampling when fine-tuning the head, where each training sample is weighted by the inverse of its subgroup frequency. Meanwhile, expert routing is handled by a training-free, text-guided router that leverages image–text similarity in the pretrained VLM to activate the appropriate DSE. Specifically, we predefine a set of demographic descriptions $\{\tau_e\}_{e\in\mathcal{E}}$ (e.g., race $\times$ gender) and compute their text embeddings $\{z_T^e\}_{e\in\mathcal{E}}$, where each z_T^e is obtained by encoding a set of prompt templates and averaging the normalized text features into a prototype. For an input image x, we extract its image

Algorithm 1 Two-stage training for OSD

Input: Training set $\mathcal{D}_{\text{train}} = \{(x, y, a)\}$
Params: Demographic experts $\{f_e\}_{e=1}^{E}$, forgery expert f, classifier head g
Hyper: λ_{orth} (orth. weight); T_{sem}, T_f, T_g (iters per stage)
Output: Trained $\{f_e\}_{e=1}^{E}, f, g$

Stage 1: Specialization
for $e = 1$ **to** E **do**
  for $t = 1$ **to** T_{sem} **do**
    $\mathcal{B} \sim \text{GroupSampling}(\mathcal{D}_{\text{train}} \mid a = e)$
    $\mathcal{L} \leftarrow L_{\text{det}}(\mathcal{B}; W_{\text{main}}, f_e, g) + \lambda_{\text{orth}} L_{\text{orth}}^{(e)}$
    $\text{Update}(\{f_e, g\},\ \mathcal{L})$
  end for
end for
for $t = 1$ **to** T_f **do**
  $\mathcal{B} \sim (\mathcal{D}_{\text{train}} \mid y = 0)$
  $\tilde{\mathcal{B}} = \text{SRBI}(\mathcal{B})$
  $\mathcal{L} \leftarrow L_{\text{det}}(\mathcal{B} \cup \tilde{\mathcal{B}}; W_{\text{main}}, f, g) + \lambda_{\text{orth}} L_{\text{orth}}^{(f)}$
  $\text{Update}(\{f, g\},\ \mathcal{L})$
end for
Stage 2: Integration
Freeze W_{main}, all f_e, and f, and fine-tune g.
for $t = 1$ **to** T_g **do**
  $\mathcal{B} \sim \mathcal{D}_{\text{train}}$
  for $(x, y, a) \in \mathcal{B}$ **do**
    $e^*(x) \leftarrow \text{Route}(x)$ using Eq. 5
  end for
  $\mathcal{L} \leftarrow L_{\text{det}}(\mathcal{B}; W_{\text{main}}, f_{e^*(\cdot)}, f, g)$
  $\text{Update}(g,\ \mathcal{L})$
end for

embedding z_I and select the expert via similarity matching:

$$e^*(x) = \arg\max_{e \in \mathcal{E}} \ \text{sim}(z_I, z_T^e), \tag{5}$$

where $\text{sim}(\cdot, \cdot)$ denotes cosine similarity between ℓ_2-normalized embeddings (optionally scaled by a temperature τ). The training procedure is summarized in Algorithm 1.

4 Experiments

4.1 Experimental Setup

Datasets. Following prior work [Yan *et al.*, 2023; Yan *et al.*, 2025], we use FaceForensics++ (FF++, c23) [Rössler *et al.*, 2019] as the training dataset, and evaluate our model on multiple public face deepfake detection benchmarks, including DFDC [Dolhansky *et al.*, 2020], DFD [Dufour *et al.*, 2019], and CDFv2 [Li *et al.*, 2020b], to assess cross-domain generalization performance. To further evaluate fairness across demographic subgroups, we follow the annotation and cross-group partitioning strategy adopted in recent fairness-oriented deepfake detection studies [Ju *et al.*, 2024; Lin *et al.*, 2024; Xu *et al.*, 2024]. Specifically, following prior fairness-oriented studies, we adopt their gender and race annotations and cross-group partitioning, forming six intersectional subgroups by combining these attributes: Asian–Male (A-M), Asian–Female (A-F), White–Male (W-M), White–Female (W-F), Black–Male (B-M), and Black–Female (B-F).

Evaluation Metrics. We use the frame-level Area Under the ROC Curve (AUC) as the primary metric to evaluate model generalization performance. For fairness, we report Skew [Geyik *et al.*, 2019; Wang and Deng, 2020; Cheng *et al.*, 2025a], which measures the maximum performance disparity across demographic subgroups; lower values indicate smaller disparities and better fairness.

Implementation Details. We adopt CLIP ViT-L/14 as the default vision–language model (VLM) backbone. For our method, all input frames are resized to 224×224. The expert fusion weights are set to $\lambda_s = 0.5$ and $\lambda_f = 0.5$, the expert rank to $r_{\text{exp}} = 8$, and the orthogonality loss weight to $\lambda_{\text{orth}} = 0.001$ by default. All experiments are conducted on a single NVIDIA H20 GPU for both training and evaluation.

4.2 Comparison Experiments

To evaluate cross-domain generalization and fairness, we compare OSD with representative deepfake detectors from two lines of research: (i) generalization-oriented methods (e.g., Xception [Rössler *et al.*, 2019], EfficientNet [Tan and Le, 2019], F^3-Net [Qian *et al.*, 2020], Face X-ray [Li *et al.*, 2020a], SBI [Shiohara and Yamasaki, 2022], RECCE [Cao *et al.*, 2022], GRU [Choi *et al.*, 2024], CADDM [Dong *et al.*, 2023], UCF [Yan *et al.*, 2023], ProDet [Cheng *et al.*, 2024], VLFFD [Sun *et al.*, 2025], and EFFORT [Yan *et al.*, 2025]) and (ii) fairness-oriented methods (e.g., DAW-FDD [Ju *et al.*, 2024], FG [Lin *et al.*, 2024], and DAID [Cheng *et al.*, 2025b]). All methods are trained on FF++ (c23) and evaluated on DFDC, DFD, and CDFv2. We report frame-level AUC for generalization and Skew across six intersectional demographic groups (race×gender) for fairness. The results are summarized in Table 1.

As shown in Table 1, OSD achieves state-of-the-art performance in both cross-domain generalization and demographic fairness under the same cross-dataset protocol. In terms of generalization, OSD attains the best Avg AUC (84.94%) across DFDC, DFD, and CDFv2, with EFFORT being the closest generalization-oriented competitor. In terms of fairness, OSD achieves the lowest Avg Skew (0.659). We also observe that Skew values on DFDC are consistently higher across methods, likely because DFDC provides the most complete and relatively balanced intersectional demographic coverage; therefore, we select DFDC for the visualization analyses in Sec. 4.4.

4.3 Ablation Experiments

We conduct ablation studies to quantify the contribution of each component in OSD. Unless otherwise specified, all ablations follow the same training protocol as in Sec. 4.1 and are evaluated under the same cross-domain setting as in Sec. 4.2.

Comparison on Modules. We ablate the three core components of OSD: Demographic Semantic Experts (DSEs), Universal Forgery Expert (UFE), and text-guided routing (R). As shown in Table 2, enabling UFE alone improves AUC (59.1%→65.0%) but increases Skew (1.50→1.61), indicating that forgery-only modeling without demographic decoupling leads to disparate generalization across subgroups. Enabling DSEs alone reduces Skew (1.50→1.17) with moderate

Method	Venue	Generalization (AUC(%)↑)				Fairness (Skew↓)			
		DFDC	DFD	CDFv2	Avg	DFDC	DFD	CDFv2	Avg
Xception [Rössler *et al.*, 2019]	ICCV'19	60.63	80.69	70.91	70.74	2.221	0.564	0.597	1.127
EfficientNet [Tan and Le, 2019]	ICML'19	60.49	83.12	75.36	72.99	2.011	0.351	0.437	0.933
F^3-Net [Qian *et al.*, 2020]	ECCV'20	60.17	77.68	74.36	70.74	2.143	0.589	0.556	1.096
Face X-ray [Li *et al.*, 2020a]	CVPR'20	62.00	80.46	74.20	72.22	1.982	0.821	0.491	1.098
SBI [Shiohara and Yamasaki, 2022]	CVPR'22	63.39	86.43	79.76	76.53	2.385	0.757	0.715	1.286
RECCE [Cao *et al.*, 2022]	CVPR'22	61.63	80.13	70.55	70.77	2.622	0.738	0.644	1.335
GRU [Choi *et al.*, 2024]	CVPR'24	62.63	86.48	76.00	75.04	2.432	0.551	0.405	1.129
CADDM [Dong *et al.*, 2023]	CVPR'23	63.77	88.59	81.75	78.04	2.183	0.547	0.391	1.040
UCF [Yan *et al.*, 2023]	ICCV'23	60.03	81.01	71.73	70.92	2.272	0.510	0.619	1.134
ProDet [Cheng *et al.*, 2024]	NeurIPS'24	65.89	89.18	82.71	79.26	2.306	0.432	0.569	1.102
VLFFD [Sun *et al.*, 2025]	CVPR'25	65.21	90.08	81.17	78.82	2.411	0.669	0.526	1.202
EFFORT [Yan *et al.*, 2025]	ICML'25	**76.04**	90.48	86.24	84.25	1.474	0.521	0.889	0.961
DAW-FDD [Ju *et al.*, 2024]	WACV'24	59.96	71.40	69.55	66.97	2.127	0.528	0.509	1.055
FG [Lin *et al.*, 2024]	CVPR'24	60.11	80.42	68.30	69.61	1.932	0.447	0.498	0.959
DAID [Cheng *et al.*, 2025b]	NeurIPS'25	66.85	91.15	84.39	80.80	1.460	**0.263**	0.289	0.671
OSD (ours)	–	73.55	**91.71**	**89.55**	**84.94**	**1.409**	0.385	**0.182**	**0.659**

Table 1: Cross-dataset evaluation (training on FF++ (c23); testing on DFDC, DFD, and CDFv2). Generalization is measured by frame-level AUC. Fairness is measured by Skew. Results of some methods are from [Cheng *et al.*, 2025b]. Best in bold, second-best underlined.

Module			DFDC		DFD		CDFv2	
DSEs	UFE	R	Skew	AUC	Skew	AUC	Skew	AUC
–	–	–	1.50	59.1	0.69	69.5	0.78	60.2
✓			1.17	64.2	0.61	81.4	0.65	67.9
	✓		1.61	65.0	0.72	83.1	1.14	71.0
✓	✓		1.46	72.6	0.40	90.6	0.18	86.8
✓	✓	✓	**1.40**	**73.5**	**0.38**	**91.7**	**0.18**	**89.5**

Table 2: Ablation study on each module of OSD. DSEs: Demographic Semantic Experts. UFE: Universal Forgery Expert. R: Text-guided Routing.

Hyperparameter		AUC↑	Skew↓
λ_{orth}	0	85.2	0.58
	0.001	**89.5**	**0.18**
	0.01	87.6	0.32
	0.1	86.1	0.21
r_{main}	$1024-16$	84.3	0.35
	$1024-12$	87.2	0.24
	$1024-8$	**89.5**	**0.18**
	$1024-4$	88.9	0.20

Table 3: Ablation on orthogonality weight λ_{orth} and principal rank r_{main} on CDFv2.

AUC gains, suggesting that subgroup-specific adaptation effectively balances cross-group performance. Combining both yields substantial improvements in both metrics, and adding text-guided routing further boosts performance, demonstrating that proper expert selection enhances the synergy between the two branches. To verify that the gains are not simply from SRBI, we adapt CLIP with LoRA using the same SRBI. Its 74.5% Avg AUC remains much lower than OSD's 84.9%.

Comparison on Hyperparameters. We examine the sensitivity to the orthogonality loss weight λ_{orth} and the principal rank r_{main} in Table 3. Removing the orthogonality constraint ($\lambda_{\text{orth}} = 0$) leads to higher Skew, indicating that expert representations become entangled without explicit decoupling. Overly aggressive constraints degrade AUC due to limited expressiveness. For the principal rank, r_{main} controls how much pretrained semantics are kept in the frozen principal subspace versus delegated to the residual experts. Too small principal ranks (e.g., $1024-16$) under-preserve pretrained semantics, while too large principal ranks (e.g., $1024-4$) leave limited residual capacity for adaptation; a moderate rank ($1024-8$) achieves the best balance.

We further analyze the effect of expert fusion weights λ_s and λ_f in Figure 4. To isolate the influence of each weight, we fix one at 0.5 while varying the other. Since the principal subspace component $W_{\text{main}}h$ is frozen and always included, overly small weights (especially when both λ_s and λ_f are small) make the prediction dominated by $W_{\text{main}}h$, resulting in limited task-specific adaptation and poor AUC. Increasing λ_s or λ_f allows the routed DSE or the UFE to contribute effective residual updates, improving performance. However, overly large weights can make the residual adaptation dominated by a single branch, weakening the complementary benefits of jointly leveraging DSE and UFE and thus hurting generalization on CDFv2.

4.4 Visualization

Feature-Space Decoupling. To validate the effect of orthogonal semantic decoupling, we visualize the learned feature representations via t-SNE. We select DFDC for visualization as it is the only test set that covers all six intersectional demographic subgroups with relatively balanced sam-

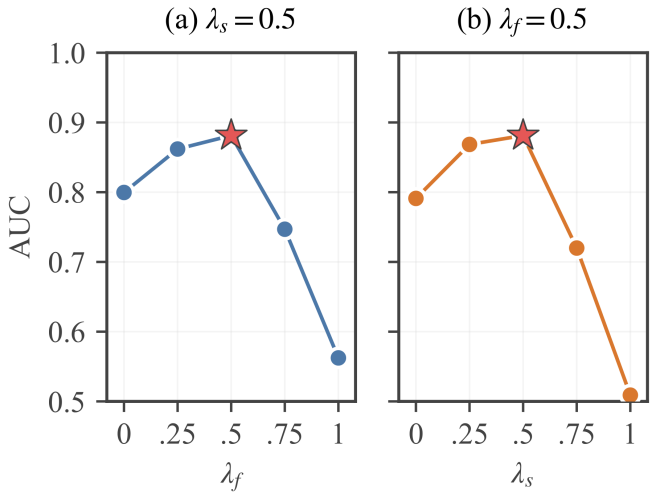

Figure 4: Effect of expert fusion weights on CDFv2. We vary one fusion weight while fixing the other at 0.5. (a) Varying λ_f with $\lambda_s = 0.5$; (b) Varying λ_s with $\lambda_f = 0.5$. The principal subspace component is always included (weight 1.0); λ_s and λ_f scale the residual updates from the routed DSE and the UFE, respectively.

ple counts and both real and fake samples in each group, enabling a complete and fair view of demographic decoupling. We compare OSD with EFFORT [Yan *et al.*, 2025], the most related baseline that also employs SVD-based orthogonal subspace adaptation on vision–language models. While EFFORT focuses solely on learning forgery patterns in the residual subspace, it does not explicitly decouple demographic semantics from forgery cues. This comparison isolates the contribution of our demographic-aware expert design.

As shown in Figure 5, each point represents a test sample; colors indicate six intersectional demographic groups: Asian–Male (A-M), Asian–Female (A-F), White–Male (W-M), White–Female (W-F), Black–Male (B-M), and Black–Female (B-F); lighter points denote fake samples and darker points denote real samples. In EFFORT (left), samples from different groups are substantially entangled, and the real/fake separation varies across groups, suggesting reliance on demographic shortcuts and cross-subgroup interference. In contrast, OSD (right) yields clearer group-wise clusters with more consistent real/fake boundaries within each subgroup, indicating that orthogonal decoupling separates demographic semantics from forgery cues and stabilizes decision boundaries across groups, improving fairness and robustness.

Attention Heatmaps. To verify that OSD focuses on forgery artifacts rather than demographic appearance, we visualize attention heatmaps on forged samples from the six demographic subgroups in DFDC. As shown in Figure 6, EFFORT (top row) tends to attend to facial morphology and facial-feature structures that correlate with demographic attributes, and its attended regions vary noticeably across subgroups. In contrast, OSD (bottom row) consistently highlights forgery-related regions such as blending boundaries and texture inconsistencies, especially in the upper face, across subgroups. This cross-group consistency indicates that orthogonal decoupling effectively suppresses demographic

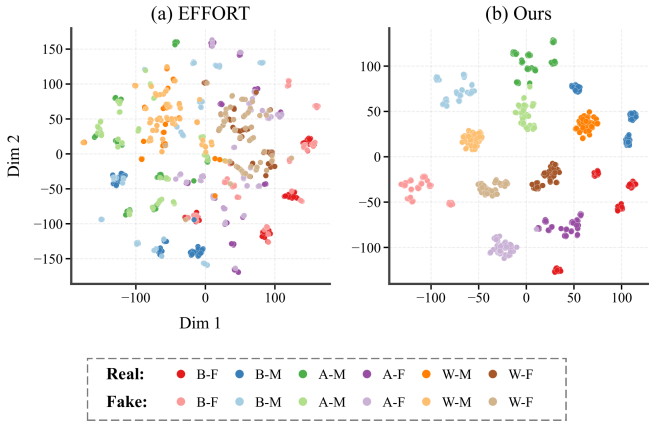

Figure 5: t-SNE visualization of feature embeddings on DFDC. Colors indicate six intersectional demographic groups; lighter points denote fake samples and darker points denote real samples. Left: EFFORT entangles demographic groups. Right: OSD produces distinct, well-separated clusters with consistent real/fake boundaries within each group.

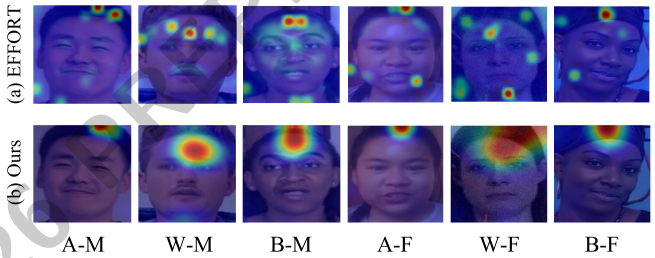

Figure 6: Attention heatmaps on samples from six demographic subgroups in DFDC. (a) EFFORT attends to demographic-correlated regions. (b) OSD highlights forgery-related regions (e.g., blending boundaries) rather than demographic-specific facial features.

shortcuts, enabling demographic-agnostic forgery detection.

5 Conclusion

This paper analyzes the practical trade-off between cross-domain generalization and demographic fairness in face deepfake detection. We propose Orthogonal Semantic Decoupling (OSD), which applies SVD to a pretrained vision–language model (VLM) to separate a principal semantic subspace and a residual subspace, and learns an orthogonal mixture of experts in the residual space to suppress demographic shortcuts and cross-subgroup interference. Extensive experiments on DFDC, DFD, and CDFv2 show that OSD achieves the best Avg AUC (84.94%) and the lowest Avg Skew (0.659), achieving state-of-the-art performance in both generalization and fairness. In real-world deployment, the text-guided, training-free router may be less reliable when demographic attributes are ambiguous or more fine-grained, and may be slightly affected by inherent biases in the VLM. Future work can explore continual residual-expert adaptation to new domains or attributes and extend the router to fine-grained demographic settings by using a larger set of tunable demographic prompts to improve image-text alignment.

Contribution Statement

Zhongyu Shi and Siyu Peng contributed equally to this paper.

References

[Cao *et al.*, 2022] Junyi Cao, Chao Ma, Taiping Yao, Shen Chen, Shouhong Ding, and Xiaokang Yang. End-to-End Reconstruction-Classification learning for face forgery detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4103–4112. IEEE, 2022.

[Chen *et al.*, 2022] Liang Chen, Yong Zhang, Yibing Song, Lingqiao Liu, and Jue Wang. Self-supervised learning of adversarial example: Towards good generalizations for Deepfake detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 18689–18698. IEEE, 2022.

[Cheng *et al.*, 2024] Jikang Cheng, Zhiyuan Yan, Ying Zhang, Yuhao Luo, Zhongyuan Wang, and Chen Li. Can we leave Deepfake data behind in training Deepfake detector? In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 21979–21998, 2024.

[Cheng *et al.*, 2025a] Harry Cheng, Yangyang Guo, Qingpei Guo, Ming Yang, Tian Gan, Weili Guan, and Liqiang Nie. Social debiasing for fair multi-modal LLMs. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 1740–1750. IEEE, 2025.

[Cheng *et al.*, 2025b] Harry Cheng, Ming-Hui Liu, Yangyang Guo, Tianyi Wang, Liqiang Nie, and Mohan Kankanhalli. Fair Deepfake detectors can generalize. In *Proceedings of the Conference on Neural Information Processing Systems (NeurIPS)*, 2025.

[Choi *et al.*, 2024] Jongwook Choi, Taehoon Kim, Yonghyun Jeong, Seungryul Baek, and Jongwon Choi. Exploiting style latent flows for generalizing Deepfake video detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1133–1143. IEEE, 2024.

[Chollet, 2017] François Chollet. Xception: Deep learning with depthwise separable convolutions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1800–1807. IEEE, 2017.

[Dolhansky *et al.*, 2020] Brian Dolhansky, Joanna Bitton, Ben Pflaum, Jikuo Lu, Russ Howes, Menglin Wang, and Cristian Canton-Ferrer. The deepfake detection challenge dataset. *CoRR*, abs/2006.07397, 2020.

[Dong *et al.*, 2023] Shichao Dong, Jin Wang, Renhe Ji, Jiajun Liang, Haoqiang Fan, and Zheng Ge. Implicit identity leakage: The stumbling block to improving Deepfake detection generalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3994–4004. IEEE, 2023.

[Dufour *et al.*, 2019] Nick Dufour, Andrew Gully, Daisy Stanton, et al. Deep fake detection dataset by google and jigsaw. https://research.google/blog/contributing-data-to-deepfake-detection-research/, September 2019.

[Geyik *et al.*, 2019] Sahin Cem Geyik, Stuart Ambler, and Krishnaram Kenthapadi. Fairness-aware ranking in search & recommendation systems with application to LinkedIn talent search. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (KDD)*, pages 2221–2231. ACM, 2019.

[Guo *et al.*, 2025] Yuncheng Guo, Junyan Ye, Chenjue Zhang, Hengrui Kang, Haohuan Fu, Conghui He, and Weijia Li. OmniAID: Decoupling semantics and artifacts for universal ai-generated image detection in the wild. *arXiv preprint arXiv:2511.08423*, 2025.

[Huang *et al.*, 2023] Baojin Huang, Zhongyuan Wang, Jifan Yang, Jiaxin Ai, Qin Zou, Qian Wang, and Dengpan Ye. Implicit identity driven Deepfake face swapping detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4490–4499. IEEE, 2023.

[Ju *et al.*, 2024] Yan Ju, Shu Hu, Shan Jia, George H. Chen, and Siwei Lyu. Improving fairness in Deepfake detection. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 4655–4665. IEEE, 2024.

[Kim *et al.*, 2025] Kihong Kim, Yunho Kim, Seokju Cho, Junyoung Seo, Jisu Nam, Kychul Lee, Seungryong Kim, and KwangHee Lee. DiffFace: Diffusion-based face swapping with facial guidance. *Pattern Recogn.*, 163, 2025.

[Li and Lyu, 2019] Yuezun Li and Siwei Lyu. Exposing DeepFake videos by detecting face warping artifacts. In *IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2019.

[Li *et al.*, 2020a] Lingzhi Li, Jianmin Bao, Ting Zhang, Hao Yang, Dong Chen, Fang Wen, and Baining Guo. Face X-Ray for more general face forgery detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5001–5010. IEEE, 2020.

[Li *et al.*, 2020b] Yuezun Li, Xin Yang, Pu Sun, Honggang Qi, and Siwei Lyu. Celeb-DF: A large-scale challenging dataset for DeepFake forensics. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3204–3213. IEEE, 2020.

[Lin *et al.*, 2024] Li Lin, Xinan He, Yan Ju, Xin Wang, Feng Ding, and Shu Hu. Preserving fairness generalization in Deepfake detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16815–16825. IEEE, 2024.

[Lin *et al.*, 2025] Li Lin, Santosh, Mingyang Wu, Xin Wang, and Shu Hu. AI-Face: A million-scale demographically annotated AI-generated face dataset and fairness benchmark. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3503–3515. IEEE, 2025.

[Liu *et al.*, 2021] Honggu Liu, Xiaodan Li, Wenbo Zhou, Yuefeng Chen, Yuan He, Hui Xue, Weiming Zhang, and Nenghai Yu. Spatial-Phase shallow learning: Rethinking

face forgery detection in frequency domain. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 772–781. IEEE, 2021.

[Luo *et al.*, 2021] Yuchen Luo, Yong Zhang, Junchi Yan, and Wei Liu. Generalizing face forgery detection with High-Frequency features. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16317–16326. IEEE, 2021.

[Ojha *et al.*, 2023] Utkarsh Ojha, Yuheng Li, and Yong Jae Lee. Towards universal fake image detectors that generalize across generative models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 24480–24489. IEEE, 2023.

[Qian *et al.*, 2020] Yuyang Qian, Guojun Yin, Lu Sheng, Zixuan Chen, and Jing Shao. Thinking in frequency: Face forgery detection by mining Frequency-Aware clues. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 86–103. Springer, 2020.

[Radford *et al.*, 2021] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the International Conference on Machine Learning (ICML)*, pages 8748–8763. PMLR, 2021.

[Rosberg *et al.*, 2023] Felix Rosberg, Eren Erdal Aksoy, Fernando Alonso-Fernandez, and Cristofer Englund. FaceDancer: Pose- and occlusion-aware high fidelity face swapping. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 3443–3452. IEEE Computer Society, 2023.

[Rössler *et al.*, 2019] Andreas Rössler, Davide Cozzolino, Luisa Verdoliva, Christian Riess, Justus Thies, and Matthias Nießner. FaceForensics++: Learning to detect manipulated facial images. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. IEEE, 2019.

[Shiohara and Yamasaki, 2022] Kaede Shiohara and Toshihiko Yamasaki. Detecting Deepfakes with Self-Blended images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 18699–18708. IEEE, 2022.

[Sun *et al.*, 2025] Ke Sun, Shen Chen, Taiping Yao, Ziyin Zhou, Jiayi Ji, Xiaoshuai Sun, Chia-Wen Lin, and Rongrong Ji. Towards general visual-linguistic face forgery detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 19576–19586. IEEE, 2025.

[Tan and Le, 2019] Mingxing Tan and Quoc V. Le. EfficientNet: Rethinking model scaling for convolutional neural networks. In *Proceedings of the International Conference on Machine Learning (ICML)*, pages 6105–6114. PMLR, 2019.

[Tan *et al.*, 2023] Chuangchuang Tan, Yao Zhao, Shikui Wei, Guanghua Gu, and Yunchao Wei. Learning on gradients: Generalized artifacts representation for GAN-generated images detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12105–12114. IEEE, 2023.

[Trinh and Liu, 2021] Loc Trinh and Yan Liu. An examination of fairness of AI models for Deepfake detection. In *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI)*. International Joint Conferences on Artificial Intelligence Organization, 2021.

[Wang and Deng, 2020] Mei Wang and Weihong Deng. Mitigating bias in face recognition using skewness-aware reinforcement learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9319–9328. IEEE, 2020.

[Wang *et al.*, 2023] Zhendong Wang, Jianmin Bao, Wengang Zhou, Weilun Wang, Hezhen Hu, Hong Chen, and Houqiang Li. DIRE for Diffusion-generated image detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 22388–22398. IEEE, 2023.

[Wang *et al.*, 2024] Tianyi Wang, Xin Liao, Kam Pui Chow, Xiaodong Lin, and Yinglong Wang. Deepfake detection: A comprehensive survey from the reliability perspective. *ACM Comput. Surv.*, 57(3), 2024.

[Wu and De La Torre, 2023] Chen Henry Wu and Fernando De La Torre. A latent space of stochastic Diffusion models for Zero-Shot image editing and guidance. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 7344–7353. IEEE, 2023.

[Wu *et al.*, 2025] Haiwei Wu, Jiantao Zhou, and Shile Zhang. Generalizable synthetic image detection via language-guided contrastive learning. *IEEE Transactions on Artificial Intelligence*, pages 1–11, 2025.

[Xia *et al.*, 2026] Ruiyang Xia, Dawei Zhou, Lin Yuan, Jie Li, Nannan Wang, and Xinbo Gao. Ssd: Making face forgery clues evident again with self-steganographic detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2026.

[Xu *et al.*, 2024] Ying Xu, Philipp Terhörst, Marius Pedersen, and Kiran Raja. Analyzing fairness in Deepfake detection with massively annotated databases. *IEEE Transactions on Technology and Society*, 5(1):93–106, 2024.

[Yan *et al.*, 2023] Zhiyuan Yan, Yong Zhang, Yanbo Fan, and Baoyuan Wu. UCF: Uncovering common features for generalizable deepfake detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 22355–22366. IEEE, 2023.

[Yan *et al.*, 2025] Zhiyuan Yan, Jiangming Wang, Peng Jin, Ke-Yue Zhang, Chengchun Liu, Shen Chen, Taiping Yao, Shouhong Ding, Baoyuan Wu, and Li Yuan. Orthogonal subspace decomposition for generalizable AI-generated image detection. In *Proceedings of the International Conference on Machine Learning (ICML)*. JMLR.org, 2025.