PURE: Purging Unrelated Representations for Content-Agnostic Forgery Detection

Xinyu Wu[1] , **Dong Li**[1] , **Minglai Shao**[2] , **Xintao Wu**[3] , **Zhong Chen**[4] , **Chen Zhao**[1]
[1]Department of Computer Science, Baylor University
[2]School of Computer Science and Technology, Tianjin University
[3]Department of Electrical Engineering and Computer Science, University of Arkansas
[4]School of Computing, Southern Illinois University
{xinyu_wu1, dong_li1, chen_zhao}@baylor.edu, shaoml@tju.edu.cn, xintaowu@uark.edu, zhong.chen@cs.siu.edu

Abstract

Existing AI-generated image (AIGI) detectors perform well in-domain but degrade severely under distribution shift. We observe that this failure is mainly caused by *content shortcuts*, where detectors spuriously couple forgery artifacts with semantic content, such as object categories or demographic attributes, learning content–label correlations instead of generalizable forgery patterns. To address this issue, we propose **PURE** (**P**urging **U**nrelated **Re**presentations for Content-Agnostic Forgery Detection), which achieves content-agnostic detection through two complementary components: a *Causal Semantic Generative* (CSG) mechanism that disentangles semantic representations from forgery-irrelevant nuisance factors, and a Gaussian Mixture Model (GMM)-based prototype alignment module that suppresses category-specific content bias. Extensive experiments on CIFAKE, GenImage, and AlFace show that PURE achieves superior generalization under spurious correlation reversal. The code is available at https://github.com/wuxinyu519/PURE.

1 Introduction

The democratization of generative models, spanning from GANs to diffusion models [Karras *et al.*, 2019; Ho *et al.*, 2020; Rombach *et al.*, 2022; Nichol *et al.*, 2021], enables unprecedented high-quality image synthesis. This newfound freedom in image generation, however, raises urgent concerns about misinformation and trust in visual media. While AI-generated image (AIGI) detectors achieve impressive performance on in-domain benchmarks, real-world deployments reveal a critical vulnerability: spurious correlations in training data systematically undermine generalization under distribution shift.

Recent efforts to improve AIGI detector generalization can be broadly organized by the types of features they exploit [Wu *et al.*, 2025a], ranging from RGB and frequency cues to noise residuals and high-level semantic representations. *RGB-based methods* [Yan *et al.*, 2025; Liu *et al.*, 2024b; Miao *et al.*, 2025] directly leverage spatial and color information from raw pixel channels, excelling at detecting boundary inconsistencies and occlusion mismatches through shallow convolutional networks. However, they struggle to capture subtle manipulation signals when forgery artifacts are imperceptible in the spatial domain. *Frequency-based approaches* [Guo *et al.*, 2023; Liu *et al.*, 2024c; Wu *et al.*, 2025b] transform images into spectral representations via DCT or Fourier transforms to identify periodic structures and compression artifacts. While effective at detecting high-frequency anomalies invisible to RGB analysis, these methods remain sensitive to preprocessing operations and may not generalize when frequency signatures vary across generators. *Noiseprint-driven methods* [Guillaro *et al.*, 2023; Guo *et al.*, 2024a] extract residual patterns through high-pass filtering to reveal device-specific traces and manipulation-related statistical inconsistencies. Despite their sensitivity to subtle perturbations, noiseprint features can be disrupted by post-processing operations that alter the underlying noise characteristics. *Representation learning methods* [Ojha *et al.*, 2023; Liu *et al.*, 2024a; Wu *et al.*, 2025b] employ vision-language models or deep networks to learn semantic abstractions, particularly valuable for cross-domain robustness. However, when semantic content correlates spuriously with authenticity labels during training, such approaches may inadvertently amplify content bias rather than mitigate it.

Despite progress across these feature paradigms, existing detectors exhibit substantial accuracy drops when distributions shift between training and testing [Lin *et al.*, 2024; Kashiani *et al.*, 2025]. This suggests a fundamental challenge remains unresolved: spurious correlations between semantic content and authenticity labels [Trinh and Liu, 2021; Fu *et al.*, 2025; Bai *et al.*, 2025]. We observe that this generalization failure stems from content bias. Detectors erroneously couple forgery artifacts with semantic attributes (*e.g.,* object categories, demographic features) during training, learning spurious content-label associations instead of generalizable forgery patterns. Consider a scenario where 80% of dogs are AI-generated while 80% of cats are real. The detector learns to associate "dog" with "fake" rather than genuine artifacts. When the spurious content-label association learned during training is reversed in the test distribution, performance col-

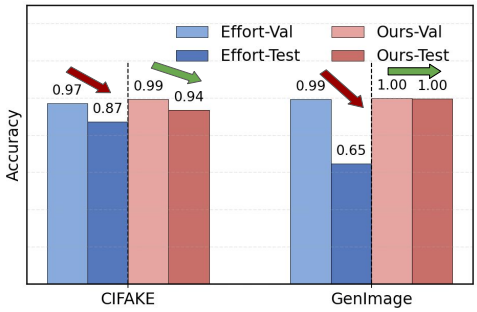

Figure 1: Accuracy under domain shift caused by changes in spurious features. Models trained on CIFAKE [Bird and Lotfi, 2024] or GenImage [Zhu *et al.*, 2023] rely on dataset-specific spurious cues, leading the baseline (Effort [Yan *et al.*, 2025]) to experience significant degradation when the spurious correlations shift at test time. Our method remains more stable and preserves higher accuracy across domains with shifted spurious features.

lapses dramatically.

Figure 1 provides empirical evidence. We train state-of-the-art detectors on CIFAKE and GenImage with deliberately constructed spurious correlations. Although both baseline (Effort [Yan *et al.*, 2025]) and our method achieve high validation accuracy (97 to 99%), the baseline's test accuracy falls to 87% on CIFAKE and 65% on GenImage when spurious correlations reverse, resulting in a 12 to 34% drop. This validation-test gap reveals that models memorize content-label associations rather than learning forgery patterns. Our method maintains stable performance, demonstrating genuine content-agnostic detection.

To address this issue, we propose **PURE** (**P**urging **U**nrelated **Re**presentations for Content-Agnostic Forgery Detection), which achieves content-agnostic detection through two synergistic components: a Causal Semantic Generative (CSG) mechanism that disentangles semantic representations from forgery-irrelevant nuisance factors, and a Gaussian Mixture Model (GMM)-based prototype alignment module that normalizes category-specific feature distributions to suppress content priors. PURE explicitly models causal relationships between content and forgery cues, introducing fairness-aware prototype balancing and gradient bridging to enable end-to-end joint optimization.

Specifically, CSG disentangles CLIP features into semantic factors (encoding content and forgery cues) and nuisance factors (capturing forgery-irrelevant style variations) via variational inference. GMM alignment then projects semantic factors onto content-normalized representations by modeling each category with Gaussian prototypes, forcing classifiers to focus on prototype-level forgery artifacts rather than content shortcuts.

Our contributions are summarized as follows:

- We propose PURE, a novel content-agnostic AIGI detection framework designed to acquire forgery-invariant knowledge under distribution shift, where target content distributions are inaccessible during training. It explicitly eliminates content bias through causal disentanglement and prototype-based normalization.
- The framework comprises two synergistic components: a CSG mechanism that isolates semantic from nuisance factors via variational inference, and a GMM alignment module with fairness-aware prototype balancing.
- Empirically, we conduct rigorous spurious correlation reversal experiments on CIFAKE, GenImage, and AI-Face to assess generalization under deliberately constructed distribution shifts. PURE consistently outperforms state-of-the-art baseline approaches.

2 Related Work

Forgery Detection. Existing image forgery detection methods can be primarily categorized based on feature-driven strategies [Wu *et al.*, 2025a], including RGB-based methods [Miao *et al.*, 2025; Liu *et al.*, 2024b], frequency-based methods [Zhu *et al.*, 2025; Guo *et al.*, 2023], noise-texture methods [Guillaro *et al.*, 2023; Guo *et al.*, 2024b], and representation learning methods [Yu *et al.*, 2024; He *et al.*, 2026]. These methods leverage various feature extraction strategies ranging from low-level pixel statistics to high-level semantic abstractions to capture diverse manipulation cues. While these methods achieve strong performance in in-domain or same-distribution evaluation settings [Lin *et al.*, 2024; Yan *et al.*, 2025], they still suffer from significant performance degradation in cross-domain scenarios involving different manipulation types, generative models, or data distributions.

Generalization in Forgery Detection. Studies [Lin *et al.*, 2024] indicate that the above generalization degradation largely stems from spurious correlations, which refer to statistical biases between authenticity labels and semantic content (*e.g.,* demographic attributes such as race and gender) in the training distribution, causing detectors to rely on content priors rather than genuine forgery artifacts at test time. To address this generalization challenge, D3 [Yang *et al.*, 2025] uncovers universal forgery artifacts through multi-generator joint training and discrepancy signal learning. Furthermore, Effort [Yan *et al.*, 2025] employs SVD to orthogonally disentangle semantic and forgery features, while UCF [Yan *et al.*, 2023] proposes a three-level disentanglement framework to extract universal forgery features. Meanwhile, vision foundation model approaches attempt to improve generalization by leveraging large-scale pre-trained knowledge, such as UniFD's [Ojha *et al.*, 2023] zero-shot detection strategy and FatFormer's [Liu *et al.*, 2024a] adaptive feature fusion mechanism. However, most of these methods fail to explicitly model and eliminate the spurious correlations introduced by content bias, leading to significant performance drops when content distributions shift at test time. Systematically eliminating content bias and achieving content-agnostic robust detection is the central focus of this paper.

3 Preliminary

Notations. Let $\mathcal{X} \subseteq \mathbb{R}^d$ denote a feature space, $\mathcal{Y} = \{0, 1\}$ the binary label space, and $\mathcal{C} = \{1, \ldots, C\}$ the content attribute space. We use random variables X, Y, and C to denote features, labels, and content attributes taking values in $\mathcal{X}$, $\mathcal{Y}$, and $\mathcal{C}$, respectively, and use x, y, and c to denote their realizations. Given a dataset $\mathcal{D} = \{(x_i, y_i, c_i)\}_{i=1}^N$ sampled i.i.d. from a joint distribution $P(X, Y, C)$, the goal is to learn

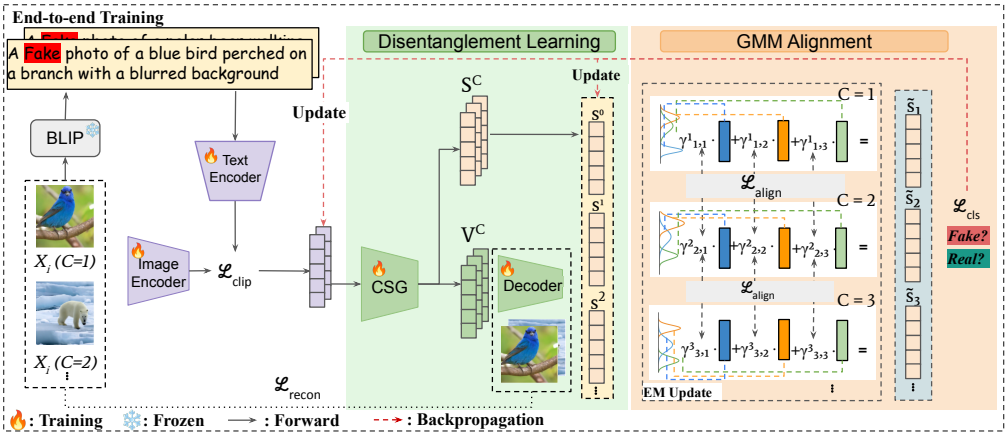

Figure 2: Overview of PURE. CLIP extracts features $\mathbf{x}$, which are disentangled by CSG into semantic $\mathbf{s}$ and nuisance $\mathbf{v}$ representations. GMM prototype alignment projects $\mathbf{s}$ to debiased $\tilde{\mathbf{s}}$ for classification. Gradient bridging enables end-to-end training. Red dashed arrows indicate backpropagation, black solid arrows indicate forward propagation.

a classifier $f_\theta : \mathcal{X} \to \mathcal{Y}$ parameterized by θ that accurately predicts authenticity labels.

Given training data $\mathcal{D}_{\text{tr}} = \{(x_i, y_i, c_i)\}_{i=1}^{N_{\text{tr}}}$ sampled from distribution $P_{\text{tr}}(X, Y, C)$, standard empirical risk minimization (ERM) solves

$$\min_{\theta} \frac{1}{N_{\text{tr}}} \sum_{i=1}^{N_{\text{tr}}} \mathcal{L}(f_\theta(x_i), y_i), \tag{1}$$

where $\mathcal{L} : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}_+$ denotes a classification loss. Under the i.i.d. assumption $P_{\text{tr}} = P_{\text{te}}$, the learned classifier generalizes to test data.

Spurious Correlation Shift. We consider a challenging scenario where training and test distributions exhibit different correlations between content and labels:

$$\begin{aligned} P_{\text{tr}}(Y, C) &\neq P_{\text{te}}(Y, C), \\ P_{\text{tr}}(X \mid Y, C) &= P_{\text{te}}(X \mid Y, C). \end{aligned} \tag{2}$$

This correlation shift preserves the underlying data generation mechanism but alters the statistical association between content attributes and labels. Classifiers trained via Eq. (1) exploit spurious shortcuts $X \to C \to Y$, learning to associate content with labels rather than genuine discriminative patterns, leading to severe degradation under distribution shift.

4 Method

To address the challenge of spurious correlations, we propose **PURE** (**P**urging **U**nrelated **Re**presentations), which achieves content-agnostic detection through three complementary components: (1) CLIP-based feature extraction that encodes both semantic content and forgery patterns, (2) Causal Semantic Generative (CSG) disentanglement that separates semantic factors from nuisance variations, and (3) GMM-based prototype alignment that normalizes category-specific distributions to suppress content priors.

4.1 CLIP Feature Extraction

We employ a pre-trained CLIP as the feature extraction backbone, leveraging its rich semantic priors from large-scale vision-language pretraining. Given an image X, CLIP produces a 768-dimensional feature embedding $\mathbf{x}$ that encodes both content semantics and visual patterns potentially associated with generative artifacts. To adapt CLIP features for forgery detection, we introduce contrastive fine-tuning with authenticity-aware descriptions. Specifically, we use BLIP [Li *et al.*, 2022] to generate a content description for each image, then construct prompts in the form "a $\{\texttt{real/fake}\}$ photo of $\{\texttt{caption}\}$" (*e.g.,* "A fake photo of a blue bird perched on a branch with a blurred background"). The CLIP text encoder maps each prompt to a text embedding $\mathbf{t}_i$. We then optimize a symmetric bidirectional contrastive loss:

$$\mathcal{L}_{\text{CLIP}} = \frac{1}{2}[\mathcal{L}_{i2t} + \mathcal{L}_{t2i}], \tag{3}$$

where $\mathcal{L}_{i2t}$ and $\mathcal{L}_{t2i}$ denote InfoNCE losses [Radford *et al.*, 2021] for image-to-text and text-to-image alignment, respectively. This mechanism encourages CLIP to not only preserve its semantic understanding but also to learn to associate visual patterns with authenticity labels, providing a foundation for subsequent disentanglement, where forgery-related patterns in $\mathbf{x}$ will be isolated from content biases.

4.2 Disentanglement Learning

Causal Forgery Generation Model. Following causal reasoning principles [Liu *et al.*, 2021], we posit that CLIP features $\mathbf{x}$ arise from two independent generative factors. Let $\mathbf{s} \in \mathbb{R}^{d_s}$ denote the *semantic factor* encoding both content semantics and forgery artifacts, and $\mathbf{v} \in \mathbb{R}^{d_v}$ denote the *nuisance factor* capturing forgery-irrelevant style variations (*e.g.,* lighting, background, artistic style). The generative process factorizes as:

$$P(x, y, s, v) = P(s, v)\, P(x \mid s, v)\, P(y \mid s). \tag{4}$$

This causal structure enforces that: (1) authenticity y is determined solely by the semantic factor $\mathbf{s}$, implying $\mathbf{v} \perp y$; and (2) nuisance variations $\mathbf{v}$ influence the observed representation only through $p(\mathbf{x} \mid \mathbf{s}, \mathbf{v})$, without directly affecting authenticity judgments. Crucially, while CSG assumes that $\mathbf{s}$ directly causes y, in our setting $\mathbf{s}$ may still entangle content semantics with forgery cues. The subsequent prototype

alignment module is designed to resolve this residual entanglement.

Variational Inference. To learn this disentangled representation, we employ variational inference with two encoders E_s and E_v, which predict the parameters of Gaussian distributions for the variational posteriors $q(\mathbf{s} \mid \mathbf{x})$ and $q(\mathbf{v} \mid \mathbf{x})$, respectively. The training objective maximizes the Evidence Lower Bound (ELBO):

$$\begin{aligned}\mathcal{L}_{\text{dis}} = \mathbb{E}_{q(\mathbf{s},\mathbf{v}|\mathbf{x})}\big[\log p(\mathbf{x} \mid \mathbf{s}, \mathbf{v})\big] \\ - \beta\, \mathrm{KL}(q(\mathbf{s}, \mathbf{v} \mid \mathbf{x}) \,\|\, \mathcal{N}(\mathbf{0}, \mathbf{I}))\,.\end{aligned} \tag{5}$$

where the first term enforces reconstruction fidelity through a decoder D, yielding $\hat{\mathbf{x}} = D(\mathbf{s}, \mathbf{v})$, and the KL terms regularize the variational posteriors toward standard Gaussian priors, with $\beta = 0.1$ controlling the strength of regularization.

4.3 GMM-based Prototype Alignment

Prototype Modeling. For each content category $c \in \{1, \ldots, C\}$, we model the distribution of semantic features using a Gaussian Mixture Model (GMM) with K prototypes $\{\boldsymbol{\mu}_c^k, \boldsymbol{\Sigma}_c^k\}_{k=1}^K$ and mixture weights $\{\pi_c^k\}_{k=1}^K$. Given the semantic representation $\mathbf{s}_i^c$ of the i-th sample from category c, its soft assignment to the k-th prototype is computed via the E-step:

$$\gamma_{i,k}^c = \frac{\pi_c^k \cdot \mathcal{N}(\mathbf{s}_i^c \mid \boldsymbol{\mu}_c^k, \boldsymbol{\Sigma}_c^k)}{\sum_{j=1}^K \pi_c^j \cdot \mathcal{N}(\mathbf{s}_i^c \mid \boldsymbol{\mu}_c^j, \boldsymbol{\Sigma}_c^j)}. \tag{6}$$

The aligned representation is obtained by projecting $\mathbf{s}_i^c$ onto prototype centers weighted by responsibilities:

$$\tilde{\mathbf{s}}_i^c = \sum_{k=1}^K \gamma_{i,k}^c \boldsymbol{\mu}_c^k. \tag{7}$$

This projection removes instance-specific variations within each category, forcing the downstream classifier to rely on prototype-level, category-normalized forgery cues.

Fairness-aware Prototype Balancing. Standard EM updates for mixture weights use $\pi_c^k \propto N_c^k$, where $N_c^k = \sum_i \gamma_{i,k}^c$ counts the effective samples assigned to prototype k in category c. However, this can result in an imbalanced prototype usage across categories: certain prototypes may become preferentially associated with specific content categories, allowing prototype identities to encode content priors.

To prevent this, we apply a fairness-aware M-step that encourages balanced prototype usage [Li *et al.*, 2024; Zhao *et al.*, 2024]. Let $\bar{\gamma}_c^k = \frac{1}{N_c} \sum_{i \in \mathcal{D}_c} \gamma_{i,k}^c$ denote the average responsibility of prototype k in category c, and $\bar{\gamma}^k = \frac{1}{C} \sum_{c=1}^C \bar{\gamma}_c^k$ the global average. We modify the mixture weight update as:

$$\pi_c^k = \frac{N_c^k}{d_c^k}, \quad \text{where} \quad d_c^k = \begin{cases} N_c + \lambda_{\text{fair}} & \text{if } \bar{\gamma}_c^k > \bar{\gamma}^k \\ N_c - \lambda_{\text{fair}} & \text{if } \bar{\gamma}_c^k \le \bar{\gamma}^k \end{cases} \tag{8}$$

and $N_c = |\mathcal{D}_c|$. This adaptive regularization penalizes prototypes that are over-represented in a category ($\bar{\gamma}_c^k > \bar{\gamma}^k$) by effectively reducing their mixture weight, while encouraging under-represented prototypes. The strength $\lambda_{\text{fair}} > 0$ controls the degree of balancing. Weights are renormalized to ensure $\sum_{k=1}^K \pi_c^k = 1$.

Joint Training via Gradient Bridging. GMM parameters are estimated via the non-differentiable EM algorithm, which prevents direct backpropagation. Prior work addresses this through staged training, decoupling feature learning from debiasing. We instead enable joint optimization via gradient bridging.

Let $\bar{\mathbf{s}}_i^c$ denote $\mathbf{s}_i^c$ with gradients detached. The classifier input is:

$$\mathbf{s}_{\text{cls}}^c = \tilde{\mathbf{s}}_i^c(\bar{\mathbf{s}}_i^c) + (\mathbf{s}_i^c - \bar{\mathbf{s}}_i^c). \tag{9}$$

During forward pass, $\mathbf{s}_i^c - \bar{\mathbf{s}}_i^c = \mathbf{0}$, so the classifier operates on aligned features $\tilde{\mathbf{s}}_i^c$. During backpropagation, gradients bypass the non-differentiable GMM: $\frac{\partial \mathcal{L}_{\text{cls}}}{\partial \mathbf{s}_i^c} = \frac{\partial \mathcal{L}_{\text{cls}}}{\partial \mathbf{s}_{\text{cls}}^c}$, allowing the encoder to adapt to the debiasing objective.

Classification Objective. Aligned features are classified via a two-layer MLP:

$$\mathcal{L}_{\text{cls}} = -\mathbb{E}_{p^*(X,y)}\big[\log p(y \mid \mathbf{s}_{\text{cls}}^c)\big]. \tag{10}$$

Stable Estimation via Memory Bank. To stabilize GMM fitting under mini-batch training, we maintain a memory bank storing recent features per category. Current batch features are merged with historical samples before EM, ensuring reliable parameter estimation.

4.4 Overall Training Objective

The complete training objective integrates contrastive feature learning, variational disentanglement, and classification:

$$\mathcal{L}_{\text{total}} = w_1 \mathcal{L}_{\text{CLIP}} + \mathcal{L}_{\text{dis}} + w_2 \mathcal{L}_{\text{cls}} \tag{11}$$

where $\mathcal{L}_{\text{CLIP}}$ (Eq. 3) is the image-text contrastive loss, $\mathcal{L}_{\text{dis}}$ (Eq. 5) is the disentanglement loss, and $\mathcal{L}_{\text{cls}}$ (Eq. 10) is the classification loss on aligned representations. We set $w_1 = 0.01$, $w_2 = 1.0$. Training alternates between updating GMM parameters via fairness-aware EM (Eq. 6) and backpropagating through $\mathcal{L}_{\text{total}}$ to optimize neural networks. The gradient bridging mechanism (Eq. 9) enables joint optimization despite the non-differentiable GMM module.

5 Theoretical Analysis

We study generalization under *spurious correlation reversal*, where the association between content categories and authenticity labels change from training to testing, while forgery patterns remain invariant.

Setup. Let P_{tr} and P_{te} be the training and test distributions over (X, Y, C).

Assumption 1 (Spurious Correlation Shift)**.** *The following conditions hold:*

$$\begin{aligned}&\textit{(i)}\ P_{\text{tr}}(y \mid c) \neq P_{\text{te}}(y \mid c) && \textit{for some } c \in \mathcal{C}, \\ &\textit{(ii)}\ P_{\text{tr}}(x \mid y, c) = P_{\text{te}}(x \mid y, c) && \textit{for all } y, c, \\ &\textit{(iii)}\ P_{\text{tr}}(c) = P_{\text{te}}(c) && \textit{for all } c, \\ &\textit{(iv)}\ P_{\text{tr}}(y) = P_{\text{te}}(y). &&\end{aligned}$$

Assumption 1 captures spurious correlation reversal: the conditional association $P(y \mid c)$ shifts across domains, while within-category forgery patterns and marginal distributions remain unchanged. Under (iii) and (iv), this is equivalent to a reversal of the category distribution within each label, *i.e.,* $P_{\text{tr}}(c \mid y) \neq P_{\text{te}}(c \mid y)$.

Disentanglement quality. We require the learned semantic representation $\mathbf{s}$ to be approximately independent of the content category given the label:

$$\mathrm{TV}(P(\mathbf{s} \mid y, c),\, P(\mathbf{s} \mid y)) \le \varepsilon, \qquad \forall\, y \in \{0,1\},\ c \in \mathcal{C}. \tag{12}$$

where $\mathrm{TV}(\cdot,\cdot)$ denotes the total variation distance between two distributions, defined as $\mathrm{TV}(P,Q) = \frac{1}{2}\int |p(z) - q(z)|\,\mathrm{d}z$. This condition ensures that, conditioned on y, the content category c provides negligible additional information about $\mathbf{s}$.

Theorem 1 (Generalization under Correlation Reversal)**.** *Let $\ell(g, Y) = |g - Y|$ and define $\Delta_y(c) = P_{\mathrm{te}}(y = 1 \mid c) - P_{\mathrm{tr}}(y = 1 \mid c)$. For any classifier $f : \mathbb{R}^{d_s} \to [0,1]$ and predictor $g(X) = f(\mathbf{s}(X))$, under Assumption 1 and* (12)*,*

$$\Big|\mathbb{E}_{\mathrm{te}}[\ell(g(X),Y)] - \mathbb{E}_{\mathrm{tr}}[\ell(g(X),Y)]\Big| \;\le\; 2\,\varepsilon \cdot \max_{c\in\mathcal{C}} |\Delta_y(c)|. \tag{13}$$

Proof sketch. By the law of total expectation and Assumption 1(iii), the train–test error difference can be decomposed as a weighted sum over content categories. Under Assumption 1(ii), the conditional expectations of the predictor given (y, c) are identical across training and test domains, so the difference arises solely from the shift in $P(y \mid c)$. Moreover, Assumption 1(iv) ensures that the label marginals cancel out globally. The remaining term depends on how much the predictor varies with c within each label. Condition (12) bounds this variation uniformly by ε for any bounded predictor, which yields the bound in (13).

Interpretation. The generalization gap is controlled by two factors: **(i)** the disentanglement quality ε, measuring deviation from $\mathbf{s} \perp c \mid y$; and **(ii)** the magnitude of spurious correlation shift $\max_c |\Delta_y(c)|$. Even under complete category reversal within each label, the error gap remains small provided ε is small.

Connection to PURE. Theorem 1 indicates that robustness follows from minimizing ε. Our method achieves this via *CSG disentanglement*, which separates forgery-relevant semantics from nuisance factors, and *GMM prototype alignment*, which further suppresses category-specific variations. If the aligned representation $\tilde{\mathbf{s}}$ satisfies $\mathrm{TV}(P(\tilde{\mathbf{s}} \mid y, c), P(\tilde{\mathbf{s}} \mid y)) \le \tilde{\varepsilon}$, the bound tightens to $2\tilde{\varepsilon} \cdot \max_c |\Delta_y(c)|$.

6 Experiments

6.1 Settings

Datasets. We utilize four real-world benchmarks to verify the effectiveness of PURE. (1) **CIFAKE** [Bird and Lotfi, 2024] is a CIFAR-10-style dataset containing 5,000 real and 5,000 GAN-generated images across 10 semantic categories. The fake images are generated using a single GAN model to create controlled artifacts. (2) **GenImage** [Zhu *et al.*, 2023] consists of 8,000 GLIDE-generated images spanning 29 object categories. This dataset evaluates detection performance on state-of-the-art diffusion models. (3) **DFD** and (4) **DFDC** are facial image subsets from the AIFace benchmark [Lin *et al.*, 2025], totaling 6,200 images annotated with the Monk Skin Tone Scale (10 levels). These datasets enable evaluation of demographic fairness under spurious correlations between skin tone and authenticity labels. To rigorously evaluate robustness against spurious correlations, we design train-test splits where class labels (real/fake) are intentionally correlated with content or demographic attributes during training, but this correlation is fully reversed at test time. This challenging protocol examines whether models rely on spurious shortcuts or genuinely learn forgery-discriminative patterns. Due to space limitations, we defer detailed dataset descriptions and split configurations to Appendix.

Baselines. We compare PURE with ten representative methods that address generalizable forgery detection from four perspectives: (1) frequency and noise-based methods attempt to capture generation artifacts in the frequency domain, including SRM [Luo *et al.*, 2021] that leverages high-frequency features via spatial rich model filters, LGrad [Tan *et al.*, 2023] that learns on gradient representations, and NPR [Tan *et al.*, 2024] that exploits up-sampling artifacts; (2) representation learning methods leverage pre-trained features for better generalization, including UniFD [Ojha *et al.*, 2023] that employs CLIP features for zero-shot detection, RINE [Koutlis and Papadopoulos, 2024] that utilizes intermediate encoder-block representations, and FatFormer [Liu *et al.*, 2024a] that proposes forgery-aware adaptive transformers; (3) disentanglement-based methods aim to separate forgery-relevant features from content, including UCF [Yan *et al.*, 2023] that extracts universal forgery features via three-level disentanglement and Effort [Yan *et al.*, 2025] that employs SVD to orthogonally separate semantic and forgery features; and (4) fairness and generalization methods explicitly address distribution shift and bias, including PFG [Lin *et al.*, 2024] that preserves fairness under domain shift and D3 [Yang *et al.*, 2025] that learns from cross-generator discrepancy. For all baselines, we use official implementations and evaluate under our spurious correlation protocol with identical train-test splits.

Implementation. We use CLIP (ViT-Large/14) [Radford *et al.*, 2021] as the feature extraction backbone in PURE, leveraging its rich semantic priors from large-scale vision-language pretraining. Image descriptions are generated using BLIP (Salesforce/blip-image-captioning-base) [Li *et al.*, 2022] for contrastive fine-tuning. The CSG encoder disentangles CLIP features into semantic representation $\mathbf{s} \in \mathbb{R}^{128}$ and nuisance representation $\mathbf{v} \in \mathbb{R}^{32}$, and a transposed-convolutional decoder reconstructs the original image from concatenated $(\mathbf{s}, \mathbf{v})$. To ensure fair comparison across different datasets, images are resized to 224×224 with standard data augmentations. The results represent the average values obtained from 5 runs for all compared methods.

6.2 Results

Performance Comparison. Table 1 shows the overall performance of our method, PURE, compared with baselines across four real-world datasets. PURE consistently outperforms all baselines, achieving 94.53% accuracy and 0.947 F1 score on average and surpassing the runner-up FatFormer by 3.55% and 3.9, respectively. On CIFAKE and GenImage, PURE achieves gains over the runner-up FatFormer by 2.6% and 2.0% in accuracy, and 0.04 and 0.01 in F1 score, respectively.

Methods	Venue	CIFAKE		GenImage		AIFace				Avg.	
						DFD		DFDC			
		Acc(%)↑	F1↑	Acc(%)↑	F1↑	Acc(%)↑	F1↑	Acc(%)↑	F1↑	Acc(%)↑	F1↑
SRM	CVPR 2021	90.0(±0.10)	0.90(±0.01)	85.1(±0.05)	0.87(±0.05)	91.3(±0.03)	0.91(±0.03)	85.3(±0.03)	0.84(±0.03)	87.93	0.880
UniFD	CVPR 2023	67.5(±0.03)	0.74(±0.03)	50.4(±0.04)	0.49(±0.04)	70.0(±0.03)	0.74(±0.03)	49.5(±0.01)	0.52(±0.01)	59.35	0.623
LGrad	CVPR 2023	50.6(±0.20)	0.50(±0.30)	81.0(±0.12)	0.78(±0.12)	87.9(±0.11)	0.88(±0.01)	66.7(±0.13)	0.667(±0.02)	71.55	0.707
UCF	ICCV 2023	90.1(±0.10)	0.90(±0.01)	92.6(±0.07)	0.93(±0.06)	92.0(±0.06)	0.92(±0.05)	81.2(±0.22)	0.79(±0.03)	89.02	0.885
PFG	CVPR 2024	88.9(±0.10)	0.89(±0.02)	86.1(±0.20)	0.86(±0.03)	50.2(±0.33)	0.53(±0.10)	45.7(±0.30)	0.29(±0.10)	67.73	0.643
NPR	CVPR 2024	85.0(±0.03)	0.86(±0.03)	97.0(±0.03)	0.97(±0.03)	85.7(±0.01)	0.87(±0.01)	74.2(±0.01)	0.77(±0.01)	85.48	0.865
RINE	ECCV2024	90.4(±0.06)	0.90(±0.06)	92.1(±0.03)	0.92(±0.03)	84.5(±0.01)	0.85(±0.01)	78.5(±0.03)	0.79(±0.03)	86.38	0.865
FatFormer	CVPR 2024	91.0(±0.03)	0.90(±0.03)	97.8(±0.03)	0.98(±0.03)	90.1(±0.03)	0.90(±0.03)	85.0(±0.03)	0.85(±0.03)	90.98	0.908
D3	CVPR 2025	87.9(±0.03)	0.88(±0.03)	51.4(±0.03)	0.33(±0.03)	92.0(±0.03)	0.92(±0.03)	87.3(±0.03)	0.87(±0.01)	79.65	0.750
Effort	ICML 2025	87.3(±0.03)	0.87(±0.03)	64.8(±0.03)	0.70(±0.03)	84.3(±0.01)	0.84(±0.01)	69.2(±0.03)	0.71(±0.03)	76.40	0.780
PURE (ours)	IJCAI 2026	**93.6**(±0.08)	**0.94**(±0.06)	**99.8**(±0.01)	**0.99**(±0.01)	**94.0**(±0.30)	**0.94**(±0.03)	**90.7**(±0.10)	**0.91**(±0.01)	**94.53**	**0.947**

Table 1: Overall performance across all experimental datasets and baselines. Bold indicates the best performance, and underlined indicates the runner-up.

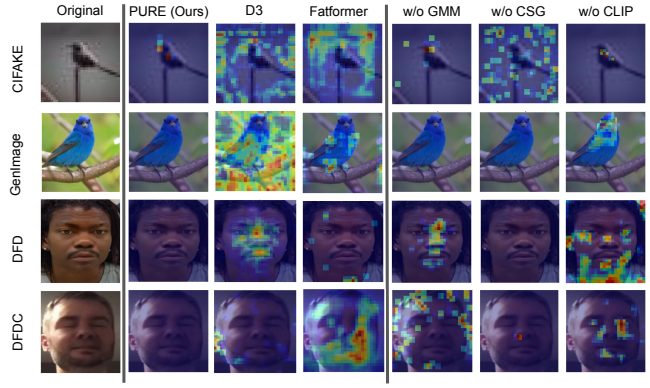

Figure 3: Occlusion sensitivity analysis across four datasets. Rows show samples from Heatmaps visualize spatial sensitivity, where red indicates high sensitivity and blue indicates low sensitivity.

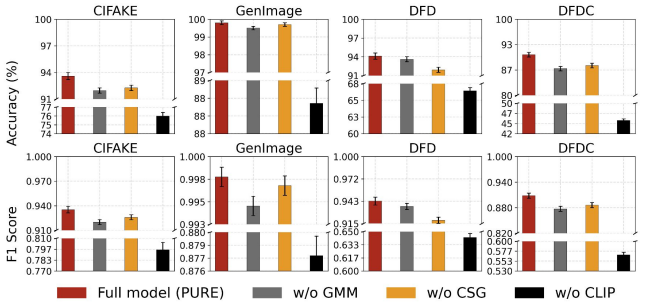

Figure 4: The ablation results for PURE highlight the complementary roles of the CLIP backbone, CSG disentanglement, and GMM-based prototype alignment in enabling robust, content-agnostic forgery detection.

CIFAKE's challenging dog-to-cat content shift validates that prototype alignment effectively decouples forgery features from category-specific patterns. The near-perfect GenImage performance demonstrates robust generalization on diffusion-generated content. On DFD and DFDC, PURE outperforms the runner-up D3 by 2.0% and 3.4% in accuracy, and 0.02 and 0.04 in F1 score. These benchmarks evaluate robustness to demographic bias under reversed skin-tone correlations. The substantial improvement on DFDC underscores the effectiveness of fairness-aware prototype balancing on challenging facial datasets with complex human-attribute biases.

Effectiveness of PURE for Content-Agnostic Detection. To understand how PURE achieves content-agnostic detection, we apply occlusion sensitivity analysis (OSA) [Zeiler and Fergus, 2014] to visualize what regions the model relies on for forgery detection. OSA systematically occludes each image patch and measures the resulting change in prediction confidence. Patches whose occlusion causes large confidence drops indicate high model sensitivity (the model relies heavily on these regions), while patches causing minimal changes indicate low sensitivity. These sensitivity scores are then mapped to a color spectrum: red for high sensitivity and blue for low sensitivity. In the context of forgery detection, red regions concentrated on semantic content (*e.g.,* faces, objects) indicate that the model is making decisions based on content cues rather than forgery artifacts, which leads to poor generalization under spurious correlation shifts. Figure 3 compares our method against runner-up baselines (D3 and FatFormer) and ablation variants of PURE (w/o GMM, w/o CSG, w/o CLIP). We observe that the baselines exhibit dense red regions concentrated on semantic areas (faces and objects), indicating strong reliance on content cues rather than forgery artifacts. In contrast, our full model presents uniformly light blue heatmaps with no red regions, demonstrating that it does not rely on any specific semantic content for prediction and instead captures forgery-relevant features distributed across the entire image. Critically, all ablation variants (w/o GMM, w/o CSG, w/o CLIP) revert to content-biased patterns with concentrated red regions similar to baselines, demonstrating that only the joint use of CSG-based semantic disentanglement and GMM prototype alignment on top of the CLIP backbone can effectively suppress content-related cues and enable content-agnostic forgery detection.

Ablation Studies. We conduct systematic ablations to quantify the contribution of each component in PURE. As shown in Figure 4, removing the GMM module (w/o GMM) leads to clear performance drops across all datasets, indicating that even after CSG-based disentanglement, forgery-related features still retain some entanglement with spurious content patterns. Without GMM's prototype-based alignment to normalize these residual category-specific variations, classifiers can still exploit spurious content-label correlations. Removing CSG (w/o CSG) also causes substantial performance degradation, highlighting that disentanglement is a critical prerequisite for effective alignment. When CLIP

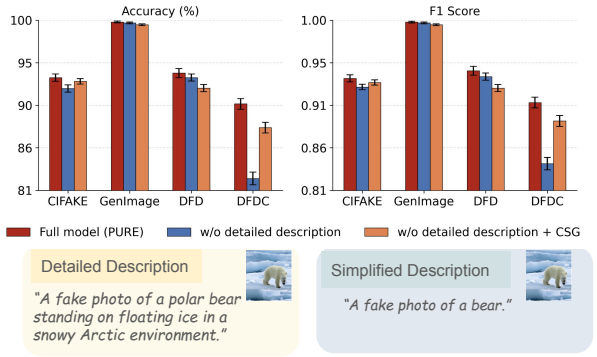

Figure 5: The top row shows Accuracy (left) and F1 score (right) on CIFAKE, GenImage, DFD, and DFDC using simplified image descriptions, with CSG (blue) and without CSG (orange). Bars denote model variants without detailed descriptions, and error bars indicate standard deviation over multiple runs. The bottom row presents examples of detailed and simplified descriptions generated by BLIP-style captioning.

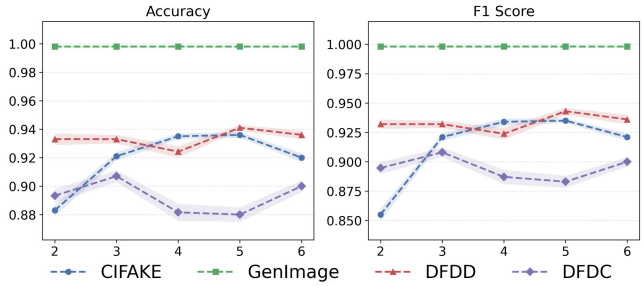

Figure 6: Performance sensitivity to the number of GMM components $K \in \{2, 3, 4, 5, 6\}$ on four datasets.

features remain entangled with semantic and style factors, GMM alignment becomes less effective because prototypes are constructed from contaminated representations that conflate content and forgery information. The complementary nature of both components is evident: removing either CSG or GMM causes substantial performance degradation, confirming their synergistic roles where CSG isolates forgery-relevant factors from semantic content and GMM normalizes residual category-specific patterns across different forgery types. Removing the CLIP encoder (w/o CLIP) results in catastrophic performance drops across all datasets, far exceeding the degradation caused by removing any individual component. This underscores that CLIP's vision-language pretraining serves as the foundational backbone of our framework, providing the semantic understanding necessary for both CSG disentanglement and GMM alignment to function effectively.

Quality of Generated Image Descriptions. We further analyze the impact of image description quality on model performance by replacing BLIP-generated detailed descriptions with simplified ones. Figure 5 compares the full model (PURE) using detailed descriptions against two variants using simplified descriptions: one that retains CSG (w/o detailed description) and one that removes CSG (w/o detailed description + CSG). Overall, using simplified descriptions consistently degrades performance across all datasets compared to the full model with detailed descriptions. This degradation occurs because simplified descriptions lack the semantic richness necessary to guide CLIP in extracting comprehensive content representations—without detailed contextual information about environment, lighting, and fine-grained object attributes, CLIP produces semantically impoverished feature representations that fail to provide a sufficient basis for effective content-forgery disentanglement. This impact is particularly pronounced on complex datasets: on DFDC, the full model achieves 90.81% F1 score, whereas using simplified descriptions reduces performance to 83.98%. Beyond this overall trend, we observe that under detailed descriptions, CSG consistently improves performance across all datasets, demonstrating its robust effectiveness when provided with sufficient semantic information. In contrast, under simplified descriptions, CSG exhibits clear dataset-dependent behavior. On GenImage and DFD, retaining CSG maintains competitive performance (*e.g.,* 93.70% F1 on DFD), while removing it causes further degradation (*e.g.,* to 92.43%), indicating that CSG can still extract meaningful disentanglement even with limited semantic signals. However, on datasets where forgery traces are highly entangled with semantic factors, such as CIFAKE and DFDC, retaining CSG under simplified descriptions actually degrades performance (*e.g.,* 83.98% F1 on DFDC versus 88.75% without CSG). This suggests that when CSG actively disentangles semantic factors but the semantic representations themselves are incomplete, the disentanglement process may inadvertently affect forgery-relevant features that are entangled with the incomplete semantic information.

Effectiveness of the GMM Component (K). We evaluate the sensitivity of PURE to the number of GMM prototypes $K \in \{2, 3, 4, 5, 6\}$ across four datasets, with results shown in Figure 6. Consistent trends are observed over two metrics. Performance peaks at $K = 3$ or $K = 5$: smaller K lacks the capacity to model intra-category content diversity, while larger K risks overfitting and capturing instance-level variations.

7 Conclusion

In this paper, we introduce a novel framework, PURE, for content-agnostic AI-generated image detection under distribution shift. The framework is designed to explicitly mitigate spurious correlations, particularly content bias, which arises when detectors couple forgery artifacts with semantic content instead of learning generalizable forgery patterns. PURE integrates a Causal Semantic Generative (CSG) mechanism and a Gaussian Mixture Model (GMM)-based prototype alignment module to achieve robust detection across content distribution shifts. The CSG mechanism disentangles semantic representations encoding forgery-relevant information from nuisance variations capturing forgery-irrelevant style. The GMM-based prototype alignment further normalizes category-specific feature distributions, suppressing content priors and enabling invariant forgery representation learning. Extensive experiments demonstrate that PURE consistently outperforms baseline methods under spurious correlation reversal settings.

Acknowledgments

This work is supported by the National Science Foundation (NSF) # 2515265. Minglai Shao did not receive any financial support for this work and contributed only by developing the research ideas, participating in discussions, and providing feedback on the manuscript.

References

[Bai *et al.*, 2025] Ningning Bai, Xiaofeng Wang, Ruidong Han, Jianpeng Hou, Qin Wang, and Shanmin Pang. Towards generalizable face forgery detection via mitigating spurious correlation. *Neural Networks*, 182:106909, 2025.

[Bird and Lotfi, 2024] Jordan J Bird and Ahmad Lotfi. Cifake: Image classification and explainable identification of ai-generated synthetic images. *IEEE Access*, 12:15642–15650, 2024.

[Fu *et al.*, 2025] Xinghe Fu, Zhiyuan Yan, Taiping Yao, Shen Chen, and Xi Li. Exploring unbiased deepfake detection via token-level shuffling and mixing. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 3040–3048, 2025.

[Guillaro *et al.*, 2023] Fabrizio Guillaro, Davide Cozzolino, Avneesh Sud, Nicholas Dufour, and Luisa Verdoliva. Trufor: Leveraging all-round clues for trustworthy image forgery detection and localization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 20606–20615, 2023.

[Guo *et al.*, 2023] Xiao Guo, Xiaohong Liu, Zhiyuan Ren, Steven Grosz, Iacopo Masi, and Xiaoming Liu. Hierarchical fine-grained image forgery detection and localization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3155–3165, 2023.

[Guo *et al.*, 2024a] Kun Guo, Haochen Zhu, and Gang Cao. Effective image tampering localization via enhanced transformer and co-attention fusion. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4895–4899. IEEE, 2024.

[Guo *et al.*, 2024b] Kun Guo, Haochen Zhu, and Gang Cao. Effective image tampering localization via enhanced transformer and co-attention fusion. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4895–4899. IEEE, 2024.

[He *et al.*, 2026] Zhixia He, Chen Zhao, Minglai Shao, Xintao Wu, Xujiang Zhao, Dong Li, Qin Tian, and Linlin Yu. Out-of-distribution detection with positive and negative prompt supervision using large language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pages 21699–21707, 2026.

[Ho *et al.*, 2020] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

[Karras *et al.*, 2019] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4401–4410, 2019.

[Kashiani *et al.*, 2025] Hossein Kashiani, Niloufar Alipour Talemi, and Fatemeh Afghah. Freqdebias: Towards generalizable deepfake detection via consistency-driven frequency debiasing. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8775–8785. IEEE, 2025.

[Koutlis and Papadopoulos, 2024] Christos Koutlis and Symeon Papadopoulos. Leveraging representations from intermediate encoder-blocks for synthetic image detection. In *European Conference on Computer Vision*, pages 394–411. Springer, 2024.

[Li *et al.*, 2022] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International conference on machine learning*, pages 12888–12900. PMLR, 2022.

[Li *et al.*, 2024] Dong Li, Chen Zhao, Minglai Shao, and Wenjun Wang. Learning fair invariant representations under covariate and correlation shifts simultaneously. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*, pages 1174–1183, 2024.

[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*, pages 16815–16825, 2024.

[Lin *et al.*, 2025] Li Lin, Santosh 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 Computer Vision and Pattern Recognition Conference*, pages 3503–3515, 2025.

[Liu *et al.*, 2021] Chang Liu, Xinwei Sun, Jindong Wang, Haoyue Tang, Tao Li, Tao Qin, Wei Chen, and Tie-Yan Liu. Learning causal semantic representation for out-of-distribution prediction. *Advances in Neural Information Processing Systems*, 34:6155–6170, 2021.

[Liu *et al.*, 2024a] Huan Liu, Zichang Tan, Chuangchuang Tan, Yunchao Wei, Jingdong Wang, and Yao Zhao. Forgery-aware adaptive transformer for generalizable synthetic image detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10770–10780, 2024.

[Liu *et al.*, 2024b] Yang Liu, Xiaofei Li, Jun Zhang, Shengze Hu, and Jun Lei. Da-hfnet: Progressive fine-grained forgery image detection and localization based on dual attention. In *2024 3rd International Conference on Image Processing and Media Computing (ICIPMC)*, pages 51–58. IEEE, 2024.

[Liu *et al.*, 2024c] Yang Liu, Xiaofei Li, Jun Zhang, Shengze Hu, and Jun Lei. Da-hfnet: Progressive fine-grained forgery image detection and localization based on dual attention. In *2024 3rd International Conference on Image Processing and Media Computing (ICIPMC)*, pages 51–58. IEEE, 2024.

[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*, pages 16317–16326, 2021.

[Miao *et al.*, 2025] Changtao Miao, Qi Chu, Tao Gong, Zhentao Tan, Zhenchao Jin, Wanyi Zhuang, Man Luo, Honggang Hu, and Nenghai Yu. Mixture-of-noises enhanced forgery-aware predictor for multi-face manipulation detection and localization. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pages 3903–3912, 2025.

[Nichol *et al.*, 2021] Alex Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and Mark Chen. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. *arXiv preprint arXiv:2112.10741*, 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*, pages 24480–24489, 2023.

[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, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR, 2021.

[Rombach *et al.*, 2022] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022.

[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*, pages 12105–12114, 2023.

[Tan *et al.*, 2024] Chuangchuang Tan, Yao Zhao, Shikui Wei, Guanghua Gu, Ping Liu, and Yunchao Wei. Rethinking the up-sampling operations in cnn-based generative network for generalizable deepfake detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 28130–28139, 2024.

[Trinh and Liu, 2021] Loc Trinh and Yan Liu. An examination of fairness of ai models for deepfake detection. *arXiv preprint arXiv:2105.00558*, 2021.

[Wu *et al.*, 2025a] Xinyu Wu, Xiaohui Chen, Xintao Wu, Dong Li, Zhong Chen, Yi He, and Chen Zhao. Explainable image-centric forgery detection: A survey. *Authorea Preprints*, 2025.

[Wu *et al.*, 2025b] Yixin Wu, Feiran Zhang, Tianyuan Shi, Ruicheng Yin, Zhenghua Wang, Zhenliang Gan, Xiaohua Wang, Changze Lv, Xiaoqing Zheng, and Xuanjing Huang. Explainable synthetic image detection through diffusion timestep ensembling. *arXiv preprint arXiv:2503.06201*, 2025.

[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*, pages 22412–22423, 2023.

[Yan *et al.*, 2025] Zhiyuan Yan, Jiangming Wang, Zhendong Wang, Peng Jin, Ke-Yue Zhang, Shen Chen, Taiping Yao, Shouhong Ding, Baoyuan Wu, and Li Yuan. Effort: Efficient orthogonal modeling for generalizable ai-generated image detection. *Proceedings of the 42nd International Conference on Machine Learning*, 2025.

[Yang *et al.*, 2025] Yongqi Yang, Zhihao Qian, Ye Zhu, Olga Russakovsky, and Yu Wu. Dˆ 3: Scaling up deepfake detection by learning from discrepancy. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 23850–23859, 2025.

[Yu *et al.*, 2024] Zeqin Yu, Jiangqun Ni, Yuzhen Lin, Haoyi Deng, and Bin Li. Diffforensics: Leveraging diffusion prior to image forgery detection and localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12765–12774, 2024.

[Zeiler and Fergus, 2014] Matthew D Zeiler and Rob Fergus. Visualizing and understanding convolutional networks. In *European conference on computer vision*, pages 818–833. Springer, 2014.

[Zhao *et al.*, 2024] Chen Zhao, Kai Jiang, Xintao Wu, Haoliang Wang, Latifur Khan, Christan Grant, and Feng Chen. Algorithmic fairness generalization under covariate and dependence shifts simultaneously. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 4419–4430, 2024.

[Zhu *et al.*, 2023] Mingjian Zhu, Hanting Chen, Qiangyu Yan, Xudong Huang, Guanyu Lin, Wei Li, Zhijun Tu, Hailin Hu, Jie Hu, and Yunhe Wang. Genimage: A million-scale benchmark for detecting ai-generated image. *Advances in Neural Information Processing Systems*, 36:77771–77782, 2023.

[Zhu *et al.*, 2025] Xuekang Zhu, Xiaochen Ma, Lei Su, Zhuohang Jiang, Bo Du, Xiwen Wang, Zeyu Lei, Wentao Feng, Chi-Man Pun, and Ji-Zhe Zhou. Mesoscopic insights: orchestrating multi-scale & hybrid architecture for image manipulation localization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 11022–11030, 2025.