FAIR: Feature-Augmented Implicit Regularization for AI-generated Fake Image Detection

Md Redwanul Haque[1(✉)], Manzur Murshed[1], Manoranjan Paul[2], and Tsz-Kwan Lee[1]

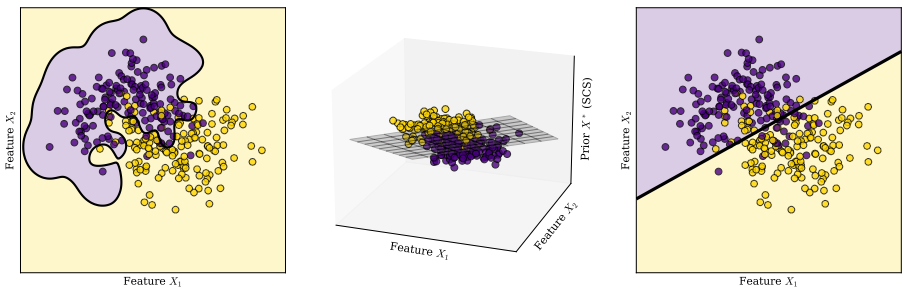
[1] Deakin University, Burwood, VIC, Australia
m.haque@research.deakin.edu.au
[2] Charles Sturt University, Bathurst, NSW, Australia

Abstract. Generalization remains a critical bottleneck in AI-generated image detection. Because many modern generators are proprietary or adversarially modified, existing detectors overfit to the low-level textural patterns of accessible training data, resulting in severe failures on unseen domains. Conventional regularization techniques (e.g., L_1/L_2 norms, Dropout) apply indiscriminate parametric constraints and fail to provide the domain-invariant structure necessary for cross-generator robustness. To address this, we propose Feature-Augmented Implicit Regularization (FAIR). FAIR introduces an orthogonal, macro-structural prior, specifically, Scene Composition Structure (SCS), during training to geometrically constrain the model's optimization trajectory. By augmenting the primary feature space with domain-invariant SCS features, FAIR explicitly penalizes texture-biased shortcut learning. Crucially, this structural prior is entirely discarded at inference, yielding a smoothed, generalized decision boundary with zero architectural or computational overhead. Extensive evaluations across five massive benchmarks demonstrate that integrating FAIR into state-of-the-art detectors significantly improves cross-generator generalization, boosting accuracy by up to 8.04% and establishing new state-of-the-art robustness in zero-shot transfer scenarios.

1 Introduction

The proliferation of highly realistic AI-Generated Content (AIGC) poses significant societal risks, necessitating robust detection methods. A primary bottleneck in this field is **cross-generator generalization**: a detector trained on artifacts from one source (e.g., Stable Diffusion [25]) must maintain high performance on entirely unseen generative manifolds (e.g., Midjourney [21]).

This crucial domain gap stems from the impossibility of comprehensive training data collection. The continuous, rapid development of new architectures and fine-tuning techniques renders any static training dataset instantly obsolete. Furthermore, real-world AIGC is frequently generated by proprietary or locally fine-tuned *hidden* models, deployed adversarially. Consequently, their specific



(a) Overfitted Boundary **(b)** Hyperplane Tilting **(c)** Smoothed Boundary

Fig. 1: Conceptual Mechanism of Feature-Augmented Implicit Regularization (FAIR). (a) Standard detectors learn highly non-linear boundaries by overfitting to local textural artifacts. (b) During training, FAIR introduces an orthogonal structural prior, allowing the optimizer to discover a simpler, tilted separating hyperplane in the augmented dimensional space. (c) Discarding the prior at inference projects this hyperplane back as a smoothed, generalized decision boundary. (X_1, X_2: primary textural features; X^*: structural prior).

artifact fingerprints remain permanently inaccessible to detection developers, making signature-based approaches futile.

Due to this data incompleteness, existing detectors exhibit severe **source-domain overfitting**. They learn to recognize low-level textural artifacts specific to the training generators, resulting in highly complex, brittle decision boundaries that collapse on unseen domains. Achieving cross-generator generalization thus demands a shift away from spectral or noise-based cues toward domain-invariant constraints.

Traditional regularization methods (L_1/L_2 penalties, Dropout [26]) impose indiscriminate, data-agnostic constraints on parameter weights, acting as blunt instruments that fail to penalize texture-biased shortcut learning. Instead, we advocate for an **informed implicit regularization** approach that exploits the physical geometry of the image. Specifically, we leverage **Scene Composition Structure (SCS)** [12]. While generators [11, 17] vary wildly in their frequency fingerprints, causing standard detectors to overfit to source-specific noise, macro-level scene composition remains fundamentally domain-invariant. By utilizing SCS, we introduce a feature space orthogonally abstract to standard texture-biased backbones, providing the stable anchor required for generalization.

To implement this idea, we introduce the Feature-Augmented Implicit Regularization (FAIR) strategy. We apply FAIR to state-of-the-art AIGC detectors, such as PatchCraft [34] and AIDE [31]. FAIR temporarily augments the detector's latent features with the domain-invariant SCS structural prior strictly during the training phase. As visualized in Fig. 1, rather than learning a highly non-linear, overfitted boundary purely within the primary textural feature space, this privileged geometric information allows the optimizer to *tilt* a lower-complexity separating hyperplane into the auxiliary structural dimensions. Crucially, the

SCS features are entirely discarded before inference. The resulting boundary, when projected back into the primary feature space, remains constrained by the structural anchor, ensuring the deployed model retains the exact architecture and computational overhead of the original base model while gaining massive cross-domain robustness.

Our main contributions are summarized as follows:

1. **A Novel Regularization Framework:** We propose **Feature-Augmented Implicit Regularization (FAIR)**. By leveraging the paradigm of Learning Using Privileged Information (LUPI) [28] to inject an orthogonal structural prior strictly during optimization, FAIR explicitly penalizes texture-biased shortcut learning without adding any computational overhead during inference.
2. **Empirical Validation and Mechanistic Insight:** We implement FAIR using Scene Composition Structure (SCS) features. Through a detailed analysis of weight distributions, loss curves, and robustness to JPEG compression, we demonstrate that FAIR encourages the network to learn a fundamentally smoother and better-calibrated decision boundary.
3. **Extensive Cross-Domain Evaluation:** We verify FAIR's architecture-agnostic nature by integrating it into two leading detectors (AIDE and PatchCraft). Across five massive benchmarks encompassing over 60 unseen target domains, FAIR consistently improves cross-generator generalization, boosting accuracy by up to 8.10% and establishing state-of-the-art robustness.

2 Related Work

2.1 AIGC Detection and Cross-Domain Generalization

The field of AI-Generated Content (AIGC) detection has rapidly shifted from model-specific forensic analysis to the pursuit of universal, cross-generator generalization. Formally, detectors are trained on a source domain $\mathcal{D}_S$ but must minimize expected error on entirely unseen target domains $\mathcal{D}_T$. Early methods focused on low-level, model-specific cues, such as periodic frequency artifacts (FreDect [7]) or cross-GAN textural consistencies (CNNSpot [29], Spec [33], F3Net [23]).

More recent approaches explore generalized perspectives: UnivFD [22] leverages vision-language models like CLIP [24]; DIRE [30] exploits diffusion reconstruction errors; and PatchCraft [34] targets inter-pixel correlations in texture-rich patches. The state-of-the-art AIDE model [31] serves as a powerful hybrid foundation, combining low-level DCT-based [1] forensics with high-level CLIP semantics. Despite these advances, existing detectors still exhibit severe source-domain overfitting, learning fragile decision boundaries that collapse under domain shift.

2.2 Regularization and Shortcut Learning

Deep neural networks are notoriously prone to *shortcut learning* [10], the tendency to exploit trivial, dataset-specific textural artifacts rather than learning robust, generalized features. In AIGC detection, this manifests as networks perfectly memorizing the high-frequency noise, spectral anomalies, and generator-specific fingerprints of the source domain [6, 22] while ignoring global semantic structures.

Traditional regularization methods, such as L_1/L_2 weight decay [5] and Dropout [26], attempt to mitigate overfitting by constraining parameter magnitudes or stochastically deactivating neurons. However, these are *parameter-level, indiscriminate* constraints. Because they operate without utilizing any domain-specific or geometric knowledge, they act as blunt instruments that fail to explicitly penalize texture-biased shortcut learning. Generalization to unseen generative manifolds demands an informed regularization strategy that explicitly guides the geometric optimization of the feature space.

2.3 Privileged Information in Representation Learning

The paradigm of Learning Using Privileged Information (LUPI), originally formulated by [28], accelerates model optimization by introducing an *intelligent teacher*. In this framework, the model is provided with auxiliary, privileged features (X^*) strictly during the training phase. While initially developed for Support Vector Machines [15] (SVM+) to help estimate the difficulty of primary data, we entirely repurpose the LUPI framework for domain generalization. Rather than using privileged information for difficulty estimation, we bridge LUPI with geometric regularization, utilizing an explicit, macro-structural coordinate system as our prior to implicitly regularize a primary textural backbone.

3 Methodology

3.1 Scene Composition Structure (SCS) Prior (X^*)

To provide an orthogonal structural constraint during optimization, we utilize Scene Composition Structure (SCS) features [12], which have been shown to effectively capture global structural regularities independently of localized pixel textures (Fig. 2).

This approach leverages a recursive partitioning technique that identifies and quantifies the most prominent structural boundaries within an image. The process begins by treating the entire image, I, as the initial segment. The structural homogeneity of any segment S is quantified using the sum of squared errors (SSE) of its pixel-level features:

$$e_S = \sum_{p_i \in S} \|p_i - \mu_S\|^2$$

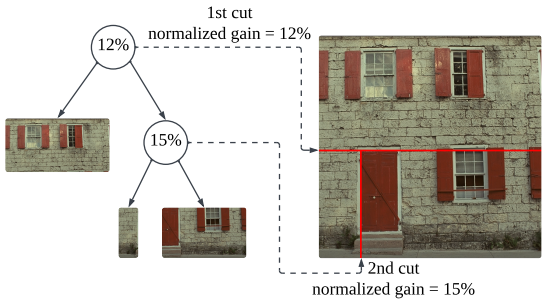


Fig. 2: SCS extraction through hierarchical partitioning: Illustrating the initial cuts made by the algorithm and their resulting binary partition tree.

where p_i is the feature vector of the i-th pixel, and μ_S is the mean feature vector of S.

The image is recursively partitioned by finding the optimal axis-parallel cut that maximally reduces the total SSE. For a segment S split into S_1 and S_2, this reduction (the *gain*, g) serves as a metric for the statistical significance of a structural boundary:

$$g = e_S - (e_{S_1} + e_{S_2})$$

By a greedy approach, the cut yielding the highest gain, $\hat{g} = \max_{\forall \text{cuts}} g$, is selected. This results in a sequence of N gain values, each corresponding to a progressively finer structural division. The final N-dimensional SCS feature vector (X^*) is formed by the normalized cumulative sum of these ordered gain values:

$$\tilde{g}_i = \frac{1}{e_I} \sum_{j=1}^{i} \hat{g}_j, \quad 1 \leq i \leq N$$

To integrate this prior, the N-dimensional vector is passed through a fully connected layer with a GELU activation function [16], creating a compact M-dimensional representation.

3.2 Base Architectures and Primary Features (X)

To demonstrate the architecture-agnostic nature of our regularization strategy, we evaluate FAIR on two distinct state-of-the-art AIGC detection baselines, abstractly denoted as extracting primary features X:

- **AIDE [31]:** A hybrid detector where X is a concatenation of high-level global semantics (via CLIP ConvNeXt [24]) and low-level frequency artifacts (via DCT patching and SRM filters [8]).
- **PatchCraft [34]:** A micro-texture-focused detector where X captures localized inter-pixel correlations within rich and poor texture patches.

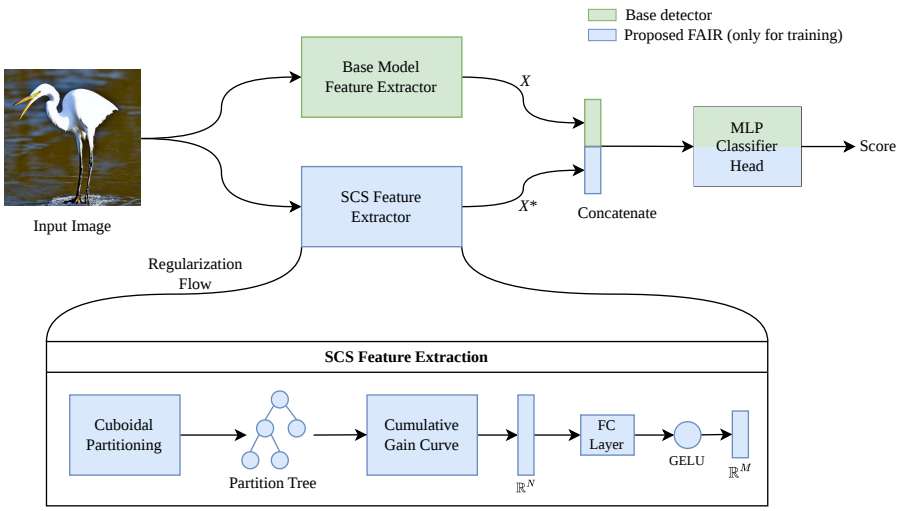


Fig. 3: FAIR Architecture Integration: The trainable Scene Composition Structure feature extractor (blue) generates the structural prior X^* which is concatenated with the primary features X strictly during training to regularize the classification head.

3.3 FAIR Implementation

Feature-Augmented Implicit Regularization (FAIR) intervenes strictly at the final classification stage, as illustrated in Fig. 3. Let the primary feature space be $X \in \mathbb{R}^D$, and let the structural prior be $X^* \in \mathbb{R}^M$. In a standard base model, the classification head would learn a weight matrix $W \in \mathbb{R}^{C \times D}$, where C is the number of classes.

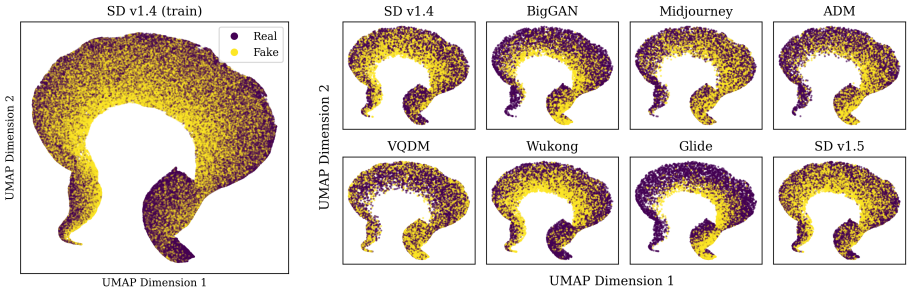
During the FAIR **training phase**, the base model's primary feature extractor remains entirely frozen. The primary features X and the structural prior X^* are concatenated into an augmented feature vector $[X, X^*]$. The classification head is expanded to accommodate this augmented space, learning a joint weight matrix $W_{\text{FAIR}} = [W_X, W_{X^*}]$:

$$\text{Logits}_{\text{train}} = [X, X^*] \cdot [W_X, W_{X^*}]^T + b$$

During the **Inference Phase**, the structural prior X^* is entirely discarded. To maintain dimensional consistency without modifying the original network architecture, we simply extract the sub-matrix W_X and the corresponding bias b from the trained regularized layer. The final, deployed model operates strictly on the primary features:

$$\text{Logits}_{\text{inference}} = X \cdot W_X^T + b$$

Consequently, the weights W_X applied to the base features are inherently constrained by the geometric optimization enforced by X^* during training, yielding a regularized model with zero added computational overhead during deployment.



(a) Low Predictability (Source) **(b)** Domain Invariance (Targets)

Fig. 4: Empirical Properties of the Structural Prior (X^*). (a) Unsupervised UMAP [14] projection of X^* on the source domain shows heavily overlapping real and fake classes ($I(X^*;Y) \approx 0$). This low marginal predictability prevents the optimizer from using the prior as a trivial discriminative shortcut. (b) Projections across diverse, unseen target domains demonstrate a highly stable topological manifold, providing the invariant anchor required for geometric relaxation.

3.4 Mechanism of Cross-Generator Generalization

While standard regularization techniques prevent intra-domain overfitting, FAIR explicitly enables cross-domain generalization through the paradigm of **Learning Using Privileged Information (LUPI)** [28]. By utilizing SCS features as a structural prior (X^*) strictly during the optimization phase, we establish a structurally invariant coordinate system that regularizes the primary backbone.

Drawing on foundational principles of shortcut learning mitigation [2,10] and domain-invariant representation theory [3,9], we establish that for this privileged information to act as an effective regularizer against domain shift, it relies on the intersection of two key properties:

- **Prevention of Shortcut Learning (via Low Marginal Predictability):** Because the SCS structural prior does not contain a trivial discriminative boundary separating real from fake images (as evidenced visually by the mixed classes in Fig. 4a and proven quantitatively in Tab. 4), the optimizer cannot simply ignore the backbone features to minimize loss. It is forced to continuously update the primary backbone weights (W_X) to find a solution.

$$I(X^*;Y) \approx 0$$

- **Invariant Anchoring (via Domain Invariance):** As shown in our unsupervised UMAP projections (Fig. 4b), the topological manifold of the SCS prior remains highly stable across entirely different generative architectures (e.g., Stable Diffusion vs. BigGAN).

$$P(X^* \mid \mathcal{D}_{\text{source}}) \approx P(X^* \mid \mathcal{D}_{\text{target}})$$

By satisfying these properties, FAIR facilitates a *geometric margin relaxation*. During optimization, the network seeks to minimize loss while maintaining low weight magnitudes. Rather than fitting a high-complexity, non-linear boundary within the X-subspace, which leads to source-domain overfitting, the optimizer discovers a smoother boundary by *tilting* the separating hyperplane into the stable X^* dimensions (as conceptualized in Fig. 1). When X^* is discarded at inference, the resulting projection onto the primary feature space X retains this structural anchoring. The learned W_X is thus conditioned on stable compositional geometry rather than fragile pixel statistics, enabling robust mapping of shifted target-domain features.

4 Benchmarks and Setup

To comprehensively evaluate the architecture-agnostic efficacy of Feature-Augmented Implicit Regularization (FAIR), we conduct extensive experiments across five diverse AIGC detection benchmarks. We structure our evaluation into two distinct phases:

1. *Standardized Cross-Generator Generalization:* We utilize the **GenImage** [36] and **AIGCDetect** [34] benchmarks. Following standard convention, models are trained on specific source domains (SDv1.4 for GenImage, ProGAN for AIGCDetect) and tested across their respective unseen target domains.
2. *In-the-Wild Cross-Benchmark Transfer:* We utilize the **UnivFD** [22], **Fake-2M** [20], and **DRCT-2M** [4] datasets. These serve exclusively as massive, out-of-distribution test suites to stress-test models trained on the GenImage SDv1.4 source domain. To highlight the scale of this domain shift, UnivFD contains 21 domains (13 GAN-based, 8 Diffusion-based), Fake2M contains 17 domains (6 GAN, 11 Diffusion), and DRCT-2M contains 16 Diffusion-based domains.

We integrate FAIR into two state-of-the-art architectures: the semantically-driven **AIDE** [31] and the micro-texture-focused **PatchCraft (PC)** [34]. Training was conducted on a single A100 GPU using the default hyperparameters of the base models. The FAIR projection dimensions (N, M) were adapted per configuration: $(1024, 256)$ for AIDE on GenImage, $(256, 64)$ for AIDE on AIGCDetect, and $(32, 1024)$ for PatchCraft. Unless otherwise specified, all reported metrics represent top-1 classification accuracy (%) on the respective unseen target domains.

5 Results

5.1 Standardized Cross-Generator Generalization

To establish the universal applicability of FAIR, we first evaluate its integration into AIDE and PC across all five benchmarks (Tab. 1). Integrating FAIR consistently elevates the mean cross-domain generalization capacity across almost

Table 1: Aggregate Mean Accuracy across all 5 benchmarks. Models for UnivFD, Fake2M, and DRCT-2M were trained strictly on the GenImage SDv1.4 source domain. N = number of unique datasets/domains in each benchmark. **Bold** values indicate an improvement by FAIR over its respective base architecture.

Method	**GenImage** ($N = 8$)	**AIGCDetect** ($N = 17$)	**UnivFD** ($N = 21$)	**Fake2M** ($N = 17$)	**DRCT-2M** ($N = 16$)
PC (Base)	82.36	89.85	83.92	**80.46**	71.39
PC + FAIR	**90.40**	**91.90**	**84.69**	78.82	**74.39**
AIDE (Base)	86.88	**93.02**	77.24	69.12	66.68
AIDE + FAIR	**91.01**	92.14	**79.72**	**76.18**	**68.83**

all datasets. The structural prior successfully regularizes highly varied architectural paradigms, preventing source-domain overfitting regardless of whether the underlying model focuses on global semantics or local micro-textures.

Table 2: Detailed Cross-Generator Generalization on GenImage. Comparison of AIDE and PC (PatchCraft) with and without the proposed FAIR. The best result and the second-best result are marked in **bold** and underline, respectively. ↑ indicates performance improvement yielded by FAIR.

Method	**Midjourney**	**SDv1.4**	**SDv1.5**	**ADM**	**GLIDE**	**Wukong**	**VQDM**	**BigGAN**	**Mean**
ResNet-50 [13]	54.90	**99.90**	99.70	53.50	61.90	98.20	56.60	52.00	72.09
DeiT-S [27]	55.60	**99.90**	99.80	49.80	58.10	98.90	56.90	53.50	71.56
Swin-T [18]	62.10	**99.90**	99.80	49.80	67.60	99.10	62.30	57.60	74.78
CNNSpot [29]	52.80	96.30	95.90	50.10	39.80	78.60	53.40	46.80	64.21
Spec [33]	52.00	99.40	99.20	49.70	49.80	94.80	55.60	49.80	68.79
F3Net [23]	50.10	**99.90**	**99.90**	49.90	50.00	99.90	49.90	49.90	68.69
GramNet [19]	54.20	99.20	99.10	50.30	54.60	98.90	50.80	51.70	69.85
DIRE [30]	60.20	**99.90**	99.80	50.90	55.00	99.20	50.10	50.20	70.66
UnivFD [22]	73.20	84.20	84.00	55.20	76.90	75.60	56.90	80.30	73.29
GenDet [35]	89.60	96.10	96.10	58.00	78.40	92.80	66.50	75.00	81.56
DRCT [4]	**94.63**	99.88	99.82	61.78	65.92	**99.91**	74.88	58.81	82.08
PC [34] (Base)	79.00	89.50	89.30	77.30	78.40	89.30	83.70	72.40	82.36
PC + FAIR	88.93↑	94.80↑	94.86↑	**91.04**↑	90.17↑	92.58↑	**88.08**↑	**82.70**↑	90.40↑
AIDE [31] (Base)	79.38	99.74	99.76	78.54	91.82	98.65	80.26	66.89	86.88
AIDE + FAIR	83.62↑	99.63	99.61	84.05↑	**96.66**↑	99.11↑	87.23↑	78.18↑	**91.01**↑

While FAIR demonstrates stable aggregate performance across legacy GAN-heavy benchmarks like AIGCDetect, its structural regularization is most profoundly impactful on modern Diffusion architectures. To explicitly demonstrate this, we provide detailed, per-domain breakdowns for the two most challenging, diffusion-centric benchmarks: GenImage (Tab. 2) and DRCT-2M (Tab. 3).

As detailed in Tab. 2, FAIR yields a massive 8.04% mean accuracy increase for PatchCraft and a 4.13% increase for AIDE, pushing the AIDE+FAIR model to a new state-of-the-art mean accuracy of 91.01% on the standardized GenImage benchmark.

Table 3: Detailed In-the-Wild Cross-Benchmark Transfer on the DRCT-2M Benchmark. Models are trained on SDv1.4 and tested zero-shot across 16 modern diffusion variations. **Bold** values indicate an improvement by FAIR over its respective base architecture.

Method	*CN-Canny-SDXL*	*LCM-LoRA-SD1.5*	*LCM-LoRA-SDXL*	*LDM-Text2Im*	*SD2.1-CN-Canny*	*SD-CN-Canny*	*SD-Turbo*	*SDXL-Turbo*	*SD-2.1*	*SD-v1.4*	*SD-v1.5*	*SD-Inpainting*	*SD-2-Inpainting*	*SDXL-Refiner*	*SDXL-Base*	*SDXL-Inpainting*	**Mean**
PC (Base)	59.03	62.63	55.53	53.39	65.17	73.87	75.36	80.38	58.00	65.97	65.96	94.45	97.89	63.86	73.02	97.65	71.39
PC + FAIR	**65.50**	**68.01**	**58.01**	**60.19**	**72.22**	**76.06**	**77.99**	79.43	**62.43**	**69.21**	**69.18**	**96.16**	**97.99**	63.35	**76.52**	**98.05**	**74.39**
AIDE (Base)	56.00	74.56	55.25	61.88	55.34	68.82	54.84	54.82	61.74	78.09	77.55	96.40	74.35	72.17	54.61	70.40	66.68
AIDE + FAIR	**56.78**	**76.79**	53.12	**72.03**	54.61	**73.29**	54.05	**55.61**	58.30	**82.49**	**82.29**	**98.70**	**84.23**	67.24	52.74	**78.96**	**68.83**

5.2 In-the-Wild Cross-Benchmark Transfer

Having established the architecture-agnostic nature of FAIR on standardized benchmarks, we further investigate its absolute robustness by subjecting the GenImage (SDv1.4) trained models to extreme cross-dataset transfer scenarios.

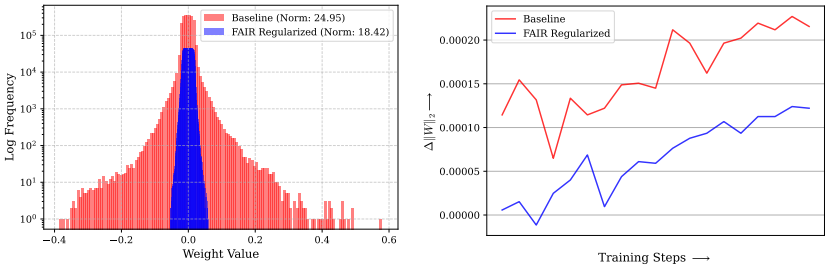
Tab. 3 details the performance on the DRCT-2M dataset, which features 16 highly sophisticated, modern diffusion pipelines (including SDXL, LoRA finetunes, and ControlNet modifications). The FAIR-regularized models consistently outperform their unregularized counterparts across almost all 16 domains. This confirms that the implicit structural constraints imposed by FAIR during source-domain optimization yield decision boundaries that remain remarkably stable, even when exposed to state-of-the-art, adversarially modified generative pipelines. Detailed breakdowns for the remaining test suites (AIGCDetect, UnivFD, Fake2M) are provided in the Supplementary Material.

5.3 The Regularization Mechanism of FAIR

To understand how FAIR achieves this generalized performance, we analyze the physical complexity of the resulting weight spaces and feature dependencies using the AIDE base model and our proposed AIDE+FAIR model, both trained on the GenImage SDv1.4 dataset.

Weight Distribution and Simplicity: Effective regularization manifests as a constraint on weight magnitude. Fig. 5a visualizes the weight distribution of the final classification layer for both the Base and FAIR-regularized models. The baseline model exhibits a heavy-tailed distribution with numerous extreme outlier weights, indicative of a highly non-linear, overfitted decision boundary. In contrast, FAIR compresses this distribution, substantially reducing the norm ($\|W\|_2$ drops from 24.95 to 18.42). This confirms FAIR successfully restricts the hypothesis space, enforcing a smoother, low-complexity decision boundary.

As shown in Fig. 5b, the FAIR-regularized model exhibits a significantly lower and more stable increase in weight magnitude per step ($\Delta\|W\|_2$). Because weight



(a) Weight distributions **(b)** Weight norm update during training

Fig. 5: Regularization Dynamics of FAIR. (a) Weight distributions of the final classification layer demonstrate that FAIR compresses the hypothesis space, eliminating extreme outlier weights. (b) Tracking the change in weight magnitude ($\Delta\|W\|_2$) during training confirms that FAIR restrains parameter growth from the onset, indicating navigation of a smoother loss landscape.

updates are directly proportional to the loss gradient, this bounded trajectory confirms that the structural prior conditions the optimizer to navigate a fundamentally smoother, flatter loss landscape, preventing the aggressive parameter updates associated with memorizing high-frequency textural noise.

Loss Trajectory and Generalization Gap: To explicitly visualize this reduction in source-domain overfitting, we conducted a post-hoc evaluation of the saved training checkpoints on a 5% sample of the target domains to reconstruct the out-of-distribution loss trajectory (Fig. 6). We emphasize that this evaluation is strictly analytical; target data was completely withheld during the actual optimization process.

The unregularized AIDE model exhibits a classic overfitting trajectory: it achieves an extremely low loss on the source domain while suffering from a high, volatile loss on the out-of-distribution target domains. Conversely, the FAIR training trajectory clearly demonstrates the anticipated bias-variance tradeoff. By constraining the optimization landscape, FAIR prevents the model from perfectly fitting the source domain (evidenced by a slightly higher source loss), but this penalty yields a significantly lower and more stable target loss.

Misclassification Confidence and Calibration: A hallmark of shortcut learning is that overfitted models become highly confident in their out-of-distribution errors. To evaluate the calibration of our smoothed decision boundary, we aggregated the prediction confidence (softmax probabilities) for every misclassified sample across the entire GenImage benchmark. As visualized in Fig. 7, the baseline model exhibits a severely top-heavy error distribution, frequently committing to incorrect predictions with extreme, unjustified certainty ($p > 0.9$). Conversely, FAIR compresses this distribution, aggressively suppressing overconfident outliers and shifting the mass toward the uncertainty threshold ($p \approx 0.5$).

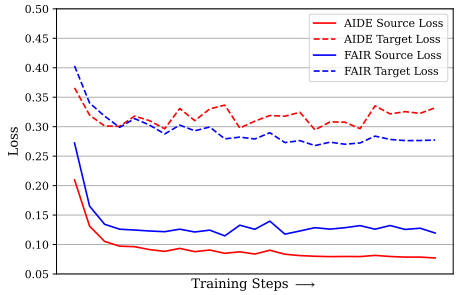


Fig. 6: Source and target domain loss trajectories. FAIR imposes a slight source-domain penalty to stabilize and reduce the out-of-distribution target loss, exemplifying the bias-variance tradeoff. Note: Target loss curves were computed post-hoc by evaluating saved epoch checkpoints on a 5% subset of the target domains. This target data was used strictly for visualization and remained completely unseen by the optimizer during training and hyperparameter selection.

Table 4: Centered Kernel Alignment (CKA) comparison on the SDv1.4 and ProGAN training sets. Standard base features exhibit high statistical dependence on the label (Y), whereas the isolated SCS features yield near-zero dependence.

CKA Score	**SDv1.4 (Training Set)**		**ProGAN (Training Set)**	
Lower is more invariant	**Base (AIDE)**	**SCS (Ours)**	**Base (AIDE)**	**SCS (Ours)**
Dependence on Label (Y)	0.82	**0.07**	0.53	**0.01**

This confirms that even when the smoothed structural boundary misclassifies a target sample, it exhibits a healthy, calibrated uncertainty, avoiding the over-confident failure modes characteristic of models that have memorized localized textural shortcuts.

Quantitative Independence of the Structural Prior: Finally, to mathematically verify that the structural prior (X^*) lacks the signal to act as a trivial discriminative shortcut ($I(X^*;Y) \approx 0$), we evaluated the Centered Kernel Alignment (CKA). Because CKA is a non-parametric measure bounded strictly to $[0, 1]$, it effectively calculates the strength of conditional dependence in high-dimensional spaces. As detailed in Tab. 4, the standard Base (AIDE) features exhibit high statistical dependence on the label. In contrast, our isolated SCS features yield near-zero dependence. This quantitatively confirms the structural prior contains virtually no predictive shortcut, forcing the network to anchor features to invariant geometry rather than bypassing the primary backbone.

5.4 Robustness to Post-Processing: JPEG Compression

In real-world deployment, AI-generated images are frequently subjected to transmission degradations, most notably JPEG compression. Because compression

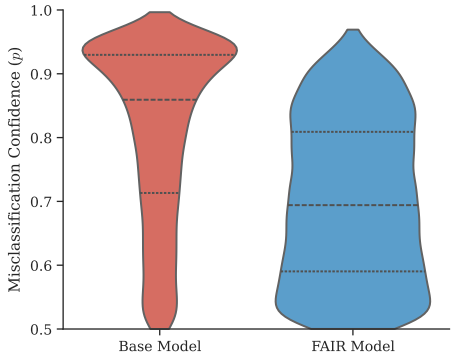


Fig. 7: Global distribution of misclassification confidence: aggregated across GenImage, FAIR suppresses the base model's overconfident errors ($p > 0.9$), yielding a fundamentally softer and better-calibrated decision boundary.

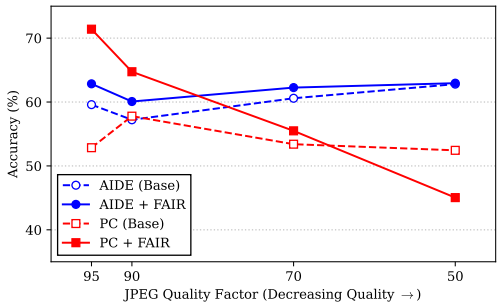


Fig. 8: Robustness to JPEG Compression: accuracy decay curves under varying compression levels show FAIR significantly enhances robustness at mild-to-moderate compression qualities.

algorithms act as low-pass filters that actively destroy the high-frequency microtextures, they provide a rigorous test for detectors that may be over-reliant on localized textural artifacts.

To evaluate resilience, we subjected the GenImage test set to varying levels of JPEG compression. As visualized in Fig. 8, FAIR significantly improves resilience at mild to moderate compression levels (e.g., yielding an +18.55% gain for PatchCraft at $Q = 95$). While AIDE+FAIR maintains superiority across all qualities, PatchCraft's performance drops below the baseline at extreme compression ($Q = 50$). We attribute this to the specific nature of heavy JPEG degradation: at low qualities, compression introduces severe 8x8 blocking artifacts. Because FAIR explicitly regularizes the weights to anchor on structural geometry, the model becomes susceptible to these false artificial edges. Conversely, the unregularized baseline simply matches localized noise and is less directly deceived by global structural distortions. However, the overall trend confirms that

anchoring decisions on macro-structural priors preserves robustness longer than relying strictly on unregularized local artifacts.

Table 5: Comprehensive ablation of the AIDE baseline on GenImage. We compare our macro-structural prior (FAIR with SCS) against traditional parameter-level regularizers (Dropout, L_1, L_2) and alternative feature priors (Random, LPIPS, Sobel, Laplacian).

Domains	Base (AIDE)	Traditional Regularization							Alternative Priors				Proposed
		Drop.	L_1 Sweep			L_2 Sweep			Rand.	LPIPS	Sobel	Lapl.	SCS
			10^{-4}	10^{-3}	10^{-2}	10^{-4}	10^{-3}	10^{-2}					
Midjourney	79.4	76.5	71.6	72.3	50.0	77.0	74.9	71.7	77.0	71.6	73.2	75.1	**83.6**
SDv1.4 (Src)	**99.7**	99.7	99.1	98.8	50.0	99.7	99.6	99.1	99.7	99.3	99.5	99.6	99.6
SDv1.5	**99.8**	99.6	99.1	98.9	50.0	99.7	99.6	99.1	99.7	99.1	99.4	99.5	99.6
ADM	78.5	75.3	77.9	79.0	50.0	74.6	76.3	76.7	74.5	70.6	72.0	72.6	**84.1**
GLIDE	91.8	90.7	86.6	92.0	50.0	92.0	89.7	87.2	90.8	85.9	85.8	88.4	**96.7**
Wukong	98.7	98.4	97.0	96.6	50.0	98.6	98.1	96.9	98.6	96.3	97.1	97.5	**99.1**
VQDM	80.3	81.7	82.0	81.7	50.0	81.6	82.3	80.6	81.9	77.4	79.0	79.1	**87.2**
BigGAN	66.9	70.4	69.1	70.9	50.0	70.6	69.8	68.5	70.7	64.8	65.6	66.7	**78.2**
Mean	86.9	86.5	85.3	86.3	50.0	86.7	86.3	85.0	86.6	83.1	83.9	84.8	**91.0**

5.5 Ablation on Alternative Regularizations and Priors

FAIR vs. Traditional Regularization: We compare FAIR against indiscriminate parameter-level regularizers (Tab. 5). While the standard Dropout baseline (at $p = 0.5$) offers a marginal improvement on radically different domains (e.g., BigGAN), it fails to improve the mean generalization. To verify if standard weight decay could solve this generalization gap, we applied full sweeps of L_1 and L_2 penalties directly to the loss. As the L_2 penalty scales from 10^{-4} to 10^{-2}, mean cross-generator performance actively drops. Similarly, L_1 regularization forces harmful sparsity; at 10^{-2}, it collapses the network entirely (50.00% random chance). This confirms that indiscriminate regularizers only shrink weights or destroy features; they do not explicitly guide the network away from artifact shortcuts, a task uniquely solved by FAIR.

SCS vs. Alternative Feature Priors: The specific choice of the structural prior is also critical. In Tab. 5, we substitute SCS with non-structural features (Random, LPIPS [32]) and classical micro-structural priors (Sobel, Laplacian). Uniform Random features provide no meaningful semantic constraint, while LPIPS features are too low-level and texture-biased, causing the model to overfit and perform worse than the baseline. Furthermore, substituting SCS with pixel-level edge maps (Sobel, 1st-order gradient; Laplacian, 2nd-order derivative) causes noticeable performance drops. Because localized micro-structures are distorted differently by different generative engines, they serve as poor generalization anchors. SCS uniquely succeeds because it provides an orthogonal,

macro-structural bottleneck that remains fundamentally consistent across diverse generators.

Table 6: Ablation of the Base AIDE model demonstrating FAIR improves generalization even when primary feature streams (CLIP-based, F_1 or Patchwise, F_2) are omitted.

Domains	F_1 only		F_2 only		Both	
	Base	FAIR	Base	FAIR	Base	FAIR
Midjourney	84.72	**86.77**	73.19	**77.84**	79.38	**83.62**
SDv1.4	**96.96**	90.08	99.16	**99.19**	**99.74**	99.63
SDv1.5	**96.90**	90.18	99.16	99.16	**99.76**	99.61
ADM	53.76	**62.46**	80.35	**85.11**	78.54	**84.05**
Glide	85.63	**88.46**	87.07	**92.10**	91.82	**96.66**
Wukong	**92.92**	89.82	97.37	**97.93**	98.65	**99.11**
VQDM	59.12	**74.81**	83.73	**86.33**	80.26	**87.23**
BigGAN	77.75	**87.82**	69.63	**74.11**	66.89	**78.18**
Mean	80.97	**83.80**	86.20	**88.97**	86.88	**91.01**

Feature Ablation and Decisiveness: By constraining the weights, FAIR forces the network to utilize its available features more decisively. Tab. 6 shows an ablation study where AIDE's primary features, CLIP semantic (F_1) and Patchwise DCT (F_2), are independently excluded. In every scenario, applying FAIR improves the mean accuracy compared to the respective baseline. This indicates that FAIR successfully guides the model toward a generalized weight distribution, preventing the network from indiscriminately co-adapting to source-specific noise in either the semantic or spectral domains.

6 Conclusion

We introduced Feature-Augmented Implicit Regularization (FAIR), a novel strategy to combat source-domain overfitting and improve cross-generator generalization in AI-generated image detection. By introducing domain-invariant Scene Composition Structure (SCS) features as an orthogonal structural prior (X^*) strictly during training, FAIR geometrically constrains the optimizer to discover a smoother, lower-complexity decision boundary. Crucially, by intentionally discarding this prior at inference, FAIR embeds these structural constraints into the primary textural backbone with zero added architectural or computational overhead. Our method establishes new state-of-the-art robustness not only on standardized benchmarks like GenImage, but across massive, unseen in-the-wild diffusion manifolds, proving the critical advantage of informed geometric priors in representation learning.

References

1. Ahmed, N., Natarajan, T., Rao, K.R.: Discrete Cosine Transform. IEEE Transactions on Computers **C-23**(1), 90–93 (Jan 1974). https://doi.org/10.1109/T-C.1974.223784
2. Arjovsky, M., Bottou, L., Gulrajani, I., Lopez-Paz, D.: Invariant risk minimization (2019), https://arxiv.org/abs/1907.02893
3. Ben-David, S., Blitzer, J., Crammer, K., Kulesza, A., Pereira, F., Vaughan, J.W.: A theory of learning from different domains. Machine Learning **79**(1), 151–175 (May 2010). https://doi.org/10.1007/s10994-009-5152-4
4. Chen, B., Zeng, J., Yang, J., Yang, R.: DRCT: Diffusion reconstruction contrastive training towards universal detection of diffusion generated images. In: Salakhutdinov, R., Kolter, Z., Heller, K., Weller, A., Oliver, N., Scarlett, J., Berkenkamp, F. (eds.) Proceedings of the 41st International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 235, pp. 7621–7639. PMLR (21–27 Jul 2024), https://proceedings.mlr.press/v235/chen24ay.html
5. Cortes, C., Mohri, M., Rostamizadeh, A.: L2 regularization for learning kernels. In: Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence. p. 109–116. UAI '09, AUAI Press, Arlington, Virginia, USA (2009)
6. Corvi, R., Cozzolino, D., Zingarini, G., Siracusa, G., Verdoliva, L.: On the detection of synthetic images generated by diffusion models. In: 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 1–5. IEEE (2023)
7. Frank, J., Eisenhofer, T., Schönherr, L., Fischer, A., Kolossa, D., Holz, T.: Leveraging Frequency Analysis for Deep Fake Image Recognition. In: Proceedings of the 37th International Conference on Machine Learning. pp. 3247–3258. PMLR (Nov 2020)
8. Fridrich, J., Kodovsky, J.: Rich models for steganalysis of digital images. IEEE Transactions on Information Forensics and Security **7**(3), 868–882 (2012). https://doi.org/10.1109/TIFS.2012.2190402
9. Ganin, Y., Ustinova, E., Ajakan, H., Germain, P., Larochelle, H., Laviolette, F., March, M., Lempitsky, V.: Domain-Adversarial Training of Neural Networks. Journal of Machine Learning Research **17**(59), 1–35 (2016), http://jmlr.org/papers/v17/15-239.html
10. Geirhos, R., Jacobsen, J.H., Michaelis, C., Zemel, R., Brendel, W., Bethge, M., Wichmann, F.A.: Shortcut learning in deep neural networks. Nature Machine Intelligence **2**(11), 665–673 (Nov 2020). https://doi.org/10.1038/s42256-020-00257-z
11. Goodfellow, I.J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial nets. In: Ghahramani, Z., Welling, M., Cortes, C., Lawrence, N., Weinberger, K. (eds.) Advances in Neural Information Processing Systems. vol. 27. Curran Associates, Inc. (2014)
12. Haque, M.R., Murshed, M., Paul, M., Lee, T.K.: A novel image similarity metric for scene composition structure. In: 2025 IEEE International Conference on Image Processing Workshops (ICIPW). pp. 446–451 (2025)
13. He, K., Zhang, X., Ren, S., Sun, J.: Deep Residual Learning for Image Recognition. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 770–778 (Jun 2016). https://doi.org/10.1109/CVPR.2016.90
14. Healy, J., McInnes, L.: Uniform manifold approximation and projection. Nature Reviews Methods Primers **4**(1), 82 (Nov 2024). https://doi.org/10.1038/s43586-024-00363-x

15. Hearst, M., Dumais, S., Osuna, E., Platt, J., Scholkopf, B.: Support vector machines. IEEE Intelligent Systems and their Applications **13**(4), 18–28 (1998). https://doi.org/10.1109/5254.708428
16. Hendrycks, D., Gimpel, K.: Gaussian error linear units (GELUs) (2016), https://arxiv.org/abs/1606.08415
17. Ho, J., Jain, A., Abbeel, P.: Denoising Diffusion Probabilistic Models. In: Advances in Neural Information Processing Systems. vol. 33, pp. 6840–6851. Curran Associates, Inc. (2020)
18. Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B.: Swin Transformer: Hierarchical Vision Transformer using Shifted Windows. In: 2021 IEEE/CVF International Conference on Computer Vision (ICCV). pp. 9992–10002. IEEE, Montreal, QC, Canada (Oct 2021). https://doi.org/10.1109/ICCV48922.2021.00986
19. Liu, Z., Qi, X., Torr, P.H.: Global Texture Enhancement for Fake Face Detection in the Wild. In: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 8057–8066. IEEE, Seattle, WA, USA (Jun 2020). https://doi.org/10.1109/CVPR42600.2020.00808
20. Lu, Z., Huang, D., Bai, L., Qu, J., Wu, C., Liu, X., Ouyang, W.: Seeing is not always believing: Benchmarking human and model perception of AI-generated images. In: Advances in Neural Information Processing Systems. vol. 36, pp. 25435–25447 (2023)
21. Midjourney, Inc.: Midjourney. Online (2022), https://www.midjourney.com, Accessed: 2025-09-24
22. Ojha, U., Li, Y., Lee, Y.J.: Towards Universal Fake Image Detectors that Generalize Across Generative Models. In: 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 24480–24489. IEEE, Vancouver, BC, Canada (Jun 2023). https://doi.org/10.1109/CVPR52729.2023.02345
23. Qian, Y., Yin, G., Sheng, L., Chen, Z., Shao, J.: Thinking in Frequency: Face Forgery Detection by Mining Frequency-Aware Clues. In: Vedaldi, A., Bischof, H., Brox, T., Frahm, J.M. (eds.) Computer Vision – ECCV 2020, vol. 12357, pp. 86–103. Springer International Publishing, Cham (2020). https://doi.org/10.1007/978-3-030-58610-2_6
24. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., Sutskever, I.: Learning Transferable Visual Models From Natural Language Supervision. In: Proceedings of the 38th International Conference on Machine Learning. pp. 8748–8763. PMLR (Jul 2021)
25. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-Resolution Image Synthesis with Latent Diffusion Models. In: 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 10674–10685 (Jun 2022). https://doi.org/10.1109/CVPR52688.2022.01042
26. Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., Salakhutdinov, R.: Dropout: A simple way to prevent neural networks from overfitting. Journal of Machine Learning Research **15**(56), 1929–1958 (2014), http://jmlr.org/papers/v15/srivastava14a.html
27. Touvron, H., Cord, M., Douze, M., Massa, F., Sablayrolles, A., Jegou, H.: Training data-efficient image transformers & distillation through attention. In: Proceedings of the 38th International Conference on Machine Learning. pp. 10347–10357. PMLR (Jul 2021)
28. Vapnik, V., Vashist, A.: A new learning paradigm: Learning using privileged information. Neural networks **22**(5-6), 544–557 (2009)

29. Wang, S.Y., Wang, O., Zhang, R., Owens, A., Efros, A.A.: CNN-Generated Images Are Surprisingly Easy to Spot... for Now. In: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 8692–8701. IEEE, Seattle, WA, USA (Jun 2020). https://doi.org/10.1109/CVPR42600.2020.00872
30. Wang, Z., Bao, J., Zhou, W., Wang, W., Hu, H., Chen, H., Li, H.: DIRE for Diffusion-Generated Image Detection. In: 2023 IEEE/CVF International Conference on Computer Vision (ICCV). pp. 22388–22398. IEEE, Paris, France (Oct 2023). https://doi.org/10.1109/ICCV51070.2023.02051
31. Yan, S., Li, O., Cai, J., Hao, Y., Jiang, X., Hu, Y., Xie, W.: A Sanity Check for AI-generated Image Detection. In: International Conference on Learning Representations. vol. 2025, pp. 70702–70720 (2025)
32. Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O.: The Unreasonable Effectiveness of Deep Features as a Perceptual Metric. In: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 586–595 (Jun 2018)
33. Zhang, X., Karaman, S., Chang, S.F.: Detecting and simulating artifacts in GAN fake images. 2019 IEEE International Workshop on Information Forensics and Security (WIFS) pp. 1–6 (2019)
34. Zhong, N., Xu, Y., Li, S., Qian, Z., Zhang, X.: PatchCraft: Exploring Texture Patch for Efficient AI-generated Image Detection. arXiv (Mar 2023). https://doi.org/10.48550/arXiv.2311.12397
35. Zhu, M., Chen, H., Huang, M., Li, W., Hu, H., Hu, J., Wang, Y.: GenDet: Towards Good Generalizations for AI-Generated Image Detection. arXiv (Dec 2023). https://doi.org/10.48550/arXiv.2312.08880
36. Zhu, M., Chen, H., Yan, Q., Huang, X., Lin, G., Li, W., Tu, Z., Hu, H., Hu, J., Wang, Y.: GenImage: A Million-Scale Benchmark for Detecting AI-Generated Image. In: Advances in Neural Information Processing Systems. vol. 36, pp. 77771–77782 (Dec 2023)