PROTO-LEAKNET: TOWARDS SIGNAL-LEAK AWARE ATTRIBUTION IN SYNTHETIC HUMAN FACE IMAGERY

A PREPRINT

Claudio Giusti
Department of Mathematics and Computer Science
University of Catania
Catania, CT 95125
claudio.giusti@studium.unict.it

Luca Guarnera
Department of Mathematics and Computer Science
University of Catania
Catania, CT 95125
luca.guarnera@unict.it

Sebastiano Battiato
Department of Mathematics and Computer Science
University of Catania
Catania, CT 95125
sebastiano.battiato@unict.it

November 7, 2025

ABSTRACT

The growing sophistication of synthetic image and deepfake generation models has turned source attribution and authenticity verification into a critical challenge for modern computer vision systems. Recent studies suggest that diffusion pipelines unintentionally imprint persistent statistical traces, known as signal leaks, within their outputs, particularly in latent representations. Building on this observation, we propose Proto-LeakNet, a signal-leak-aware and interpretable attribution framework that integrates closed-set classification with a density-based open-set evaluation on the learned embeddings, enabling analysis of unseen generators without retraining. Operating in the latent domain of diffusion models, our method re-simulates partial forward diffusion to expose residual generator-specific cues. A temporal attention encoder aggregates multi-step latent features, while a feature-weighted prototype head structures the embedding space and enables transparent attribution. Trained solely on closed data and achieving a Macro AUC of 98.13%, Proto-LeakNet learns a latent geometry that remains robust under post-processing, surpassing state-of-the-art methods, and achieves strong separability between known and unseen generators. These results demonstrate that modeling signal-leak bias in latent space enables reliable and interpretable AI-image and deepfake forensics. The code for the whole work will be available upon submission.

1 Introduction

The rapid progress of generative models has transformed digital content creation, enabling the synthesis of highly realistic images and videos, also called deepfakes, that are often indistinguishable from authentic ones [1, 2, 3]. New biases such as 'impostor bias' [4] further complicate deepfake detection. While these advances have fostered creativity and accessibility, they have also blurred the boundary between real and artificial content, posing serious challenges to media forensics and public trust. As deepfakes proliferate across social, political, and creative domains, distinguishing and attributing their origin has become critical for security, accountability, and digital evidence validation [5, 6]. Early research in multimedia forensics focused primarily on classifying whether an image is real or generated [7, 8]. However, a deeper forensic question lies in attribution, namely identifying which generative model produced a given image.

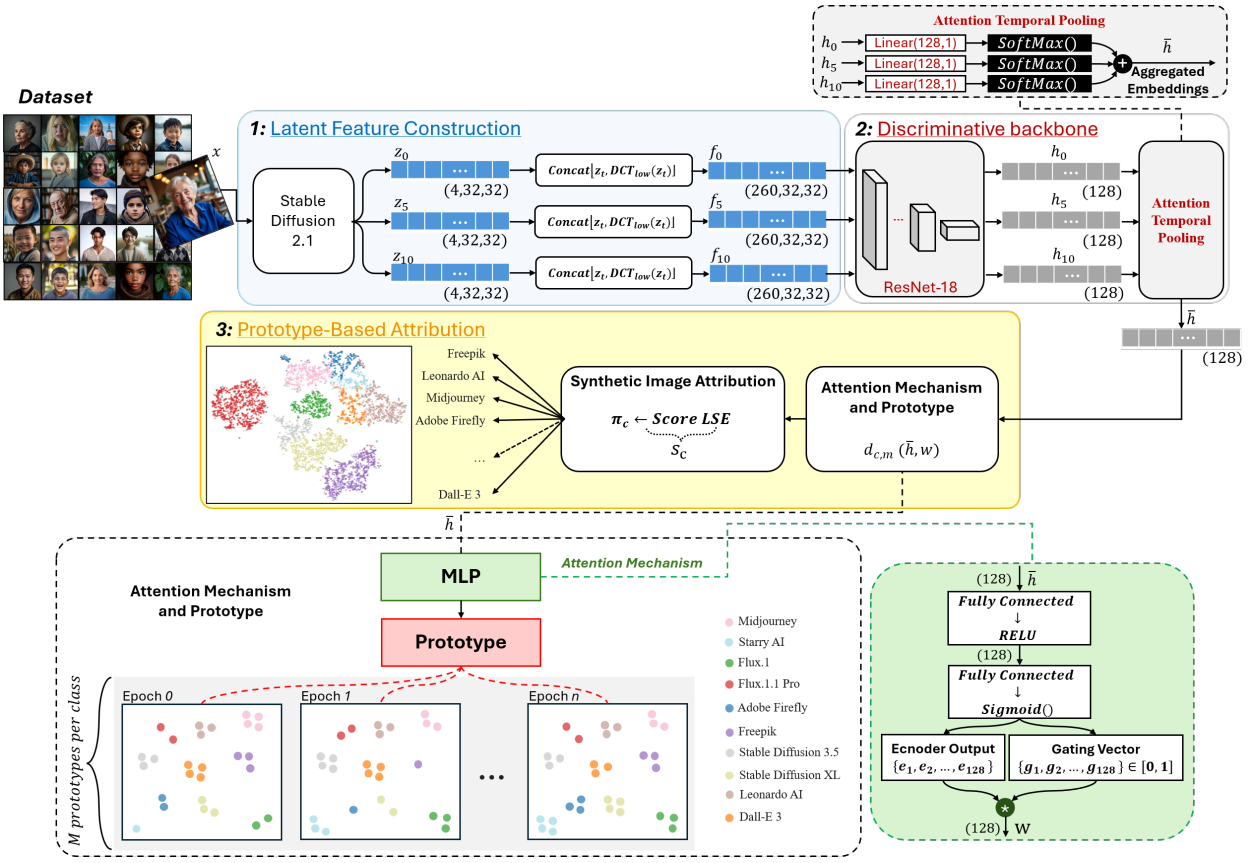

Figure 1: Proto-LeakNet: given an input image x, latent features are extracted from the pretrained Stable Diffusion 2.1 VAE in **Block 1 (Latent Feature Construction)**. For each diffusion step $t \in \{0, 5, 10\}$, the latent z_t is concatenated with its low-frequency *DCT* map $\mathrm{DCT}_{\mathrm{low}}(z_t)$ to form f_t, preserving generator-specific structural cues. In **Block 2 (Discriminative Backbone)**, each f_t is encoded by a ResNet-18, producing embeddings $\{h_t\}$ that are temporally aggregated through the **Attention Temporal Pooling** module to yield a single representation $\bar{h}$. **Block 3 (Prototype-Based Attribution)** computes distances between $\bar{h}$ and class prototypes $p_{c,m}$, modulated by a feature-wise gating vector w obtained from a small MLP. The resulting attention-weighted distances are aggregated via a *LogSumExp* scoring function to produce class probabilities π_c. Symbols: "**+**" denotes the weighted sum over attention coefficients across timesteps, and "*****" indicates the element-wise product between the encoder output and the gating vector.

This task, essential for tracing provenance and assessing responsibility, remains extremely challenging, especially in open-set conditions where unknown generators appear at test time [9, 10]. Recent studies have shown that even advanced detectors struggle to generalize beyond the closed domain or to maintain interpretability when facing unseen architectures [11, 12]. Diffusion models have recently equaled most generative models in terms of image quality, yet they introduce subtle statistical artifacts in their latent representations, known as *signal leaks*, caused by residual low-frequency information that survives the denoising process [13]. These traces, although imperceptible, encode model-specific biases and can serve as reliable forensic cues for source attribution. While prior works have improved detection accuracy, they typically lack robustness to domain shifts and offer limited interpretability, especially under open-set or heavily post-processed conditions. To overcome these issues, we aim to exploit the intrinsic statistical biases embedded in diffusion latents as stable, model-specific signatures. In detail, Proto-LeakNet encodes image x through Stable Diffusion latents and temporal attention, aggregates timestep embeddings via ResNet-18, and performs attribution using prototype-based distances modulated by per-feature attention and gating. This full pipeline can be observed in Figure 1. Our main contributions are the following:

- We introduce **Proto-LeakNet**, a signal-leak-aware and interpretable attribution framework that operates entirely in the latent domain of diffusion models, learning generator-specific biases as stable forensic cues.

- We design a **temporal attention pooling mechanism** that aggregates latent representations across diffusion timesteps, enhancing discriminative power and interpretability by revealing which steps contribute most to attribution.
- We propose a **prototype-based attribution head** that shapes the latent geometry through learnable class prototypes and per-feature attention, enabling both compact cluster formation and feature-level interpretability.
- We develop a **density-based open-set evaluation** using kernel density estimation on the learned embeddings to assess separability between known and unseen generators without retraining.
- We demonstrate that modeling **signal-leak bias in latent space** leads to robust attribution under heavy post-processing and provides transparent prototype-level explanations.

This work is organized as follows: Section 2 presents the current state-of-the-art and the works we will be comparing ourselves against. In Section 3 we explain in detail our proposed methodology. Section 4 focuses on explaining the employed dataset and all its specifications. Section 5 presents the results of our experiments and the various ablations we performed. In Section 6 we discuss the results obtained and the limitation of our framework. Finally, Section 7 we evaluate our work and propose future upgrades to enhance the proposed pipeline, establishing where it places itself among the literature.

2 Related Work

Research on synthetic image detection and attribution has rapidly evolved with the growing realism of generative models. Early detectors focused on binary real-vs-fake discrimination through spatial or frequency inconsistencies. Some early works have established strong baselines that served as foundations for the topic of deepfake detection and attribution. For example Rössler et al. [14] introduced FaceForensics++, an in-depth benchmark for deepfake detection with various image forgery tools. While Nguyen et al. [15] proposes Capsule-Forensics, a method that uses a capsule network to detect various kinds of image modifications and forgeries. Recent works have brought deepfake detection and attribution to the video domain, like StyleFlow (Choi et al. [16]) which leverages temporal dynamics of style latent vectors for video deepfake detection. Works like FreqNet (Tan et al. [17]) introduced frequency-space learning within CNNs to extract source-agnostic cues from amplitude and phase spectra, while SuSy (Bernabeu-Pérez et al. [18]) provided a large-scale benchmark highlighting the fragility and dataset bias of current detectors under real-world deployment. Recent efforts have shifted toward explicit source attribution. LatentTracer (Wang et al. [19]) performs inversion-based tracing of latent generative models without artificial watermarks, whereas OCC-CLIP (Liu et al. [20]) reformulates attribution as few-shot one-class classification using CLIP embeddings, enabling model recognition even with limited samples. Alternative approaches aim to generalize by modeling intrinsic generation artifacts: NPR (Tan et al. [21]) captures local pixel correlations induced by up-sampling operations. More recent works have started to employ diffusion latents for synthetic images detection like LATTE [22] which introduces a transformer-based architecture operating directly in the latent space of diffusion models to detect generated content. By modeling long-range dependencies among latent tokens, it effectively captures generator-specific artifacts. Finally, Everaert et al. [13] identified a signal-leak bias in diffusion latents low-frequency mismatches that encode generator-specific traces. Building on this insight, our work introduces **Proto-LeakNet**, a framework that explicitly models and exploits these latent residuals for robust and interpretable generator attribution under both closed- and open-set conditions. Unlike LATTE, which focuses on discriminative detection through self-attention among latent tokens, **Proto-LeakNet** explicitly structures the latent space through prototype supervision and temporal attention across multiple diffusion steps. This design aggregates generator-specific cues over time, enabling interpretable attribution rather than binary detection. While LATTE captures static latent correlations, our approach leverages the temporal evolution of signal-leak residuals to model generator-specific biases with higher forensic transparency and open-set separability.

3 Proposed Method

This section presents the architecture of the proposed Proto-LeakNet framework, divided into three functional blocks. The first, Latent Feature Construction, extracts informative representations from the diffusion latents of Stable Diffusion 2.1. The second, Discriminative Backbone, aggregates these multi-scale features through a temporal attention encoder. Finally, the Prototype-Based Attribution Head interprets the learned representation using class-specific prototypes for interpretable generator attribution.

3.1 Latent Feature Construction

We extract features directly from the latent domain of Stable Diffusion 2.1 (SD2.1) [3]. Each image $x \in \mathbb{R}^{3\times H\times W}$ is resized to 256×256 and encoded through the pretrained SD2.1's Variational Autoencoder (VAE) into a latent z_0 scaled by the constant $s = 0.18215$, as in

$$z_0 = \frac{\mathcal{E}(x)}{s}, \tag{1}$$

where $\mathcal{E}(\cdot)$ is the VAE encoder. To reveal residual diffusion traces, we reapply the forward diffusion process at discrete steps $t \in \mathcal{T} = \{0, 5, 10\}$, sampling

$$z_t = \alpha_t z_0 + \sigma_t\, \varepsilon, \qquad \varepsilon \sim \mathcal{N}(0, I), \tag{2}$$

where K is the number of diffusion steps and (α_t, σ_t) follow the cosine schedule with

$$\alpha_t^2 + \sigma_t^2 = 1. \tag{3}$$

Each z_t is normalized by σ_t to maintain consistent scale, producing $\{z_t/\sigma_t\}_{t\in\mathcal{T}}$. To preserve low-frequency structural information, each latent z_t is augmented with a low-frequency Discrete Cosine Transform (DCT) map $\mathrm{DCT}_{\text{low}}(z_t)$ containing the first 8×8 coefficients per channel. The resulting feature tensor is defined as

$$f_t = \mathrm{concat}[z_t,\, \mathrm{DCT}_{\text{low}}(z_t)]\,, \tag{4}$$

where each f_t is standardized per channel using statistics computed from the training set. These tensors $\{f_t\}$ constitute the multi-scale latent features provided to the temporal attention module.

3.2 Discriminative Backbone

The feature tensors $\{f_t\}_{t\in\mathcal{T}}$ are encoded by a ResNet18 [23] backbone $\phi(\cdot;\theta)$, with the first convolution adapted to match the channel dimensionality of f_t. For each timestep t, the backbone produces a latent embedding

$$h_t = \phi(f_t;\, \theta) \in \mathbb{R}^D, \tag{5}$$

where θ denotes the trainable backbone parameters and D is the embedding dimensionality. A learnable attention module assigns relevance to each timestep via

$$a_t = \frac{\exp(q^\top u_t)}{\sum_{t'\in\mathcal{T}} \exp(q^\top u_{t'})}, \qquad u_t = W_a h_t + b_a, \tag{6}$$

where q, W_a, and b_a are learned parameters. The temporally aggregated embedding is obtained as

$$\bar{h} = \sum_{t\in\mathcal{T}} a_t\, h_t, \tag{7}$$

where $\sum_t a_t = 1$. The weights $\{a_t\}$ provide temporal interpretability by quantifying the contribution of each diffusion step to the final embedding $\bar{h}$.

3.3 Prototype-Based Attribution Head

Each class $c\in\{1,\ldots,C\}$ is represented by M learnable prototypes $p_{c,m}\in\mathbb{R}^D$, which serve as representative points in latent space. Empirically, four prototypes yielded compact yet well-separated latent clusters. An attention feature-wise gating vector $w\in(0,1)^D$ is computed using a small MLP:

$$w = \sigma(A\bar{h} + b), \tag{8}$$

where A and b are learnable parameters. The attention-weighted distance between $\bar{h}$ and each prototype $p_{c,m}$ is defined as

$$d_{c,m}(\bar{h}, w) = \sum_{i=1}^{D} w_i\, (\bar{h}_i - p_{c,m,i})^2. \tag{9}$$

Per-class scores aggregate distances using a temperature-controlled LogSumExp:

$$s_c = -\tau \log \sum_{m=1}^{M} \exp\Big(-\tfrac{1}{\tau} d_{c,m}(\bar{h}, w)\Big), \tag{10}$$

where $\tau > 0$ determines the aggregation smoothness. Posterior probabilities are computed as

$$\pi_c = \frac{\exp(s_c)}{\sum_{c'} \exp(s_{c'})}, \tag{11}$$

where c' denotes the index iterating over all classes in the denominator. The model is trained via cross-entropy loss

$$\mathcal{L}_{\text{CE}} = -\frac{1}{B} \sum_{b=1}^{B} \log \pi_{y^{(b)}}, \tag{12}$$

where B is the mini-batch size and $y^{(b)}$ are ground-truth labels. All parameters $\{\theta, A, b, p_{c,m}\}$ are optimized jointly with Adam and weight decay. The main training pipeline can be visualized in full detail in Figure 1.

Mahalanobis Scoring and Interpretability

During evaluation, embeddings are scored using a diagonal Mahalanobis comparator fitted on training embeddings. For each class c, with empirical mean μ_c and diagonal covariance $\Sigma_c = \text{diag}(\sigma_{c,1}^2, \ldots, \sigma_{c,D}^2)$, the score is

$$s_c^{\text{maha}}(\bar{h}) = -\sum_{i=1}^{D} \frac{(\bar{h}_i - \mu_{c,i})^2}{\sigma_{c,i}^2 + \epsilon}, \tag{13}$$

where $\sigma_{c,i}$ represents the empirical standard deviation of feature i within class c, capturing the intra-class variance along each embedding dimension and ϵ prevents numerical instability. These scores provide calibrated likelihoods for attribution and open-set evaluation.

Interpretability of Proto-LeakNet

Proto-LeakNet is inherently interpretable. Each distance in Eq. 9 decomposes feature-wise as

$$d_{c,m}(\bar{h}, w) = \sum_{i=1}^{D} \underbrace{w_i(\bar{h}_i - p_{c,m,i})^2}_{r_{c,m,i}}, \tag{14}$$

where $r_{c,m,i}$ measures the contribution of feature i. Prototype responsibilities are obtained as

$$\pi_{c,m} = \frac{\exp(-d_{c,m}/\tau)}{\sum_{m'} \exp(-d_{c,m'}/\tau)}. \tag{15}$$

The most activated prototype $\arg\max_m \pi_{c,m}$ identifies the latent region that best matches the input. Together, the feature gates w, temporal weights $\{a_t\}$, and prototype responsibilities $\{\pi_{c,m}\}$ provide a three-level interpretability hierarchy visible in Figure 2 , describing which prototype, which features, and which diffusion steps drive each attribution decision from layer, as shown in Figure 3.

3.4 Representation-Level Generalization via Density Estimation

While Proto-LeakNet is trained only on closed-set generators, its latent encoder learns a structured representation that can be evaluated for generalization without requiring explicit supervision on unseen classes. In this setting, the goal is not to classify open samples, but to assess whether the learned latent geometry consistently separates embeddings of known generators from those of unseen ones. After training, we discard the prototype-based classifier and use only the frozen ResNet18 backbone-based encoder to produce pooled embeddings $\bar{h} \in \mathbb{R}^D$ for both closed and open samples, forming the sets $\mathcal{H}_{\text{c}} = \{\bar{h}_i^{(\text{c})}\}_{i=1}^{N_{\text{c}}}$ and $\mathcal{H}_{\text{o}} = \{\bar{h}_j^{(\text{o})}\}_{j=1}^{N_{\text{o}}}$. A Gaussian kernel density estimator (KDE) is fitted on $\mathcal{H}_{\text{c}}$ to model the manifold of closed embeddings:

$$p_{\text{KDE}}(h) = \frac{1}{N_{\text{c}}(2\pi\sigma^2)^{\frac{D}{2}}} \sum_{i=1}^{N_{\text{c}}} \exp\left(-\frac{\|h - \bar{h}_i^{(\text{c})}\|^2}{2\sigma^2}\right), \tag{16}$$

where σ is the kernel bandwidth and D is the embedding dimensionality. For each sample h, we compute its log-likelihood score

$$s(h) = \log p_{\text{KDE}}(h), \tag{17}$$

which measures how likely h lies within the distribution of known generators. High $s(h)$ values correspond to familiar latent regions, while low scores indicate that the sample is far from any known manifold, suggesting an unseen generator.

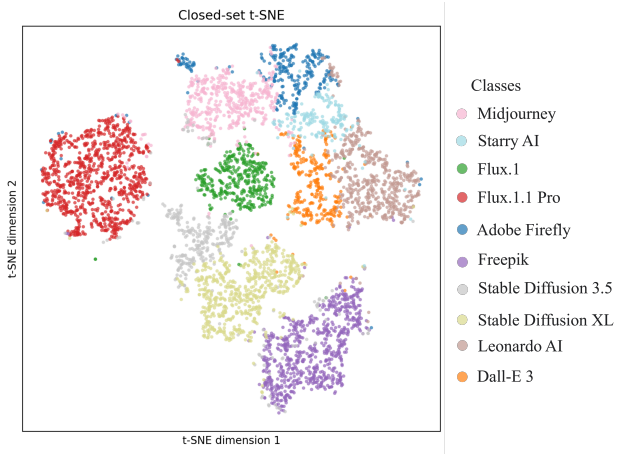

Figure 2: Plot of the final clusters shaped by the prototypes and attention mechanism.

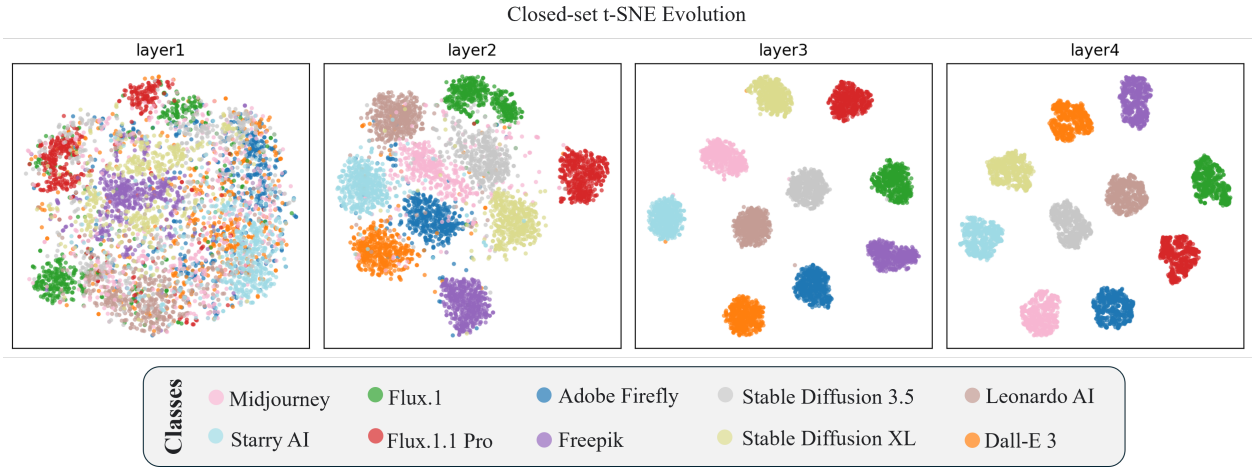

Figure 3: Plot showing the evolution from layer to layer of the clusters, displaying how separation is emphasized due to prototypes.

This approach defines a form of representation-level generalization: the model is never trained on open samples or labels, yet its latent space forms structured low-density regions that naturally reject out-of-distribution inputs. Rather than predicting unseen categories, the KDE analysis evaluates whether the learned representation preserves discriminative geometry under domain shifts. Open generators, though unlabeled, consistently occupy regions separated from the closed manifold, showing that the encoder captures signal-leak statistics that generalize beyond training sources and enable unsupervised detection of unseen generators.

3.5 Evaluation Metrics

We report two complementary metrics to evaluate both closed- and open-set performance: Macro AUC, Equal Error Rate (EER) and Overlap Coefficient. These jointly measure ranking consistency and separability of confidence scores.

- **Macro AUC:** For closed- and open-set scoring, we compute the per-class area under the ROC curve (AUC) (Eq. 18) and average over all C classes:

$$\text{MacroAUC} = \frac{1}{C}\sum_{c=1}^{C}\text{AUC}_c, \tag{18}$$

  where AUC_c measures the ranking quality for class c. This metric evaluates discriminative consistency across classes, independent of the decision threshold.

- **Equal Error Rate (EER):** In open-set evaluation, EER is the point at which the false-acceptance rate (FAR) equals the false-rejection rate (FRR), defined in Eq. 19:

$$\text{EER} = \min_{\tau} \left| \text{FAR}(\tau) - \text{FRR}(\tau) \right|, \tag{19}$$

  where τ is the decision threshold. Lower EER indicates better separation between known and unknown samples.

- **Overlap Coefficient (OVL):** As shown in Eq. 20, quantifies the empirical intersection between the KDE score distributions of closed and open samples:

$$\text{OVL} = \int \min \left(P_{\text{closed}}(s),\, P_{\text{open}}(s) \right) ds, \tag{20}$$

  where P_{closed} and P_{open} are normalized density estimates of the log-scores s. Values close to zero indicate non-overlapping, perfectly separable distributions.

4 Dataset

All experiments are conducted on the WILD dataset [24], a large-scale benchmark for synthetic image attribution explicitly designed for closed- and open-set evaluation. It contains 20,000 high-quality head-and-shoulder portraits generated by twenty distinct text-to-image models, including diffusion- and GAN-based systems such as Stable Diffusion, Midjourney, DALL-E, and StyleGAN. No image, prompt, or generator overlaps across splits, ensuring complete separation and preventing any data leakage. This dataset was selected for its diversity of generated samples, its unbiased distribution across generative models, and its inclusion of a rigorously designed post-processed test set for robustness evaluation.

4.1 Closed-Set

The closed set comprises 10,000 images from ten generators, each contributing 1,000 samples created from unique randomized prompts controlling facial and contextual attributes. Data are split into 5,000 training, 2,000 validation, and 3,000 test samples, partitioned strictly at the prompt level to avoid cross-split repetition. Examples of raw images are shown in Figure 4.

4.2 Post-Processed Closed-Set

To evaluate robustness, half of the closed-set samples undergo up to three random transformations (compression, geometric, or photometric adjustments). Parameters are stochastically sampled to prevent duplicate augmentations. These post-processed images are used exclusively at inference, providing a controlled robustness benchmark against real-world degradation. An example of these post-processed samples can be visualized in Figure 4.

4.3 Open-Set

The open set includes 10,000 additional portraits generated by ten unseen models, fully disjoint from those used for training. This split evaluates the model's ability to generalize across architectures and prompt domains, measuring attribution consistency on previously unseen generators.

5 Experimental Results

This section presents the experiments conducted to evaluate the effectiveness of the proposed Proto-LeakNet framework in Section 3 across different scenarios. We compare our method against state-of-the-art approaches introduced in Section 2 using the same dataset and experimental conditions. Finally, we include ablation studies to assess the contribution of each component in our pipeline.

5.1 Results on Closed Set

We first evaluated Proto-LeakNet under the closed-set configuration to verify its ability to learn and capture generator-specific signal-leak patterns. This experiment establishes whether the model forms compact, well-separated latent clusters when all generator classes are known during training. Among all other methods we additionally reported

Figure 4: Samples from the WILD dataset, for each image in the first row, its column represent its various steps of post processing.

the best three models reported on the WILD dataset [24] all trained in our exact conditions. As shown in the first block of Table 1, Proto-LeakNet achieves the highest Macro AUC among all compared methods on the raw closed-set, confirming its superior latent discriminability. In the first block of Table 1 we can closely inspect the single AUCs for each class to observe how our model model achieves superior performances overall. The combination of temporal attention and prototype supervision proves essential for this structured separation. Attention emphasizes the most informative diffusion timesteps, while prototypes enforce geometric consistency by pulling samples toward class centroids and repelling those of different generators. To further support these findings, Figure 5 reports the distributions of the Top-1 Accuracy on the raw Closed Set showing that our models still achieves a competitive accuracy compared to other methods. Together, these mechanisms yield a compact and interpretable latent manifold, as illustrated in Figure 2, demonstrating that Proto-LeakNet effectively captures intrinsic signal-leak biases that underpin robust generator attribution.

5.2 Results on Post-Processed Closed Set

To evaluate robustness against real-world degradations, we tested all models on progressively post-processed closed-set samples, applying up to three levels of perturbations described in Section 4. The experiment used pretrained checkpoints from the models trained on raw closed-set configuration described before. During this experiment, the models are trained exclusively on the raw closed-set samples, while post-processed images from Steps 1 to 3 are introduced only at inference time, without any additional fine-tuning. As shown in the second, third and fourth block in Table 1, all models performance decay due to the higher number of perturbance introduced but Proto-LeakNet consistently maintains the highest Macro AUC across all perturbation levels, confirming its resilience to visual degradation. This robustness arises from the model's reliance on latent-domain signal-leak cues rather than pixel-level textures. Since signal-leak originates from generator-specific residuals embedded in the latent representation, these cues persist even after post-processing operations that distort pixel information. In contrast, competing methods operating in the image domain exhibit progressive performance collapse as perturbations increase, indicating their dependence on surface-level artifacts. Table 1 allows us to closely inspect single AUCs per class to assess proper attribution and in all three steps our proposed method achieves best performances. To support the results obtained in this experiment, Figure 5 reports the distribution of Top-1 Accuracy throughout all post processing steps, where Proto-LeakNet visibly achieves the highest accuracy proving to be the most robust to post-process attacks. Overall, these results demonstrate that Proto-LeakNet effectively captures stable generator-specific biases at the latent level, preserving attribution accuracy even under aggressive degradations that typically erode pixel-based forensic evidence.

5.3 Representation-Level Generalization Analysis

To assess how well the learned latent space separates known and unseen generators, we evaluate the frozen embeddings of Proto-LeakNet's ResNet18 encoder rather than its classifier head. This isolates the representational geometry induced by signal-leak cues, independent of class supervision. A kernel density estimator, introduced in Section 3, is fitted on the closed embeddings and used to measure how unseen generators align with, or deviate from, the learned manifold under three attention configurations. As shown in Table 2 and visualized in Figure 6, symmetric configurations (Both Off and Both On) yield overlapping score distributions between closed and open samples. Without attention, the encoder loses its ability to emphasize generator-specific latent dimensions, resulting in mixed distributions. When attention is applied to both domains, the same weighting pattern is projected onto unseen data, falsely aligning them with the

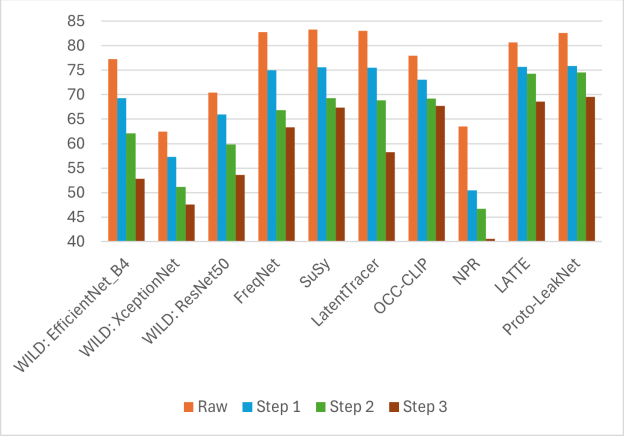

Figure 5: Histogram of the Top-1 accuracies distributions of each method per class portraying how each class accuracy behaves with more degrees of degradation to the images.

Table 1: AUC results (%) for individual class per method including Macro AUC starting from results on raw closed-set until individual results for each post-processed experiment from step 1 to 3.

		Closed-set Classes										
Type	**Methodologies**	**Adobe Firefly**	**Dall-E 3**	**Flux.1**	**Flux.1.1 Pro**	**Freepik**	**Leonardo AI**	**Midjourney**	**Stable Diffusion 3.5**	**Stable Diffusion XL**	**Starry AI**	**Macro AUC**
RAW	WILD: EfficientNet_B4 [24]	95.90	**96.53**	96.59	98.80	**99.88**	97.28	94.83	95.55	**99.99**	95.69	97.10
	WILD: XceptionNet [24]	94.58	94.90	96.94	98.81	99.16	94.04	91.41	91.85	99.90	91.60	95.32
	WILD: ResNet50 [24]	94.28	90.79	97.20	97.05	99.34	98.05	89.02	96.53	97.57	94.13	95.40
	FreqNet [17]	96.82	95.46	99.39	99.58	99.43	97.87	95.44	97.11	97.90	96.60	97.56
	SuSy [18]	**98.10**	96.06	99.80	99.74	99.63	98.56	96.04	97.68	97.27	97.32	98.02
	LatentTracer [19]	97.29	95.92	99.87	99.80	99.71	98.57	95.66	97.50	97.18	97.30	97.88
	OCC-CLIP [20]	94.39	93.12	98.54	97.42	98.25	96.21	93.45	94.60	95.68	95.34	95.70
	NPR [21]	88.34	86.12	94.10	93.54	93.82	90.53	86.05	87.91	89.02	88.71	89.81
	LATTE [22]	96.39	94.98	99.05	98.96	98.83	97.70	94.95	96.88	97.35	96.48	97.16
	Proto-LeakNet	97.32	96.17	**99.91**	**99.85**	99.74	**98.67**	**96.15**	**97.79**	98.38	**97.35**	**98.13**
Step 1	WILD: EfficientNet_B4 [24]	**96.46**	**94.43**	94.97	92.43	97.63	94.05	**94.49**	94.84	89.84	94.83	94.40
	WILD: XceptionNet [24]	91.79	94.30	96.01	97.62	98.12	93.06	90.98	88.92	89.10	93.71	93.36
	WILD: ResNet50 [24]	93.19	91.62	97.63	93.81	**98.35**	96.51	87.66	94.93	**98.01**	92.09	94.38
	FreqNet [17]	94.98	93.29	97.90	97.78	97.62	**96.74**	93.15	94.62	95.40	94.62	95.61
	SuSy [18]	95.17	93.61	98.12	97.98	97.84	96.27	93.57	94.95	95.71	**95.78**	95.90
	LatentTracer [19]	95.33	93.87	98.29	98.15	97.99	96.44	93.84	**95.21**	95.98	95.36	96.05
	OCC-CLIP [20]	93.18	91.21	96.74	96.58	96.39	94.36	91.17	92.68	93.79	92.96	93.91
	NPR [21]	79.73	74.55	87.15	86.66	87.34	79.12	75.43	76.00	80.76	78.05	80.48
	LATTE [22]	95.21	93.79	98.16	98.19	98.02	96.43	93.76	95.14	95.81	95.43	95.99
	Proto-LeakNet	95.51	93.72	**98.43**	**98.31**	98.27	96.54	93.88	95.15	95.93	95.69	**96.14**
Step 2	WILD: EfficientNet_B4 [24]	92.58	90.53	92.59	91.88	96.89	93.49	**92.04**	92.62	84.97	90.60	91.82
	WILD: XceptionNet [24]	92.90	**92.94**	94.66	**97.33**	**96.91**	93.21	91.12	87.41	89.54	92.26	92.83
	WILD: ResNet50 [24]	89.61	90.96	96.17	90.28	95.23	**97.55**	86.80	**94.13**	**96.59**	89.14	92.65
	FreqNet [17]	88.74	86.58	94.53	94.70	94.46	91.07	85.52	88.61	89.89	88.82	90.29
	SuSy [18]	92.13	89.74	96.19	96.03	95.86	93.01	89.54	91.20	92.37	91.63	92.77
	LatentTracer [19]	91.89	89.49	95.08	95.79	95.60	93.11	89.43	91.26	92.09	91.57	92.53
	OCC-CLIP [20]	92.03	89.62	96.12	95.91	95.49	93.30	89.57	91.17	92.28	91.21	92.67
	NPR [21]	75.71	70.39	83.37	82.89	83.64	75.22	69.34	72.15	75.11	77.01	76.48
	LATTE [22]	**95.73**	91.82	96.99	96.92	96.87	94.79	91.71	93.13	94.03	92.09	94.41
	Proto-LeakNet	93.95	91.98	**97.17**	97.01	96.86	95.46	91.96	93.46	94.53	**93.80**	**94.62**
Step 3	WILD: EfficientNet_B4 [24]	90.07	88.12	89.65	88.88	92.54	89.03	86.97	89.89	71.86	86.40	87.34
	WILD: XceptionNet [24]	89.57	**90.88**	93.84	**97.05**	94.73	**93.66**	85.69	86.06	83.95	**92.24**	90.77
	WILD: ResNet50 [24]	88.40	90.46	91.92	89.79	94.00	92.45	82.48	89.03	**95.10**	86.59	90.02
	FreqNet [17]	88.72	85.51	93.91	93.74	93.52	90.06	84.81	87.55	88.19	87.38	89.34
	SuSy [18]	89.82	86.68	94.95	94.79	94.58	91.24	86.01	87.13	89.99	88.83	90.40
	LatentTracer [19]	87.43	83.57	93.18	92.85	92.51	89.24	83.49	81.27	88.73	88.63	88.09
	OCC-CLIP [20]	91.37	88.79	95.48	95.23	94.75	92.58	**90.54**	**91.41**	91.13	88.63	91.99
	NPR [21]	71.77	64.29	78.23	77.34	78.71	70.51	65.38	66.42	69.94	68.91	71.15
	LATTE [22]	90.32	87.68	94.45	94.20	94.01	91.52	87.61	89.50	90.36	89.95	90.96
	Proto-LeakNet	**91.62**	89.24	**95.63**	95.46	**95.29**	92.84	89.18	90.35	91.61	91.30	**92.25**

closed distribution and reducing separability. The asymmetric Dual Attention configuration, where attention is active only for closed embeddings, preserves generator-specific latent biases and produces distinct, non-overlapping densities indicative of true representation-level generalization. All sanity checks, including label shuffling, seed variation, bandwidth cross-validation, and per-channel normalization, confirmed that the observed separation originates from genuine distributional differences rather than data leakage or training artifacts.

Table 2: Open-set evaluation under different attention configurations. We report AUROC, Equal Error Rate (EER), and Overlap Coefficient (OVL). Lower EER and OVL indicate better separation between closed and open domains.

Configuration	AUC (%)	EER	OVL
Both Off	57.24	0.44	0.89
Both On	56.62	0.45	0.90
Dual	**100.00**	**0.00**	**0.00**

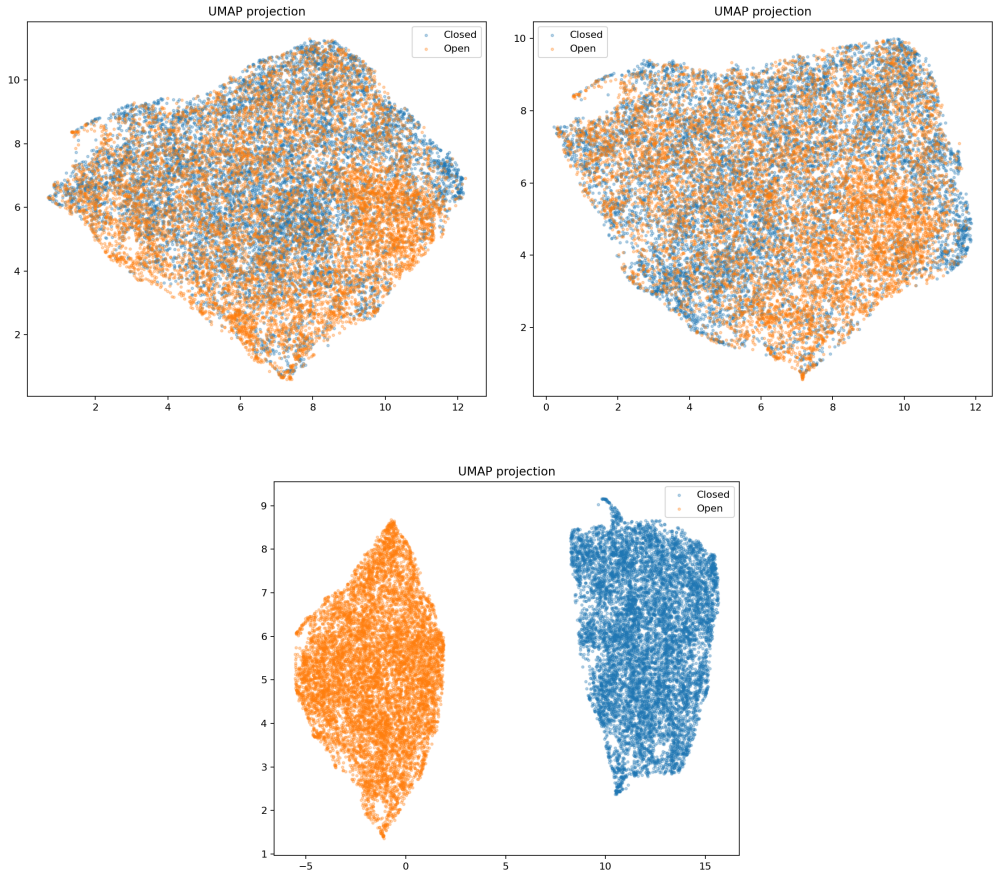

Figure 6: Plots of both off, both on and dual settings clusters, showing how separation differs from setting to setting.

5.4 Ablation Study

To assess the contribution of each key component in Proto-LeakNet, we performed three ablation experiments targeting the modules that define the distinct behavior of the pipeline. First, we analyzed the impact of prototypes and feature-attention on latent-space organization and discriminability. As reported in Table 3, removing prototypes substantially degrades the latent structure, reducing both Top-1 Accuracy and Macro AUC, while removing attention diminishes the model's ability to emphasize the most informative latent dimensions. The combination of both components yields the best overall results, confirming that prototypes and attention jointly enforce structured and discriminative embeddings that better capture signal-leak patterns. We then evaluated the influence of the backbone architecture (Table 4). Replacing the ResNet18 encoder with larger models such as EfficientNet-B4, ViT-B16. ResNet50 or ResNet101 results in lower accuracy and AUC, suggesting that overly deep or transformer-based architectures may overfit local latent

variations and weaken generalization. ResNet18 achieves the best balance between representational compactness and discriminative power, providing the most stable feature extraction across post-processed samples. Finally, we investigated the effect of varying the number of prototypes per class (M) in Table 5. Using only two prototypes limits intra-class flexibility, while increasing to six introduces redundancy and overlapping prototype regions. Empirically, $M = 4$ offered the optimal compromise, yielding compact yet well-separated clusters and the highest closed-set performance. It is important to note that temporal attention pooling, used to aggregate the diffusion-step features, was not ablated. Unlike the feature-attention module within the classifier, the temporal attention pooling provides the only mechanism for combining multi-step latent representations. Removing it would eliminate any temporal aggregation and collapse the entire data flow, making comparison with the original architecture meaningless. Finally, in the first row of Table 3 we reported the results of switching to Stable Diffusion XL instead of Stable Diffusion 2.1; the results reported are almost identical proving that the signal-leak is directly correlated to the forward diffusion process typical of Stable Diffusion. Together, these results validate the architectural choices adopted in Proto-LeakNet and demonstrate how each component contributes to robust and interpretable latent-space attribution.

Table 3: Ablation study on Proto-LeakNet components. We report Top-1 Accuracy and Macro AUC to quantify the effect of removing prototypes and attention mechanisms or changing the Stable Diffusion model.

Experiment	**Top-1 Acc (%)**	**Macro AUC (%)**
SDXL	82.55	98.09
No prototypes	72.63	93.09
No attention	81.23	97.49
No prototypes & no attention	79.80	96.67
Full Proto-LeakNet	**82.60**	**98.13**

Table 4: Ablation on different backbone architectures. We report Top-1 Accuracy and Macro AUC on the closed-set configuration to evaluate the impact of the feature extractor on attribution performance.

Backbone	**Top-1 Acc. (%)**	**Macro AUC (%)**
EfficientNet_B4	65.80	91.82
ViT-B16	72.07	94.53
ResNet50	81.83	97.39
ResNet101	76.53	95.36
ResNet18	**82.60**	**98.13**

Table 5: Ablation on the number of prototypes per class (M). We report Top-1 Accuracy and Macro AUC on the closed-set configuration.

Prototypes (M)	**Top-1 Acc. (%)**	**Macro AUC (%)**
$M = 2$	81.17	97.41
$M = 6$	81.33	97.58
$M = 4$	**82.60**	**98.13**

6 Discussion

Our study shows that the *signal-leak bias* is a stable and exploitable forensic cue for source attribution across generative models. Proto-LeakNet is, to our knowledge, the first framework to leverage these latent residuals as discriminative features, shifting the focus from bias correction to exploitation. Across all experiments, the model maintained stable attribution under post-processing and perturbations, confirming that latent-domain methods [19, 20, 22] better preserve generator-specific cues than pixel-space approaches. Despite its robustness, Proto-LeakNet presents several limitations. It requires a distinct evaluation phase for unseen generators and is sensitive to hyperparameter settings such as the loss weighting, attention configuration, and kernel bandwidth in the KDE analysis. Furthermore, the model relies on the assumption that latent signal-leak traces persist across diffusion steps, which may vary across architectures or training regimes. Its prototype–attention head, although interpretable, increases both memory usage and inference time compared to lightweight convolutional baselines. Finally, while the method demonstrates clear separability in representation space, it does not yet provide end-to-end open-set discrimination, which remains an open direction for future research.

7 Conclusions and Future Works

We introduced Proto-LeakNet, a signal-leak-aware attribution framework that combines robustness and interpretability within a latent-space formulation. By operating directly on diffusion latents and modeling residual generator-specific cues, the model learns a structured embedding space shaped by prototypes and attention, enabling reliable attribution even under strong post-processing. Experiments showed that signal-leak bias is a stable and discriminative forensic cue across both diffusion and non-diffusion models, while density-based evaluation revealed clear separability between known and unseen generators. Beyond accuracy, Proto-LeakNet provides transparent prototype- and feature-level explanations linking latent statistics to interpretable evidence. Future work will explore open-set discriminative training, contrastive regularization for improved disentanglement, and lightweight runtime variants for real-time forensics. Treating signal-leak bias as an interpretable fingerprint rather than a nuisance establishes a foundation for robust and interpretable attribution of generative media.

References

[1] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative Adversarial Networks, 2014. URL `https://arxiv.org/abs/1406.2661`.

[2] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*. Curran Associates Inc., 2020. ISBN 9781713829546.

[3] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-Resolution Image Synthesis with Latent Diffusion Models. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. doi:10.1109/CVPR52688.2022.01042.

[4] Mirko Casu, Luca Guarnera, Pasquale Caponnetto, and Sebastiano Battiato. GenAI mirage: The impostor bias and the deepfake detection challenge in the era of artificial illusions. *Forensic Science International: Digital Investigation*, 50:301795, 2024. ISSN 2666-2817. doi:https://doi.org/10.1016/j.fsidi.2024.301795.

[5] Irene Amerini, Mauro Barni, Sebastiano Battiato, Paolo Bestagini, Giulia Boato, Vittoria Bruni, Roberto Caldelli, Francesco De Natale, Rocco De Nicola, Luca Guarnera, et al. Deepfake media forensics: Status and future challenges. *Journal of Imaging*, 11(3):73, 2025.

[6] Luisa Verdoliva. Media Forensics and DeepFakes: An Overview. *IEEE Journal of Selected Topics in Signal Processing*, 14, 2020. ISSN 1941-0484. doi:10.1109/jstsp.2020.3002101.

[7] Heng Zhang, Wen Yang, Huai Yu, Haijian Zhang, and Gui-Song Xia. Detecting Power Lines in UAV Images with Convolutional Features and Structured Constraints. *Remote Sensing*, 11, 2019. ISSN 2072-4292. doi:10.3390/rs11111342.

[8] Run Wang, Felix Juefei-Xu, Lei Ma, Xiaofei Xie, Yihao Huang, Jian Wang, and Yang Liu. FakeSpotter: A Simple yet Robust Baseline for Spotting AI-Synthesized Fake Faces, 2020. URL `https://arxiv.org/abs/1909.06122`.

[9] Brandon Khoo, Raphaël C.-W. Phan, and Chern-Hong Lim. Deepfake attribution: On the source identification of artificially generated images. *WIREs Data Mining and Knowledge Discovery*, 12(3):e1438, 2022. doi:https://doi.org/10.1002/widm.1438.

[10] Luca Bindini, Giulia Bertazzini, Daniele Baracchi, Dasara Shullani, Paolo Frasconi, and Alessandro Piva. Tiny Autoencoders are Effective Few-Shot Generative Model Detectors. In *2024 IEEE International Workshop on Information Forensics and Security (WIFS)*, pages 1–6, 2024. doi:10.1109/WIFS61860.2024.10810686.

[11] Riccardo Corvi, Davide Cozzolino, Giada Zingarini, Giovanni Poggi, Koki Nagano, and Luisa Verdoliva. On the detection of synthetic images generated by diffusion models, 2022. URL `https://arxiv.org/abs/2211.00680`.

[12] Jun Wang, Benedetta Tondi, and Mauro Barni. BOSC: A Backdoor-Based Framework for Open Set Synthetic Image Attribution. *IEEE Transactions on Information Forensics and Security*, 20:8043–8058, 2025. ISSN 1556-6021. doi:10.1109/tifs.2025.3592531. URL `http://dx.doi.org/10.1109/TIFS.2025.3592531`.

[13] Everaert, Martin Nicolas and Fitsios, Athanasios and Bocchio, Marco and Arpa, Sami and Süsstrunk, Sabine and Achanta, Radhakrishna. Exploiting the signal-leak bias in diffusion models. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 4025–4034, January 2024.

[14] Andreas Rössler, Davide Cozzolino, Luisa Verdoliva, Christian Riess, Justus Thies, and Matthias Nießner. FaceForensics++: Learning to Detect Manipulated Facial Images, 2019.

[15] Huy H. Nguyen, Junichi Yamagishi, and Isao Echizen. Capsule-Forensics: Using Capsule Networks to Detect Forged Images and Videos, 2018.

[16] Jongwook Choi, Taehoon Kim, Yonghyun Jeong, Seungryul Baek, and Jongwon Choi. Exploiting Style Latent Flows for Generalizing Deepfake Video Detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1133–1143, 2024.

[17] Chuangchuang Tan, Yao Zhao, Shikui Wei, Guanghua Gu, Ping Liu, and Yunchao Wei. Frequency-Aware Deepfake Detection: Improving Generalizability through Frequency Space Learning, 2024.

[18] Pablo Bernabeu-Perez, Enrique Lopez-Cuena, and Dario Garcia-Gasulla. Present and Future Generalization of Synthetic Image Detectors, 2024. URL `https://arxiv.org/abs/2409.14128`.

[19] Zhenting Wang, Chen Sehwag Vikash, Chen, Lingjuan Lyu, Dimitris N Metaxas, and Shiqing Ma. How to Trace Latent Generative Model Generated Images without Artificial Watermark? In *International Conference on Machine Learning*, 2024.

[20] Fengyuan Liu, Haochen Luo, Yiming Li, Philip Torr, and Jindong Gu. Which Model Generated This Image? A Model-Agnostic Approach for Origin Attribution, 2024.

[21] Chuangchuang Tan, Huan Liu, 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, 2023.

[22] Ana Vasilcoiu, Ivona Najdenkoska, Zeno Geradts, and Marcel Worring. Latte: Latent trajectory embedding for diffusion-generated image detection. *arXiv preprint arXiv:2507.03054*, 2025.

[23] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition, 2015. URL `https://arxiv.org/abs/1512.03385`.

[24] Pietro Bongini, Sara Mandelli, Andrea Montibeller, Mirko Casu, Orazio Pontorno, Claudio Vittorio Ragaglia, Luca Zanchetta, Mattia Aquilina, Taiba Majid Wani, Luca Guarnera, Benedetta Tondi, Giulia Boato, Paolo Bestagini, Irene Amerini, Francesco De Natale, Sebastiano Battiato, and Mauro Barni. WILD: a new in-the-Wild Image Linkage Dataset for synthetic image attribution, 2025. URL `https://arxiv.org/abs/2504.19595`.