FORENSIC TWINS: SELF-SUPERVISED RESIDUAL LEARNING FOR AI-GENERATED IMAGE FORENSICS

Anonymous authors
Paper under double-blind review

ABSTRACT

Detectors of AI-generated images are typically trained using samples from all Generative AI architectures they must catch, and struggle as soon as a new architecture emerges. Recent approaches have explored self-supervised pre-training as an alternative solution, yet standard frameworks work against the forensic task, e.g., their augmentations overwrite the micro-statistics of image formation. This paper introduces *Forensic Twins*, a Self-Supervised Residual Learning (SSRL) framework whose pretext task suppresses macroscopic content availability. Each image is mapped through a frozen, off-the-shelf forensic residual extractor, from which two spatially disjoint crops are drawn. Sharing no pixel, the two views retain minimal semantic structure to align, leaving a redundancy-reduction objective with a predominant common signal: the stationary fingerprint of the image acquisition pipeline. Additionally, *Forensic Twins* is trained exclusively on real images; no AI-generated image is observed at any stage. Experiments show that *Forensic Twins* attributes AI generator sources with 56.61% accuracy, i.e., 6.13% above the previous state-of-the-art zero-shot method at $375\times$ lower latency. We also demonstrate that fitting a Gaussian Mixture Model (GMM) offline using only the real image embeddings extracted from *Forensic Twins* turns it into a state-of-the-art zero-shot detector, reaching 97.99% AUC across 27 unseen AI generators, including GANs, diffusion models and commercial systems. Code, weights and exact splits will be made publicly available[1].

1 INTRODUCTION

Generative models now synthesize images with unprecedented detail (Croitoru et al., 2023; Sauer et al., 2022; Alaluf et al., 2022; Tolosana et al., 2020) and increasing physical and semantic coherence (Avrahami et al., 2022), so realistic synthetic media can be deployed at scale. Condition-guided generation, from text prompts (Rombach et al., 2022; Dhariwal & Nichol, 2021) or reference images (Baliah et al., 2025), further lets them replicate structural layouts (Cao et al., 2025; Zhang et al., 2024), emulate artistic styles (Gao et al., 2026), or inject novel entities into authentic contexts (Chen et al., 2023; 2025; Zhao et al., 2024).

Supervised detectors (Tan et al., 2024; Wang et al., 2020b; Ojha et al., 2023) degrade in Out-of-Distribution Detection (OOD) scenarios (Park et al., 2024) and fail to generalize to unseen generators: handcrafted forensic priors and binary objectives push them to overfit to source-specific artifacts. Recent methods instead *i)* treat semantics and forensic traces as decorrelated spaces (Yan et al., 2025; Guo et al., 2026), or *ii)* model the manifold (Zhang et al., 2025) or descriptors (Nguyen et al., 2025) of natural images, relying on massively parameterized backbones such as DINOv2 (Oquab et al., 2024) and CLIP (Radford et al., 2021), or on costly Test-Time Optimization (TTO).

Self-Supervised Learning (SSL) (Gui et al., 2024) is thus a compelling alternative, learning generalizable representations from unannotated data without costly, biased labeling. Yet standard frameworks such as Barlow Twins (BT) (Zbontar et al., 2021), Momentum Contrast (MoCo) (He et al., 2020; Chen et al., 2020) and Self DIstillation with NO Labels (DINO) (Caron et al., 2021; Oquab et al., 2024; Simeoni et al., 2025) transfer poorly to forensics: they minimize the cross-view loss through semantic shortcuts, discarding the low-amplitude, micro-statistical noise intrinsic to natural image formation (Wang et al., 2020b; Marra et al., 2019; Sinitsa & Fried, 2024).

[1]The link will be provided upon acceptance.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

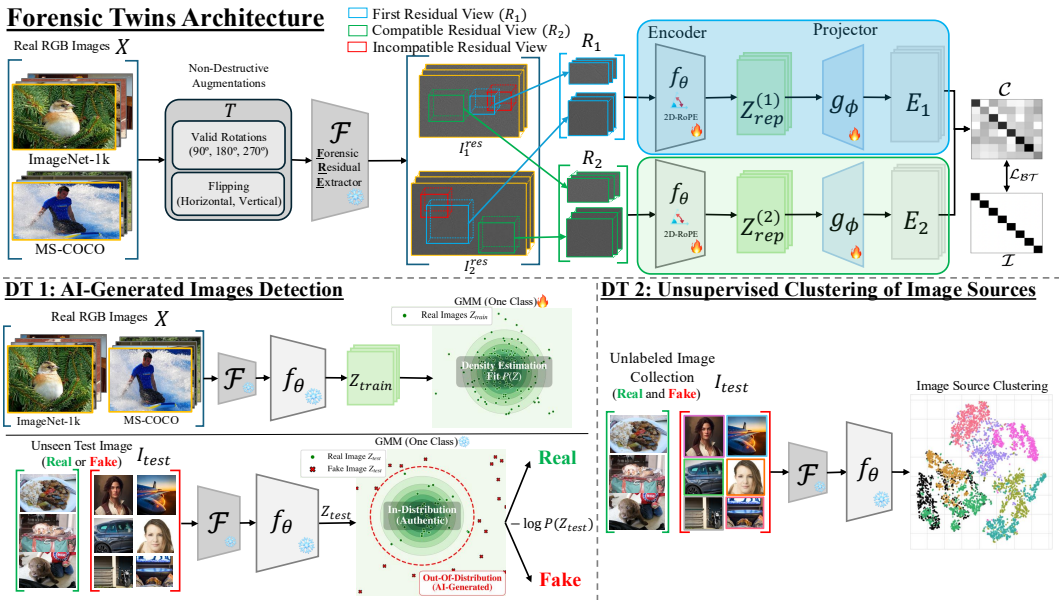

Figure 1: **Overview of *Forensic Twins* and applications. Top:** Real RGB images (X) undergo non-destructive augmentations ($\mathcal{T}$) and residual filtering ($\mathcal{F}$). Then, two strictly disjoint crops (R_1, R_2) pass through a ViT (f_θ) and a projector (g_ϕ), trained with a redundancy-reduction objective on their cross-correlation matrix $\mathcal{C}$. **Bottom-left:** a one-class GMM is fitted on frozen embeddings of real images (Z_{train}) and scores each unseen test image (I_{test}) by $-\log P(Z_{test})$. **Bottom-right:** k-Means groups the frozen embeddings of an unlabeled real/AI-generated collection by image source.

In this paper we propose *Forensic Twins*, a Self-Supervised Residual Learning (SSRL) framework that decouples structural semantics from forensic noise. Figure 1 provides an overview of the proposed method and applications. Following the redundancy-reduction principle of BT (Zbontar et al., 2021), *Forensic Twins* is designed so that the low-level fingerprint of image formation is the only factor the objective can exploit. A frozen, pre-trained forensic residual extractor (Nguyen et al., 2025) suppresses macroscopic content while preserving high-frequency residuals, and two *disjoint* crops of these residuals, which break macro-spatial coherence, serve as the two views. A Vision Transformer (ViT) encodes both, and driving the cross-correlation matrix of their embeddings towards the identity enforces cross-view invariance while decorrelating feature dimensions. The dominant signal both crops still share is the stationary, microscopic trace that image formation spreads uniformly over the image plane.

Crucially, *Forensic Twins* is trained exclusively on real images. Generators leave fingerprints such as up-sampling artifacts and unnatural pixel-level correlations (Marra et al., 2019; Wang et al., 2023; Corvi et al., 2023a; Tan et al., 2024), but these are generator-specific and shift with every new architecture. Real images, in contrast, form a single, physically constrained class produced by the same family of acquisition and processing pipelines (e.g., sensors, color filter arrays, color corrections, etc.). Modeling only that class makes any departure from it, including generators that did not exist at training time, OOD by construction. As can be seen in Figure 1, we explore the potential of *Forensic Twins* on two downstream forensic tasks: *i)* AI-generated image detection, and *ii)* unsupervised clustering of image sources. In summary, the main contributions are:

- ***Forensic Twins* architecture.** We propose a novel SSRL architecture that mitigates the bias towards semantic representations present in standard SSL architectures. We showcase that applying a residual extractor along with spatially disjoint crops, favors the stationary micro-statistics of the image acquisition pipeline as the dominant signal the objective can exploit. Furthermore, *Forensic Twins* is trained exclusively on real images, using a lightweight encoder of 3.0M parameters. We release the code, weights and data splits for reproducible research.
- **Proficiency at Downstream Forensic Tasks.** We showcase the utility of *Forensic Twins* in two downstream forensic tasks: Detection of AI-Generated Images and Image Source

Clustering. For the detection task, a single GMM is fitted offline using only real image embeddings, and a test real/AI-generated image is flagged by its density under that model, achieving 97.99% mean AUC across 27 generative architectures. For image source clustering, a k-Means algorithm is fitted using the real/AI-generated image representations extracted using *Forensic Twins*. Our proposed *Forensic Twins* improves previous state-of-the-art zero-shot method by 6.13% in terms of accuracy.

2 Related Work

AI-Generated Image Detection. Generative pipelines leave subtle forensic traces, largely inherited from the architectural priors of the decoders that map latent representations back to pixel space (Marra et al., 2019; Corvi et al., 2023a; Neves et al., 2020). Early detectors exploit them through supervised binary classification on a single generator family (Wang et al., 2020b), sometimes under an explicit prior: LGrad (Tan et al., 2023) uses per-channel RGB gradients, and NPR (Tan et al., 2024) targets the pixel correlations left by up-sampling. These models are strong in-domain but degrade sharply on unseen generators. Generalization has since been pursued by adapting vision-language backbones (Ojha et al., 2023): EFFORT (Yan et al., 2025) and OmniAID (Guo et al., 2026) both steer CLIP-L/14 towards forensic patterns while preserving its semantic knowledge, the latter through a Mixture of Experts (MoE) decomposition. DIRE (Wang et al., 2023) scores images by their diffusion reconstruction error before a binary classifier, while Zhang et al. (2025) dispense with training entirely, flagging an image by its distance to the natural-image manifold in DINOv2 space. The latter shares our premise that real data alone suffices to define the decision rule, yet builds it on a representation optimized for semantics. *Forensic Twins* is trained so that the representations steer away from semantic content towards capturing the statistical fingerprint of image formation.

Unsupervised Clustering of Image Sources. Beyond binary detection, attributing a synthetic image to its source in an open-set scenario remains a formidable challenge. Nguyen et al. (2025) extract forensic self-descriptions for zero-shot detection and unsupervised clustering, but refine them per image through TTO, reporting approximately 0.11 images per second. We adopt their residual extractor and remove that bottleneck.

Self-Supervised Representation Learning. SSL learns from unlabelled corpora by making an encoder agree on two augmented views of an image. The different approaches differ mainly in how collapse is avoided: repelling negative pairs (He et al., 2020; Chen et al., 2020), matching a student to a momentum teacher (Caron et al., 2021; Oquab et al., 2024; Simeoni et al., 2025), or driving the cross-correlation matrix of the embeddings towards the identity (Zbontar et al., 2021). All rest on the same premise, that the augmentations declare what the representation must discard, and they are chosen so that semantic content survives. Under a forensic objective this premise inverts: rescaling, blurring and re-compression are precisely the operations that overwrite the micro-statistical fingerprint of image formation, and whenever the two views overlap spatially, macroscopic structure remains the cheapest signal to align. SPAI (Karageorgiou et al., 2025) is the closest attempt, pretraining with a masked spectral objective (Xie et al., 2023) before fitting a supervised binary head shaped by the generators seen in training, while NoisePrint++ (Guillaro et al., 2023) used a supervised contrastive objective (Khosla et al., 2020), where they carefully selected positive and negative pairs considering different forensic attributes. In contrast, our proposed *Forensic Twins* architecture changes the input signal from standard RGB images to residuals along with spatially disjoint views, steering the optimization towards forensic representations and away from semantics without the need for contrastive negative samples.

3 Proposed Method: *Forensic Twins*

We introduce *Forensic Twins*, a novel SSRL framework designed to extract label-free forensic representations of the image generation process using only a corpus of real images. In this section, we formulate the underlying objective, detail our feed-forward architecture, and explain how strict spatial and sampling constraints enforce the decoupling of semantics from physical fingerprints.

3.1 BACKGROUND

Forensic Residuals. Any digital image $I \in \mathbb{R}^{H\times W}$ can be conceptualized as a combination of its macroscopic semantic content C and its microscopic, high-frequency noise fingerprint Ψ, e.g., color filter arrays, camera sensor Photo-Response Non-Uniformity (PRNU) or generative upsampling/quantized artifacts. As isolating Ψ is a very difficult task, Nguyen et al. (2025) proposed to model an image as an additive operation between both terms, $I(x, y) = C(x, y) + \Psi(x, y)$, where x and y are the 2D pixel coordinates. Therefore, to isolate Ψ, they introduced a constrained learnable filter $\mathcal{K}$ which is optimized to estimate the semantic content $\hat{C}$, hence extracting the residual by subtracting the estimation $\hat{C}$ from the original image I. However, this approximation does not always capture the micro-statistical forensic patterns from an image, as approximating semantic content is challenging. As a result, the authors proposed to train a collection of 8 different filters denoted as $\mathcal{F}$ which are specialized in capturing different patterns of the micro-structures. The resulting tensor $I_{res} = \mathcal{F}(I) \in \mathbb{R}^{H\times W\times 8}$ theoretically suppresses low-frequency semantics while amplifying Ψ. However, even with these elaborated residual extractors, macroscopic edges often survive this filtering, which may lead to semantic entanglement when processed by standard deep architectures.

Mutual Information and Redundancy Reduction (Barlow Twins). To learn meaningful representations without labels, we build upon the Barlow Twins objective (Zbontar et al., 2021). Given two views of an image, a neural network composed of an encoder f_θ and a projector g_ϕ produces the embeddings $E_1, E_2 \in \mathbb{R}^{B\times D}$ for a batch of size B. The embeddings are batch-normalized, and each element $\mathcal{C}_{ij}$ of the cross-correlation matrix $\mathcal{C} \in \mathbb{R}^{D\times D}$ across a batch of views is computed as:

$$\mathcal{C}_{ij} = \frac{1}{B}\sum_{b=1}^{B} E_{1,b,i}E_{2,b,j} \tag{1}$$

where i and j correspond to the index along the activations in the embedding layer across the batch B. The loss function minimizes the difference between $\mathcal{C}$ and the identity matrix $\mathcal{I}$:

$$\mathcal{L}_{\mathcal{BT}} = \sum_i (1 - \mathcal{C}_{ii})^2 + \lambda \sum_i \sum_{j\neq i} \mathcal{C}_{ij}^2 \tag{2}$$

By driving the diagonal elements to 1 (invariance) and off-diagonal elements to 0 (redundancy reduction), the network is forced to find the underlying invariant factors of variation between the two views, without the need for negative samples.

3.2 FORENSIC TWINS: A FEED-FORWARD ARCHITECTURE FOR SEMANTIC DECOUPLING

Figure 1 provides an overview of our proposed *Forensic Twins*. Given a batch of color images $X_{RGB} = \{I_i^{RGB}\}_{i=1}^N$, where $I_i^{RGB} \in \mathbb{R}^{H\times W\times 3}$, we first apply a non-invasive transformation pipeline that preserves as much forensic information from the pristine pixel distribution as possible. This deliberately excludes transformations that alter the forensic micro-structures of the image. Therefore, we define the collection of valid transformations as *i)* random orthogonal rotations (i.e., 90º, 180º and 270º), and *ii)* random flipping in the vertical and horizontal axes. Both transformations are drawn once per image and shared by the two views, so that the location from which each view is cut remains the only factor distinguishing them.

The transformed color images are then converted to grayscale, denoted as $X_g = \{I_i^g\}_{i=1}^N$, where $I_i^g \in \mathbb{R}^{H\times W}$, and normalized to be in the $[0, 1]$ range by dividing each pixel value by 255. The semantic content C is then heavily suppressed by applying a Forensic Residual Extractor (FRE) $\mathcal{F}$ from (Nguyen et al., 2025), which is kept frozen, yielding $X_{res} = \{I_i^{res}\}_{i=1}^N$, where $I_i^{res} \in \mathbb{R}^{H\times W\times 8}$, where 8 is the number of channels of the FRE. These residual images contain multiple complementary channels capturing different patterns of residual high-frequency information (Nguyen et al., 2025). However, the residual images X_{res} inherently retain structural semantic edges that can be leveraged by the encoder to create an optimization shortcut, steering representations toward semantics rather than the underlying forensics.

To ensure that the invariant factor maximized by $\mathcal{L}_{\mathcal{BT}}$ is this micro-statistical noise Ψ and not the semantic content C left by the FRE, we propose to avoid standard overlapping views in favor of spatially disjoint ones. The statistical fingerprint left by physical acquisition devices in real images and generative artifacts in AI-generated images is correlated, as stationary patterns occur across the entire image plane (Marra et al., 2019). This means that spatially disjoint views carry cross-correlated fingerprints from the image while restricting semantic information. Given an arbitrary residual image from the batch I_i^{res}, we sample two spatial coordinates (x_1, y_1) and (x_2, y_2) to extract two disjoint crops (Chai et al., 2020), $R_1 \in \mathbb{R}^{H_1 \times W_1 \times 8}$ and $R_2 \in \mathbb{R}^{H_2 \times W_2 \times 8}$, which may or may not have different resolutions, and use them as input views for the encoder f_θ.

Both batches of disjoint views R_1 and R_2 are then fed into our representation learning pipeline, composed of a feature encoder f_θ and a non-linear projector g_ϕ. For batched tensor operations, all R_1 crops within a mini-batch share a common spatial size $H_1 \times W_1$, and independently all R_2 crops share a common size $H_2 \times W_2$. However, $H_1 \times W_1$ need not equal $H_2 \times W_2$. Both crops sizes are arbitrarily re-sampled independently at every training iteration, which motivates using 2D rotary positional embeddings (Su et al., 2024; Simeoni et al., 2025) in the encoder f_θ, so that a single set of weights handles crops of arbitrary size for both views. The representation produced by the encoder is denoted as $Z_{rep}^{(v)}$, where v denotes the view (i.e., R_1 or R_2) from which the representation is extracted. Finally, to compute the self-supervised objective, these dense representations are mapped through the projector network g_ϕ to obtain the embeddings $E_v = g_\phi(Z_{rep}^{(v)})$. The cross-correlation matrix $\mathcal{C}$ is then constructed from E_1 and E_2. Since the underlying micro-statistical noise Ψ remains the dominant shared signal between the disjoint views, the Barlow Twins loss $\mathcal{L}_{\mathcal{BT}}$ is coerced to align forensic cues rather than semantics.

3.3 Downstream Forensic Applications

Following standard practices in self-supervised learning (Zbontar et al., 2021), once *Forensic Twins* is pre-trained exclusively on real images, the non-linear projector g_ϕ is discarded. The frozen encoder f_θ is retained to extract the robust micro-statistical representations Z_{rep} directly from the summary representation (i.e., the `[CLS]` token in the ViT). These embeddings can then be leveraged for various downstream forensic tasks without requiring any fine-tuning or test-time optimization.

AI-Generated Image Detection. Since real image formation fingerprints exhibit a bounded and stationary distribution, we approach AI-generated image detection as an anomaly detection problem in the latent space. Given a training dataset of N real images, we extract their representations $\mathcal{Z}_{train} = \{Z_{rep}^1, \ldots, Z_{rep}^N\}$. We then fit a GMM to the real image distribution, estimating the density $P(Z)$. During inference, for any unseen test image (whether real or AI-generated by an out-of-distribution model), we extract its representation Z_{test} and compute its anomaly score as the negative log-likelihood $\mathcal{S}_{anomaly}(Z_{test}) = -\log P(Z_{test})$. We hypothesize that AI-generated images, lacking the physical process of acquiring an image and containing generative structural artifacts, fall in low-density regions of the GMM, yielding higher anomaly scores.

Clustering of Image Sources. Beyond binary detection, the structural topology of the learned latent space allows clustering distinct Generative AI architectures based on their unique up-sampling or decoding artifacts. For source attribution, we extract representations Z_{rep} from an unlabeled mixture of real and diverse AI-generated images. To perform unsupervised clustering of real/AI-generated image sources, we apply the k-Means algorithm directly on the encoder's output.

4 Experimental Results

Optimization Details. *Forensic Twins* optimization was conducted on a local compute node equipped with four NVIDIA RTX 4090 GPUs. Models were trained for 80 epochs using the AdamW optimizer, with a base learning rate of 0.0015 and a weight decay of 0.0025. We employ a learning rate schedule consisting of a 15-epoch linear warm-up, followed by a cosine annealing decay down to 0.000001. Given the VRAM constraints of the GPUs, models were trained in `bfloat16` and using FlashAttentionV2 (Dao, 2024) for optimized memory consumption, with a batch size of 320. The output dimension of the projector g_θ is set to 8,192 across all the experiments. Regarding the training of the GMM, we use the same real images as the encoder pre-training, with a number of components set to $K = 1$ using full covariance matrix.

Training Dataset. Following recent approaches in the field of AI-generated image detection (Karageorgiou et al., 2025; Nguyen et al., 2025; Zhang et al., 2025; Corvi et al., 2023a), our proposed *Forensic Twins* is trained using only real images from ImageNet-1k (Deng et al., 2009) and MS-COCO (Lin et al., 2014). Specifically, 250,000 images are gathered from ImageNet-1k training set and 241,690 images from MS-COCO training data, resulting in 491,690 images in total.

4.1 AI-Generated Image Detection

Experimental Set-Up. We evaluate *Forensic Twins* against state-of-the-art AI-generated image detection methodologies. This involves the following methods: CNNDet (Wang et al., 2020b), LGrad (Tan et al., 2023), UFD (Ojha et al., 2023), NPR (Tan et al., 2024), ConV (Zhang et al., 2025), FSD (Nguyen et al., 2025), SPAI (Karageorgiou et al., 2025), EFFORT (Yan et al., 2025) and OmniAID (Guo et al., 2026). As in (Nguyen et al., 2025), we leverage AI-generated images from different well-known datasets: OSSIA (Fang et al., 2023), DMID (Corvi et al., 2023b) and SynthBuster (Bammey, 2023). Some notable generators are: StyleGAN family (Sauer et al., 2022; Alaluf et al., 2022), ProGAN (Karras et al., 2018), EG3D (Chan et al., 2022), GLIDE (Nichol et al., 2022), Stable Diffusion family (Rombach et al., 2022; Esser et al., 2024; Podell et al., 2024) or DiT (Peebles & Xie, 2023). A full breakdown of the different generators considered in the evaluation is detailed in Appendix F. Furthermore, we add AI-generated images from commercial generators (i.e., Firefly (Adobe Inc., 2024), Midjourney v6 (Midjourney, Inc., 2024) or Flux (Black Forest Labs, 2024)) extracted from online resources [2],[3],[4] (e.g., Kaggle or HuggingFace), evaluating both *Forensic Twins* and previous state-of-the-art methods in 27 total AI generators, which we divide into three fundamental subsets: GAN-based generators, Diffusion-based generators and Commercial generators. For the real images, following previous approaches in the literature (Nguyen et al., 2025; Corvi et al., 2023a; Zhu et al., 2023), 1,000 images are selected from the held-out validation and test splits of ImageNet's and MS-COCO, disjoint from the 491,690 images used for pre-training. For the AI-generated images, a total of 1,000 images per generator were considered across the datasets (i.e., OSSIA, DMID, SynthBuster), balancing across each real vs. generator evaluation. Also, to avoid any confound regarding image resolution, we take a 256×256 center crop in inference.

Results. Table 1 reports the performance. Our proposed method achieves state-of-the-art results on average (97.99% AUC), ranking third among considered detectors and within 1% of the best one. A simple ranking over the column of global averages might hide the important advantages of *Forensic Twins*. First, seven of the nine state-of-the-art methods are calibrated on AI-generated images (see *Real-Only?* column), and the cost of a decision (latency) spans three orders of magnitude. Nevertheless, *Forensic Twins* is only trained using real images. Considering similar state-of-the-art methods that never observe AI-generated images (i.e., ConV and FSD), *Forensic Twins* outperforms ConV by 21% (77.13 vs. 97.99% AUC) with a backbone that is two orders of magnitude smaller (3.0M vs. 304.4M parameters). FSD achieves slightly better performance (98.52% vs. 97.99% AUC), at the price of a considerably higher per-image optimization that takes 2,415.27 ms per decision against our 6.44 ms.

The per-category performance columns also provide interesting insights. *Forensic Twins* achieves its lowest performance on commercial generators (96.10%), although 2.38% above FSD, which falls from 99.88% on GANs to 93.72% there. Other methods trained using AI-generated images degrade far more sharply: NPR drops from 94.45% on diffusion models to 74.50% on commercial ones, and CNNDet, LGrad and UFD span 28.17%, 24.56% and 14.13% between their best and worst category. Our own gap is just 2.46%. Having only seen real images, *Forensic Twins* is not tied to a specific AI generator family, and its performance is correspondingly flat across the three of them.

To summarize, two methods rank above us on the global average. FSD does so at 375 times our latency, as noted above. EFFORT leads by 0.97% with a 504.6M-parameter backbone trained with supervision on AI-generated images, while OmniAID (a 508.8M-parameter mixture of experts trained the same way) falls 0.32% behind *Forensic Twins*. Neither of the two above can be compared with *Forensic Twins* directly: EFFORT considers supervised learning with a model 168 times the size of

[2] https://www.kaggle.com/datasets/aryankaushik005/firefly

[3] https://huggingface.co/datasets/brivangl/midjourney-v6-llava

[4] https://huggingface.co/datasets/Rapidata/Flux_SD3_MJ_Dalle_Human_Alignment_Dataset

Table 1: AI-generated image detection performance (AUC %), averaged per category across 27 generative architectures and 2 real-image distributions (ImageNet-1k, MS-COCO). Parameter counts and latency are reproduced from the full cost breakdown in Appendix C (for *Forensic Twins* we include FRE's module from (Nguyen et al., 2025) in the computational cost and complexity statistics); latency is a median over 500 ImageNet-1k images on a single RTX 4090. Full per-generator breakdown in App. D. We also indicate whether the method was trained only with real images and the per-generator family average and the global average. Our results are in **bold**, top-3 models are underlined.

Method	Real Only?	#Params	Lat. (ms)	GANs (8)	Diffusion (14)	Commercial (5)	Global (27)
CNNDet[CVPR20]	✗	23.5M	3.09	93.79	65.62	68.29	74.46
UFD[CVPR23]	✗	304.0M	8.28	97.73	83.60	84.80	88.01
LGrad[CVPR23]	✗	46.6M	10.43	78.52	62.90	53.96	65.87
NPR[CVPR24]	✗	1.4M	2.17	92.83	94.45	74.50	90.27
SPAI[CVPR25]	✗	139.9M	26.49	94.94	95.00	94.48	94.88
EFFORT[ICML25]	✗	504.6M	12.76	99.53	98.73	98.68	98.96
OmniAID[ICLR26]	✗	508.8M	40.25	99.51	97.25	95.92	97.67
ConV[NeurIPS25]	✓	304.4M	16.82	83.18	75.33	72.46	77.13
FSD[CVPR25]	✓	6.2M	2,415.27	99.88	99.47	93.72	98.52
Ours	✓	**3.0M**	**6.44**	**98.17**	**98.56**	**96.10**	**97.99**

Forensic Twins, and FSD pays for its margin with a per-image optimization. Our claim is that the proposed representation learned from real images alone comes very close to the strongest supervised detector, at only 3.0M parameters and a single forward pass (6.44 ms). Additional details regarding the computational advantages of *Forensic Twins* are included in Appendix C.

4.2 Unsupervised Clustering of Image Sources

Experimental Set-Up. We also evaluate the capabilities of *Forensic Twins* for image source clustering against similar self-supervised methods that only consider real images for training: FSD (Nguyen et al., 2025) and ConV (Zhang et al., 2025). Furthermore, we include CLIP's vision backbone, given its many uses in supervised AI-generated detection methods (Ojha et al., 2023; Cozzolino et al., 2024; Yan et al., 2025; Guo et al., 2026). We follow the unsupervised clustering protocol from (Nguyen et al., 2025) and use real images from ImageNet-1k test set —which is disjoint from the training set used to train the encoder— and AI-generated images from ProGAN (Karras et al., 2018), StyleGAN3 (Alaluf et al., 2022), GLIDE (Nichol et al., 2022), SD1 (Rombach et al., 2022), DALLE-3 (Betker et al., 2023), Midjourney v6 and Firefly. As in (Nguyen et al., 2025), we explore different numbers of clusters as factors of the true number of sources N=8 (i.e., N, $2N$ and $4N$).

Table 2: Comparison for unsupervised clustering of image sources in terms of accuracy. The number of sources N is 8. Our results are in **bold**. Best result is underlined.

Method	# *N*	# 2*N*	# 4*N*
CLIP-L/14[ICML21]	51.73	72.03	75.00
ConV[NeurIPS25]	43.98	54.16	65.95
FSD[CVPR25]	50.48	66.05	73.58
Ours	**56.61**	**83.17**	**84.34**

Results. Table 2 reports unsupervised clustering of image sources. *Forensic Twins* outperforms FSD by 6.13 points of accuracy at k=*N*, by 17.12 at k=2*N* and by 10.76 at k=4*N*. ConV leverages DINOv2 representations which seem to not be enough to properly identify sources, as it falls behind all compared methods (43.98% at k=*N*). We attribute this mainly because DINOv2 is oriented to extract semantic representations. Similarly, CLIP's visual encoder appears to perform better than FSD (51.73% acc. at k=*N*) but worse than *Forensic Twins*.

Figure 2 projects the same t-SNE representation of *Forensic Twins* over seven semantic superclasses of held-out ImageNet images (left) and over seven AI generators (right), with the same number of images per class and identical t-SNE settings. As can be seen, the feature space is well organized in terms of source identity (right) but not semantic content (left), evidencing that the representation is dominated by image-formation structure, capturing the subtle forensic differences across generators, which is the main purpose of *Forensic Twins*. App. B further details the entanglement between semantics and forensics in the representation space of different methods.

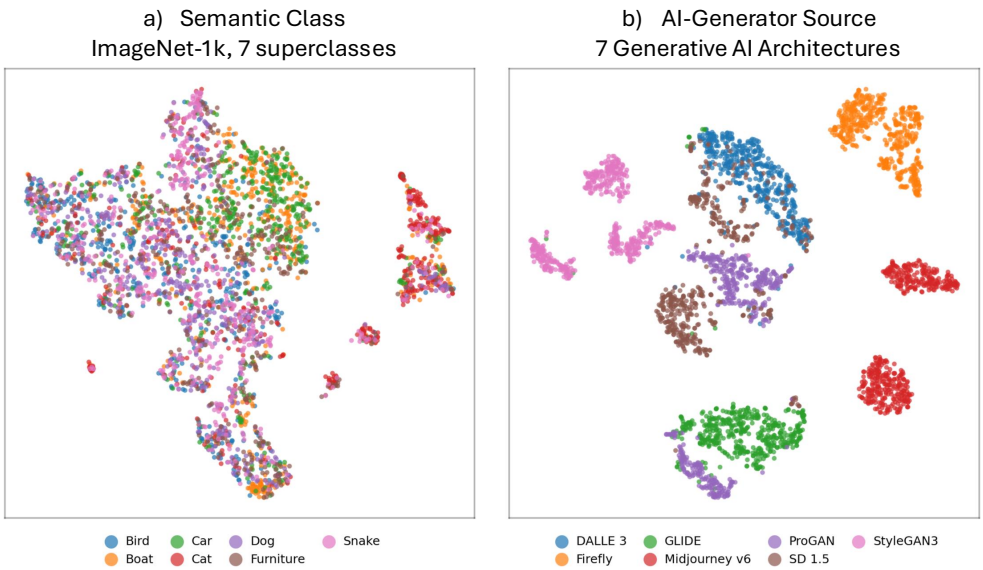

Figure 2: t-SNE visualization of images regarding semantic classes on ImageNet-1k (left) and different generators with different architectures (right). Note that *Forensic Twins* has never seen AI-generated images at training time; however, it is able to cluster them. The semantic visualization showcases mild aggregation per semantic class, but no clear structure.

4.3 Ablation Studies

All ablations vary different aspects of *Forensic Twins* starting from a base ViT-Tiny (six blocks, patch-16 linear stem, `[CLS]` token) on RGB data with $\lambda = 0.005$, trained for 80 epochs (batch size 320) on the 491,690 ImageNet and MS-COCO images. Views are disjoint crops of the training images, and non-destructive augmentations are applied (i.e., flips and orthogonal rotations). The GMM ($K = 1$) is fitted on real image embeddings alone, and we report global AUC (%) over the 27 AI generators and both real domains. Table 3 shows the AUC (%) results achieved in each experiment for the task of AI-generated image detection. Additional ablations can be found in App. A and App. E regarding robustness to different compressions and unseen real image distributions, respectively.

Residual Extractor. When using raw RGB images as input, the baseline aligns views through semantics and barely detects any AI-generated image (60.95% AUC), with GANs at chance (51.76%). Placing a residual extractor before the encoder shifts what it learns: fixed SRM filters (Fridrich & Kodovsky, 2012) recover most of the gap (84.14%), and the learned FRE from (Nguyen et al., 2025) adds a further 9.63 points (93.77%). Suppressing content, rather than the specific filter, is what enables detection, with the FRE as its strongest instance. We use the FRE for the rest of the ablations as it is a key component for our architecture.

Augmentations. We ablate how the two views are built. Overlapping instead of disjoint crops lowers AUC to 89.49 (-4.28 points compared to the base configuration with FRE), as the shared pixels likely let the encoder match views through content rather than forensic traces. DINOv2-style (Oquab et al., 2024) destructive augmentations are far more harmful (57.84%, decreasing performance 35.93 points): they overwrite the residual signal both views share, and training fails to align them. Additional ablations regarding compression robustness are explored in App. A.

Architecture. We observe that patch size is the dominant factor as changing it to 8 is worth 1.70% AUC. We also replace the patch stem with the convolutional stem of Xiao et al. (2021), to assess if the inherent high-frequency bias of CNNs (Wang et al., 2020a; Abello et al., 2021) helps the ViT to extract better forensic patterns from the residuals; in isolation it costs 1.37%. Summary statistics never beat the `[CLS]` token. Global Average Pooling (GAP) loses 2.67%, and concatenating the per-token Global Standard Deviation Pooling (STD), motivated by the stationarity of forensic residuals (Corvi et al., 2023a), recovers 1.07% of them but stays 1.60% below the base, while the same concatenation applied to `[CLS]` costs 2.70%. The `[CLS]` token already encodes what the second moment adds to an average.

Table 3: Ablation study of the encoder (491k training images, ImageNet-1k + MS-COCO) regarding the architecture and different optimization hyper-parameters changing one factor at a time. The factor composition contemplates ablations combining individual factors that improved the model performance in AI-generated image detection. Performance is shown in terms of AUC (%). We show the per-generator family average and the global average.

Factor	Config	Changed value	GANs (8)	Diffusion (14)	Commercial (5)	Global
—	Base	—	51.76	66.14	61.13	60.95
		Residual Extractor				
Extractor	SRM	RGB → SRM	80.60	85.88	84.95	84.14
	FRE	RGB → FRE	92.28	94.68	93.59	93.77
		Augmentations (with FRE)				
Img. Trans.	Destructive Aug.	Transformation Pipeline	59.56	57.91	54.90	57.84
	Overlapped Crops	View Strategy	87.53	90.68	89.31	89.49
		Architecture (with FRE)				
Patch size	Patch=8	16 → 8	94.38	96.30	94.88	95.47
Stem	Stem=Conv	linear → conv	91.92	93.09	91.26	92.40
Representation	`[CLS]`+STD	`[CLS]` → `[CLS]`+STD	88.83	92.28	91.26	91.07
	GAP	`[CLS]` → GAP	90.36	91.68	90.63	91.10
	GAP+STD	`[CLS]` → GAP+STD	90.13	93.53	91.66	92.17
		Regularization (with FRE)				
BT λ	λ=0.001	λ: 0.005 → 0.001	94.29	95.90	94.69	95.20
		Factor composition (all rows start from patch=8, with FRE)				
Interaction	P8 + Conv	stem linear → conv, λ=0.005	92.59	94.81	92.18	93.66
BT λ (at P8)	P8 + λ=0.001	λ=0.001	95.96	97.51	95.57	96.69
	P8 + λ=2e-4	λ=0.0002	97.10	98.42	95.77	97.54
	P8 + λ=5e-5	λ=0.00005	98.17	98.56	96.10	97.99
	P8 + λ=1e-5	λ=0.00001	98.13	98.16	96.11	97.77

Regularization. The redundancy term has an optimum rather than a direction: $\lambda = 0.001$ gains 1.43% at patch 16, and at patch 8 the sweep achieves an optimum at 5×10^{-5} (97.99%). The factors are not independent either. Patch 8 with $\lambda = 0.001$ gains 2.92% over the base using FRE, below the 3.13% their separate gains predict, and patch 8 with the convolutional stem (93.66%) exceeds the convolutional stem alone (92.40%). The final configuration improves on the base configuration using FRE (93.77% AUC) by 4.22% AUC.

5 Conclusion

This paper has presented *Forensic Twins*, a self-supervised residual framework that learns the statistical fingerprint of real image formation without ever observing an AI-generated image. The design principle relies on preventing the encoder from taking the easy path towards semantic content by *i)* applying a residual filter that suppresses the majority of the semantic content from a given image, *ii)* extracting disjoint crops to alleviate potential semantic shortcuts, as statistical fingerprints occur along the whole image plane and *iii)* removing destructive augmentation pipelines present in standard SSL pipelines as they remove most of the forensic signal from real images.

We have tested the learned representations in two challenging tasks: AI-generated image detection, evaluating the performance across 27 different Generative AI architecture across different architectures (i.e., GANs, Diffusion Models and Commercial) and Clustering of Image Sources, where *Forensic Twins* is capable of attributing image sources without *i)* ever having seen any AI-generated image and *ii)* or relying on semantic information.

Future work will address *Forensic Twins*' limitations. First, commercial architectures remain the weakest family for almost every method, ours included, as their non-public technology makes it hard to pinpoint which modules leave exploitable traces. Second, one-class density estimation itself is limited: App. A shows that GMM robustness to re-compression degrades performance, though a handful of labeled examples largely restores it, suggesting few-shot calibration as a natural complement to our zero-shot setting. Designing an objective that leverages dense ViT features or exploring alternative one-class modeling such as Normalizing Flows (Kobyzev et al., 2021) is also in our plans.

AI USE STATEMENT

In this work, we used generative AI tools to aid or polish writing, for retrieval and discovery, and to draft sections of the paper. We have not used generative AI tools for research ideation or execution, proving mathematical claims or generating synthetic datasets, as all synthetic datasets were already available from prior work in the area. We have reviewed all AI-assisted work. We have checked that the references from the literature provided by the AI tool were real and of relevance for our work. Furthermore, all code generated using AI tools was executed and tested by the first author. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This work aims to improve the reliability of AI-generated image detection, a capability with clear defensive value against misinformation, non-consensual synthetic media, and fraud. Because our method learns exclusively from real images, it does not require curating or distributing datasets of harmful synthetic content for training, which we view as a positive property relative to supervised alternatives.

We acknowledge a dual-use concern common to all forensic detection work: characterizing the statistical fingerprint that separates real from synthetic images could, in principle, inform generators or post-processing pipelines designed to evade detection. App. A shows that our one-class density model is sensitive to JPEG re-encoding; we report this limitation transparently, believing that honest characterization of failure modes is more valuable to defenders than withholding it would be costly to evaders, who already have access to the underlying literature we build on.

We caution against deploying detectors such as ours as a sole, automated arbiter of authenticity in high-stakes settings (content moderation, legal evidence, fact-checking) without human review, given the generator —and compression— dependent performance gaps documented throughout Sec. 4 and App. A and D.

REPRODUCIBILITY STATEMENT

Our code will be available upon acceptance, which contains training and evaluation scripts. Sec. 4, along with App. A to E provide the required implementation details to reproduce our results.

REFERENCES

Antonio A Abello, Roberto Hirata, and Zhangyang Wang. Dissecting the High-Frequency Bias in Convolutional Neural Networks. In *Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 863–871, 2021.

Adobe Inc. Adobe firefly. `https://www.adobe.com/es/products/firefly.html`, 2024. Accessed: 2026-09-17.

Yuval Alaluf et al. Third Time's the Charm? Image and Video Editing with StyleGan3. In *Proc. European Conference on Computer Vision (ECCV)*, pp. 204–220. Springer, 2022.

Omri Avrahami, Dani Lischinski, and Ohad Fried. Blended Diffusion for Text-Driven Editing of Natural Images. In *Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 18208–18218, June 2022.

Sanoojan Baliah et al. Realistic and Efficient Face Swapping: A Unified Approach with Diffusion Models. In *Proc. IEEE/CVF Winter Conf. on Applications of Computer Vision (WACV)*, pp. 1062–1071, 2025.

Quentin Bammey. Synthbuster: Towards Detection of Diffusion Model Generated Images. *IEEE Open Journal of Signal Processing*, 5:1–9, 2023.

James Betker, Gabriel Goh, Li Jing, Tim Brooks, Jianfeng Wang, Linjie Li, Long Ouyang, Juntang Zhuang, Joyce Lee, Yufei Guo, et al. Improving Image Generation with Better Captions. *Computer Science. https://cdn.openai.com/papers/dall-e-3.pdf*, 2(3):8, 2023.

Black Forest Labs. Flux. https://bfl.ai/research, 2024. Accessed: 2026-09-17.

Yukang Cao et al. FreeMorph: Tuning-Free Generalized Image Morphing with Diffusion Model. In *Proc. IEEE/CVF Intl. Conf. on Computer Vision (ICCV)*, pp. 18111–18120, 2025.

Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging Properties in Self-Supervised Vision Transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 9650–9660, 2021.

Lucy Chai, David Bau, Ser-Nam Lim, and Phillip Isola. What Makes Fake Images Detectable? Understanding Properties That Generalize. In *Proc. of the European Conference on Computer Vision (ECCV)*, pp. 103–120, 2020.

Eric R Chan, Connor Z Lin, Matthew A Chan, Koki Nagano, Boxiao Pan, Shalini De Mello, Orazio Gallo, Leonidas J Guibas, Jonathan Tremblay, Sameh Khamis, et al. Efficient Geometry-Aware 3D Generative Adversarial Networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 16123–16133, 2022.

Jingye Chen, Yupan Huang, Tengchao Lv, Lei Cui, Qifeng Chen, and Furu Wei. TextDiffuser: Diffusion Models as Text Painters. In *Proceedings of Advances in Neural Information Processing Systems (NeurIPS)*, volume 36, pp. 9353–9387, 2023.

Jingye Chen et al. TextDiffuser-2: Unleashing the Power of Language Models for Text Rendering. In *Proc. European Conf. on Computer Vision (ECCV)*, pp. 386–402, 2025. ISBN 978-3-031-72652-1.

Xinlei Chen, Haoqi Fan, Ross Girshick, and Kaiming He. Improved Baselines with Momentum Contrastive Learning. *arXiv preprint arXiv:2003.04297*, 2020.

Riccardo Corvi, Davide Cozzolino, Giovanni Poggi, Koki Nagano, and Luisa Verdoliva. Intriguing properties of synthetic images: From generative adversarial networks to diffusion models. In *Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pp. 973–982, June 2023a.

Riccardo Corvi, Davide Cozzolino, Giada Zingarini, Giovanni Poggi, Koki Nagano, and Luisa Verdoliva. On the Detection of Synthetic Images Generated by Diffusion Models. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5, 2023b.

Davide Cozzolino, Giovanni Poggi, Riccardo Corvi, Matthias Nießner, and Luisa Verdoliva. Raising the bar of ai-generated image detection with clip. In *Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops)*, 2024.

Florinel-Alin Croitoru et al. Diffusion Models in Vision: A Survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(9):10850–10869, 2023.

Tri Dao. FlashAttention-2: Faster Attention with Better Parallelism and Work Partitioning. In *Proc. of the International Conference on Learning Representations (ICLR)*, pp. 35549–35562, 2024.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: A Large-Scale Hierarchical Image Database. In *Proc. of the IEEE/CVF Computer Vision and Pattern Recognition (CVPR)*, pp. 248–255, 2009.

Prafulla Dhariwal and Alexander Nichol. Diffusion Models Beat GANs on Image Synthesis. In *Proc. Advances on Neural Information Processing Systems (NeurIPS)*, volume 34, pp. 8780–8794, 2021.

Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, Dustin Podell, Tim Dockhorn, Zion English, and Robin Rombach. Scaling Rectified Flow Transformers for High-Resolution Image Synthesis. In *Proc. of the International Conference on Machine Learning*, 2024.

Shengbang Fang, Tai D Nguyen, and Matthew C. Stamm. Open Set Synthetic Image Source Attribution. In *Proc. of the British Machine Vision Conference (BMVC)*, 2023.

Jessica Fridrich and Jan Kodovsky. Rich Models for Steganalysis of Digital Images. *IEEE Transactions on Information Forensics and Security*, 7(3):868–882, 2012.

Junyao Gao, Yanan Sun, Yanchen Liu, Yinhao Tang, Yanhong Zeng, Ding Qi, Kai Chen, and Cairong Zhao. StyleShot: A Snapshot on Any Style. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 48(2):1215–1228, 2026.

Jie Gui, Tuo Chen, Jing Zhang, Qiong Cao, Zhenan Sun, Hao Luo, and Dacheng Tao. A Survey on Self-Supervised Learning: Algorithms, Applications, and Future Trends. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(12):9052–9071, 2024.

Fabrizio Guillaro, Davide Cozzolino, Avneesh Sud, Nicholas Dufour, and Luisa Verdoliva. TruFor: Leveraging All-Round Clues for Trustworthy Image Forgery Detection and Localization. In *Proc. of the IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, 2023.

Yuncheng Guo, Junyan Ye, Chenjue Zhang, Hengrui Kang, Haohuan Fu, Conghui He, and Weijia Li. OmniAID: Decoupling Semantics and Artifacts for Universal AI-Generated Image Detection in the Wild. In *Forty-third International Conference on Machine Learning*, 2026.

Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum Contrast for Unsupervised Visual Representation Learning. In *Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9726–9735. IEEE, 2020.

Dimitrios Karageorgiou, Symeon Papadopoulos, Ioannis Kompatsiaris, and Efstratios Gavves. Any-resolution AI-Generated Image Detection by Spectral Learning. In *Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 18706–18717, 2025.

Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive Growing of GANs for Improved Quality, Stability, and Variation. In *Proc. of the International Conference on Learning Representations (ICLR)*, 2018.

Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. Supervised contrastive learning. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (eds.), *Proc. of Advances in Neural Information Processing Systems (NeurIPS)*, pp. 18661–18673, 2020.

Ivan Kobyzev, Simon J.D. Prince, and Marcus A. Brubaker. Normalizing Flows: An Introduction and Review of Current Methods. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(11):3964–3979, 2021.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO: Common Objects in Context. In *Proc. of the European Conference on Computer Vision (ECCV)*, pp. 740–755, 2014.

Francesco Marra et al. Do GANs Leave Artificial Fingerprints? In *2019 IEEE Conf. on Multimedia Information Processing and Retrieval (MIPR)*, pp. 506–511. IEEE, 2019.

Midjourney, Inc. Midjourney version 6.1. `https://updates.midjourney.com/version-6-1/`, 2024. Accessed: 2026-09-17.

João C. Neves, Ruben Tolosana, Ruben Vera-Rodriguez, Vasco Lopes, Hugo Proença, and Julian Fierrez. GANprintR: Improved Fakes and Evaluation of the State of the Art in Face Manipulation Detection. *IEEE Journal of Selected Topics in Signal Processing*, 14(5):1038–1048, 2020.

Tai D Nguyen, Aref Azizpour, and Matthew C Stamm. Forensic Self-Descriptions Are All You Need for Zero-Shot Detection, Open-Set Source Attribution, and Clustering of AI-Generated Images. In *Proc. of the IEEE/CVF Computer Vision and Pattern Recognition Conference (CVPR)*, pp. 3040–3050, 2025.

Alexander Quinn 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. In *Proc. of the International Conference on Machine Learning (ICML)*, 2022.

Utkarsh Ojha, Yuheng Li, and Yong Jae Lee. Towards Universal Fake Image Detectors that Generalize Across Generative Models. In *Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.

Maxime Oquab, Timothee Darcet, Theo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. DinoV2: Learning Robust Visual Features Without Supervision. *Transactions on Machine Learning Research Journal*, 2024.

Daeeol Park, Hyunsik Na, and Daeseon Choi. Performance Comparison and Visualization of AI-Generated-Image Detection Methods. *IEEE Access*, 12:62609–62627, 2024.

William Peebles and Saining Xie. Scalable Diffusion Models with Transformers. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 4172–4182, 2023.

Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. Sdxl: Improving Latent Diffusion Models for High-Resolution Image Synthesis. In *Proc of the International Conference on Learning Representations*, volume 2024, pp. 1862–1874, 2024.

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 *Proc. of the International Conference on Machine Learning*, pp. 8748–8763, 2021.

Robin Rombach et al. High-Resolution Image Synthesis With Latent Diffusion Models. In *Proc. IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, pp. 10684–10695, June 2022.

Axel Sauer et al. StyleGan-XL: Scaling Stylegan to Large Diverse Datasets. In *Proc. ACM SIGGRAPH Conf.*, pp. 1–10, 2022.

Oriane Simeoni, Huy V Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michael Ramamonjisoa, et al. DinoV3. *arXiv preprint arXiv:2508.10104*, 2025.

Sergey Sinitsa and Ohad Fried. Deep Image Fingerprint: Towards Low Budget Synthetic Image Detection and Model Lineage Analysis. In *Proc. IEEE/CVF Winter Conf. on Applications of Computer Vision (WACV)*, pp. 4067–4076, 2024.

Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. Roformer: Enhanced Transformer with Rotary Position Embedding. *Neurocomputing*, 568:127063, 2024.

Chuangchuang Tan, Yao Zhao, Shikui Wei, Guanghua Gu, and Yunchao Wei. Learning on Gradients: Generalized Artifacts Representation for GAN-Generated Images Detection. In *Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 12105–12114, 2023.

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. In *Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 28130–28139, 2024.

Ruben Tolosana et al. Deepfakes and Beyond: A Survey of Face Manipulation and Fake Detection. *Information Fusion*, 64:131–148, 2020.

Haohan Wang, Xindi Wu, Zeyi Huang, and Eric P Xing. High-Frequency Component Helps Explain the Generalization of Convolutional Neural Networks. In *Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8681–8691, 2020a.

Sheng-Yu Wang et al. CNN-Generated Images Are Surprisingly Easy to Spot... for Now. In *Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8695–8704, 2020b.

Zhendong Wang, Jianmin Bao, Wengang Zhou, Weilun Wang, Hezhen Hu, Hong Chen, and Houqiang Li. DIRE for Diffusion-Generated Image Detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 22445–22455, October 2023.

Tete Xiao, Mannat Singh, Eric Mintun, Trevor Darrell, Piotr Dollár, and Ross Girshick. Early Convolutions Help Transformers See Better. *Proc. of Advances in Neural Information Processing Systems (NeurIPS)*, 34:30392–30400, 2021.

Jiahao Xie, Wei Li, Xiaohang Zhan, Ziwei Liu, Yew Soon Ong, and Chen Change Loy. Masked frequency modeling for self-supervised visual pre-training. In *Proc. of the International Conference on Learning Representations (ICLR)*, 2023.

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. In *Proc. of the International Conference on Machine Learning*, 2025.

Jure Zbontar, Li Jing, Ishan Misra, Yann LeCun, and Stephane Deny. Barlow Twins: Self-Supervised Learning via Redundancy Reduction. In *Proc. of the International Conference on Machine Learning (ICML)*, 2021.

Kaiwen Zhang et al. DiffMorpher: Unleashing the Capability of Diffusion Models for Image Morphing. In *Proc. IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, pp. 7912–7921, 2024.

Yonggang Zhang, Jun Nie, Xinmei Tian, Mingming Gong, Kun Zhang, and Bo Han. Detecting Generated Images by Fitting Natural Image Distributions. In *Proc. of Advances of Neural Information Processing Systems (NeurIPS)*, 2025.

Yiming Zhao et al. UDiffText: A Unified Framework for High-Quality Text Synthesis in Arbitrary Images via Character-Aware Diffusion Models. In *Proc. European Conf. on Computer Vision (ECCV)*, pp. 217–233, 2024. ISBN 978-3-031-72750-4.

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. *Proc of the Conference on Advances in neural information processing systems (NeurIPS)*, 36:77771–77782, 2023.

Table 4: Global AUC (%) under JPEG re-encoding of the test images. GMM rows are averaged over the 27 generators; k-NN rows over the 25 generators never selected for calibration (Guid.D and LDM excluded). *Fakes*: labelled fake images used to fit the decision rule, all from a single generator. *Re-enc.*: fitting images also included re-encoded (GMM: QF $\sim \mathcal{U}[50, 100]$; k-NN: the same images at native, QF95, 80, 65, 50). k-NN rows: mean$_{\pm \text{sd}}$ over 10 draws of the calibration set. †Outside the calibrated range. Bottom: calibration size N (N real + N fake images, each at the 5 qualities); at $N < 400$, the generator selected in a draw is also excluded from that draw.

Decision rule	Fakes	Re-enc.	Native	QF95	QF80	QF65	QF50	QF35†
GMM (paper)	✗	✗	97.99	71.93	66.35	63.78	59.92	61.63
GMM	✗	✓	93.50	80.85	70.63	71.46	71.18	64.10
k-NN ($N = 400$)	✓	✗	**99.27**$_{\pm 0.09}$	79.34$_{\pm 2.48}$	71.86$_{\pm 4.64}$	56.88$_{\pm 6.24}$	63.05$_{\pm 6.25}$	56.84$_{\pm 10.6}$
k-NN ($N = 400$)	✓	✓	99.03$_{\pm 0.17}$	**94.26**$_{\pm 0.40}$	**95.26**$_{\pm 0.24}$	**94.00**$_{\pm 0.19}$	**92.22**$_{\pm 0.36}$	58.95$_{\pm 15.5}$
k-NN, fakes ✓, re-enc. ✓: calibration size								
$N = 10$			97.63$_{\pm 0.91}$	89.23$_{\pm 3.85}$	90.32$_{\pm 2.91}$	87.67$_{\pm 2.40}$	85.69$_{\pm 1.30}$	81.91$_{\pm 14.4}$
$N = 25$			98.49$_{\pm 0.41}$	91.67$_{\pm 1.49}$	91.24$_{\pm 5.32}$	89.91$_{\pm 2.28}$	88.16$_{\pm 1.17}$	84.11$_{\pm 4.16}$
$N = 50$			98.63$_{\pm 0.36}$	92.87$_{\pm 0.58}$	93.52$_{\pm 0.66}$	91.33$_{\pm 0.69}$	89.14$_{\pm 0.95}$	78.92$_{\pm 4.54}$
$N = 100$			98.89$_{\pm 0.31}$	93.89$_{\pm 0.34}$	94.29$_{\pm 0.66}$	92.73$_{\pm 0.47}$	90.41$_{\pm 0.61}$	72.95$_{\pm 13.1}$

A Compression Robustness Analysis

A limitation of one-class detection. Images shared online are routinely re-encoded. We re-encode every test image (real and fake) at a fixed quality factor (QF) and report global AUC (%) over the 25 generators never used for calibration (Table 4). Because *Forensic Twins* is one-class, its Gaussian Mixture Model (GMM) can only model the post-processing present in its real reference: re-encoding is an unseen shift, and AUC drops from 97.99 to 61–71. Adding the same real images re-encoded at QF $\sim \mathcal{U}[50, 100]$ to the fitting pool recovers part of the loss (64–81) without any fake sample.

The representation is not the bottleneck. To separate what the frozen encoder encodes from what the one-class rule can exploit, we calibrate a cosine k-NN on N real and N fake images of a *single* generator. The generator and k are selected by leave-one-generator-out validation on held-out images of the other candidates, never on test; the selected generators (Guid.D or LDM) are excluded from evaluation. All calibration images are disjoint from the test set. Labels alone do not help: fitted on native images, the k-NN reaches 99.27 without re-encoding but stays at 57–79 under JPEG. Fitting the same images at native, QF95, 80, 65 and 50 closes the gap: 92.2–95.3 at every in-range QF, with 99.03 on native images and a standard deviation below 0.4 over 10 draws. A handful of examples suffices: with $N = 10$, AUC is already 81.9–90.3 under re-encoding and $N = 25$ goes up to 84.1–91.6(Table 4, bottom).

Remaining limits. Calibration does not extrapolate reliably beyond the calibrated range (QF35: 58.95 ± 15.46), and commercial generators remain the hardest under strong compression (e.g. Midjourney v6 at QF50: 57.71). However, when the number of training samples is reduced, the extrapolation improves. For example, in a Few-Shot scenario with number of fitting samples N is 10 or 25, observe that robustness to high compressions emerge, specially with $N = 25$, where AUC under QF35 goes up to 84.11 ± 4.16. This pattern continues, since at $N = 100$ the performance in QF35 drops to 72.95 ± 13.1. We attribute this to a clear case of bias-variance trade-off, as the k-NN model over-fits to the training data as more samples are given.

B Detailed Image Source Clustering Analysis

Motivation. A representation can separate image sources for two very different reasons. It may encode the low-level traces that each generation pipeline leaves in the pixels, which is what a forensic detector is meant to capture. It may also simply encode *what* the images depict: generators differ not only in their artifacts but also in their content distribution (the datasets they were trained on, the prompts or classes used to sample them), so a purely semantic embedding can attribute sources through content alone. Source clustering and attribution accuracies cannot tell these two explana-

Table 5: Content-versus-source k-NN probe. Accuracy (%) of a cosine k-NN classifier on 7 ImageNet-1k superclasses (*semantic*) and on 7 generators (*source*), with 480 images per class, half used for fitting and half for evaluation. Mean $\pm$ standard deviation over 10 random splits; chance level is 14.29%. Δ is the 1-NN source accuracy minus the 1-NN semantic accuracy.

		Semantic (7 superclasses) $\downarrow$		**Source (7 generators)** $\uparrow$		
Method	**Dim.**	**1-NN**	**5-NN**	**1-NN**	**5-NN**	Δ
NPR	512	24.32 ± 0.66	26.57 ± 0.93	80.60 ± 0.88	81.04 ± 0.95	+56.28
FSD	960	34.17 ± 0.68	39.09 ± 1.32	80.65 ± 0.49	78.29 ± 0.71	+46.48
ConV	1024	99.35 ± 0.15	99.51 ± 0.12	78.35 ± 1.08	77.34 ± 0.72	-21.00
EFFORT	1024	99.21 ± 0.19	99.39 ± 0.10	97.27 ± 0.21	97.38 ± 0.32	-1.94
OmniAID	1024	99.29 ± 0.19	99.50 ± 0.11	95.15 ± 0.65	94.98 ± 0.55	-4.14
Ours	**192**	$\mathbf{39.03 \pm 0.98}$	$\mathbf{42.64 \pm 0.67}$	$\mathbf{95.57 \pm 0.58}$	$\mathbf{95.07 \pm 0.42}$	**+56.54**

tions apart. To disentangle them, we measure every representation twice with the same probe: once on a task that is *only* solvable through content, and once on a source-attribution task.

Protocol. For the semantic task we take seven coarse, visually unambiguous ImageNet-1k superclasses (bird, boat, car, cat, dog, furniture and snake), each formed by 12 ILSVRC synsets. All images come from the ILSVRC validation split, so none of them was seen during the pre-training of our encoder, which only uses the training split. Images whose shorter side is below 256 pixels are discarded up front, since our pipeline center-crops at 256×256; after this filter the smallest superclass retains 480 images, and every class is balanced to that size. The resulting list of 3,360 images is fixed and shared by all methods, so that every representation is evaluated on exactly the same pictures. For the source task we use seven generators spanning the three families of the benchmark: ProGAN and StyleGAN3 (GANs), GLIDE and Stable Diffusion 1.5 (diffusion) and Firefly, DALLE 3 and Midjourney v6 (commercial), again with 480 held-out test images per generator. Both tasks therefore have seven balanced classes and a chance level of $1/7 = 14.29\%$.

Each method is represented by the embedding immediately preceding its decision stage: the 192-dimensional output of our encoder, the 960-dimensional projected descriptor that FSD feeds to its GMM, the 1024-dimensional CLIP ViT-L/14 feature consumed by the EFFORT classifier, the 1024-dimensional pooled CLS token consumed by the OmniAID head (taken after the mixture-of-experts routing), and the 512-dimensional pooled feature preceding NPR's final linear layer. Embeddings are ℓ_2-normalised, and a k-nearest-neighbour classifier with cosine distance is fitted on half of the images of each class and evaluated on the other half. The k-NN probe has no trainable parameters, so its accuracy reflects how the embedding space is organised locally rather than what a classifier could learn on top of it. Because a single fit/probe split moves the semantic accuracy by up to one point, every value is the mean $\pm$ standard deviation over 10 random splits. The t-SNE panels in Figure 3 use one fixed split of the same images (perplexity 30, PCA initialisation, cosine metric) and are annotated with the 10-split 1-NN mean, so the figure and Table 5 report the same quantity.

Foundation-model detectors are semantic spaces. EFFORT and OmniAID classify the seven superclasses almost perfectly (99.21% and 99.29% with 1-NN), and their semantic t-SNE panels are organised by superclass, with every group well separated from the others (a few superclasses, such as bird and cat, split into two or three sub-clusters). This is expected: both are built on a CLIP ViT-L/14 backbone, and neither fine-tuning procedure removes the content structure that CLIP was trained to produce. Their source accuracy is also high (97.27% and 95.15%), but it is *lower* than their semantic accuracy ($\Delta = -1.94$ and -4.14). Since these embeddings encode content almost perfectly, and the seven generators differ in the content they depict, the probe cannot establish how much of their source separation is due to forensic evidence and how much to content. The source panels are consistent with a sizeable content component: in both methods Firefly does not form a single group but breaks into many small, scattered islands, the pattern one would expect if images were grouped by what they depict within each generator. For OmniAID in particular, the representation we probe is taken after the router has mixed in its semantic experts, so a strong content component is part of its design rather than a side effect. ConV illustrates that training on real images alone does not

suffice: its self-supervised DINOv2 space is the most semantic of all (99.35% 1-NN) and the least source-discriminative (78.35%), giving the most negative gap ($\Delta = -21.00$).

Forensic representations discard content. NPR and FSD show the opposite profile. Their semantic accuracies (24.32% and 34.17%) are much closer to chance, and their semantic t-SNE panels show no class structure: NPR produces a single intermixed cloud, and FSD produces several elongated structures in which all superclasses are mixed, so its geometry is driven by factors other than content. Both still attribute sources at about 80% (80.60% and 80.65%). For these methods the source signal is unlikely to be a by-product of content, since the embeddings retain little content to exploit. The price is a markedly lower attribution accuracy than that of the semantic detectors: neither reaches the 95–97% of EFFORT and OmniAID. Their source panels show where this loss comes from. In NPR, Firefly and Midjourney v6 occupy a separate region but overlap with each other, while the remaining five generators share a largely intermixed cloud; in FSD, Firefly and DALL·E 3 are isolated and Midjourney v6 is fragmented into many small groups, but GLIDE and Stable Diffusion 1.5 collapse, together with parts of ProGAN and StyleGAN3, into a single large cluster.

Our representation. Our encoder combines the two desirable properties. On the semantic task it stays at 39.03% (1-NN), much closer to the forensic detectors than to the foundation-based ones, and its semantic t-SNE panel is dominated by one large cloud in which all superclasses are mixed, with only a few small islands at its periphery. On the source task it reaches 95.57%, about 15 points above FSD and NPR and on par with OmniAID (95.15%), while its semantic accuracy is 60.26 points lower than OmniAID's. As a result it has the largest gap between source and semantic accuracy of all five methods ($\Delta = +56.54$), marginally above NPR ($+56.28$), but at an attribution accuracy about 15 points higher. In the source t-SNE, Firefly and Midjourney v6 form compact isolated groups, and the remaining confusions are limited to pairs of adjacent groups (DALL·E 3 with Stable Diffusion 1.5, and part of ProGAN with GLIDE). This structure arises even though no generated image, and no source label, was used at any point of training: it emerges from the Barlow Twins objective on real images only. This supports the interpretation that the embedding is organised by how an image was produced rather than by what it depicts, and it achieves this with a representation of only 192 dimensions, 2.7 to 5.3 times smaller than those of the state-of-the-art.

Residual content and robustness of the conclusions. Our semantic accuracy is clearly above chance (39.03% against 14.29%), so the representation is not fully content-invariant. Some dependence on content is to be expected from any pixel-level forensic cue, since low-level image statistics such as texture density, high-frequency energy or the amount of flat background covary with the depicted category (FSD, whose descriptors are explicitly residual-based, shows a comparable 34.17%). The conclusions do not depend on the number of neighbours: moving from 1-NN to 5-NN changes every source accuracy by less than 2.4 points and preserves the ordering of the methods, except for FSD and NPR, which are within 0.05 points of each other with 1-NN, and the semantic accuracies of the forensic representations increase by only 2.2–4.9 points, indicating that the little content structure they carry is diffuse rather than organised into clusters.

Local separability versus global clustering. The k-NN probe and the unsupervised clustering analysis of Section 4.2 answer different questions. The probe measures whether images from the same source are *neighbours*, whereas k-Means additionally requires each source to form a compact, roughly spherical cluster. Our representation attains 95.57% 1-NN source accuracy but a lower k-Means accuracy, which is consistent with its t-SNE panel: several sources are multi-modal (StyleGAN3 appears as three separate groups, and Midjourney v6 and ProGAN as two each), and some groups of different sources lie next to each other. Such sources are locally well separated, and therefore easy for a nearest-neighbour rule, but a k-Means partition with one centroid per source is forced to spend several centroids on the sub-clusters of a multi-modal source and, to compensate, to merge adjacent sources under a single centroid. This also explains why the clustering accuracy of our representation rises sharply when k-Means is allowed more centroids than sources. The t-SNE visualisations should, in general, be read qualitatively; distances between clusters in a t-SNE embedding are not meaningful, which is why all quantitative claims in this section rely on the k-NN probe computed in the original embedding space.

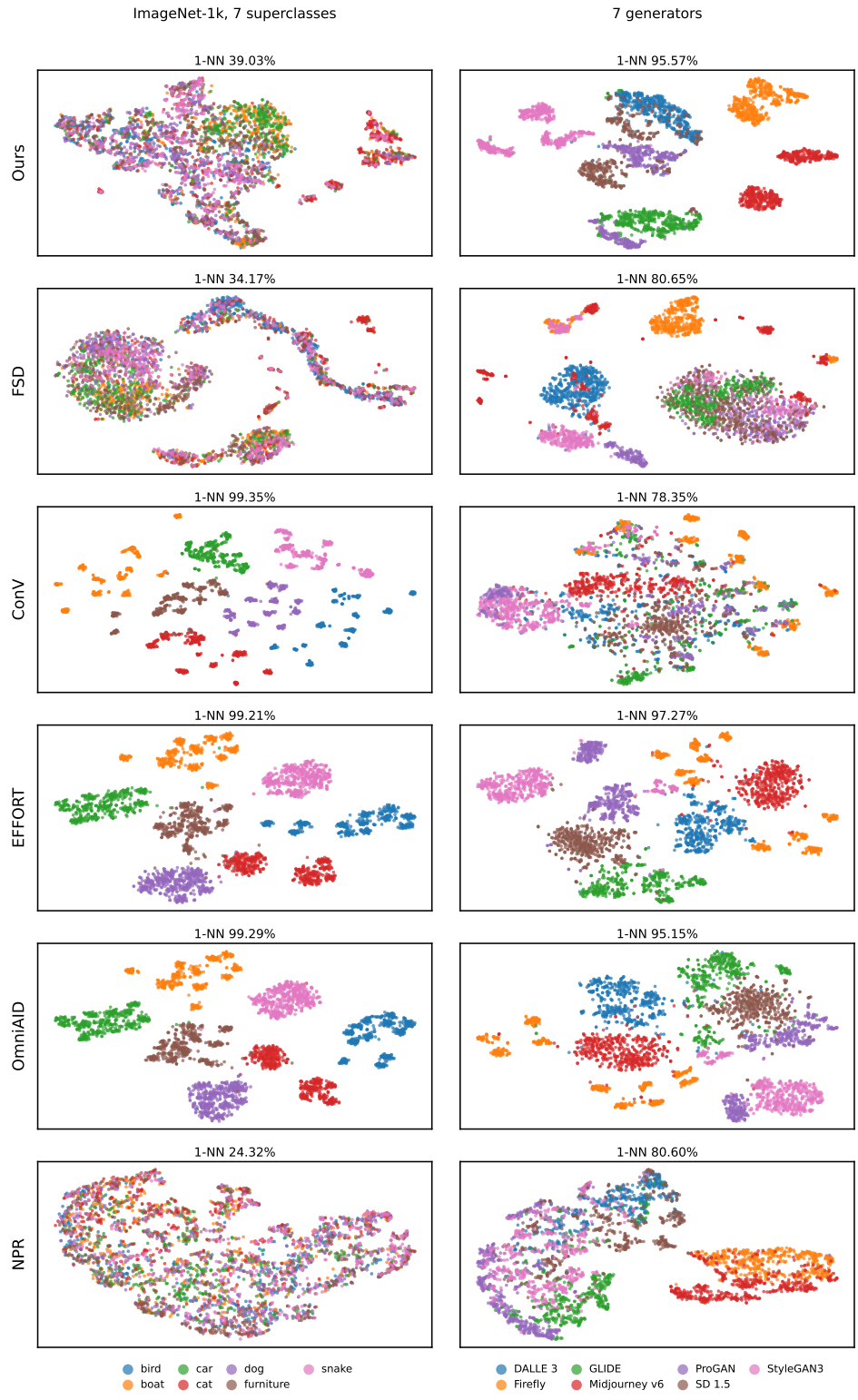

Figure 3: t-SNE comparison between 7 semantic classes from the ImageNet-1k dataset (left) and images from different generators, across GANs, Diffusion Models and Commercial Models (right) from *Forensic Twins* and different methods considered in Sec. 4.2.

Table 6: Computational cost of inference, measured on a fixed 500-image sample of ImageNet-1k – the same images for every column. **GFLOPs** and **latency** start from an input tensor that is already decoded and already preprocessed by each method's own pipeline, so they isolate the detector from file I/O. Latency is a median over the 500 images on one RTX 4090; memory is the peak GPU footprint over the model's whole lifetime.

Method	Inference Type	#Params	GFLOPs ↓	Latency (ms) ↓	GPU (MB) ↓
CNNDet[CVPR20]	Forward (native res.)	23.5M	31.4	3.09	2230
UFD[CVPR23]	Forward	304.0M	162.0	8.28	2176
LGrad[CVPR23]	Forward + backward	46.6M	102.0	10.43	1006
SPAI[CVPR25]	Forward (native res.)	139.9M	656.8	26.49	8316
NPR[CVPR24]	Forward (native res.)	1.4M	12.9	2.17	2130
FSD[CVPR25]	Per-image solve	6.2M	2537.4	2415.27	12038
ConV[NeurIPS25]	Forward (10 crops)	304.4M	1620.2	16.82	1176
EFFORT[ICML25]	Forward	504.6M	162.2	12.76	2448
OmniAID[ICLR26]	Forward	508.8M	765.7	40.25	3672
Ours	**Forward**	**3.0M**	**10.7**	**6.44**	**516**

C Computational Cost

Computational Cost Assessment. Table 6 reports the inference cost of every detector. In terms of parameters, ForensicTwins (3.0M) is the second smallest model, behind only NPR (1.4M) and ahead of FSD (6.2M). It is at least $100\times$ smaller than the detectors built on foundation backbones (UFD, ConV, EFFORT and OmniAID, 304–509M parameters), and it is trained on fewer than 0.5M real images, yet it outperforms UFD and OmniAID and remains within 0.96 AUC points of OmniAID and EFFORT (Table 1). In terms of latency, NPR (2.17 ms) and CNNDet (3.09 ms) are the fastest detectors, followed by ForensicTwins (6.44 ms), which is $1.3\times$ faster than UFD, $2\times$ faster than EFFORT, $2.6\times$ faster than ConV and $6.3\times$ faster than OmniAID. ForensicTwins has the lowest cost of all detectors in GFLOPs (10.7, against 12.9 for NPR) and in GPU memory (516 MB, about half that of the next most frugal model, LGrad, with 1006 MB). FSD is the clear outlier: its per-image solve makes it the most expensive detector in GFLOPs (2537.4), latency (2415.27 ms, $375\times$ slower than ForensicTwins) and memory (12038 MB), which severely limits its use in production settings. Overall, *Forensic Twins* is the only detector that ranks among the three cheapest on all four cost metrics.

D Detailed Results on Zero-Shot Detection

Protocol. Tables 7 and 8 break down the category averages of Table 1 by real-image distribution. Each cell is the AUC of one detector on a balanced set of real images from the stated distribution against images from a single generator. Block averages are flat means over the generators in the block, and the global figures in Table 1 are the mean of the 27-family averages of the two tables. AUC values are reported as measured and are never folded: a score below 50 means that the detector systematically ranks generated images as *more* real than real ones, which is a failure mode in its own right rather than an absence of signal.

The real reference distribution changes the ranking. Most detectors are markedly easier to satisfy when the real images come from MS-COCO. Averaged over the 27 families, CNNDet gains 8.86 points when moving from ImageNet-1k to MS-COCO, UFD 4.03, SPAI 3.06, NPR 2.87 and our method 2.10 (96.94,→,99.04). The effect is not uniform in sign: EFFORT loses 0.73, OmniAID 1.19, LGrad 3.80 and ConV 12.73, and FSD is nearly invariant (+0.29). As a consequence the ordering of the strongest methods depends on which real distribution is used. Against ImageNet-1k the top four are EFFORT (99.32), FSD (98.38), OmniAID (98.27) and ours (96.94); against MS-COCO they are ours (99.04), FSD (98.67), EFFORT (98.59) and OmniAID (97.08). A benchmark built on a single real distribution would therefore report a ranking that is partly an artefact of that choice, which is why the main table averages both.

Table 7: Comprehensive Zero-Shot Detection Performance (AUC %) across 27 distinct generative architectures using ImageNet-1k as real images. Results are evaluated using official pre-trained weights on identical test splits. Best result per column is underlined.

Generative Adversarial Networks (GANs)

Method	ProGAN	SG2	SG3	BigGAN	StarGAN	EG3D	GauGAN	GigaGAN	Avg.
CNNDet Wang et al. (2020b)	100.00	96.21	94.54	87.07	93.22	95.70	98.14	72.61	92.19
UFD Ojha et al. (2023)	99.32	95.03	95.29	96.64	96.98	97.33	99.97	95.13	96.96
LGrad Tan et al. (2023)	96.05	89.36	91.91	63.72	48.99	84.49	84.73	78.55	79.73
NPR Tan et al. (2024)	99.58	99.51	99.53	82.95	58.64	99.60	99.18	98.20	92.15
SPAI Karageorgiou et al. (2025)	90.86	94.28	96.32	89.04	97.97	98.43	89.99	89.61	93.31
EFFORT Yan et al. (2025)	99.97	99.87	99.92	99.36	99.74	99.76	99.81	99.45	99.73
OmniAID Guo et al. (2026)	100.00	99.96	99.95	99.17	99.52	99.99	99.98	98.75	99.66
ConV Zhang et al. (2025)	96.23	87.80	89.01	93.93	68.21	83.13	97.71	91.72	88.47
FSD Nguyen et al. (2025)	99.87	99.80	99.83	99.89	99.68	99.79	99.91	99.88	99.83
Ours	**97.66**	**99.18**	**99.49**	**94.47**	**91.79**	**98.95**	**97.79**	**98.87**	**97.27**

Diffusion & Autoregressive Models

Method	DDPM	Guid.D	LDM	SD1	SD2	SD3	SDXL	Flux	GLIDE	DeepF	DiT	DiffG	LSGM	TamTr	Avg.
CNNDet Wang et al. (2020b)	59.44	45.57	57.68	60.79	48.17	63.27	76.65	50.04	57.18	69.63	58.08	72.53	61.13	60.36	60.04
UFD Ojha et al. (2023)	92.86	79.14	89.69	96.74	78.69	36.89	75.93	46.73	85.73	90.20	77.35	92.72	95.36	97.47	81.11
LGrad Tan et al. (2023)	27.70	41.21	76.75	61.64	68.85	33.59	90.43	64.72	82.64	74.05	57.10	71.12	89.84	73.13	65.20
NPR Tan et al. (2024)	83.86	91.02	98.87	95.37	97.55	63.71	99.60	89.22	99.16	92.02	97.61	99.32	99.59	95.74	93.05
SPAI Karageorgiou et al. (2025)	98.68	76.65	99.67	99.36	97.99	83.24	98.44	85.23	94.43	97.03	97.13	97.20	99.70	86.10	93.63
EFFORT Yan et al. (2025)	98.97	96.69	99.28	99.97	99.91	99.61	99.94	99.51	99.69	99.90	95.61	99.77	99.99	99.04	99.13
OmniAID Guo et al. (2026)	99.13	86.28	99.02	100.00	99.84	99.17	98.01	97.30	99.89	98.30	96.76	100.00	99.99	96.85	97.89
ConV Zhang et al. (2025)	65.66	76.58	85.54	96.57	89.55	84.52	94.34	68.62	90.53	96.44	44.65	96.15	63.06	96.00	82.01
FSD Nguyen et al. (2025)	99.74	99.78	99.74	99.85	99.49	99.35	99.58	94.95	99.95	99.50	99.80	99.92	99.72	99.65	99.36
Ours	**90.10**	**95.59**	**98.79**	**98.82**	**98.68**	**97.86**	**98.92**	**98.46**	**99.85**	**98.72**	**98.74**	**95.30**	**98.15**	**99.15**	**97.65**

Commercial / Black-Box APIs

Method	Midjourney v6	Firefly	DALLE-2	DALLE-3	DALLE-M	Avg.
CNNDet Wang et al. (2020b)	58.44	59.21	82.93	56.53	55.76	62.57
UFD Ojha et al. (2023)	63.35	72.97	95.58	82.99	95.82	82.14
LGrad Tan et al. (2023)	25.13	9.51	81.95	89.94	72.68	55.84
NPR Tan et al. (2024)	26.50	35.95	99.44	99.17	97.74	71.76
SPAI Karageorgiou et al. (2025)	88.17	83.32	95.29	98.98	97.39	92.63
EFFORT Yan et al. (2025)	97.65	98.76	99.74	99.93	99.88	99.19
OmniAID Guo et al. (2026)	97.58	91.76	97.35	99.70	99.01	97.08
ConV Zhang et al. (2025)	71.15	50.19	97.75	91.30	87.94	79.67
FSD Nguyen et al. (2025)	94.26	73.78	99.25	99.48	99.77	93.31
Ours	**86.53**	**88.24**	**99.66**	**98.76**	**98.87**	**94.41**

Large-scale supervised detectors. EFFORT is the most uniform detector in the benchmark: its worst family is DiT (95.61 on ImageNet-1k, 93.31 on MS-COCO) and its per-family standard deviation is 1.06 and 1.78 points. It is also the only method that does not show the commercial-model gap, with a commercial average of 99.19 and 98.16. OmniAID is close to it on GANs (99.66 and 99.36), but has a clear weak point on Guided Diffusion (86.28 and 81.72) and on Firefly (91.76 and 85.85). SPAI sits in between: it has no catastrophic failure (minimum 76.65 and 84.37, both on Guided Diffusion) and similar averages across the three categories, but it rarely approaches saturation.

Comparison with the other zero-shot method. FSD obtains the highest GAN and diffusion averages in both tables (99.83/99.36 on ImageNet-1k, 99.92/99.57 on MS-COCO), but it has a single, severe failure on Firefly (73.78 and 75.96) and a secondary drop on Flux (94.95 and 95.92). Our method is 14.46 and 19.73 points better on Firefly and 3.51 and 3.89 points better on Flux. As a result our method obtains the higher commercial average on both real distributions (94.41 vs. 93.31 on ImageNet-1k, 97.78 vs. 94.13 on MS-COCO). Against MS-COCO our method obtains the best score of all nine detectors on ten of the 27 families, including Flux (99.81) and GLIDE (100.00), and it lies within one point of the best detector on 21 of the 27 families; against ImageNet-1k this holds f or 6 families. These results are obtained with roughly half of FSD's parameters (3.0M vs. 6.2M) and at a latency about 375 times lower (6.44,ms vs. 2415.27,ms per image, Appendix C).

Table 8: Comprehensive Zero-Shot Detection Performance (AUC %) across 27 distinct generative architectures using MS-COCO as real images. Results are evaluated using official pre-trained weights on identical test splits. Best result per column is underlined.

Generative Adversarial Networks (GANs)									
Method	**ProGAN**	**SG2**	**SG3**	**BigGAN**	**StarGAN**	**EG3D**	**GauGAN**	**GigaGAN**	**Avg.**
CNNDet Wang et al. (2020b)	100.00	97.93	97.15	92.09	96.49	98.12	99.09	82.30	95.40
UFD Ojha et al. (2023)	99.78	97.52	97.51	98.16	98.47	99.08	100.00	97.42	98.49
LGrad Tan et al. (2023)	95.87	87.81	90.89	59.54	45.52	81.86	82.02	74.96	77.31
NPR Tan et al. (2024)	99.84	99.83	99.83	85.96	63.73	99.85	99.75	99.30	93.51
SPAI Karageorgiou et al. (2025)	95.01	97.34	98.48	93.78	99.39	99.56	94.58	94.32	96.56
EFFORT Yan et al. (2025)	99.86	99.67	99.77	98.54	99.17	99.45	99.34	98.86	99.33
OmniAID Guo et al. (2026)	99.96	99.84	99.82	98.52	98.96	99.90	99.93	97.93	99.36
ConV Zhang et al. (2025)	91.25	75.48	79.49	87.55	47.25	63.34	94.50	84.30	77.89
FSD Nguyen et al. (2025)	99.94	99.91	99.92	99.95	99.84	99.90	99.95	99.94	99.92
Ours	**99.56**	**99.91**	**99.95**	**97.43**	**96.19**	**99.88**	**99.68**	**99.85**	**99.06**

Diffusion & Autoregressive Models															
Method	**DDPM**	**Guid.D**	**LDM**	**SD1**	**SD2**	**SD3**	**SDXL**	**Flux**	**GLIDE**	**DeepF**	**DiT**	**DiffG**	**LSGM**	**TamTr**	**Avg.**
CNNDet Wang et al. (2020b)	71.54	57.02	69.22	71.75	60.48	75.73	86.21	61.51	67.92	81.07	68.96	81.69	72.54	71.31	71.21
UFD Ojha et al. (2023)	96.38	85.85	93.85	98.38	85.13	44.83	83.14	54.88	91.14	94.52	84.58	95.98	97.80	98.91	86.10
LGrad Tan et al. (2023)	20.07	35.18	72.92	55.85	63.95	25.77	89.05	60.60	80.82	69.74	50.67	66.06	88.37	69.25	60.59
NPR Tan et al. (2024)	85.47	95.57	99.56	97.77	99.07	78.11	99.85	93.74	99.67	96.48	99.08	99.76	99.85	97.90	95.85
SPAI Karageorgiou et al. (2025)	99.68	84.37	99.92	99.84	99.19	91.73	99.56	92.48	97.37	98.60	98.79	98.72	99.94	88.84	96.36
EFFORT Yan et al. (2025)	97.32	94.08	98.59	99.88	99.76	98.81	99.81	98.72	99.14	99.67	93.31	99.30	99.95	98.19	98.32
OmniAID Guo et al. (2026)	98.14	81.72	98.39	99.98	99.60	98.27	95.80	94.90	99.74	96.60	94.74	99.97	99.88	94.61	96.60
ConV Zhang et al. (2025)	44.35	59.38	73.63	91.13	76.37	66.30	85.43	47.28	79.26	90.44	26.19	89.87	40.83	90.66	68.65
FSD Nguyen et al. (2025)	99.87	99.89	99.88	99.92	99.69	99.73	99.82	95.92	99.97	99.78	99.90	99.96	99.86	99.84	99.57
Ours	**96.59**	**98.72**	**99.85**	**99.86**	**99.84**	**99.71**	**99.88**	**99.81**	**100.00**	**99.84**	**99.84**	**99.04**	**99.75**	**99.89**	**99.47**

Commercial / Black-Box APIs						
Method	**Midjourney v6**	**Firefly**	**DALLE-2**	**DALLE-3**	**DALLE-M**	**Avg.**
CNNDet Wang et al. (2020b)	71.03	73.99	90.61	68.68	65.69	74.00
UFD Ojha et al. (2023)	70.98	82.02	97.75	88.74	97.78	87.45
LGrad Tan et al. (2023)	19.14	5.31	79.24	88.49	68.25	52.08
NPR Tan et al. (2024)	37.57	49.81	99.80	99.74	99.27	77.24
SPAI Karageorgiou et al. (2025)	93.74	91.06	98.28	99.72	98.88	96.34
EFFORT Yan et al. (2025)	94.99	97.10	99.29	99.76	99.68	98.16
OmniAID Guo et al. (2026)	95.41	85.85	95.06	99.33	98.19	94.77
ConV Zhang et al. (2025)	49.77	28.64	94.04	78.52	75.28	65.25
FSD Nguyen et al. (2025)	95.31	75.96	99.70	99.78	99.89	94.13
Ours	**93.68**	**95.69**	**99.94**	**99.78**	**99.82**	**97.78**

Worst-case behaviour. Averages hide the failures that matter most in deployment, so we also inspect the minimum AUC over the 27 families. Against ImageNet-1k the minima are EFFORT 95.61, ours 86.53, OmniAID 86.28, SPAI 76.65, FSD 73.78, CNNDet 45.57, UFD 36.89, NPR 26.50 and LGrad 9.51. Against MS-COCO they are ours 93.68, EFFORT 93.31, SPAI 84.37, OmniAID 81.72, FSD 75.96, CNNDet 57.02, UFD 44.83, NPR 37.57 and LGrad 5.31. Our method is thus the only detector besides EFFORT with no family below 90 against MS-COCO, where it also attains the highest minimum of the benchmark, and its per-family standard deviation (3.62 on ImageNet-1k, 1.64 on MS-COCO) is the lowest of all nine detectors on MS-COCO. Notably, this worst-case profile is obtained without any generated image during training, whereas EFFORT and OmniAID are trained with generated images.

Limitations of our method. Our weakest families are consistent across the two real distributions: Midjourney (86.53 and 93.68), Firefly (88.24 and 95.69), DDPM (90.10 and 96.59), StarGAN (91.79 and 96.19), BigGAN (94.47 and 97.43) and Diffusion-GAN (95.30 and 99.04). Native resolution alone does not explain this group. DDPM and StarGAN are generated at 256×256, but so are LDM, GLIDE, LSGM and Taming Transformers, on which our method scores at least 98.15 on ImageNet-1k and 99.75 on MS-COCO. On ImageNet-1k, EFFORT, FSD and OmniAID score higher than our method on 24, 24 and 19 of the 27 families respectively; closing this gap while still training on real images only is the main open direction for this line of work.

E Robustness Against Unseen Real Images

Since the GMM is fitted on real image embeddings alone, its coverage of the test-time real domain matters. We retrain the encoder on a single real source and fit the GMM on that same source alone, so the other source is never observed at any stage, and we then evaluate both real axes. Against the main model —96.94% AUC on ImageNet and 99.04% on MS-COCO— the ImageNet-only scenario (500k images) drops the in-distribution ImageNet axis to 95.36% and the now-unseen MS-COCO axis to 93.44%, for a global mean of 94.40%. The MS-COCO-only scenario behaves differently despite training on 242k images, less than half the corpus: the unseen ImageNet axis gives up only 1.39% at 95.55%, while its own axis rises to 99.47%, for a global mean of 97.51%. Detection therefore survives a wholly unseen real domain in both directions, but the cost is markedly asymmetric: removing MS-COCO costs 3.59% globally, whereas removing ImageNet costs 0.48% with less than half the training images.

F Detailed Breakdown on Generators

Overview. The detection benchmark comprises 27 generator families and two real-image distributions. Table 9 describes every family as it enters the evaluation: the architecture and conditioning of the generator, the content domain it was trained on or prompted with, the resolution and file format of the evaluated images, and the number of images used. A family groups all the checkpoints and configurations of the same generator. For example, StyleGAN3 includes both the translation- (T) and rotation-equivariant (R) configurations trained on FFHQ, FFHQ-U and AFHQv2 at three resolutions. Grouping at this level means that each AUC in Tables 7 and 8 measures generalisation to a generator rather than to one particular checkpoint. Images are drawn from three public collections of synthetic images, DMID (Corvi et al., 2023b), SynthBuster (Bammey, 2023) and OSSIA (Fang et al., 2023); the sub-sets merged into each family are listed in the last column.

Image selection. For each family we use a fixed test split and evaluate its first 1,000 images, or all of them when fewer are available (Firefly, 800 images) (Nguyen et al., 2025). The same list of images is used for every detector, so all methods are scored on identical generated images. Real images follow the same rule. ImageNet-1k images are taken from the ILSVRC2012 validation split, which is disjoint from the training split used to pre-train our encoder, and MS-COCO images from the 2017 test split. Each generator is paired with the first N real images, where N is the number of generated images of that family, so every AUC is computed on a balanced set. No image is resized, cropped or re-encoded before it reaches a detector's own pre-processing pipeline. Because our pipeline center-crops at 256×256, real images whose shorter side is below 256 pixels are skipped by our method and replaced by the next ones in the list; this affects 26 of the 1,000 ImageNet-1k images and 7 of the 1,000 MS-COCO images. All generated images are at least 256 pixels on their shorter side, so the generated sets are unaffected.

Diversity of the benchmark. The families cover the three main generation paradigms (adversarial, diffusion and autoregressive) and the main forms of conditioning: unconditional sampling, class labels, semantic layouts, image-to-image translation and free-form text. Content ranges from narrow, aligned domains such as FFHQ and CelebA-HQ faces, through LSUN scene categories and the 1,000 ImageNet classes, to the unrestricted prompts of the commercial systems. Resolutions range from 256×256 to 2048 pixels on the shorter side, and several families mix resolutions within themselves (ProGAN, StyleGAN2, StyleGAN3, BigGAN, GigaGAN, Stable Diffusion 2.1, DiT, Flux, Midjourney), so a detector cannot rely on a single fixed image size to recognise a generator. Finally, a large part of the benchmark is conditioned on MS-COCO annotations (GauGAN layouts and, as captions, GigaGAN, LDM, Stable Diffusion 1 and 2.1, SDXL, GLIDE, DeepFloyd IF and the three DALL·E models). For these families the generated images depict the same kind of scenes as the MS-COCO real images, so the MS-COCO comparison cannot be solved on content alone, whereas the ImageNet-1k comparison additionally differs in content; this is one more reason to report both real distributions.

Table 9: Generator families of the detection benchmark. *Uncond.*: unconditional sampling; *class*: class-conditional; *text*: text-to-image; *layout*: semantic-layout-to-image; *i2i*: image-to-image translation. *COCO captions* denotes prompts taken from MS-COCO validation captions. Resolution is width×height of the evaluated images; the proportion of each resolution or content domain within the family is given in parentheses. N is the number of evaluated images. n/s: content not specified by the source collection.

Family	Conditioning	Content domain	Resolution	N	Sources
Generative Adversarial Networks					
ProGAN	uncond.	LSUN (67%), CelebA-HQ (33%)	256^2 (67%), 1024^2 (33%)	1000	DMID, SynthBuster, OSSIA
StyleGAN2	uncond.	FFHQ (62%), LSUN Dog (16%), AFHQv2 (13%), LSUN Church (9%)	1024^2 (45%), 256^2 (42%), 512^2 (13%)	1000	DMID, SynthBuster, OSSIA (incl. StyleGAN2-ADA)
StyleGAN3	uncond.	FFHQ / FFHQ-U (84%), AFHQv2 (16%); R and T configs	1024^2 (56%), 256^2 (28%), 512^2 (15%)	1000	DMID, SynthBuster, OSSIA
BigGAN	class / uncond.	ImageNet (79%), CelebA (21%, U-Net BigGAN)	256^2 (60%), 512^2 (40%)	1000	DMID, SynthBuster, OSSIA
StarGAN	i2i	Face attribute / style translation, CelebA-HQ (StarGAN, StarGAN v2)	256^2	1000	SynthBuster, OSSIA
EG3D	uncond.	3D-aware synthesis (n/s)	512^2	1000	DMID
GauGAN	layout	MS-COCO semantic layouts	256^2	1000	SynthBuster
GigaGAN	class / text	ImageNet (50%), COCO captions (50%)	256^2 (50%), 512^2 (50%)	1000	SynthBuster
Diffusion and autoregressive models					
DDPM	uncond.	CelebA-HQ (with and without EMA weights)	256^2	1000	SynthBuster, OSSIA
Guided Diff. (ADM)	class / uncond.	ImageNet (29%), LSUN Horses / Cats / Bedrooms (71%)	256^2	1000	DMID, SynthBuster
LDM	text / class / uncond.	FFHQ (24%), COCO captions (20%), ImageNet (17%), LSUN Bedrooms (14%), LSUN Churches (14%), CelebA-HQ (10%)	256^2	1000	DMID, SynthBuster, OSSIA
Stable Diffusion 1	text	COCO captions	256^2	1000	DMID, SynthBuster
Stable Diffusion 2.1	text	COCO captions	512^2 (40%), 256^2 (30%), 768^2 (30%)	1000	SynthBuster
Stable Diffusion 3	text	n/s	1024^2	1000	OSSIA
SDXL	text	COCO captions	1024^2	1000	SynthBuster
Flux	text	n/s; varied samplers, guidance scales and steps	1024×768, 768×1024, 1024^2	1000	Online
GLIDE	text	COCO captions	256^2	1000	DMID, SynthBuster
DeepFloyd IF	text	COCO captions (stage III output)	1024^2	1000	SynthBuster
DiT	class	ImageNet	256^2 (50%), 512^2 (50%)	1000	SynthBuster
Diffusion-GAN	uncond.	LSUN Bedrooms (51%), LSUN Churches (49%)	256^2	1000	SynthBuster
LSGM	uncond.	CelebA-HQ (two sampling settings)	256^2	1000	OSSIA
Taming Transformers	uncond. / class / layout	FFHQ (50%), semantic layouts (26%), ImageNet (24%)	256^2	1000	DMID
Commercial / black-box APIs					
Midjourney v6	text	n/s	Variable aspect ratio, shorter side 260–2048	1000	Online
Adobe Firefly	text	Six prompt categories: artistic, biological, geometry, lighting, colour, material	512×427 (80%), 512^2 (15%), 672×299 (5%)	800	Online
DALL·E 2	text	COCO captions	1024^2	1000	DMID, SynthBuster
DALL·E 3	text	COCO captions	1024^2	1000	SynthBuster
DALL·E Mini	text	COCO captions	256^2	1000	DMID
Real images					
ImageNet-1k	–	ILSVRC2012 validation split	Variable, median shorter side 375	1000	ILSVRC2012
MS-COCO	–	MS-COCO 2017 test split	Variable, median shorter side 428	1000	MS-COCO 2017