Forensic Self-Descriptions Are All You Need for Zero-Shot Detection, Open-Set Source Attribution, and Clustering of AI-generated Images

Tai D. Nguyen, Aref Azizpour, Matthew C. Stamm
Drexel University
Philadelphia, PA, USA
tdn47,aa4639,mcs382@drexel.edu

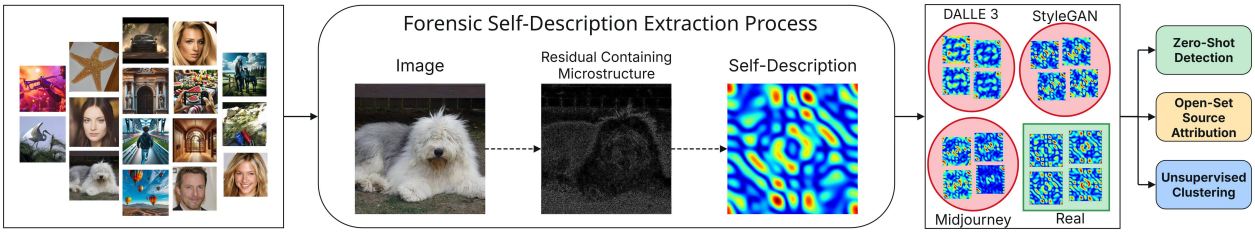

Figure 1. We extract the self-description of the forensic microstructures in each image and use them to accurately perform a variety of challenging tasks, including: zero-shot detection, open-set source attribution, and unsupervised clustering of image sources.

Abstract

The emergence of advanced AI-based tools to generate realistic images poses significant challenges for forensic detection and source attribution, especially as new generative techniques appear rapidly. Traditional methods often fail to generalize to unseen generators due to reliance on features specific to known sources during training. To address this problem, we propose a novel approach that explicitly models forensic microstructures—subtle, pixel-level patterns unique to the image creation process. Using only real images in a self-supervised manner, we learn a set of diverse predictive filters to extract residuals that capture different aspects of these microstructures. By jointly modeling these residuals across multiple scales, we obtain a compact model whose parameters constitute a unique forensic self-description for each image. This self-description enables us to perform zero-shot detection of synthetic images, open-set source attribution of images, and clustering based on source without prior knowledge. Extensive experiments demonstrate that our method achieves superior accuracy and adaptability compared to competing techniques, advancing the state of the art in synthetic media forensics.

1. Introduction

The rapid improvement in AI-generated image quality has made synthetic images increasingly difficult to distinguish from real ones [26, 68]. While traditional detection methods can be trained to identify these images, they struggle to generalize to content produced by new or unseen generators. As new generative models emerge at a rapid pace, there is an urgent need for detection methods that can reliably identify images from novel sources without prior exposure [40, 55].

Conventional approaches to synthetic image detection and source attribution typically rely on learning embeddings that are discriminative between real and synthetic images, or between real and a number of specific synthetic sources [14, 29, 46, 65, 70]. While these methods are effective for sources similar to those in training, they often fail to adapt to new generative models [19, 54]. This occurs because their objective functions tend to make them learn features that are only useful to discriminate between known sources in the training data. Consequently, these methods often overlook features that would be critical for identifying images from new, unseen generators.

To address this problem, we propose an alternative approach (as illustrated in Fig. 1 and detailed in Fig. 3) that is both more effective and general for detecting synthetic images and attributing them to their source. Instead of learning a discriminative embedding space, we focus on explicitly modeling the forensic microstructures embedded in images. It is well-established that both cameras and synthetic image generators imprint unique forensic traces in the form of statistical microstructures—subtle, pixel-level relationships that can serve as identifying features [45, 47, 73, 74]. To isolate these microstructures from the image content, relying on only real images, we employ a self-supervised process that learns a set of diverse predictive filters to approximate the scene content. By applying these filters, we obtain multiple distinct residuals, each captures a different as-

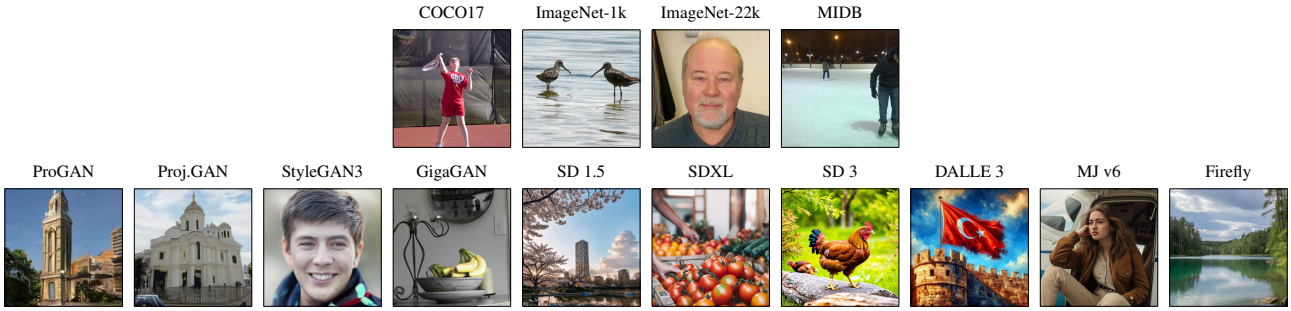

Figure 2. Visualization of real (top row) and synthetic (bottom row) images in the datasets used in this paper.

pect of the forensic microstructures. We then jointly model these residuals across multiple scales using a compact parametric model, whose parameters constitute a unique **forensic self-description** for each image. This self-description effectively encapsulates the intrinsic forensic properties of an image, allowing us to perform several challenging tasks: **(1)** zero-shot detection of synthetic images, **(2)** attribute images to their source generators in an open-set manner, and **(3)** cluster images based on their sources without any prior knowledge of the generators involved.

Through extensive experiments and ablation studies, we demonstrate that our method achieves high accuracy in zero-shot detection, open-set source attribution, and clustering, consistently outperforming competing techniques in robustness and adaptability.

Our main contributions are summarized as follows:

1. We introduce forensic self-descriptions as a way to capture intrinsic properties of the forensic microstructures in an image. We then use these descriptions to accurately perform several critical tasks related to detecting and attributing the source of synthetic images.
2. We demonstrate that these forensic self-descriptions enable accurate zero-shot detection of synthetic images without ever seeing them.
3. We show that forensic self-descriptions are also well-suited to perform open-set attribution and clustering, allowing precise source identification and organization of images from unknown generators.
4. We provide comprehensive experimental validation, highlighting the robustness and generalizability of our approach across a broad set of real and synthetic sources.

2. Background and Related Work

The rise of realistic AI-generated images has posed significant challenges for detection and source attribution, prompting the development of supervised, open-set, and zero-shot approaches.

Forensic Microstructures. It is well-established that different design choices in a generator's neural architecture induce specific statistical microstructures into AI generated images [19, 68, 73]. Leveraging this, researchers initially built handcrafted filters or explicit mathematical models to extract these microstructures for detecting synthetic images [7, 18, 20, 43, 51]. However, recent approaches often leverage CNNs to learn these models from data, enabling more generalized detection systems.

Supervised Methods. Supervised methods to detect synthetic images [5, 13, 47, 65, 70, 72] often train their models on binary labeled datasets. While these methods perform well on data sources similar to those in the training set, prior work has shown that they struggle with images from unseen generative models [19, 54]. This is because their learned features are specific to the training data, and may not capture the unique artifacts of new generators [40, 55].

Open-Set Source Attribution. To address the limitations of supervised methods, researchers have recently explored adapting open-set recognition techniques developed from other computer vision areas [10, 16, 39, 52, 76] to synthetic image source attribution. Notable works are POSE [71], Fang et al. [28], and Abady et al [1]. While these methods have better generalization than supervised ones, they still heavily rely on feature representations learned from known sources, which may not generalize well to unseen ones.

Zero-Shot Detection. Recent work has developed approaches to detect synthetic images without requiring exposure to specific generative models. These methods typically rely on non-forensic features that differ between real and synthetic images. For instance, some methods [22, 59] use autoencoders (i.e., diffusion model, image compression network) for reconstruction error analysis, while others [54, 63] leverage CLIP embeddings to detect inconsistencies in general visual features. Few others [66, 67] use a limited set of forensic features for generalized detection.

While promising, as we show later, these zero-shot methods often yield inconsistent performance, which varies depending on the real-vs-synthetic dataset pairs used for benchmarking. This variability arises because non-forensic features may be influenced by the specific content characteristics of the datasets. Furthermore, there is no guarantee that these features will remain effective as generative technologies continue to improve and evolve.

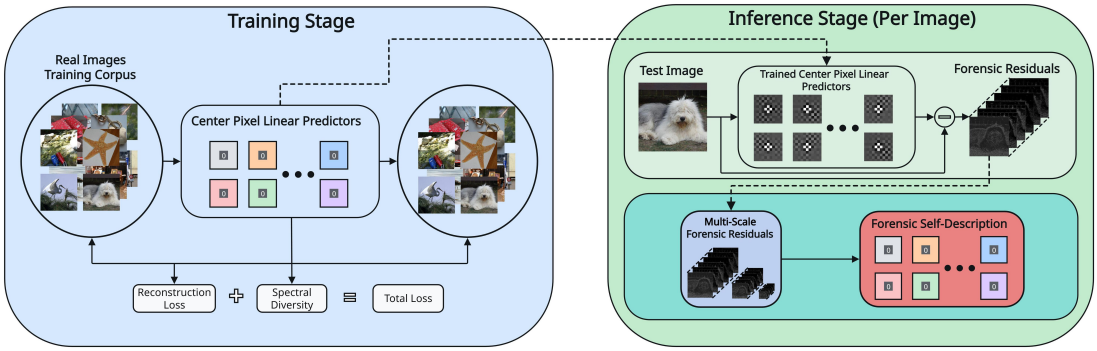

Figure 3. Our method can detect and attribute synthetic images without prior knowledge of the source. We do this by extracting residuals containing forensic microstructures from a single image and jointly modeling them across scales as a forensic self-description.

Unsupervised Clustering. Research to accomplish this task for synthetic images has been largely under-explored. Girish et al. [31] proposed a way to discover new GAN generators by over-clustering embeddings from a simple CNN. Yang et al. [71] proposed a new open-set method that can be leveraged to perform clustering. Overall, without any supervision, accurately clustering images based on their source remains very challenging.

3. Proposed Method

In this paper, we propose a novel approach for detecting and attributing synthetic images without any exposure to them. As illustrated in Fig. 3, we first learn a set of diverse predictive filters using only real images to approximate scene content. We then apply these filters and extract residuals containing forensic microstructures from a single image. Finally, we jointly model these residuals across multiple scales with a parametric model to derive a unique forensic self-description for each image. This self-description captures intrinsic forensic properties, enabling precise distinction of image sources. More details are presented below.

3.1. Forensic Microstructures Extraction

Prior research has shown that the process used to form an image leaves behind unique forensic microstructures [47]. This holds true for both cameras and AI image generators [45, 73]. While a common strategy to identify synthetic images is to utilize the differences in these microstructures [66, 67], they are not directly observable. However, we can estimate them using the procedure below.

We begin by modeling an image I as the sum of two independent components: the scene content S and the forensic microstructures Ψ, such that:

$$I(x,y) = S(x,y) + \Psi(x,y), \tag{1}$$

where (x, y) are the 2D pixel coordinates.

Using this model, we can estimate Ψ by approximating S and subtracting $\hat{S}$ from I. This subtraction results in a residual which contains forensic microstructures and estimation noise ϵ. In practice, however, it is challenging to perfectly approximate the scene content, which means the estimate of the microstructures will be imperfect.

To address this problem, we use a series of K distinct scene predictions to produce a set of unique residuals $\{r_k\}_{k=1}^{K}$, such that:

$$r_k(x,y) = I(x,y) - \hat{S}_k(x,y) = \Psi_k(x,y) + \epsilon_k(x,y). \tag{2}$$

Since each residual captures a different aspect of the microstructures, the collection of these residuals fully describes the microstructures present.

To produce scene content estimates, we use a series of K learnable linear predictive filters $\mathbf{w} = \{w_k\}_{k=1}^{K}$ that predict the value of each pixel based on its surrounding neighborhood, such that:

$$\hat{S}_k(x,y) = \sum_{(i,j)\in\mathcal{M}} w_k(i,j) \cdot I(x+i, y+j), \tag{3}$$

where $\mathcal{M}$ is the set of offsets in the $M \times M$ neighborhood around (x, y) excluding $(0, 0)$. We implement these filters by constraining a convolutional layer such that the center kernel weight is always set to 0 and the sum of all kernel weights is 1 to preserve the energy of the output prediction.

To learn $\mathbf{w}$, we minimize the total energy across all residuals, which results in the following loss term $\mathcal{L}_E$:

$$\mathcal{L}_E(\mathbf{w}) = \sum_{k=1}^{K} \sum_{x,y} \Big(I(x,y) - \hat{S}_k(x,y) \Big)^2. \tag{4}$$

However, this loss term alone may produce filters that are redundant. To prevent this, we introduce a novel spectral diversity regularization term that encourages the filters to be as linearly independent as possible, maximizing the diversity of information captured.

To do this, we first construct a matrix $\mathbf{W} \in \mathbb{R}^{K\times(M^2)}$ by reorienting the weights of each filter into a vector:

$$\mathbf{W} = \begin{bmatrix} \mathrm{vec}(w_1)^\top \\ \mathrm{vec}(w_2)^\top \\ \vdots \\ \mathrm{vec}(w_K)^\top \end{bmatrix} \tag{5}$$

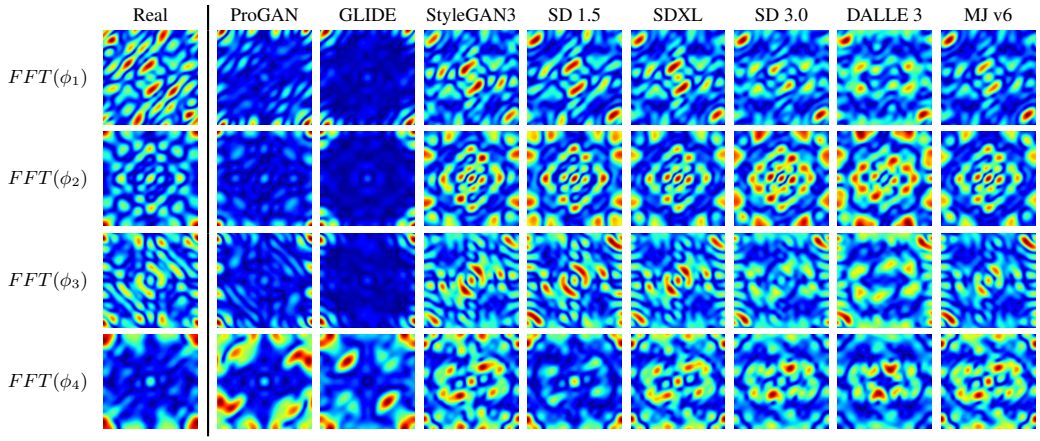

Figure 4. Visualization of the average power spectrum of different filters in the forensic self-descriptions obtained from various sources.

We then perform the singular value decomposition on $\mathbf{W}$ to obtain the set of singular values $\{\sigma_i\}$. Finally, the spectral diversity regularization term is defined as:

$$\mathcal{L}_{\text{diversity}}(\mathbf{w}) = -\sum_{i=1}^{\min(K,M^2)} \log(\sigma_i + \alpha), \tag{6}$$

where α is a small constant to prevent numerical instability. This term penalizes filter configurations where singular values are small, which would indicate greater degrees of linear dependence among filters. By minimizing $\mathcal{L}_{\text{diversity}}$, we encourage the filters to be as diverse as possible.

We combine the two terms to obtain the overall objective for learning the predictive filters:

$$\mathbf{w}^* = \arg\min_{\mathbf{w}} \left[\mathcal{L}_E(\mathbf{w}) + \lambda \mathcal{L}_{\text{diversity}}(\mathbf{w})\right], \tag{7}$$

where λ is a hyperparameter that balances the two terms. We note that $\mathbf{w}$ is learned from a training set consisting of only real images.

3.2. Forensic Self-Description

After $\mathbf{w}$ is learned, we use it to extract a set of residuals $\{r_k\}_{k=1}^K$ for a single image, irrespective of whether the image is real or synthetic. To capture structures present in these residuals, we build a parametric model of these residuals and use its parameters to describe the forensic microstructures. We refer to these parameters as the **forensic self-description** of an image.

To do this, we model the k-th residual $r_k(x, y)$ on the basis of residual values in a $B \times B$ neighborhood around (x, y), similar to an autoregressive model. Additionally, to capture structures present across different scales, we define the residual at scale l as:

$$r_k^{(l)} = \text{Downsample}\left(r_k, 2^{l-1}\right), \tag{8}$$

where $\text{Downsample}(X, Y)$ reduces the spatial resolution of the input X by a factor of Y.

Then, the model of the residuals at scale l is defined as:

$$\hat{r}_k^{(l)} = \sum_{(m,n)\in\mathcal{B}} \phi_k(m, n) \cdot r_k^{(l)}(x + m, y + n), \tag{9}$$

where ϕ_k are the parameters of a linear convolutional filter that models r_k at scale l, and $\mathcal{B}$ is the set of offsets in the $B \times B$ neighborhood excluding $(0, 0)$.

Although we model each residual $r_k^{(l)}$ separately with its own filter ϕ_k, we optimize all filters $\{\phi_k\}$ jointly across all residuals and scales. This joint optimization ensures that the filters collectively capture the interdependent forensic microstructures present in the image.

Hence, the collection of all filters in the model $\Phi = \{\phi_1, \phi_2, \ldots, \phi_K\}$ corresponds to an image's forensic self-description.

To learn Φ jointly across all scales, we first define the total model error at a location (x, y) as:

$$\varepsilon(x, y) = \sum_{k=1}^{K}\sum_{l=1}^{L} r_k^{(l)}(x, y) - \hat{r}_k^{(l)}(x, y). \tag{10}$$

Then, we optimize the parameters Φ by minimizing the total model error power across all locations in the image:

$$\Phi^* = \arg\min_{\Phi} \sum_x \sum_y \left|\varepsilon(x, y)\right|^2. \tag{11}$$

The final parameter set Φ constitutes the forensic self-description of the image.

4. Applications of Forensic Self-Description

Forensic self-descriptions can be used to perform a number of critical tasks related to synthetic image detection and source attribution, such as: zero-shot detection, open-set source attribution, and unsupervised clustering.

4.1. Zero-Shot Synthetic Image Detection

Zero-shot detection refers to the task of determining whether an image is real or AI-generated without prior exposure to images from the generator in question. Super-

Table 1. Zero-shot synthetic image detection performance, measured in average AUC over all pairs of a real dataset vs each synthetic generator source.

Method	COCO17	IN-1k	IN-22k	MIDB	Average
CNNDet [70]	0.756	0.714	0.733	0.683	0.722
PatchFor [14]	0.833	0.823	0.845	0.790	0.823
UFD [54]	0.903	0.862	0.815	0.612	0.798
LGrad [66]	0.819	0.770	0.866	0.824	0.820
DE-FAKE [63]	0.765	0.749	0.617	0.791	0.731
Aeroblade [59]	0.728	0.741	0.582	0.646	0.674
ZED [22]	0.751	0.676	0.716	0.747	0.723
NPR [67]	0.945	0.900	0.900	0.957	0.926
Ours	**0.968**	**0.962**	**0.941**	**0.971**	**0.960**

Table 2. Worst case zero-shot detection performance across all pairs of a real dataset vs each synthetic generator source. Metrics are reported in AUC.

Method	COCO17	IN-1k	IN-22k	MIDB
CNNDet [70]	0.477 (DALLE 3)	0.424 (DALLE 3)	0.439 (DALLE3)	0.373 (DALLE 3)
PatchFor [14]	0.547 (SD 2.1)	0.543 (SD 2.1)	0.565 (SD2.1)	0.536 (SD 2.1)
UFD [54]	0.680 (DALLE 3)	0.607 (DALLE 3)	0.527 (DALLE 3)	0.244 (MJ v6)
LGrad [66]	0.617 (SD 2.1)	0.625 (Firefly)	**0.776** (Firefly)	0.606 (SD 2.1)
DE-FAKE [63]	0.534 (BigGAN)	0.487 (BigGAN)	0.383 (BigGAN)	0.563 (BigGAN)
Aeroblade [59]	0.425 (BigGAN)	0.458 (BigGAN)	0.336 (BigGAN)	0.360 (BigGAN)
ZED [22]	0.462 (ProGAN)	0.402 (ProGAN)	0.375 (ProGAN)	0.331 (ProGAN)
NPR [67]	0.396 (Firefly)	0.239 (Firefly)	0.295 (Firefly)	0.449 (Firefly)
Ours	**0.892** (SD 1.5)	**0.903** (GigaGAN)	0.714 (GLIDE)	**0.896** (MJ v6)

vised detectors struggle in this task as they typically learn representations optimized to discriminate between known sources during training.

We can perform zero-shot detection using forensic self-descriptions because they capture all aspects of the forensic microstructures in an image, not just features discriminative among known sources. By modeling the distribution of forensic self-descriptions from real images, we can flag images whose self-descriptions deviate from this distribution. This ability is qualitatively demonstrated in Fig. 4, which shows the power spectra of forensic self-description filters learned from images of different sources. The figure reveals substantial differences between the self-descriptions of real images and those of AI-generated images.

We perform zero-shot detection by first using a Gaussian Mixture Model (GMM) [49] to model the distribution of the self-descriptions obtained from a set of real images. Detection is performed by computing the likelihood that an image is real, defined as: $p(\Phi|\text{Real}) = \sum_\ell \pi_\ell \mathcal{N}(\boldsymbol{\mu}\ell, \boldsymbol{\Sigma}\ell)$, where Φ is the self-description of the image, and π_ℓ, $\boldsymbol{\mu}\ell$, $\boldsymbol{\Sigma}\ell$ are the GMM's parameters. If $p(\Phi|\text{Real}) \geq \tau_{\text{real}}$, the image is classified as real; otherwise, it is flagged as synthetic.

4.2. Open-Set Synthetic Image Source Attribution

Open-set source attribution refers to the task of identifying the source of an image amongst a set of known source generators, or determining if the image originates from an unknown source.

We can leverage forensic self-descriptions to perform this task as images from common sources share similar forensic microstructures, while those from different sources do not [19]. To accomplish this, we can model the distribution of forensic self-descriptions from each source separately. Then, we can attribute an image by assigning it to the most likely source. If this likelihood is sufficiently low, we designate the source to be unknown.

We perform open-set source attribution by first collecting a set of images from known sources. Then, for images from source S, we model the distribution of their corresponding self-descriptions using a GMM as follows: $p(\Phi|S) = \sum_\ell \pi_\ell \mathcal{N}(\boldsymbol{\mu}_\ell, \boldsymbol{\Sigma}_\ell)$. This will result in one GMM for each known source. After training the GMMs, we can then use them to attribute the source of an image by computing the likelihood of its embedding under each GMM. The generator source with the highest likelihood is considered to be the candidate source of the image:

$$S^* = \arg\max_S p(\Phi|S). \tag{12}$$

If $p(\Phi|S^*) < \tau_{\text{reject}}$, the image's source is unknown, otherwise, the candidate source is accepted.

4.3. Unsupervised Clustering of Image Sources

In many practical scenarios, we need to identify common sources in an unlabeled image dataset by applying a clustering algorithm on to the features extracted for each image. In these cases, we can also use the forensic self-descriptions of images as their features.

Particularly, in this paper, we show that we can successfully apply K-means [4] to the set of forensic self-descriptions produced from individual images to group them based on their description's similarity. The number of clusters can be set based on the expected number of sources or via the elbow method [12] or silhouette analysis [64].

5. Experiments and Results

5.1. Implementation Details

Extracting Forensic Residuals. Following Sec. 3.1, we trained a scene content approximator with $K = 8$ learnable linear predictive filters of neighborhood size $\mathcal{M} = 11 \times 11$ on gray-scaled real images. We used the AdamW optimizer [44] (learning rate 0.001) for 10 epochs. A balance factor of $\lambda = 1.0$ optimized the two loss terms.

Extracting Forensic Self-Descriptions. For each image, we modeled the $K = 8$ forensic residuals with 8 corresponding predictive filters of neighborhood size $\mathcal{B} = 11 \times 11$, across $L = 3$ scales (obtained via bilinear downsampling). The filters are optimized over multi-scale residuals using the AdamW optimizer with a learning rate of 0.1, decaying by half on plateau, for up to 10,000 iterations.

Table 3. Open-set source attribution performance comparisons with various techniques.

Category	Method	Accuracy	AU-CRR	AU-OSCR
Transferable Embeddings	CLIP [57]	0.570	0.543	0.304
	ResNet-50 [32]	0.538	0.605	0.372
Supervised	DCTCNN [29]	0.855	0.452	0.406
	RepMix [13]	0.982	0.746	0.741
Metric-learning	FSM [48]	0.422	0.565	0.207
	EXIFNet [75]	0.186	0.412	0.064
Open-set	Abady et al. [1]	0.828	0.640	0.555
	POSE [71]	0.913	0.629	0.608
	Fang et al. [28]	**0.988**	0.856	0.852
	Ours	0.964	**0.933**	**0.913**

5.2. Datasets

To conduct our experiments, we pooled together a large composite dataset of real and synthetic images from various publicly available sources. Real images are drawn from: (1) COCO2017 [41], (2) ImageNet-1k [24], (3), ImageNet-22k [61], and (4) MISL Image Database (MIDB) [8, 9]. Synthetic images come from: (1) OSSIA dataset [28], (2) DMID dataset [19], and (3) Synthbuster dataset [6]. Overall, our set of synthetic images includes 24 generators across diverse architectures. Some notable ones are: ProGAN [34], StyleGAN [1 to 3] [35–37], GigaGAN [33], EG3D [15], GLIDE [53], Stable Diffusion (SD) [1.3 to 3.0] [27, 56, 60], DALLE [M, 2, 3] [11, 23, 58], Midjourney (MJ) [5, 6] [50], and Adobe Firefly [2]. Data composition details are available in the supplemental materials.

5.3. Zero-Shot Detection Evaluation

Setup. To assess zero-shot detection performance, we divided the composite dataset, described in Sec. 5.2, into a training set of real images and a test set of both real and synthetic images. We measured performance across 96 real-synthetic dataset pairs and report the average result over all real-vs-synthetic dataset pairs per real source. A detailed breakdown of the results by generator is provided in the supplemental materials.

Metrics. We report the average AUC (Area Under the ROC curve) for direct comparison with prior works.

Competing Methods. We compared our method to 2 traditional approaches: CNNDet [70], PatchFor [14], and 6 state-of-the-art zero-shot methods: LGrad [66], UFD [54], DE-FAKE [63], Aeroblade [59], ZED [22], and NPR [67].

Results. This experiment's results are provided in Tab. 1 and 2. These results show that our method achieves the highest zero-shot detection performance, with an overall average AUC of 0.960 across all datasets. In contrast, supervised methods like CNNDet and PatchFor obtain lower performance because the features they learned during training do not transfer well to new generators.

While zero-shot methods such as ZED, DE-FAKE, and NPR show strong performance on some generators, they struggle on others. Tab. 2 shows the worst-case performance of each method across all real-versus-synthetic dataset pairs. The table reveals that ZED consistently struggled with detecting ProGAN, DE-FAKE with BigGAN, and NPR with Firefly. In contrast, by using forensic self-descriptions, we achieve consistently strong performance, with an overall worst-case AUC of 0.89 or greater, substantially higher the other methods. The only exception is IN22k, where we are slightly behind LGrad. These results show that forensic self-descriptions offer reliable detection capability across a wide range of real and synthetic sources.

Table 4. Clustering performance comparisons with various techniques. Here, the ground-truth number of sources is $N = 8$.

Method	# Clusters = N			# Clusters = 2N			# Clusters = 4N		
	Avg. Acc.	Purity	NMI	Avg. Acc.	Purity	NMI	Avg. Acc.	Purity	NMI
CLIP [57]	0.68	0.68	0.60	0.72	0.72	0.59	0.73	0.74	0.52
ResNet-50 [32]	0.50	0.51	0.38	0.56	0.59	0.40	0.60	0.59	0.37
FSM [48]	0.16	0.16	0.01	0.18	0.18	0.02	0.20	0.20	0.03
EXIFNet [75]	0.21	0.22	0.06	0.24	0.26	0.08	0.32	0.28	0.09
Abady et al. [1]	0.45	0.40	0.30	0.46	0.46	0.30	0.51	0.48	0.28
POSE [71]	0.57	0.49	0.36	0.56	0.50	0.32	0.49	0.52	0.32
CNNDet [70]	0.47	0.36	0.28	0.49	0.38	0.27	0.52	0.42	0.26
NPR [67]	0.46	0.39	0.34	0.57	0.48	0.33	0.63	0.51	0.32
DE-FAKE [63]	0.32	0.25	0.16	0.24	0.25	0.14	0.22	0.25	0.12
UFD [54]	**0.78**	0.71	0.68	0.67	0.69	0.55	0.71	0.72	0.50
Ours	**0.78**	**0.77**	**0.69**	**0.80**	**0.81**	**0.65**	**0.83**	**0.85**	**0.61**

5.4. Open-Set Source Attribution Evaluation

Setup. To evaluate open-set source attribution performance, we selected 9 sources (1 real and 8 synthetic) from our pooled dataset (described in Sec. 5.2), dividing them into five known (ImageNet-1k, StyleGAN, StyleGAN3, SD 1.4, ProGAN) and four unknown sources (StyleGAN2, SD 3, DALLE 3, Firefly). Supervised and open-set methods were trained on known sources and tested on both known and unknown sources.

Metrics. Following other open-set works [17, 25, 28, 52, 71], we show (1) the average accuracy across all known sources, and (2) the Area Under the Correct Rejection Rate curve (AU-CRR) [28, 71], and (3) the Area Under the Open Set Classification Rate curve (AU-OSCR) [25, 71].

Competing Methods. We compared our method against three state-of-the-art methods designed for this task: Abady et al. [1], Fang et al. [28], POSE [71]; two supervised methods: DCTCNN [29], and RepMix [13]; two metric-learning methods designed for image forensics: FSM [48], EXIFNet [75]; and two methods which produce generic visual embeddings: CLIP [57], and a ResNet-50 [32] trained on ImageNet1k. For methods which only produce a generic embedding, we apply the same open-set procedure proposed in Sec. 4.2 to their produced embeddings.

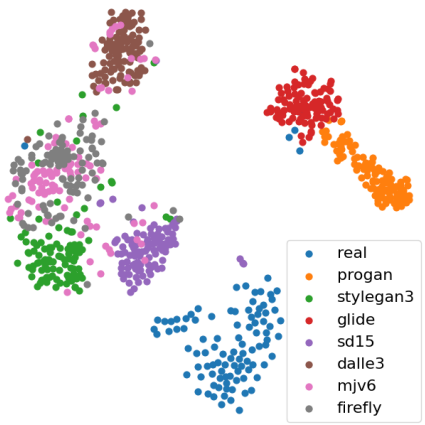

Figure 5. 2D t-SNE plot showing the distribution of the self-descriptions among real and synthetic sources.

Results. Tab. 3 shows the results of this experiment. These results show that leveraging forensic self-descriptions leads to the highest AU-CRR (0.933) and AU-OSCR (0.913). We also obtained near-best known source accuracy (0.964), behind Fang et al.'s 0.988 and RepMix's 0.982. These results indicate that forensic self-descriptions enable both accurate attribution of images to their sources and reliable detection of images from unknown sources. This is also qualitatively demonstrated in Fig. 4, where we can see that the forensic self-descriptions of each source differ from one another.

Both supervised methods like RepMix and dedicated open-set methods like POSE, Abady et al., and Fang et al. achieve moderate to strong known source accuracies but fall short in AU-CRR and AU-OSCR compared to our method. This is because they rely on embedding spaces learned from known generators to generalize to new and unknown generators, which is challenging in practice. In contrast, forensic self-descriptions capture all aspects of forensic microstructures, not just those useful for discriminating between known sources during training. This enables us to perform accurate open-set attribution of image sources.

5.5. Unsupervised Clustering Evaluation

Setup. To evaluate clustering, we used 8 sources representing distinct generation techniques from our composite dataset described in Sec. 5.2 (Real: ImageNet-1k; Synthetic: ProGAN, StyleGAN3, GLIDE, SD 1.5, DALLE 3, MJ v6, Firefly). Our method, applied in an unsupervised manner, does not have training data. For other methods that require synthetic images in their training data, we retrained them on other sources not seen during testing.

Metrics. We present clustering accuracy, purity, and Normalized Mutual Information (NMI), measured across integer multiples of the true number of sources (N, 2N, and 4N) to benchmark performance under different scenarios.

Competing Methods. We evaluated our method against four methods in the zero-shot experiment: NPR [67], UFD [54], DE-FAKE [63] & CNNDet [70], four metric-learning-based methods: FSM [48], EXIFNet [75], Abady et al. [1] & POSE [71], as well as general vision embeddings: CLIP [57], and ResNet-50 [32] trained on ImageNet1k. For each method, we extracted embeddings from either the specified embedder network or the penultimate layer and applied K-means clustering using Euclidean distance or the method's provided distance metric.

Results. We present the results in Tab. 4, which show that clustering based on forensic self-descriptions achieves the highest performance across all metrics and cluster sizes. This is because these descriptions effectively capture forensic microstructures, causing images from the same source to cluster naturally. This behavior is further illustrated in Fig. 5, where the t-SNE plot [69] reveals clear separation between real and synthetic images, with each synthetic generator forming a tight, well-defined cluster.

Notably, when the number of clusters equals the number of sources, UFD performs competitively and CLIP shows moderate clustering ability. This is not surprising, as UFD was designed for enhanced source-separability and CLIP was demonstrated in recent works to have promising detection capabilities [3, 21, 54].

In more realistic scenarios where the number of sources is unknown, clustering is often performed with an overestimated number of clusters followed by merging. Under these conditions, our method continues to improve with larger cluster counts, whereas others show modest gains (Abady et al., CLIP) or performance declines (UFD, POSE). This trend highlights the suitability of forensic self-descriptions for accurate, unsupervised source clustering.

6. Ablation Study

We conducted an ablation study to understand the impact of different design choices on the performance of forensic self-descriptions. To do this, we measured the performance of the zero-shot detection task in terms of average AUC over a subset of real-vs-synthetic dataset pairs (ImageNet-1k versus ProGAN, SDXL, DALLE 3, MJ v6, and Firefly). We also calculated the relative error reduction (RER) in detection AUC of our method compared to alternative design choices. The results are provided in Tab. 5.

Residual Extraction Method. We examined the detection performance impact of various design choices in the forensic residual extraction process. Results in Tab. 5 show that our method of learning a set of diverse linear predictive filters from a corpus of real images is essential for optimal performance. Nonetheless, we observe that even with a simple high-pass filter to extract residuals, our forensic self-descriptions still achieve strong performance.

Table 5. Zero-shot detection performance of our proposed forensic self-description and its alternative design choices.

Component	Method	AUC	RER%
	Proposed	**0.986**	–
Residual Extraction	5×5 high-pass filter [30, 38]	0.913	83.38
	3×3 high-pass filter [30, 38]	0.955	67.70
	Neighbor Pixel Relations [67]	0.952	70.22
	No spectral diversity	0.969	53.34
Obtaining Self-Descriptions	No multi-scale	0.956	67.51
	1 learnable filter	0.951	70.47
	4 learnable filters	0.931	79.08
	7×7 neighborhood	0.961	63.28
	5×5 neighborhood	0.897	85.98
Utilizing Self-Descriptions	One-Class SVM [62]	0.968	55.00
	Isolation Forest [42]	0.968	55.00

Obtaining Self-Descriptions. We explored different design choices and their impact on obtaining forensic self-descriptions. Tab. 5's results show that using multiple filters to capture underlying structures in the forensic residuals is essential for optimal performance. Additionally, we observe that the self-description extracted from multi-scaled residuals yielded significant performance gains. Overall, these findings highlight that the combination of multi-scale modeling, an adequate number of learnable filters, and an appropriate neighborhood size is vital for obtaining effective forensic self-descriptions.

Utilizing Self-Descriptions. We analyzed several out-of-distribution detection methods using forensic self-descriptions. This is important because different approaches offer unique trade-offs between space-time complexity, practicality, and performance. The results in Tab. 5 show that forensic self-descriptions are versatile and can also be used with a One-Class SVM or an Isolation Forest with minimal performance loss.

7. Discussion

Qualitative Analysis. To qualitatively analyze the characteristics of the microstructures captured by forensic self-descriptions, we visualize the average power spectrum of each filter, computed from 100 images across various sources. The resulting power spectra are presented in Fig. 4

As shown in Fig. 4, the power spectra of all filters in the self-descriptions of real images are significantly distinct from those of synthetic images. Among synthetic sources, each generator exhibits at least one unique spectral characteristic that differ from others. For instance, StyleGAN3 and SD 1.5 have similar spectral responses in filter 1-3 but differ in filter 4. This property of the forensic self-descriptions is confirmed by our experimental results above and further illustrated in the t-SNE plot in Fig. 5. In this plot, we observe the same property: real images cluster distinctly apart from synthetic images, with each synthetic source forming tight, easily distinguishable clusters.

Table 6. Average Zero-Shot AUC of our method over different JPEG quality factors.

JPEG Quality	None	100	90	80	70	60	50	Avg.
Our method	0.986	0.968	0.963	0.960	0.979	0.972	0.979	**0.972**

JPEG Robustness. To assess the robustness of forensic self-descriptions to compression at various JPEG quality factors, we evaluated our method's zero-shot detection performance by measuring the average AUC across quality factors ranging from 50 to 100. This was done on the same subset of real-vs-synthetic dataset pairs used in Sec. 6.

As shown in Tab. 6, our method consistently achieves high AUC scores across all JPEG quality factors with an overall average AUC of 0.972. Even at a low quality factor of 60, our method maintains an AUC of 0.972, showing minimal degradation in detection performance. These results show that forensic microstructures of real and synthetic images still remain distinct and detectable even after compression. This demonstrates that forensic self-descriptions are highly robust and suitable for practical use.

Limitations and Future Work. One possible limitation of forensic self-descriptions is their reliance on accurate and diverse forensic residuals, which in turn depend on training the scene content predictors with a high-quality, diverse set of real images. Future work could explore adaptive filter learning to accommodate new data distributions or develop domain-specific filters for targeted forensic tasks. Extending the approach to handle more complex scenarios, such as post-processed or social media–shared images, could further improve its robustness in real-world settings.

8. Conclusion

We introduced forensic self-descriptions as a robust approach for zero-shot detection, open-set attribution, and unsupervised clustering of synthetic images. By using a self-supervised process to extract residuals containing forensic microstructures, our approach constructs a compact, representative model, that accurately distinguishes real from synthetic images, identifies unknown sources, and clusters images by origin without any supervision. Experimental results confirm forensic self-descriptions resilience to compression artifacts and adaptability across diverse generative models, establishing them as a powerful tool for combating the proliferation of AI-generated fake images.

9. Acknowledgement

This material is based upon work supported by the National Science Foundation under Grant No. 2320600. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

References

[1] Lydia Abady, Jun Wang, Benedetta Tondi, and Mauro Barni. A siamese-based verification system for open-set architecture attribution of synthetic images. *Pattern Recognition Letters*, 180:75–81, 2024. 2, 6, 7, 1

[2] Adobe. Adobe Firefly. https://www.adobe.com/products/firefly.html. Accessed: November 12, 2024. 6

[3] Roberto Amoroso, Davide Morelli, Marcella Cornia, Lorenzo Baraldi, Alberto Del Bimbo, and Rita Cucchiara. Parents and children: Distinguishing multimodal deepfakes from natural images. *ACM Trans. Multimedia Comput. Commun. Appl.*, 2024. Just Accepted. 7

[4] David Arthur and Sergei Vassilvitskii. k-means++: The advantages of careful seeding. Technical report, Stanford, 2006. 5

[5] Aref Azizpour, Tai D. Nguyen, Manil Shrestha, Kaidi Xu, Edward Kim, and Matthew C. Stamm. E3: Ensemble of expert embedders for adapting synthetic image detectors to new generators using limited data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 4334–4344, 2024. 2

[6] Quentin Bammey. Synthbuster: Towards detection of diffusion model generated images. *IEEE Open Journal of Signal Processing*, 5:1–9, 2024. 6, 1, 3

[7] Mauro Barni, Kassem Kallas, Ehsan Nowroozi, and Benedetta Tondi. Cnn detection of gan-generated face images based on cross-band co-occurrences analysis. In *2020 IEEE International Workshop on Information Forensics and Security (WIFS)*, pages 1–6, 2020. 2

[8] Belhassen Bayar and Matthew C. Stamm. Towards open set camera model identification using a deep learning framework. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2007–2011, 2018. 6, 1, 2

[9] Belhassen Bayar and Matthew C. Stamm. Constrained convolutional neural networks: A new approach towards general purpose image manipulation detection. *IEEE Transactions on Information Forensics and Security*, 13(11):2691–2706, 2018. 6, 1, 2

[10] Abhijit Bendale and Terrance E Boult. Towards open set deep networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1563–1572, 2016. 2

[11] 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. 6

[12] Purnima Bholowalia and Arvind Kumar. Ebk-means: A clustering technique based on elbow method and k-means in wsn. *International Journal of Computer Applications*, 105 (9), 2014. 5

[13] Tu Bui, Ning Yu, and John Collomosse. Repmix: Representation mixing for robust attribution of synthesized images. In *European Conference on Computer Vision*, pages 146–163. Springer, 2022. 2, 6, 1

[14] Lucy Chai, David Bau, Ser-Nam Lim, and Phillip Isola. What makes fake images detectable? understanding properties that generalize. In *European Conference on Computer Vision*, pages 103–120. Springer, 2020. 1, 5, 6, 2

[15] 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*, pages 16123–16133, 2022. 6

[16] Guangyao Chen, Limeng Qiao, Yemin Shi, Peixi Peng, Jia Li, Tiejun Huang, Shiliang Pu, and Yonghong Tian. Learning open set network with discriminative reciprocal points. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part III 16*, pages 507–522. Springer, 2020. 2

[17] Guangyao Chen, Peixi Peng, Xiangqian Wang, and Yonghong Tian. Adversarial reciprocal points learning for open set recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(11):8065–8081, 2022. 6

[18] Umur Aybars Ciftci, Ilke Demir, and Lijun Yin. Fakecatcher: Detection of synthetic portrait videos using biological signals. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 1–1, 2020. 2

[19] 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)*, pages 1–5. IEEE, 2023. 1, 2, 5, 6, 3

[20] Davide Cozzolino, Diego Gragnaniello, and Luisa Verdoliva. Image forgery detection through residual-based local descriptors and block-matching. In *2014 IEEE International Conference on Image Processing (ICIP)*, pages 5297–5301, 2014. 2

[21] Davide Cozzolino, Giovanni Poggi, Riccardo Corvi, Matthias Nießner, and Luisa Verdoliva. Raising the bar of ai-generated image detection with clip. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 4356–4366, 2024. 7

[22] Davide Cozzolino, Giovanni Poggi, Matthias Nießner, and Luisa Verdoliva. Zero-shot detection of ai-generated images. In *European Conference on Computer Vision*, pages 54–72. Springer, 2025. 2, 5, 6, 1

[23] Boris Dayma, Suraj Patil, Pedro Cuenca, Khalid Saifullah, Tanishq Abraham, Phúc Lê Khac, Luke Melas, and Ritobrata Ghosh. Dall·e mini, 2021. 6

[24] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, 2009. 6, 1, 2

[25] Akshay Raj Dhamija, Manuel Günther, and Terrance Boult. Reducing network agnostophobia. In *Advances in Neural Information Processing Systems*. Curran Associates, Inc., 2018. 6

[26] David C Epstein, Ishan Jain, Oliver Wang, and Richard Zhang. Online detection of ai-generated images. In *Proceed-*

ings of the IEEE/CVF International Conference on Computer Vision, pages 382–392, 2023. 1

[27] 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 *Forty-first International Conference on Machine Learning*, 2024. 6

[28] Shengbang Fang, Tai D Nguyen, and Matthew c Stamm. Open set synthetic image source attribution. In *34th British Machine Vision Conference 2023, BMVC 2023, Aberdeen, UK, November 20-24, 2023*. BMVA, 2023. 2, 6, 1, 3

[29] Joel Frank, Thorsten Eisenhofer, Lea Schönherr, Asja Fischer, Dorothea Kolossa, and Thorsten Holz. Leveraging frequency analysis for deep fake image recognition. In *International conference on machine learning*, pages 3247–3258. PMLR, 2020. 1, 6, 2

[30] Jessica Fridrich and Jan Kodovsky. Rich models for steganalysis of digital images. *IEEE Transactions on Information Forensics and Security*, 7(3):868–882, 2012. 8

[31] Sharath Girish, Saksham Suri, Sai Saketh Rambhatla, and Abhinav Shrivastava. Towards discovery and attribution of open-world gan generated images. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 14094–14103, 2021. 3

[32] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 6, 7

[33] Minguk Kang, Jun-Yan Zhu, Richard Zhang, Jaesik Park, Eli Shechtman, Sylvain Paris, and Taesung Park. Scaling up gans for text-to-image synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10124–10134, 2023. 6

[34] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of GANs for improved quality, stability, and variation. In *International Conference on Learning Representations*, 2018. 6

[35] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4401–4410, 2019. 6

[36] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of stylegan. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8110–8119, 2020.

[37] Tero Karras, Miika Aittala, Samuli Laine, Erik Härkönen, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Alias-free generative adversarial networks. *Advances in neural information processing systems*, 34:852–863, 2021. 6

[38] Jan Kodovsky, Jessica Fridrich, and Vojtech Holub. On dangers of overtraining steganography to incomplete cover model. In *Proceedings of the thirteenth ACM multimedia workshop on Multimedia and security*, pages 69–76, 2011. 8

[39] Shu Kong and Deva Ramanan. Opengan: Open-set recognition via open data generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 813–822, 2021. 2

[40] Li Lin, Neeraj Gupta, Yue Zhang, Hainan Ren, Chun-Hao Liu, Feng Ding, Xin Wang, Xin Li, Luisa Verdoliva, and Shu Hu. Detecting multimedia generated by large ai models: A survey. *arXiv preprint arXiv:2402.00045*, 2024. 1, 2

[41] 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 *European Conference on Computer Vision*, pages 740–755. Springer, 2014. 6, 1, 2

[42] Fei Tony Liu, Kai Ming Ting, and Zhi-Hua Zhou. Isolation forest. In *2008 eighth ieee international conference on data mining*, pages 413–422. IEEE, 2008. 8

[43] Yun Liu, Zuliang Wan, Xiaohua Yin, Guanghui Yue, Aiping Tan, and Zhi Zheng. Detection of gan generated image using color gradient representation. *Journal of Visual Communication and Image Representation*, 95:103876, 2023. 2

[44] I Loshchilov. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 5

[45] Jan Lukas, Jessica Fridrich, and Miroslav Goljan. Digital camera identification from sensor pattern noise. *IEEE Transactions on Information Forensics and Security*, 1(2): 205–214, 2006. 1, 3

[46] Francesco Marra, Diego Gragnaniello, Davide Cozzolino, and Luisa Verdoliva. Detection of gan-generated fake images over social networks. In *2018 IEEE conference on multimedia information processing and retrieval (MIPR)*, pages 384–389. IEEE, 2018. 1

[47] Francesco Marra, Diego Gragnaniello, Luisa Verdoliva, and Giovanni Poggi. Do gans leave artificial fingerprints? In *2019 IEEE conference on multimedia information processing and retrieval (MIPR)*, pages 506–511. IEEE, 2019. 1, 2, 3

[48] Owen Mayer and Matthew C Stamm. Forensic similarity for digital images. *IEEE Transactions on Information Forensics and Security*, 15:1331–1346, 2019. 6, 7, 2

[49] G. Mclachlan and K. Basford. *Mixture Models: Inference and Applications to Clustering*. 1988. 5

[50] Midjourney. Midjourney. https://www.midjourney.com/home. Accessed: November 12, 2024. 6

[51] Lakshmanan Nataraj, Tajuddin Manhar Mohammed, Shivkumar Chandrasekaran, Arjuna Flenner, Jawadul H. Bappy, Amit K. Roy-Chowdhury, and B. S. Manjunath. Detecting gan generated fake images using co-occurrence matrices. *Electronic Imaging*, 2019. 2

[52] Lawrence Neal, Matthew Olson, Xiaoli Fern, Weng-Keen Wong, and Fuxin Li. Open set learning with counterfactual images. In *Proceedings of the European conference on computer vision (ECCV)*, pages 613–628, 2018. 2, 6

[53] 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 *Proceedings of the 39th International Conference on Machine Learning*, pages 16784–16804. PMLR, 2022. 6

[54] Utkarsh Ojha, Yuheng Li, and Yong Jae Lee. Towards universal fake image detectors that generalize across generative models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 24480–24489, 2023. 1, 2, 5, 6, 7

[55] Daeeol Park, Hyunsik Na, and Daeseon Choi. Performance comparison and visualization of ai-generated-image detection methods. *IEEE Access*, 12:62609–62627, 2024. 1, 2

[56] 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. *arXiv preprint arXiv:2307.01952*, 2023. 6

[57] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 6, 7, 2

[58] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 1 (2):3, 2022. 6

[59] Jonas Ricker, Denis Lukovnikov, and Asja Fischer. Aeroblade: Training-free detection of latent diffusion images using autoencoder reconstruction error. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9130–9140, 2024. 2, 5, 6, 1

[60] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022. 6

[61] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)*, 115 (3):211–252, 2015. 6, 1, 2

[62] Bernhard Schölkopf, John C Platt, John Shawe-Taylor, Alex J Smola, and Robert C Williamson. Estimating the support of a high-dimensional distribution. *Neural computation*, 13(7):1443–1471, 2001. 8

[63] Zeyang Sha, Zheng Li, Ning Yu, and Yang Zhang. De-fake: Detection and attribution of fake images generated by text-to-image generation models. In *Proceedings of the 2023 ACM SIGSAC Conference on Computer and Communications Security*, page 3418–3432, New York, NY, USA, 2023. Association for Computing Machinery. 2, 5, 6, 7, 1

[64] Ketan Rajshekhar Shahapure and Charles Nicholas. Cluster quality analysis using silhouette score. In *2020 IEEE 7th international conference on data science and advanced analytics (DSAA)*, pages 747–748. IEEE, 2020. 5

[65] Sergey Sinitsa and Ohad Fried. Deep image fingerprint: Towards low budget synthetic image detection and model lineage analysis. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 4067–4076, 2024. 1, 2

[66] Chuangchuang Tan, Yao Zhao, Shikui Wei, Guanghua Gu, and Yunchao Wei. Learning on gradients: Generalized artifacts representation for gan-generated images detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12105–12114, 2023. 2, 3, 5, 6, 1

[67] Chuangchuang Tan, Yao Zhao, Shikui Wei, Guanghua Gu, Ping Liu, and Yunchao Wei. Rethinking the up-sampling operations in cnn-based generative network for generalizable deepfake detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 28130–28139, 2024. 2, 3, 5, 6, 7, 8, 1

[68] Diangarti Tariang, Riccardo Corvi, Davide Cozzolino, Giovanni Poggi, Koki Nagano, and Luisa Verdoliva. Synthetic image verification in the era of generative artificial intelligence: What works and what isn't there yet. *IEEE Security & Privacy*, 22(3):37–49, 2024. 1, 2

[69] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9 (86):2579–2605, 2008. 7

[70] Sheng-Yu Wang, Oliver Wang, Richard Zhang, Andrew Owens, and Alexei A. Efros. Cnn-generated images are surprisingly easy to spot... for now. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 1, 2, 5, 6, 7

[71] Tianyun Yang, Danding Wang, Fan Tang, Xinying Zhao, Juan Cao, and Sheng Tang. Progressive open space expansion for open-set model attribution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15856–15865, 2023. 2, 3, 6, 7, 1

[72] Ning Yu, Larry S. Davis, and Mario Fritz. Attributing fake images to gans: Learning and analyzing gan fingerprints. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019. 2

[73] Xu Zhang, Svebor Karaman, and Shih-Fu Chang. Detecting and simulating artifacts in gan fake images. In *2019 IEEE International Workshop on Information Forensics and Security (WIFS)*, pages 1–6, 2019. 1, 2, 3

[74] Xinwei Zhao and Matthew C Stamm. Computationally efficient demosaicing filter estimation for forensic camera model identification. In *2016 IEEE international conference on image processing (ICIP)*, pages 151–155. IEEE, 2016. 1

[75] Chenhao Zheng, Ayush Shrivastava, and Andrew Owens. Exif as language: Learning cross-modal associations between images and camera metadata. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6945–6956, 2023. 6, 7, 2

[76] Da-Wei Zhou, Han-Jia Ye, and De-Chuan Zhan. Learning placeholders for open-set recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4401–4410, 2021. 2