Attribution as Retrieval: Model-Agnostic AI-Generated Image Attribution

Hongsong Wang[1,2], Renxi Cheng[3], Chaolei Han[3], Jie Gui[3,4,5]*
[1]School of Computer Science and Engineering, Southeast University, Nanjing 210096, China
[2]Key Laboratory of New Generation Artificial Intelligence Technology and Its Interdisciplinary Applications (Southeast University), Ministry of Education, China
[3]School of Cyber Science and Engineering, Southeast University, Nanjing 210096, China
[4]Purple Mountain Laboratories, Nanjing 210000, China
[5]Engineering Research Center of Blockchain Application, Supervision And Management (Southeast University), Ministry of Education, China
{hongsongwang, renxi, chaoleihan, guijie}@seu.edu.cn

Abstract

With the rapid advancement of AIGC technologies, image forensics will encounter unprecedented challenges. Traditional methods are incapable of dealing with increasingly realistic images generated by rapidly evolving image generation techniques. To facilitate the identification of AI-generated images and the attribution of their source models, generative image watermarking and AI-generated image attribution have emerged as key research focuses in recent years. However, existing methods are model-dependent, requiring access to the generative models and lacking generality and scalability to new and unseen generators. To address these limitations, this work presents a new paradigm for AI-generated image attribution by formulating it as an instance retrieval problem instead of a conventional image classification problem. We propose an efficient model-agnostic framework, called Low-bIt-plane-based Deepfake Attribution (LIDA). The input to LIDA is produced by Low-Bit Fingerprint Generation module, while the training involves Unsupervised Pre-Training followed by subsequent Few-Shot Attribution Adaptation. Comprehensive experiments demonstrate that LIDA achieves state-of-the-art performance for both Deepfake detection and image attribution under zero- and few-shot settings. The code is at https://github.com/hongsong-wang/LIDA.

1. Introduction

With the rapid advancement of AI-Generated Content (AIGC) technologies, such as image generation [1], motion generation [39, 40, 43, 47] and video generation [51], synthetic media has become increasingly realistic and widely accessible. While AIGC brings significant benefits to entertainment and productivity, it also raises critical concerns regarding authenticity and potential misuse [59]. As a result, AIGC forensics [50] has emerged as an essential research area, aiming to detect, attribute, and trace AI-generated or AI-manipulated content. Reliable AIGC forensics techniques are crucial for safeguarding digital media integrity and preventing malicious abuse in the era of generative AI.

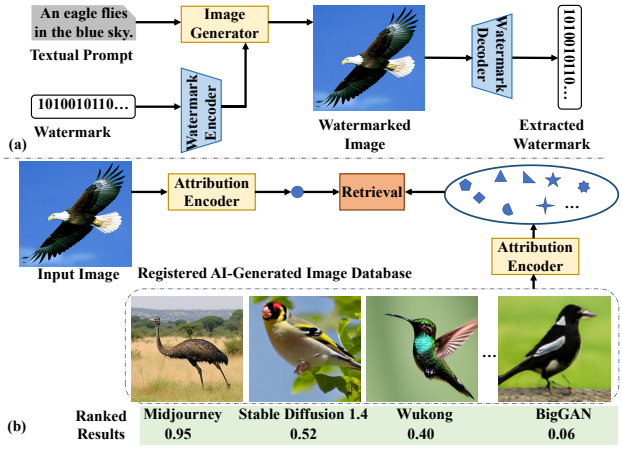


Figure 1. **Comparison between generative image watermarking and our retrieval-based AI-generated image attribution.** Our framework is versatile and easily adapted to new generators.

Traditional media forensics methods [31] struggle to adapt to the challenges posed by AIGC, as AI-generated image or video does not contain camera-based physical traces and exhibits far fewer inconsistency artifacts that conventional forensics rely on. However, although modern generative models can produce highly realistic content, they still leave distinctive generative fingerprints or traces that differ from those found in natural images [36, 57].

*Corresponding author

Recently, the detection of AI-generated images has become an increasingly important problem. The research focus has shifted from early approaches that relied on spotting visible artifacts to more robust strategies that emphasize generalization across diverse and unseen generation models [15, 52]. A number of generator-agnostic methods [15, 21, 41] have been proposed, which are capable of accurately and efficiently distinguishing real images from fake ones. However, Deepfake or AI-generated image detection only determines whether an image is real or fake, without providing any additional forensic information.

To enable the attribution of AI-generated images, two main research directions have emerged: generative image watermarking [16, 58] and AI-generated image attribution [13, 56]. The former embeds invisible watermarks into the image during the generation process, while the latter is independent of the image generation step. Although approaches of generative image watermarking achieve high accuracy in attributing AI-generated images, they require full access to the image generation model and often do not generalize across different generators (see Figure 1(a)).

Most studies on AI-generated image attribution target the close-set scenario [4, 53, 56], which is less applicable to modern generative models that are rapidly evolving. Few works [13, 38, 54] address the open-set scenario. However, all existing works treat AI-generated image attribution as a classification problem and require labeled or unlabeled AI-generated images from different generators during training. Therefore, these approaches are not flexible enough to adapt to numerous new or unseen image generators.

To address the above limitations, we study AI-generated image attribution from the novel perspective of instance retrieval, and introduce a model-agnostic framework. As illustrated in Figure 1(b), this framework requires only the training of an attribution encoder and is readily scalable to unseen image generators. To guarantee retrieval-based attribution, a registered AI-generated image database is maintained, containing only a few images for each generator.

More specifically, we introduce a method called LIDA, which consists low-bit fingerprint generation, unsupervised pre-training and few-shot attribution adaptation. Low bit-planes of each RGB channel are used to compose the fingerprint image. During unsupervised pre-training, a pretext task with a corresponding side loss is employed to train a lightweight network on large-scale real images to enhance generalization. Few-shot attribution adaptation uses only a limited number of AI-generated images from the registered dataset, along with an equal number of real images. The adaptation is supervised by the image attribution loss and the Deepfake detection loss. Comprehensive evaluations on the GenImage [60] dataset and WildFake [20] dataset, covering zero-shot and few-shot Deepfake detection as well as cross-architecture and cross-generator image attribution, are conducted to validate the effectiveness and robustness of our bit-plane-based forensic technique.

Our contributions are summarized as follows:

- **Novel solution for AI-generated image attribution:** We formulate AI-generated image attribution as an instance retrieval problem and address it using bit-planes.
- **Versatile and efficient pipeline design:** We propose a simple yet effective pipeline consisting of three modules: low-bit fingerprint generation, unsupervised pre-training and few-shot attribution adaptation.
- **Superior zero- and few-shot image forensics results:** Our method achieves state-of-the-art performance on two popular AI-generated image datasets for zero- and few-shot detection and attribution.

2. Related Work

Generative Image Watermarking: Generative image watermarking [16, 58] is a technique that embeds identifiable signatures into images produced by generative models, enabling the attribution of AI-generated images. Traditional image watermarking protects the copyright of individual images, whereas generative image watermarking safeguards the copyright of the generative model itself. To embed watermarks, existing methods either fine-tune the image decoder of generative models [11] or modify their latent representations [29]. For example, Tree-Ring [46] embeds an invisible watermark into the initial noise vector of a diffusion model in Fourier space. Gaussian Shading [55] enables plug-and-play watermarking by embedding watermarks into the diffusion model's latent representations following a Gaussian distribution. Although watermarking techniques achieve high accuracy in attributing AI-generated images, they require access to the generative models and involve modifying them, which limits the flexibility of these methods.

Closed-Set AI-Generated Image Attribution: Closed-set AI-generated image attribution is a supervised image classification task that assumes all image generators are known during training. Yu et al. [56] present a systematic study showing that images generated by Generative Adversarial Networks (GANs) carry distinct model fingerprints that enable attribution. Frank et al. [12] demonstrate that GAN-generated images produce consistent artifacts in the frequency domain, which can also be used for source identification. RepMix [4] attributes generated images to their GAN architecture regardless of semantic content and under benign transformations. Yang et al. [53] show that even when GANs are fine-tuned or retrained, the underlying architecture leaves globally consistent fingerprints that enable attribution. These closed-set approaches, which focus on GAN-based image generators, are less practical and less applicable to modern generative models.

Open-Set AI-Generated Image Attribution: Open-set

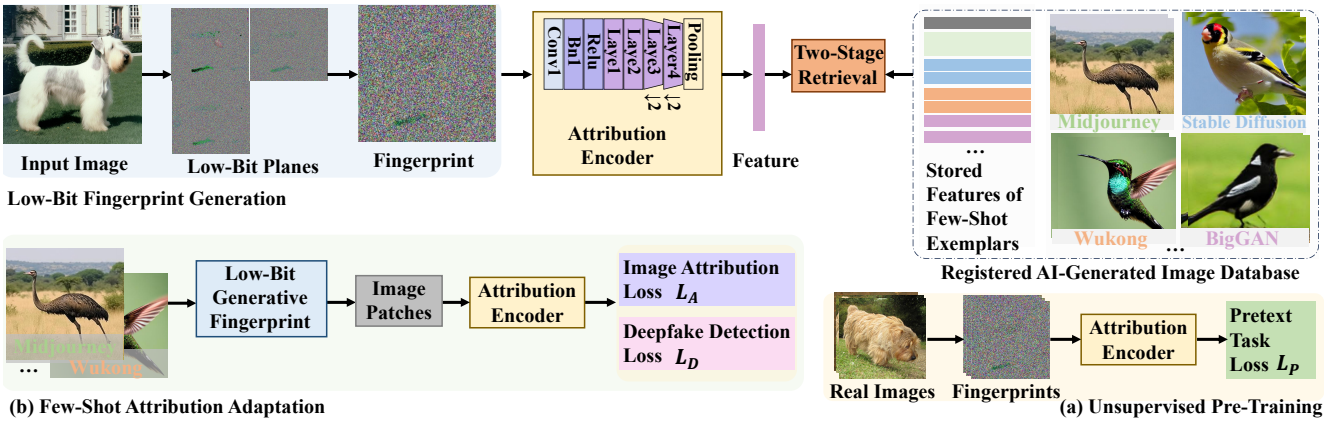


Figure 2. **Pipeline of the proposed model-agnostic framework LIDA for AI-generated image attribution.** LIDA treats image attribution as an instance retrieval problem, and uses low-bit-plane-based generative fingerprint as the input. The training stage consists of two consecutive steps: (a) unsupervised pre-training and (b) few-shot attribution adaptation.

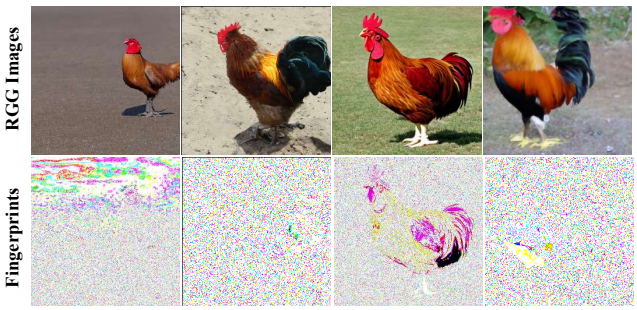


Figure 3. **Comparison of AI-generated images and low-bit generative fingerprints from different image generators.** Generators include Stable Diffusion [32], ADM [10], and Wukong [49].

approaches aim to address the attribution of images generated by new and unseen models. Girish et al. [13] present an algorithm that discovers images generated by unseen GANs and simultaneously attributes them to source models. Yang et al. [54] simulate open-set samples via lightweight models to enable attribution of both known and unknown generative models. Sun et al. [38] combine a voting module and confidence-based pseudo-labels to attribute forged faces in an open-world scenario. De-fake [35] attributes images from various text-to-image generation models. Li et al. [22] show that high-pass handcrafted filters improve attribution performance for both closed-set and open-set scenarios. These approaches also treat attribution as a classification task and leverage both labeled and unlabeled images during training, where unseen classes exist in the unlabeled set. However, this setting is still not flexible enough in real-world scenarios, as it requires a large number of unlabeled AI-generated images from new generators for training. Accordingly, this work aims to devise a more flexible and practical paradigm for open-set AI-generated image attribution.

3. Retrieval Perspective for Image Attribution

Rather than treating AI-generated image attribution as a classification task, we formulate it as a retrieval task, which naturally supports open-set scenarios by requiring only a well-trained image-based feature extractor for attribution.

A registered database of AI-generated images, $\mathcal{D} = \{x_1^1, \ldots, x_i^j, \ldots, x_N^J\}$, needs to be maintained, where j denotes the index of the image generator G_j and i denotes the index of the image. To accommodate new image generators, only one or a few example images need to be added to the registered database.

A feature encoder $f(\cdot)$ maps both query and database images into a unified feature space. Given a query image q generated by an arbitrary image generator, the similarity between q and x_i^j is measured using a similarity function based on their extracted features. The most similar images in the registered database are then retrieved.

The retrieval model ranks all database images according to similarity scores, and the top-K retrieved neighbors are defined as:

$$\text{Top-K}(q) = \underset{x_i \in \mathcal{D}}{\arg\,\text{top-K}}\ \text{sim}(q, x_i). \tag{1}$$

The attribution decision is based on the generator labels of the retrieved neighbors, e.g., by assigning the label of the top-ranked retrieved image to the query. The main focus is to train or fine-tune the feature encoder $f(\cdot)$, specifically designed for image attribution.

This retrieval-based attribution paradigm eliminates the need for retraining when encountering new generators, making it inherently open-set friendly, as the model can directly incorporate samples from unseen generators into the registered database. This paradigm also provides evidence-based attribution, as the retrieved images justify the predicted source.

4. Low-Bit-Plane-Based Deepfake Attribution

We study AI-generated image attribution (i.e., Deepfake attribution) from the bit-plane perspective, and present a model-agnostic approach called Low-bIt-plane-based Deepfake Attribution (LIDA). LIDA does not need to access models of any image generators. Given only a few AI-generated images as exemplars, it can quickly gain the ability to predict the corresponding image generator of an arbitrary image. The pipeline of LIDA is shown in Figure 2. Details are described as follows.

4.1. Low-Bit Fingerprint Generation

Similar to works on camera fingerprints [6, 24], recent studies demonstrate that AI-generated images also contain distinctive generative fingerprints that enable model attribution and source tracing [56, 57]. Generative fingerprints refer to inherent and consistent artifacts unintentionally embedded by a generative model during the image synthesis process. These artifacts are model-specific and remain stable regardless of the image content.

The low-bit-plane-based AI-generated image detection method [41, 42] demonstrates that low-bit planes inherently contain intrinsic artifacts that can be exploited to distinguish real from AI-generated images. Motivated by this observation, we hypothesize that such low-bit-plane noise images can also be leveraged for AI-generated image attribution. Thus, we term such noise image low-bit generative fingerprint, which can be quickly obtained via the following simple operations.

For an RGB image $\mathbf{x} \in \mathbb{R}^{H\times W\times 3}$, let $\mathbf{x}_c(i,j)$ denotes the pixel value at position (i,j) in channel $c \in \{R,G,B\}$. The bit-plane decomposition for each channel is:

$$\mathbf{x}_c = \sum_{k=0}^{7} 2^k \cdot \mathbf{b}_c^k \tag{2}$$

where $\mathbf{b}_c^k$ represents the k-th bit plane of the c-th channel.

We combine the three least significant bit-planes of each channel and employ the thresholding strategy [41] to construct the generative fingerprint $\tilde{\mathbf{x}}_c$:

$$\tilde{\mathbf{x}}_c = 255 \cdot \mathrm{sgn}(\sum_{k=0}^{2} 2^k \cdot \mathbf{b}_c^k) \tag{3}$$

where $\mathrm{sgn}(\cdot)$ is the sign function which maps elements greater than zero to one, and elements equal to zero to zero.

We visualize in Figure 3 the low-bit generative fingerprints of AI-generated images with the same content, produced by different image generators, including Stable Diffusion [32], ADM [10], Wukong [49], and GLIDE [27]. Compared with the original RGB images, the low-bit generative fingerprints, which discard most of the image content, better reveal distinctive traits for model attribution.

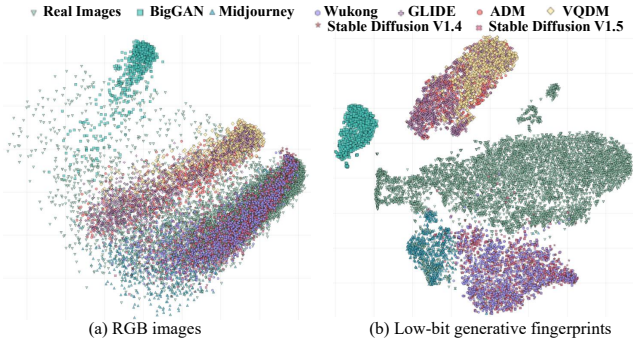


Figure 4. **Feature distribution of images from different sources** for (a) RGB images and (b) low-bit generative fingerprints.

To further analyze the distinctive capabilities of the low-bit generative fingerprint, we directly extract features using a pretrained ResNet-50 [17] on ImageNet [8], and visualize the PCA-reduced features of thousands of samples from different image generators in Figure 4. For RGB images, real images are mixed with AI-generated images, and the distribution differences among images from different generators are almost indistinguishable. In contrast, for low-bit generative fingerprints, real and AI-generated images are clearly separated, and images from the same generator are relatively clustered. Therefore, we use low-bit generative fingerprints as input to train the network to learn features for AI-generated image attribution.

4.2. Unsupervised Pre-Training

To enhance generalization, we adopt unsupervised pre-training on a large-scale real image dataset using the fingerprint input computed in Eq. (3). After pre-training, the network learns to capture intrinsic noise structures that are transferable to downstream tasks of generative image forensics. Moreover, the unsupervised pre-training provides a robust weight initialization, leading to faster convergence and improved performance during fine-tuning.

For both simplicity and effectiveness, we adopt ResNet-50 [17] as the attribution encoder. To better preserve spatial information, we modify the network by removing the downsampling operations in the lower layers. This design maintains high-resolution feature maps, which are essential for capturing subtle structural details in forensic analysis.

We use the pretext task training strategy, and train the attribution encoder on the ImageNet [8]. As an example, the image classification is used as the side task to train the network. The pretext task training loss $\mathcal{L}_P$ is formulated as:

$$\mathcal{L}_P = -\sum_{b=1}^{B}\sum_{c=1}^{C} s_b^c \log q_b^c, \tag{4}$$

where b is the index of fingerprints of real images, q_b is the corresponding ImageNet category label, s_b is the predicted

class probabilities, c is the category index, B and C represent the total number of images and categories, respectively.

4.3. Few-Shot Attribution Adaptation

The registered AI-generated image database contains a limited number of samples for each image generator, including new and unseen ones. The proposed few-shot attribution adaptation leverages this database to efficiently adapt the pretrained model to these unseen generators.

We first define the image attribution loss. We do not use the commonly applied cross-entropy loss, as it may disrupt the feature representations learned during pretraining. Cross-entropy focuses solely on maximizing classification accuracy and does not explicitly preserve the structure of the feature space. As a result, previously learned discriminative features may be altered, even when fine-tuning with only a few samples. Instead, we incorporate the center loss [45] to encourage samples of the same class to cluster around their corresponding class centers. Mathematically, the image attribution loss $\mathcal{L}_A$ is given by:

$$\mathcal{L}_A = \sum_{i=1}^{m} \|x_i - c_{y_i}\|_2^2, \tag{5}$$

where x_i denotes the learned feature of the i-th sample, y_i is its corresponding attribution category label, and c_{y_i} represents the center of the y_i-th category. During training, the center c_j is updated within each mini-batch as follows:

$$c_j^{t+1} = c_j^t - \alpha \cdot \frac{\sum_{i=1}^{m} \delta(y_i = j) \cdot (c_j^t - x_i)}{1 + \sum_{i=1}^{m} \delta(y_i = j)}, \tag{6}$$

where α is the learning rate, and $\delta(\cdot)$ is the indicator function that equals one if the condition is true and zero otherwise. The center loss is considered as a regularization which encourages intra-class compactness, constraining the drift of learned features and helping preserve the structure of the pretrained feature space.

We employ a two-stage paradigm for image attribution, in which Deepfake detection is conducted first, followed by assigning images to their respective generators. Thus, the Deepfake detection loss is also defined. To enable the model to better distinguish the feature discrepancies between real and fake images, we adopt a real-prototype-based contrastive loss, which pulls the features of real images closer to the real prototype while pushing the features of AI-generated images away from it. Formally, the real-prototype-based contrastive loss is defined as:

$$\begin{aligned}\mathcal{L}_D = &-\frac{1}{N_r}\sum_{i=1}^{N_r} \log \sigma\Big(\frac{\mathrm{sim}(x_i^r, p_r)}{\tau}\Big) \\ &-\frac{1}{N_f}\sum_{j=1}^{N_f} \log\Big(1 - \sigma\Big(\frac{\mathrm{sim}(x_i^f, p_r)}{\tau}\Big)\Big),\end{aligned} \tag{7}$$

where N_r and N_f denote the numbers of real and AI-generated images, respectively; x_i^r and x_i^f represent the learned features of real and AI-generated images, respectively; p_r is the prototype of the real class, which is the averaged feature of all images on the ImageNet; $\sigma(\cdot)$ is the sigmoid function; τ is a temperature parameter; and $\mathrm{sim}(\cdot,\cdot)$ denotes the cosine similarity. Note that we also avoid using the cross-entropy loss for Deepfake detection in order to preserve the structure of the pretrained feature space.

The final few-shot attribution adaptation loss is:

$$\mathcal{L} = \mathcal{L}_A + \lambda \mathcal{L}_D, \tag{8}$$

where λ is the weight parameter.

5. Experiments

5.1. Experimental Setup

Dataset: This study evaluates image attribution on two large-scale benchmarks for AI-generated image detection, namely GenImage [60] and WildFake [20]. The GenImage dataset comprises 1,331,167 real images and 1,350,000 synthetic images. The real images are derived from ImageNet, while the synthetic images are generated by eight representative diffusion and GAN models, including Midjourney [26], Stable Diffusion (v1.4/1.5) [32], ADM [10], GLIDE [27], Wukong [49], VQDM [14], and BigGAN [2]. The WildFake dataset contains 1,013,446 real images and 2,557,278 synthetic images. The real images are collected from public sources consistent with the training distributions of mainstream generative models, such as COCO [23] and ImageNet [8]. The synthetic images include those generated by the authors using GANs, diffusion models, and other generation mechanisms, as well as those gathered from open platforms such as Civitai [7] and Midjourney [26]. The dataset specifies five evaluation levels, and we focus on the cross-generator and cross-architecture settings to assess attribution performance across different levels of granularity, ranging from coarse to fine.

Evaluation Metrics: For Deepfake detection, accuracy serves as the evaluation metric for this binary classification task. For image attribution, Rank 1 and mean Average Precision (mAP) are reported for evaluation. Rank-1 is the proportion of queries whose top prediction matches the ground truth, while mAP is the mean of Average Precision scores across all queries.

Implementation Details: Following the train-test split protocols of GenImage and WildFake, we construct a registered database by randomly selecting 1, 5, and 10 synthetic images per generator, respectively, from the training set. All images in the test set are then used as queries to evaluate the performance. The temperature parameter τ and the weight parameter λ are set to 0.1 and 0.9, respectively, for both datasets. The pretrained model is adapted to the image

Shot	Method	Real		Big		Mid		Wuk		SDV4		SDV5		ADM		GLI		VQ		Avg	
		Rank-1	mAP	Rank-1	mAP	Rank-1	mAP	Rank-1	mAP	Rank-1	mAP	Rank-1	mAP	Rank-1	mAP	Rank-1	mAP	Rank-1	mAP	Rank-1	mAP
1-shot	ResNet [17]	2.5	21.5	19.6	40.6	35.7	53.2	**46.7**	**68.8**	28.6	54.3	9.5	32.7	2.0	17.6	0.5	18.6	11.1	30.1	17.4	37.5
	DIRE [44]	13.1	**38.8**	11.6	26.8	16.1	28.0	14.6	41.7	13.6	29.3	15.1	33.9	7.0	28.0	34.2	59.8	4.0	26.8	14.3	34.8
	ESSP [5]	6.0	23.4	26.1	45.0	8.5	19.1	18.1	39.1	22.6	52.4	**22.6**	**40.7**	9.0	28.2	36.2	53.6	4.0	22.1	17.0	36.0
	Ours	**21.5**	22.7	**97.0**	**88.3**	**74.4**	**91.6**	30.2	54.3	**32.2**	**63.2**	1.5	31.0	**23.6**	**53.7**	**40.2**	**63.6**	**52.8**	**75.4**	**40.4**	**61.5**
5-shot	ResNet [17]	25.1	27.5	10.1	20.4	32.2	25.2	30.7	32.3	17.6	22.6	27.6	33.4	6.5	19.8	15.6	24.8	9.0	19.5	19.4	25.0
	DIRE [44]	30.2	34.1	13.1	18.5	36.7	38.5	20.1	27.7	**26.1**	25.5	8.0	20.0	16.6	21.9	14.6	21.4	3.0	15.4	18.7	24.8
	ESSP [5]	16.1	22.6	14.6	17.6	30.2	30.3	20.6	24.9	11.6	21.5	10.6	23.2	11.1	20.4	**32.7**	31.8	10.1	21.1	17.5	23.7
	Ours	**76.9**	**54.5**	**98.5**	**98.6**	**73.9**	**57.3**	**32.2**	**38.4**	23.6	**47.3**	**43.7**	**46.1**	**36.7**	**56.7**	32.2	**47.1**	**69.3**	**60.5**	**54.1**	**56.3**
10-shot	ResNet [17]	16.1	19.5	10.6	15.4	56.3	30.0	**27.6**	27.7	12.1	23.6	22.6	24.2	17.1	19.8	20.1	25.4	10.1	16.2	21.4	22.4
	DIRE [44]	22.6	22.9	11.6	19.2	0.0	**70.2**	26.6	27.0	22.1	27.0	20.6	29.2	18.1	21.9	22.6	23.9	10.6	17.7	17.2	28.8
	ESSP [5]	17.6	22.1	13.6	18.0	49.7	29.4	27.1	27.9	19.6	22.7	17.6	22.3	16.6	21.6	27.1	25.5	13.1	17.1	22.4	23.0
	Ours	**83.4**	**50.8**	**98.5**	**97.9**	**69.3**	40.4	13.1	**30.7**	**23.6**	**36.1**	**50.3**	**46.8**	**47.2**	**48.7**	**55.3**	**54.7**	**45.7**	**58.5**	**54.0**	**51.6**

Table 1. Performance comparison of AI-generated image attribution on the GenImage dataset under the cross-architecture setting with different numbers of shots. The best score for each shot setting is highlighted in bold.

Shot	Method	Real		VQVAE		COM		DD		VQDM		BigGAN		StyleGAN		StarGAN		DF-GAN		GALIP		GigaGAN		Avg	
		Rank-1	mAP	Rank-1	mAP	Rank-1	mAP	Rank-1	mAP	Rank-1	mAP	Rank-1	mAP	Rank-1	mAP	Rank-1	mAP	Rank-1	mAP	Rank-1	mAP	Rank-1	mAP	Rank-1	mAP
1-shot	ResNet [17]	23.1	40.0	36.7	60.1	**31.7**	**45.5**	2.0	14.3	0.5	16.9	12.6	33.1	5.0	36.9	19.1	45.8	10.1	21.2	45.7	55.6	5.0	33.5	17.4	36.6
	DIRE [44]	9.0	26.5	20.6	37.2	4.5	23.4	**31.2**	**57.9**	9.0	24.1	4.5	16.6	10.1	27.5	**32.2**	55.8	17.1	42.2	47.7	58.7	1.5	15.9	17.0	35.1
	ESSP [5]	27.6	47.7	33.7	50.3	9.5	24.5	6.0	31.5	30.2	41.2	7.5	33.6	13.1	34.1	22.6	46.5	10.1	23.1	45.7	53.6	15.1	39.4	20.1	38.7
	Ours	**68.8**	**79.4**	**64.8**	**81.7**	11.6	32.3	15.1	50.2	**60.8**	**74.6**	**100.0**	**100.0**	**35.7**	**64.6**	29.6	**52.0**	**43.7**	**71.9**	**56.8**	**72.3**	**69.8**	**81.5**	**50.6**	**69.1**
5-shot	ResNet [17]	2.5	12.9	48.2	41.8	0.0	10.0	21.6	31.8	12.6	19.6	21.1	24.4	26.6	25.9	37.7	45.3	22.6	25.9	47.2	55.1	1.0	9.6	21.9	27.5
	DIRE [44]	9.5	17.1	32.7	32.2	11.6	19.3	27.6	24.2	5.0	12.5	13.6	20.7	**36.2**	27.1	36.7	35.5	**27.6**	24.8	48.2	54.8	30.7	26.6	25.4	26.8
	ESSP [5]	19.1	18.9	47.7	49.4	22.1	27.8	10.1	20.8	16.6	22.2	22.6	23.7	28.6	26.8	25.6	27.4	**27.6**	23.0	47.2	50.0	26.1	26.7	26.7	28.8
	Ours	**69.3**	**71.5**	**50.3**	**65.9**	**30.2**	**31.6**	**44.7**	**53.8**	**43.2**	**54.8**	**99.5**	**86.6**	25.6	**34.2**	**51.8**	**51.0**	23.6	**48.1**	**84.4**	**66.1**	**86.9**	**74.6**	**55.4**	**58.0**
10-shot	ResNet [17]	25.1	26.0	**57.8**	35.9	30.2	21.9	13.1	15.1	18.6	15.6	25.6	22.4	30.7	26.1	39.7	28.4	17.6	15.3	49.2	46.8	21.1	20.7	29.9	24.9
	DIRE [44]	31.7	20.4	56.3	43.8	22.6	22.5	19.1	18.4	28.6	20.5	10.6	14.4	**51.8**	30.9	26.1	29.4	19.1	21.1	51.8	57.5	28.6	22.4	31.5	27.4
	ESSP [5]	25.6	25.3	28.6	28.9	20.6	18.0	20.6	17.0	15.6	18.3	27.6	18.8	39.7	26.5	40.7	34.3	19.1	19.0	50.3	60.7	23.6	20.2	28.4	26.1
	Ours	**67.8**	**63.0**	46.7	**52.1**	**54.3**	**44.3**	**62.8**	**45.1**	**52.8**	**60.6**	**97.5**	**99.0**	39.2	**34.4**	**55.3**	**54.3**	**50.8**	**53.6**	**83.4**	**69.2**	**74.4**	**52.6**	**62.3**	**57.1**

Table 2. Performance comparison of AI-generated image attribution on the WildFake dataset under the cross-architecture setting with different numbers of shots. The best score for each shot setting is highlighted in bold.

Shot	Method	Real		GAN-based		Midjourney		Diffusion-based		Avg	
		Rank-1	mAP	Rank-1	mAP	Rank-1	mAP	Rank-1	mAP	Rank-1	mAP
1-shot	ResNet [17]	44.7	71.3	6.0	34.5	**79.9**	**89.0**	19.1	46.9	37.4	60.4
	DIRE [44]	27.6	55.5	18.1	48.1	26.1	49.0	**55.3**	**77.2**	31.8	57.5
	ESSP [5]	54.8	74.5	51.3	71.2	31.7	62.0	5.5	32.4	35.8	60.0
	Ours	**88.4**	**93.0**	**96.0**	**97.9**	74.9	87.4	50.8	72.6	**77.5**	**87.7**
5-shot	ResNet [17]	27.1	40.9	25.6	37.0	68.3	48.8	36.7	50.3	39.4	44.3
	DIRE [44]	36.2	46.3	44.7	55.5	53.3	53.0	35.2	46.3	42.3	50.3
	ESSP [5]	24.1	35.6	51.3	51.0	49.2	45.0	39.7	57.1	41.1	47.2
	Ours	**79.4**	**81.2**	**98.0**	**99.1**	**84.9**	**78.2**	**81.9**	**63.9**	**86.1**	**80.6**
10-shot	ResNet [17]	48.7	47.5	43.2	41.6	66.8	51.0	18.1	34.6	44.2	43.7
	DIRE [44]	35.7	47.9	32.2	35.4	65.3	55.1	22.6	31.2	38.9	42.4
	ESSP [5]	34.7	38.8	32.2	40.7	76.9	58.8	28.1	36.4	43.0	43.7
	Ours	**89.4**	**80.4**	**99.0**	**98.8**	**94.0**	**63.6**	**72.4**	**56.2**	**88.7**	**74.7**

Table 3. Performance comparison of AI-generated image attribution on the GenImage dataset under the cross-generator setting with different numbers of shots. The best score for each shot setting is highlighted in bold.

database using a batch size of 32 and an initial learning rate of 1×10^{-4}, trained for a total of 100 epochs. All experiments are conducted on an Ubuntu 22.04 system with an RTX 4090 GPU, implemented using PyTorch 2.0.1.

5.2. Evaluation of AI-generated Image Attribution

For performance comparison, we construct three baselines, including ResNet-50 [17] and two models specifically tailored for AI-generated image detection, DIRE [44] and ESSP [5]. All these attribution extractors are pre-trained on RGB images and employed to extract features from both the query and registered database images, followed by detection and attribution based on cosine similarity.

Results on GenImage: For the GenImage dataset, we first perform image attribution across eight different generative architectures, with Rank-1 and mAP reported in Table 1. It can be observed that the random guess probability is 11.1%, whereas our method achieves a Rank-1 exceeding 50%. Compared with other attribution extractors, LIDA outperforms ResNet, DIRE, and ESSP in the 10-shot setting in Rank-1 by 32.6%, 36.8%, and 31.6%, respectively, demonstrating that low-bit generative fingerprints effectively capture the noise structures characteristic of different generative architectures. As the number of shots per generative architecture increases from 1 to 5 and 10, the Rank-1 for our method improves by 4.8% and 11.7%, respectively. This suggests that accuracy is positively correlated with the size of the database, as a larger number of reference samples provides richer and more reliable class information, leading to more stable and accurate attribution by the model. Note that the trend of mAP is opposite to that of Rank-1, as having more retrieval samples makes it more challenging to maintain high-quality ranking. Nevertheless, our method consistently outperforms other attribution methods across different shot settings.

Results on WildFake: We observe that some models share similar noise patterns due to their underlying architectures, which makes it challenging to differentiate among them. Therefore, we merge DDPM [19], DDIM [37], and ADM [9] into the DD subset, and DALL·E [28], Imagen [34], Stable Diffusion [33], and Midjourney [26] into

Shot	Method	Big	Mid	WuK	SDV4	SDV5	ADM	GLI	VQ	Avg
1-shot	ResNet [17]	53.8	48.7	56.5	54.3	57.5	48.5	47.0	49.0	51.9
	DIRE [44]	52.0	3.5	52.0	54.0	51.3	53.0	52.5	51.5	46.2
	ESSP [5]	56.8	38.7	51.0	51.5	51.3	52.7	49.0	49.5	50.1
	Ours	**86.8**	**85.4**	**87.2**	**85.6**	**84.5**	**86.5**	**89.1**	**88.1**	**86.5**
5-shot	ResNet [17]	55.3	57.8	55.5	52.3	53.5	52.5	56.0	56.0	54.9
	DIRE [44]	55.8	38.7	54.3	56.3	54.0	52.5	44.5	52.0	51.0
	ESSP [5]	53.3	55.8	49.3	52.0	51.8	53.3	52.5	49.3	52.1
	Ours	**87.2**	**84.8**	**89.4**	**83.7**	**85.8**	**85.7**	**88.8**	**88.8**	**86.8**
10-shot	ResNet [17]	52.3	58.8	55.8	55.0	60.5	54.0	60.8	60.1	57.1
	DIRE [44]	59.8	51.7	51.5	53.3	52.3	56.5	60.8	52.5	54.8
	ESSP [5]	54.7	58.5	53.7	58.5	52.4	54.2	51.9	57.5	55.2
	LARE2* [25]	72.0	62.7	79.6	79.6	79.6	63.5	80.2	76.9	72.5
	FSD* [48]	82.2	80.9	**88.8**	88.8	**88.8**	79.2	**97.1**	76.2	84.1
	Ours	**88.1**	**89.4**	87.4	**89.7**	85.1	**86.1**	90.7	**90.2**	**88.3**

Table 4. Accuracy of AI-generated image detection on the GenImage dataset with different numbers of shots. ($\star$) denotes results taken from [48] and best score for each shot setting is highlighted in bold.

the COM subset. As a result, a total of 10 different generative images need to be attributed. All experimental results exceed the random guess probability of 9.1%. Under the 10-shot setting, our method achieves a Rank-1 accuracy of 62.3%, surpassing ResNet, DIRE, and ESSP by substantial margins of 32.4%, 30.8%, and 33.9%, respectively. Across all subsets, our method achieves the highest attribution performance on BigGAN, attaining a Rank-1 of 100% in the 1-shot setting, greatly outperforming other attribution extractors. All these results demonstrate the effectiveness of the low-bit generative fingerprints we constructed.

Generator-Level Image Attribution: We then combine the six subsets, excluding the BigGAN and Midjourney subsets, to form a diffusion-based set and evaluate performance at the generator level, as shown in Table 3. Without the need to further differentiate between specific diffusion models, only distinguishing between different generative paradigms becomes simpler, resulting in improvements of more than 30% in Rank-1 and 20% in mAP, respectively. It can also be observed that the low-bit generative fingerprint is particularly effective for GAN-based methods, as it boosts the mAP from 41.6% with ResNet to 98.8% with our method.

5.3. Evaluation of AI-generated Image Detection

Few-Shot Detection: For AI-generated image detection, our method consistently achieves the highest average accuracy across different shot settings, as shown in Table 4. Under the 10-shot setting, it outperforms the state-of-the-art few-shot Deepfake detection method FSD [48] by 4.2% on the GenImage dataset. Unlike FSD, which merges WuK, SDV4, and SDV5 into a single subset, our approach preserves the original fine-grained subset partition and still achieves over 85% accuracy across all subsets. This highlights the forensic effectiveness of our low-bit generative fingerprints and the strength of our adaptation strategy.

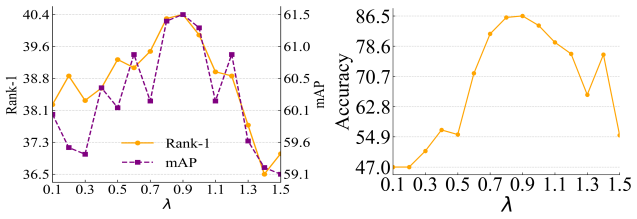
Zero-Shot Detection: We further evaluate the model's performance under the zero-shot setting, which can be regarded as the lower bound of its detection capability. Specifically, we first create low-bit generative fingerprints from ImageNet to pretrain an adapted ResNet-50, as described in Eq. (4). This pretrained model is then used to extract features from query images, which are subsequently compared with the mean feature vector of all real images used during pretraining. By manually selecting a classification threshold of 0.85, queries with similarity above this value are considered real, while those below are classified as fake. As shown in Table 5, even without any prior knowledge of fake images, our method achieves an accuracy of 86.3%, surpassing RIGID, AEROBLADE, Manifold, and FSD by 16.5%, 37.5%, 18.1%, and 9.2%, respectively. All of these competitors are specifically designed for zero-shot Deepfake detection. The high accuracy achieved by our method under zero-shot settings indicates that the extracted features are sufficiently discriminative.

Method	Big	Mid	WuK	SDV4	SDV5	ADM	GLI	VQ	Avg
RIGID [18]	53.0	94.1	87.8	87.0	87.2	51.4	45.9	52.2	69.8
AEROBLADE [30]	58.3	40.2	51.4	52.6	55.1	50.7	29.4	52.8	48.8
Manifold [3]	77.6	55.5	65.4	62.0	63.0	57.3	88.3	76.9	68.2
FSD [48]	62.1	75.1	88.0	88.0	88.0	74.1	93.9	69.1	77.1
Ours	91.0	85.9	86.2	86.3	86.8	85.5	83.9	84.5	86.3

Table 5. Accuracy of zero-shot AI-generated image detection on the GenImage dataset.

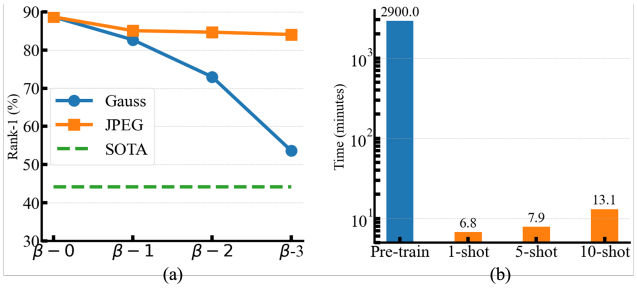
5.4. Ablation Studies and Analyses

Ablation Studies: In Table 6, we discuss the effectiveness of bit-plane-based fingerprints (BF) and three types of loss functions: **(1) Effectiveness of BF**: Comparing results between rows 1 and 2, when using raw images or low-bit generative fingerprints from the registered database as input to a ResNet-50 pretrained on ImageNet, the latter achieves an average mAP that is 10.6% mAP higher than the former. This demonstrates the effectiveness of BF in capturing generator-specific noise structures. **(2) Effectiveness of unsupervised pre-training**: Comparing results between rows 2 and 3, by training the model with BF of real images under supervision of the pretext task training loss L_P, the model achieves an additional 1.5% mAP, highlighting the importance of unsupervised pre-training. **(3) Effectiveness of attribution loss L_A**: Comparing results between rows 3 and 4, adapting the model to the registered database with only one fake image per generator under supervision of the attribution loss L_A achieves an mAP of 53.3%, outperforming the model without adaptation by 3.7%. This indicates that the attribution loss effectively guides sample features to cluster around their corresponding class centers. **(4) Effectiveness of Deepfake detection loss L_D**: Comparing results between rows 4 and 5, incorporating the Deepfake detection loss L_D further improves overall perfor-



(a) (b)

Figure 5. **Impact of loss weight** for (a) AI-generated image attribution and (b) AI-generated image detection.



(a) (b)

Figure 6. **Robustness analysis and comparison of adaptation time:** (a) Robustness to image degradation under Gaussian blur and JPEG compression, where β represents the degree of blurring and image quality. (b) Comparison of the running time for unsupervised pre-training and few-shot attribution adaptation.

mance by 8.2%. This suggests that the real-prototype-based contrastive loss effectively increases the separation between real and fake images in the feature space.

Comparison of Different Losses during Adaptation: In Sec. 4.3, we choose the center loss and the contrastive loss as the attribution loss L_A and the Deepfake detection loss L_D, respectively. Table 7 presents the results of replacing these losses with the cross-entropy loss. Replacing either L_A or L_D with cross-entropy results in performance degradations of 1.8% and 0.8%, respectively, while substituting both leads to a drop of 3.9%.

Impact of Loss Weight λ: In Figure 5, we analyze the impact of the loss-balancing hyperparameter λ in Eq. (8). Both Rank-1 and mAP increase as λ grows, reaching their peak at $\lambda = 0.9$ (ACC = 40.4%, mAP = 61.5%), while accuracy also attains its maximum of 86.5%. Beyond this value, performance declines, indicating that placing excessive emphasis on separating real and fake samples can hinder the discrimination among different fake generators.

Robustness of Image Degradation: To assess the robustness of our forensic method, we apply Gaussian blur ($\theta = 0, 1, 2, 3$) and JPEG compression (quality = 100%, 95%, 90%, 85%) to the raw images during testing, as shown in Figure 6(a). A larger value of the horizontal-axis parameter β indicates stronger degradation or lower JPEG quality. Our method exhibits strong robustness under varying levels of JPEG compression. Gaussian blur directly distorts the distribution of low-bit generative fingerprints. However, even under such degradation, the resulting features still preserve generator-specific noise patterns far more effectively than RGB-based features from unaltered images.

BF	L_P	L_A	L_D	Real	Big	Mid	Wuk	SDV4	SDV5	ADM	GLI	VQ	Avg
✗	✗	✗	✗	21.5	40.6	53.2	68.8	54.3	32.7	17.6	18.6	30.1	37.5
✓	✗	✗	✗	4.8	87.9	22.7	52.7	56.8	33.6	78.2	46.6	49.6	48.1
✓	✓	✗	✗	46.0	97.1	38.6	30.8	42.7	43.3	55.7	45.5	46.7	49.6
✓	✓	✓	✗	56.1	97.5	47.5	34.7	54.7	40.4	44.6	55.4	48.5	53.3
✓	✓	✓	✓	22.7	88.3	91.6	54.3	63.2	31.0	53.7	63.6	75.4	61.5

Table 6. Ablation study on the effectiveness of bit-plane-based fingerprints (BF) and different loss functions for one-shot AI-generated image attribution on the GenImage dataset. L_p, L_A, and L_D denote the pretext task training loss, image attribution loss, and Deepfake detection loss, respectively.

L_A	L_D	Real	Big	Mid	Wuk	SDV4	SDV5	ADM	GLI	VQ	Avg
CE	CE	68.8	94.7	33.5	41.5	32.5	71.2	45.6	50.6	79.9	57.6
CE	–	72.2	93.4	57.7	38.9	53.3	64.6	42.1	50.3	64.6	59.7
–	CE	47.8	93.5	65.2	67.2	60.3	28.3	76.1	58.7	49.3	60.7
–	–	22.7	88.3	91.6	54.3	63.2	31.0	53.7	63.6	75.4	61.5

Table 7. Effects of replacing the losses with cross-entropy in few-shot attribution adaptation. CE denotes the cross-entropy loss and '–' denotes the original loss.

Practical Efficiency Analysis: We report the training time for both unsupervised pre-training and attribution adaptation in Figure 6(b). The running time overhead introduced by our few-shot attribution adaptation is negligible compared to pre-training. When encountering unseen AI-generated images, the proposed retrieval-based attribution paradigm enables accurate attribution through rapid adaptation, overcoming the limitations of conventional classification-based approaches that require full retraining. Since low-bit fingerprint generation relies on efficient binary operations and the attribution encoder is based on ResNet-50, our model operates at millisecond-level inference speed.

6. Conclusion

We propose a versatile and efficient AI-generated image attribution framework called LIDA, which treats attribution as instance retrieval and leverages bit-planes for generative fingerprints extraction. LIDA only trains an attribution encoder using an adapted ResNet-50, and the training involves unsupervised pre-training and few-shot attribution adaptation. Our forensic technology succeeds in Deepfake attribution and detection under both zero-shot and few-shot settings on two popular AI-generated image datasets. Further experiments demonstrate the effectiveness of each component and the robustness to perturbations and degradations. By relying solely on efficient binary operations and a lightweight encoder, our approach achieves low computational complexity for both training and inference.

Acknowledgments

This work was supported by National Science Foundation of China (62302093, 52441503), Jiangsu Province Natural Science Fund (BK20230833), the CIPS-SMP-Zhipu Large Model Fund, the Fundamental Research Funds for the Central Universities (2242025K30024), the Open Research Fund of the State Key Laboratory of Multimodal Artificial Intelligence Systems (E5SP060116) and the Big Data Computing Center of Southeast University.

References

[1] Fengxiang Bie, Yibo Yang, Zhongzhu Zhou, Adam Ghanem, Minjia Zhang, Zhewei Yao, Xiaoxia Wu, Connor Holmes, Pareesa Golnari, David A Clifton, et al. RenAIssance: A survey into AI text-to-image generation in the era of large model. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47(3):2212–2231, 2025. 1

[2] Andrew Brock, Jeff Donahue, and Karen Simonyan. Large scale GAN training for high fidelity natural image synthesis. In *ICLR*, 2019. 5

[3] Jonathan Brokman, Amit Giloni, Omer Hofman, Roman Vainshtein, Hisashi Kojima, and Guy Gilboa. Manifold induced biases for zero-shot and few-shot detection of generated images. *arXiv preprint arXiv:2504.15470*, 2025. 7

[4] 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

[5] Jiaxuan Chen, Jieteng Yao, and Li Niu. A single simple patch is all you need for ai-generated image detection. *arXiv preprint arXiv:2402.01123*, 2024. 6, 7

[6] Mo Chen, Jessica Fridrich, Miroslav Goljan, and Jan Lukás. Determining image origin and integrity using sensor noise. *IEEE Transactions on Information Forensics and Security*, 3 (1):74–90, 2008. 4

[7] Civitai. https://civitai.com/, 2022. 5

[8] 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. Ieee, 2009. 4, 5

[9] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in neural information processing systems*, 34:8780–8794, 2021. 6

[10] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in neural information processing systems*, 34:8780–8794, 2021. 3, 4, 5

[11] Pierre Fernandez, Guillaume Couairon, Hervé Jégou, Matthijs Douze, and Teddy Furon. The stable signature: Rooting watermarks in latent diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 22466–22477, 2023. 2

[12] 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. 2

[13] 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*, pages 14094–14103, 2021. 2, 3

[14] Shuyang Gu, Dong Chen, Jianmin Bao, Fang Wen, Bo Zhang, Dongdong Chen, Lu Yuan, and Baining Guo. Vector quantized diffusion model for text-to-image synthesis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10696–10706, 2022. 5

[15] Fabrizio Guillaro, Giada Zingarini, Ben Usman, Avneesh Sud, Davide Cozzolino, and Luisa Verdoliva. A bias-free training paradigm for more general ai-generated image detection. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 18685–18694, 2025. 2

[16] Sam Gunn, Xuandong Zhao, and Dawn Song. An undetectable watermark for generative image models. In *International Conference on Learning Representations*, 2025. 2

[17] 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. 4, 6, 7

[18] Zhiyuan He, Pin-Yu Chen, and Tsung-Yi Ho. Rigid: A training-free and model-agnostic framework for robust ai-generated image detection. *arXiv preprint arXiv:2405.20112*, 2024. 7

[19] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020. 6

[20] Yan Hong, Jianming Feng, Haoxing Chen, Jun Lan, Huijia Zhu, Weiqiang Wang, and Jianfu Zhang. Wildfake: A large-scale and hierarchical dataset for ai-generated images detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 3500–3508, 2025. 2, 5

[21] Zexi Jia, Chuanwei Huang, Yeshuang Zhu, Hongyan Fei, Xiaoyue Duan, Zhiqiang Yuan, Ying Deng, Jiapei Zhang, Jinchao Zhang, and Jie Zhou. Secret lies in color: Enhancing ai-generated images detection with color distribution analysis. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 13445–13454, 2025. 2

[22] Jialiang Li, Haoyue Wang, Sheng Li, Zhenxing Qian, Xinpeng Zhang, and Athanasios V Vasilakos. Are handcrafted filters helpful for attributing ai-generated images? In *Proceedings of the ACM International Conference on Multimedia*, pages 10698–10706, 2024. 3

[23] 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. 5

[24] 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. 4

[25] Yunpeng Luo, Junlong Du, Ke Yan, and Shouhong Ding. Lareˆ 2: Latent reconstruction error based method for diffusion-generated image detection. In *Proceedings of*

the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 17006–17015, 2024. 7

[26] Midjourney. https://www.midjourney.com/home/, 2022.5. 5, 6

[27] Alex Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and Mark Chen. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. *arXiv preprint arXiv:2112.10741*, 2021. 4, 5

[28] 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

[29] Ahmad Rezaei, Mohammad Akbari, Saeed Ranjbar Alvar, Arezou Fatemi, and Yong Zhang. Lawa: Using latent space for in-generation image watermarking. In *European Conference on Computer Vision*, pages 118–136. Springer, 2024. 2

[30] 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*, pages 9130–9140, 2024. 7

[31] Anderson Rocha, Walter Scheirer, Terrance Boult, and Siome Goldenstein. Vision of the unseen: Current trends and challenges in digital image and video forensics. *ACM Computing Surveys*, 43(4):1–42, 2011. 1

[32] 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. 3, 4, 5

[33] 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

[34] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in neural information processing systems*, 35:36479–36494, 2022. 6

[35] 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 ACM SIGSAC Conference on Computer and Communications Security*, pages 3418–3432, 2023. 3

[36] Hae Jin Song, Mahyar Khayatkhoei, and Wael AbdAlmageed. ManiFPT: Defining and analyzing fingerprints of generative models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10791–10801, 2024. 1

[37] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020. 6

[38] Zhimin Sun, Shen Chen, Taiping Yao, Bangjie Yin, Ran Yi, Shouhong Ding, and Lizhuang Ma. Contrastive pseudo learning for open-world deepfake attribution. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 20882–20892, 2023. 2, 3

[39] Xiaofeng Tan, Hongsong Wang, Xin Geng, and Pan Zhou. SoPo: Text-to-motion generation using semi-online preference optimization. In *Annual Conference on Neural Information Processing Systems*, 2025. 1

[40] Xiaofeng Tan, Wanjiang Weng, Haodong Lei, and Hongsong Wang. Easytune: Efficient step-aware fine-tuning for diffusion-based motion generation. In *International Conference on Learning Representations*, 2026. 1

[41] Hongsong Wang, Renxi Cheng, Yang Zhang, Chaolei Han, and Jie Gui. LOTA: Bit-planes guided ai-generated image detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 17246–17255, 2025. 2, 4

[42] Hongsong Wang, Renxi Cheng, Linjiang Huang, Fang Zhao, and Jie Gui. RAID: Towards robust AI-generated image detection with bit reversed images, 2026. 4

[43] Hongsong Wang, Wenjing Yan, Qiuxia Lai, and Xin Geng. Temporal consistency-aware text-to-motion generation. *Visual Intelligence*, 4(1):7, 2026. 1

[44] 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*, pages 22445–22455, 2023. 6, 7

[45] Yandong Wen, Kaipeng Zhang, Zhifeng Li, and Yu Qiao. A discriminative feature learning approach for deep face recognition. In *European conference on computer vision*, pages 499–515. Springer, 2016. 5

[46] Yuxin Wen, John Kirchenbauer, Jonas Geiping, and Tom Goldstein. Tree-rings watermarks: Invisible fingerprints for diffusion images. *Advances in Neural Information Processing Systems*, 36:58047–58063, 2023. 2

[47] Wanjiang Weng, Xiaofeng Tan, Junbo Wang, Guo-Sen Xie, Pan Zhou, and Hongsong Wang. ReAlign: Text-to-motion generation via step-aware reward-guided alignment. In *Annual AAAI Conference on Artificial Intelligence*, 2026. 1

[48] Shiyu Wu, Jing Liu, Jing Li, and Yequan Wang. Few-shot learner generalizes across ai-generated image detection. *arXiv preprint arXiv:2501.08763*, 2025. 7

[49] Wukong. https://xihe.mindspore.cn/, 2022.5. 3, 4, 5

[50] Qiang Xu, Wenpeng Mu, Jianing Li, Tanfeng Sun, and Xinghao Jiang. Advancements in ai-generated content forensics: A systematic literature review. *ACM Computing Surveys*, 58 (3):1–36, 2025. 1

[51] Haiwei Xue, Xiangyang Luo, Zhanghao Hu, Xin Zhang, Xunzhi Xiang, Yuqin Dai, Jianzhuang Liu, Zhensong Zhang, Minglei Li, Jian Yang, Fei Ma, Zhiyong Wu, Changpeng Yang, Zonghong Dai, and Fei Richard Yu. Human motion video generation: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47(11):10709–10730, 2025. 1

[52] Zhiyuan Yan, Yuhao Luo, Siwei Lyu, Qingshan Liu, and Baoyuan Wu. Transcending forgery specificity with latent

space augmentation for generalizable deepfake detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8984–8994, 2024. 2

[53] Tianyun Yang, Ziyao Huang, Juan Cao, Lei Li, and Xirong Li. Deepfake network architecture attribution. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 4662–4670, 2022. 2

[54] 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

[55] Zijin Yang, Kai Zeng, Kejiang Chen, Han Fang, Weiming Zhang, and Nenghai Yu. Gaussian shading: Provable performance-lossless image watermarking for diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12162–12171, 2024. 2

[56] 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*, pages 7556–7566, 2019. 2, 4

[57] Ning Yu, Vladislav Skripniuk, Sahar Abdelnabi, and Mario Fritz. Artificial fingerprinting for generative models: Rooting deepfake attribution in training data. In *Proceedings of the IEEE/CVF International conference on computer vision*, pages 14448–14457, 2021. 1, 4

[58] Hanlin Zhang, Benjamin L Edelman, Danilo Francati, Daniele Venturi, Giuseppe Ateniese, and Boaz Barak. Watermarks in the sand: Impossibility of strong watermarking for generative models. In *ICLR Workshop on Secure and Trustworthy Large Language Models*, 2024. 2

[59] Shoulong Zhang, Haomin Li, Kaiwen Sun, Hejia Chen, Yan Wang, and Shuai Li. Security and privacy challenges of aigc in metaverse: A comprehensive survey. *ACM Computing Surveys*, 57(10):1–37, 2025. 1

[60] Mingjian Zhu, Hanting Chen, Qiangyu Yan, Xudong Huang, Guanyu Lin, Wei Li, Zhijun Tu, Hailin Hu, Jie Hu, and Yunhe Wang. Genimage: A million-scale benchmark for detecting ai-generated image. *Advances in Neural Information Processing Systems*, 36:77771–77782, 2023. 2, 5