Few-Shot Synthetic Image Attribution: Identifying Unseen Generators with Limited Samples

Shiyu Wu*[1,2,3], Shuyan Li*[4], Jing Li[5], Jing Liu†[1,3], and Yequan Wang†[2,6]

[1] Institute of Automation, Chinese Academy of Sciences, Beijing, China
[2] Beijing Academy of Artificial Intelligence, Beijing, China
[3] University of Chinese Academy of Sciences, Beijing, China
[4] Tsinghua Shenzhen International Graduate School, Tsinghua University, Shenzhen
[5] Harbin Institute of Technology, Shenzhen, China
[6] Peking University, Beijing, China
wushiyu2022@ia.ac.cn, Lishuyan@sz.tsinghua.edu.cn, jingli.phd@hotmail.com, jliu@nlpr.ia.ac.cn, tshwangyequan@gmail.com

Abstract. AI-generated image (AIGI) attribution presents a pressing challenge that goes beyond mere AIGI detection, aiming to identify the source model or technique responsible for a synthetic image. However, most previous source attribution methods operate in a closed-set manner, which necessitates retraining to recognize any novel category, preventing adaptation to the rapid evolution of image generation. In this work, we propose a new paradigm for synthetic image attribution, termed few-shot attribution. This paradigm targets the reliable identification of unseen generators using only limited samples, making it highly suitable for real-world applications. To facilitate this work, we construct OmniFake, a large-scale, well-categorized synthetic image dataset that contains 1.17 million images from 45 distinct generators. We further introduce OmniDFA (Omni Detector and Few-shot Attributor), a few-shot attribution baseline that not only assesses the authenticity of images but also determines their synthesis origins. Experiments demonstrate that OmniDFA exhibits excellent capability in few-shot attribution and achieves state-of-the-art generalization performance in AIGI detection. Our dataset and code are available at https://github.com/teheperinko541/OmniDFA.

Keywords: Image Attribution · Synthetic Image Detection · Few-Shot Learning

1 Introduction

Generative models [21] now forge photorealistic images that defy visual scrutiny, collapsing the boundary between authentic and synthetic. The growing threat of widespread misuse requires operational methods to not only distinguish AI-generated images (AIGI) from real ones but also trace their generative origins,

*Equal contribution. †Corresponding Authors.

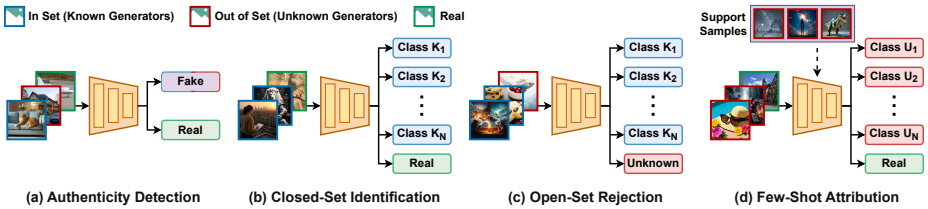


Fig. 1: Comparison of task-specific pipelines for synthetic analysis. Our new attribution paradigm (d) offers dramatically improved scalability over previous works (b) and (c).

which is critical for understanding model-specific vulnerabilities. However, existing tools for detecting and analyzing synthetic content struggle to match the pace of advancing generation methods. This poses a significant challenge for such countermeasures in an open-set scenario, where they must handle data from generative models unseen during training, necessitating robust generalization ability and strong discrimination capability.

Synthetic image attribution represents a broader and more fine-grained task in forgery analysis, moving past the limitations of coarse-grained authenticity detection. It aims to identify the specific generative source among multiple candidates, consequently providing deeper insights into image provenance. However, existing image attribution paradigms are confined to identifying solely known generator types, rendering them ineffective against the rapidly evolving landscape of image generation. Specifically, closed-set methods, as illustrated in Fig. 1 (b), can only trace images to generators seen during the training phase, which prevents them from handling unknown sources [16]. In contrast, open-set methods [9, 51] categorize unseen samples by collapsing all unknown sources into a single class, as shown in Fig. 1 (c). Adapting these approaches to incorporate new categories necessitates retraining the entire network, which is resource-intensive and unsustainable. This inherent inflexibility renders them quickly obsolete as new generative models emerge. The question of how to efficiently adapt to novel generators therefore represents a critical challenge for image attribution.

In this work, we propose a new paradigm termed few-shot attribution to tackle synthetic image attribution from a novel perspective. This paradigm operates in a fully open-set scenario, requiring the model to identify image sources unseen during training, as depicted in Fig. 1 (d). In contrast to prior tasks, our focus lies in enhancing generalization and scalability to novel categories. We argue that the advantages of few-shot learning, such as rapid generalization from limited data, are also applicable to image attribution. Thus, we formulate this novel task, which challenges models to identify unknown generators from a minimal number of support samples. The model learns to rapidly extract critical artifact evidence from a limited number of samples, and once trained, it can be swiftly adapted to novel categories at test time. This training-free adaptation offers a substantial advantage over conventional methods: it is highly scalable, cost-effective, and well-suited for real-world deployment.

Table 1: Comparison with existing AIGI datasets. "Attribution" indicates that the dataset is designed for attribution task, ensuring pairwise separability among classes.

AIGI Dataset	Year	Generators	Fake Images	Attribution
CNNSpot [52]	CVPR 2020	11	362 K	✓
DiffusionForensics [53]	ICCV 2023	11	439 K	✓
GenImage [63]	NeurIPS 2023	8	1.33 M	✗
DE-FAKE [45]	CSS 2023	4	222 K	✓
UniversalFakeDetect [38]	CVPR 2023	19	400 K	✗
Artifact [41]	ICIP 2023	25	2.49 M	✗
AIGCBenchmark [62]	Arxiv 2024	17	360 K	✗
DRCT-2M [7]	ICML 2024	16	2.00 M	✗
ImagiNet [4]	Arxiv 2025	8	100 K	✓
WildFake [22]	AAAI 2025	23	2.55 M	✗
OmniFake	ECCV 2026	**45**	1.17 M	✓

To facilitate our work, we propose OmniFake, a well-categorized dataset specifically designed for multi-class attribution. We construct this dataset as our experiments require a large number of categories for both training and testing, which most previous datasets cannot provide. Specifically, we collect generated images from 45 distinct generative models spanning GANs [18], diffusion models [21], autoregressive models [57], and hybrid architectures [31]. We maintain a substantial collection of synthetic images for each class, ensuring comprehensive coverage and diversity. OmniFake significantly surpasses prior datasets [62,63] in both model diversity and up-to-date coverage, with a clear comparison presented in Tab. 1. Crucially, none of the generators employ parameter-efficient fine-tuning (PEFT) techniques (e.g., DreamBooth [42] and LoRA [23]), ensuring inherently distinct generative mechanisms rather than feature-similar variants derived from minor adaptations. Such diversity enriches the fundamental resources available for both AIGI detection and attribution research.

Leveraging our comprehensive dataset, we propose OmniDFA (Omni Detector and Few-shot Attributor), an innovative framework that simultaneously addresses AIGI detection and open-set few-shot attribution. We adopt a dual-path architecture that extracts image characteristics from both low-level and high-level perspectives to effectively capture fine-grained details while maintaining global representations. We integrate supervised contrastive learning [27] with center loss [54] to simultaneously address both the attribution and detection tasks. Experimental results show that OmniDFA achieves state-of-the-art performance on our proposed OmniFake dataset, surpassing previous methods by 2.03% in 5-way attribution and 5.83% in AIGI detection. Additionally, it exhibits robust zero-shot detection performance on GenImage [63] and Chameleon [58], confirming the generalizability and effectiveness of our approach.

Our contributions are summarized as follows:

– We introduce the new task of few-shot attribution, which leverages limited samples to identify unseen generators in an open-set scenario.

- We construct the OmniFake dataset, a large-scale, well-categorized synthetic image dataset that enables in-depth study of model-specific patterns for image attribution.
- We propose OmniDFA, a novel framework integrating AIGI detection with few-shot attribution, establishing a strong baseline with broad applicability.

2 Related Work

2.1 Synthetic Datasets

Early datasets for AI-generated image detection primarily relied on Generative Adversarial Networks (GANs) [18]. CNNSpot [52] introduces a widely-used dataset and establishes an evaluation paradigm where detectors are trained on images from ProGAN [26] and LSUN [61], then tested across various other models. As diffusion models [21] advance in image synthesis, many datasets [13, 38] have incorporated images from these models to evaluate and improve detection methods. GenImage [63] introduces the first million-scale synthetic dataset, paired with real images from ImageNet [43]. However, these datasets only cover a limited number of generators, restricting the generalization capability of detectors. Recent studies [2] have begun emphasizing synthetic model diversity in dataset construction. WildFake [22] introduces an in-the-wild collected dataset featuring diverse fake images from various generative models, covering a wide range of content and styles. Community Forensics [40] addresses the diversity limitation by aggregating samples from thousands of generative models. However, these large-scale datasets prioritize the detection task and contain numerous variants of generative models with consistent architectures. The high feature similarity among them can adversely affect both training and testing for the attribution task. Moreover, such datasets rely heavily on diffusion models, which is misaligned with the current paradigm shift toward autoregressive architectures.

2.2 Detection Methods

Previous research on synthetic image detection has been dedicated to feature extraction from multiple perspectives, such as frequency [49], semantics [48], and reconstruction difficulty [36, 53]. Although they have achieved promising results on previous datasets, recent studies [19, 28] suggest that data augmentation techniques such as resizing and JPEG compression can significantly degrade the performance of detectors. To extract more robust features, OOC-CLIP [33] employs the pre-trained CLIP model as its feature extractor, while FakeReasoning [17] utilizes vision-language models, which not only effectively incorporate textual prompts but also produce human-interpretable feature descriptions. Recent studies [10] have introduced more sophisticated learning objectives to overcome the limitations of binary classifiers. DetectByReal [3] introduces learning with real images only by analyzing pixel-level distributions. NTF [30] employs self-supervised feature mapping to enhance transfer learning, while FSD [56]

learns a specialized metric space to distinguish unseen fake images with given samples. Inspired by these advances, we aim to explore the potential of metric learning not only for detection but also for out-of-distribution attribution.

2.3 Attribution Methods

Image attribution aims to identify the source model of generated images. Early studies [5, 16] focus on closed-set classification over GANs, where all generators encountered during testing are assumed to be known at training time. They demonstrate that different generative models exhibit distinct fingerprints, which can be distinguished by a neural network. However, the closed-set setting has limited practicality in real-world scenarios, as new generative models are continuously being released. To address this limitation, the open-set rejection task has been introduced [15], where a model is required to categorize fake images from unseen generators into an additional rejection class, thereby distinguishing them from known categories. For instance, DNA-Det [59] captures globally consistent architectural traces through patch-based contrastive learning. CPL [46] introduces a global-local voting module to evaluate attribution performance on various GAN-generated face images. DE-FAKE [45] employs the CLIP model to attribute fake images created by text-to-image diffusion models.

Nevertheless, most existing methods are inherently limited to attributing images to generative models seen during training. When confronted with a newly emerging generator, they typically require retraining to adapt. In this work, we aim to explore rapid model adaptability to continuously emerging generative models—a practical demand that remains largely unaddressed. We propose an open-set few-shot paradigm to evaluate model performance in detecting forgery clues from unseen, novel generators. This necessitates that the model rapidly learns to identify salient features from limited samples, while also distinguishing them from features characteristic of other generative models.

3 Construction of OmniFake

As presented in Tab. 1, most million-scale datasets do not emphasize the uniqueness of each category, collecting data from generators with similar model architectures. To support our investigation in multi-class attribution, we carefully construct a dataset comprising fake images sampled from a diverse set of models. For a visual overview, a snapshot of our dataset is shown in Fig. 2.

3.1 Fake Image Collection

A key limitation of prior synthetic image datasets for attribution is the scarcity of categories, which considerably restricts research progress. To ensure comprehensive data diversity, we gather fake images through three distinct channels: (1) established datasets and benchmarks, (2) community-shared collections, and (3) synthetic images generated using open-source models. We curate data from

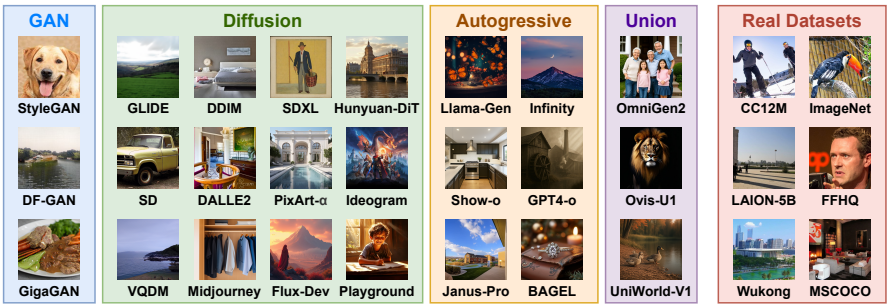


Fig. 2: Samples of the OmniFake dataset. Our dataset covers a broad spectrum of generative models, with real images sourced from multiple datasets for comprehensive coverage.

open-source datasets including GenImage [63], WildFake [22], and MPBench [35]. The final collection of this channel comprises images produced by 19 distinct generators spanning GANs, diffusion models, and VAEs. To enrich our dataset with data from closed-source models, we select 8 synthetic datasets from Hugging Face [24], such as DALLE3 [39], Ideogram [25], and Midjourney V6 [37]. We also select 18 state-of-the-art open-source models from Hugging Face or the models' official repositories, and generate corresponding images based on a comprehensive collection of prompts. These include flow-matching models such as Hunyuan-DiT [29] and SD3-Medium [14], as well as autoregressive models like Janus-Pro [8], BAGEL [12], and Show-o [57]. Additionally, we incorporate several cutting-edge multimodal unified models that utilize diffusion-based decoders for high-fidelity image generation, including OmniGen2 [55], Ovis-U1 [50], and UniWorld-V1 [31], among others. Details on the image generators and the synthesis pipeline are provided in the Supplementary Material.

3.2 Real Image Collection

To establish a comprehensive benchmark for image forensics research, we curate a diverse collection of authentic images from 10 publicly available datasets spanning multiple domains. We carefully structure our dataset into two distinct categories: (1) Large-scale raw datasets including LAION [44], Wukong [20], and CC12M [6], providing vast amounts of in-the-wild data; (2) Carefully constructed datasets such as ImageNet [43] and COCO [32], offering high-quality images from specific domains. This approach ensures a broad and representative set of real images, facilitating robust evaluation of forensic techniques across varied contexts. The final curated collection covers diverse image characteristics including resolution, scene complexity, and photographic styles.

3.3 Analyses of OmniFake

The OmniFake dataset comprises a total of 2.34 M images in the training set, with a balanced distribution of 1.17 M real images and 1.17 M synthetic images.

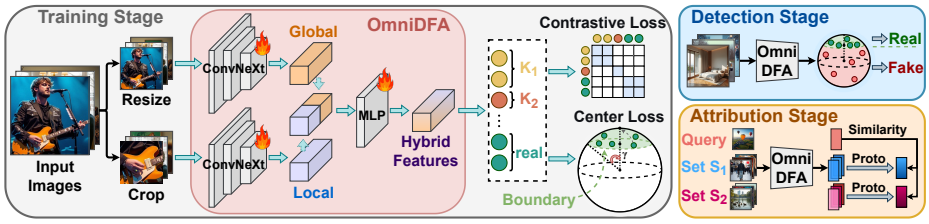


Fig. 3: Overview of OmniDFA. Our dual-path architecture captures both low-level and high-level image features, balancing fine details and global representations. We use contrastive learning to enhance feature discrimination and apply sphere center loss to compactly cluster real samples. Our model is capable of tackling both the authenticity detection and image attribution tasks.

The synthetic images originate from 45 distinct generative models that span a broad spectrum of architectures. This establishes a level of granularity and scale for few-shot attribution, which remains unattainable in existing benchmarks. To ensure sufficient intra-class diversity, we guarantee that each synthetic image category in the training set contains at least 20 K samples. We also construct a separate test set comprising 90 K synthetic images and an equal number of real images for evaluation. To maintain the conceptual clarity of the open-set attribution task, our dataset comprises only architecturally distinct base models, explicitly excluding fine-tuned or LoRA-adapted variants. We omit them to focus our evaluation on rigorous cross-structure generalization, avoiding noise from trivial parameter shifts. OmniFake provides a rich foundation for investigating the impact of model architectures on generalization, thereby advancing synthetic image detection and attribution.

4 Method

In this section, we present in detail our OmniDFA, which is a novel AI-generated image (AIGI) framework that jointly addresses authenticity detection and few-shot open-set attribution. Fig. 3 illustrates the overall architecture of our proposed method.

4.1 Few-Shot Image Attribution

Few-shot learning provides a robust framework for our attribution task, enabling models to generalize to new categories from only a limited number of examples. This setting is particularly suitable for AIGI attribution, as generative models are numerous and continuously evolving, making extensive per-model data collection impractical. Building upon this concept, we formally introduce the task of few-shot attribution, as illustrated in Fig. 1 (d).

In the context of attribution, each distinct generative model is treated as a unique class, and authentic images serve as a separate reference category.

Specifically, the attribution task is structured as an N-way K-shot problem. We are provided with a support set $\mathcal{S} = \cup_{i=1}^{N} \mathcal{S}_i$, which comprises samples from N distinct generator classes, serving as the definitive reference library for source identification. Each subset $\mathcal{S}_i = \{(\mathbf{x}_{i1}, y_{i1}), (\mathbf{x}_{i2}, y_{i2}), \ldots, (\mathbf{x}_{iK}, y_{iK})\}$ consists of K sample images produced by the i-th generator, where $\mathbf{x}_{ij} \in \mathbb{R}^D$ is an image with D dimensions and $y_{ij} = i$ identifies the source generator. Given a query set of unlabeled images, the objective is to leverage the limited information in $\mathcal{S}$ to accurately attribute each query sample to its originating generator.

4.2 Structure of OmniDFA

Since current generative models are able to produce high-resolution images, resizing inputs to a fixed scale as in prior studies [63] tends to degrade fine-grained details. Conversely, maintaining the original resolution without resizing often results in information loss due to computational constraints or input size limitations of the model input. To address this problem, we propose a dual sampling strategy that captures both local and global features, thereby maximizing the retention of critical visual information across varying image resolutions.

As illustrated in Fig. 3, OmniDFA processes an image $\mathbf{x}_i$ through two complementary pathways: (1) an aspect-ratio-preserving resize, where the shorter edge is scaled to the target input size followed by center cropping, to retain holistic global features; and (2) a direct high-resolution crop from the original image, which maximally preserves fine-grained textures and local details. The distinct global and local features from the high-level and low-level branches are channel-wise concatenated and subsequently processed through a multi-layer perceptron (MLP) to produce the final feature representation $f_\theta(\mathbf{x}_i)$, where θ denotes the set of all trainable parameters in the network.

4.3 Unified Attribution and Detection Objective

To facilitate both image attribution and detection, we leverage supervised contrastive learning [27] integrated with sphere center loss, which jointly enforces angular margin maximization and feature centroid convergence. Specifically, for a batch of N images $\{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N\}$ with known categories, we first extract their features $f_\theta(\mathbf{x}_i)$ and then apply L2-normalization to obtain the corresponding vectors $\mathbf{z}_i = f_\theta(\mathbf{x}_i)/||f_\theta(\mathbf{x}_i)||_2$. The supervised contrastive loss can be formulated as follows:

$$\mathcal{L}_{sup} = \sum_{i=1}^{N} \frac{-1}{|P(i)|} \sum_{p \in P(i)} \log \frac{e^{(\mathbf{z}_i \cdot \mathbf{z}_p/\tau)}}{\sum_{a \in P(i) \setminus \{p\}} e^{(\mathbf{z}_i \cdot \mathbf{z}_a/\tau)}}, \tag{1}$$

where $P(i)$ denotes the set of samples in the same class as $\mathbf{x}_i$, and τ denotes the temperature factor that scales the similarity scores.

Given that the substantial quantity of real images may lead to feature dispersion in the embedding space, we implement a sphere center loss specifically

for the real category to better regularize its feature distribution, formulated as:

$$\mathcal{L}_{cen} = \frac{1}{|P_r|} \sum_{p \in P_r} (1 - \mathbf{z}_p \cdot \mathbf{c}_r), \tag{2}$$

where $\mathbf{c}_r$ is a learnable L2-normalized vector representing the feature center of the real class, and P_r denotes the set of all real samples in the current mini-batch. By unifying these constraints, we formulate the final learning objective with the scaling factor λ:

$$\mathcal{L} = \mathcal{L}_{sup} + \lambda \mathcal{L}_{cen}. \tag{3}$$

To enable direct measurement of image authenticity, we introduce a boundary threshold γ defined as the maximum angular separation between real image features and the real center. We compute this bound using Tukey's fences and update it via momentum-based adjustment, with full details provided in the Supplementary Material.

5 Experiment

In this section, we first present the benchmarks involved and our experimental configurations. We then conduct comprehensive comparisons between OmniDFA and several state-of-the-art synthetic image detection and attribution methods.

5.1 Benchmarks and Evaluation Metrics

Datasets and benchmarks. We first conduct comprehensive experiments on our OmniFake dataset. To adapt our framework for the few-shot attribution task, we perform our experiments using 3-fold cross-validation by randomly dividing the OmniFake dataset into three balanced parts. In each validation round, we perform training on two parts while using the remaining part for testing. The evaluation is conducted under both 5-way 10-shot and 15-way 10-shot scenarios. To comprehensively evaluate the detection capability of our model, we conduct extensive zero-shot experiments on the GenImage dataset [63] and the Chameleon dataset [58], demonstrating its generalization ability in both dataset-specific and real-world contexts.
Evaluation metrics. Following established practices in previous research [56, 63], we adopt accuracy (ACC) and macro-averaged F1 score (Macro-F1) as the evaluation metrics for the attribution task. For the detection task, we additionally report average precision (AP), computed with a threshold step of 0.05.

5.2 Experimental Settings

We adopt the ConvNeXt-Small [34] pretrained on ImageNet as the feature extractor of our model, which outputs a vector of 512 dimensions. These features

Table 2: Results of open-set few-shot attribution on all the three parts of OmniFake. Each task is evaluated with 10 support samples. We report Accuracy / Macro-F1 in percentage, with the best results highlighted in boldface.

Methods	OmniFake Dataset						Averages (%)	
	Part I		Part II		Part III			
	5-way	15-way	5-way	15-way	5-way	15-way	5-way	15-way
DNA-Det	49.06 / 47.58	28.41 / 26.69	48.63 / 46.60	26.65 / 23.84	50.73 / 49.31	29.70 / 28.24	49.47 / 47.83	28.25 / 26.26
CPL	46.54 / 44.52	26.82 / 24.34	48.42 / 46.63	28.09 / 26.19	55.73 / 53.68	38.31 / 34.89	50.23 / 48.28	31.07 / 28.47
SiameseNet	45.32 / 43.83	28.30 / 24.89	50.52 / 47.65	30.51 / 27.18	51.04 / 48.95	36.34 / 32.17	48.96 / 46.81	31.72 / 28.08
POSE	47.27 / 45.82	29.61 / 26.52	52.13 / 48.97	32.47 / 28.85	53.20 / 50.11	38.52 / 34.68	50.87 / 48.30	33.53 / 30.02
OOC-CLIP	54.62 / 53.01	36.14 / 33.59	58.68 / 57.40	37.07 / 35.23	63.13 / 62.45	44.30 / 42.96	58.81 / 57.62	39.17 / 37.26
UnivAttr	54.51 / 52.89	34.47 / 32.77	57.95 / 56.53	35.45 / 33.99	61.38 / 60.16	42.99 / 41.15	57.95 / 56.53	37.64 / 35.97
ComFor	58.82 / 57.86	37.82 / 36.78	60.35 / 59.28	37.98 / 36.53	60.40 / 59.32	40.45 / 38.90	59.86 / 58.82	38.75 / 37.40
FSD	67.32 / 66.07	42.47 / 40.88	75.22 / 74.53	53.55 / 52.40	77.40 / 76.96	**60.81** / **59.93**	73.31 / 72.52	52.28 / 51.07
OmniDFA	**69.04** / **67.93**	**44.25** / **42.09**	**78.00** / **77.19**	**56.11** / **54.39**	**78.98** / **78.24**	60.26 / 58.73	**75.34** / **74.45**	**53.54** / **51.74**

are then processed through an MLP to produce the final 128-dimensional embeddings. Following ComFor [40], we employ RandAugment [11], Gaussian blur, and JPEG compression during training to effectively mitigate potential biases while enhancing the robustness of our model.

During training, we sample fake images with a per-GPU batch size of 128 and real images with a per-GPU batch size of 16, resulting in a total batch size of 1152 across 8 GPUs to meet the requirements of contrastive learning. We use AdamW as our optimizer with a base learning rate of 2×10^{-5} and employ a cosine annealing learning rate scheduler for a total of 20 epochs. We set the temperature parameter $\tau = 0.07$ and loss coefficient $\lambda = 0.01$. The proposed method is implemented with the PyTorch library and all the experiments are conducted on 8 A100 with 40 GB memory. Additional implementation details can be found in the Supplementary Material.

5.3 Open-Set Few-Shot Attribution

The out-of-distribution classification task serves as a crucial testbed for evaluating whether a model genuinely learns essential features of unseen categories, rather than simply memorizing training patterns. Our study evaluates a selection of existing methods, including six attribution approaches and two synthetic image detection techniques, as summarized in Tab. 2. The attribution methods are DNA-Det [59], CPL [46], SiameseNet [1], POSE [60], OOC-CLIP [33], and UniversalAttr [9]. We also include AIGI detection methods FSD [56] and ComFor [40], motivated by their extensive training across a large number of categories, which suggests a strong potential for capturing diverse feature distributions. Although these models are not specifically designed for this task, we extract features from the last layer of their backbones and employ prototypical classification. Specifically, we compute the prototype centers from the support set and classify each query sample based on its distance to these centers. We evaluate each model on OmniFake using both 5-way 10-shot and 15-way 10-shot settings across 15 unseen categories from each test fold. To ensure reliability, we

Table 3: Results of authenticity detection on OmniFake. We evaluate zero-shot detection performance across different dataset partitions, reporting both real and fake accuracy (Acc) and average precision (AP) in percentage. The best results are highlighted in boldface.

Methods	OmniFake Dataset												Averages (%)	
	Part I				Part II				Part III					
	F-Acc	R-Acc	Acc	AP	F-Acc	R-Acc	Acc	AP	F-Acc	R-Acc	Acc	AP	Acc	AP
UnivFD	7.39	98.66	53.03	56.07	24.08	98.22	61.15	67.78	16.43	98.64	57.54	62.14	57.24	62.00
NPR	74.13	82.18	78.16	76.98	75.99	82.21	79.10	78.38	79.26	81.93	80.60	80.35	79.28	78.57
AIDE	77.51	96.28	86.90	94.01	83.20	96.32	89.76	94.62	78.53	96.20	87.37	93.68	88.01	94.10
SAFE	76.93	97.76	87.35	91.46	73.73	97.61	85.67	92.01	65.06	97.67	81.37	88.94	84.79	90.80
ComFor	75.58	**98.96**	87.27	89.78	80.78	**99.06**	89.92	92.50	85.06	**99.07**	92.07	97.10	89.75	93.13
FSD	90.66	77.15	83.91	-	83.33	77.80	80.57	-	90.51	75.63	83.07	-	82.51	-
OmniDFA	**97.43**	95.32	**96.38**	**97.56**	**96.36**	93.63	**95.00**	**95.93**	**97.06**	93.65	**95.36**	**97.42**	**95.58**	**96.97**

conduct $10,000$ independent test episodes for each experimental configuration following the episodic testing protocol.

As shown in Tab. 2, our model exhibits strong few-shot learning capabilities, confirming its effectiveness in identifying subtle inter-category features. These characteristics make our model a powerful starting point for future research in this newly proposed task. FSD also demonstrates competitive performance, underscoring the effectiveness of metric learning for this task. To our surprise, the standard binary classifier ComFor, though untrained for multi-class tasks, inherently exhibits some classification capability. This stems from its extensive exposure to a wide array of models during training, which yields relatively robust feature extraction capabilities. Other attribution methods perform inadequately in this open-set few-shot attribution task, primarily because they focus excessively on seen categories during training and fail to generalize to unseen ones. Our results significantly surpass previous methods, demonstrating the effectiveness of our framework and highlighting promising directions for future research.

5.4 Cross-Generator Authenticity Detection

Our authenticity detection evaluation prioritizes the zero-shot capability of classifiers, defined as the generalization performance over previously unseen categories. Our experiments use the OmniFake dataset and compare against state-of-the-art methods, including UnivFD [38], NPR [47], AIDE [58], SAFE [28], ComFor [40], and FSD [56]. While each compared method was optimized on its respective training set, our evaluation intentionally focuses on their generalization to unseen, novel data. This design stems from the fundamental objective of synthetic detection, where robustness and performance beyond the training distribution are of central importance. Accordingly, we use their officially released weights to ensure optimal generalization. For thorough benchmarking against FSD based on metric learning, we retrain this model on our dataset. We include ComFor in our comparisons to benchmark against large-scale pretrained models. Although it is trained on a series of models built upon Stable Diffusion, poten-

Table 4: Zero-shot detection results on GenImage and Chameleon datasets. The best results are highlighted in boldface, while the second-optimal results are marked with underline. Our method demonstrates remarkable improvements in detection accuracy.

Methods	GenImage									Chameleon			
	ADM	Glide	Midjourney	SD v1.4	SD v1.5	VQDM	Wukong	BigGAN	Average	F-Acc	R-Acc	Acc	F1
CNNSpot	60.39	58.07	51.39	50.57	50.53	56.46	51.03	71.17	56.20	9.86	99.55	60.89	17.85
UnivFD	66.87	62.46	56.13	63.66	63.49	85.31	70.93	95.08	70.49	**85.52**	41.56	60.42	64.97
DIRE	75.78	71.75	58.01	49.74	49.83	53.68	54.46	70.12	60.42	2.09	99.73	57.83	4.08
PatchCraft	82.17	83.79	90.12	95.38	95.30	88.91	91.07	95.80	90.32	1.39	96.52	55.70	2.62
NPR	69.69	78.36	77.85	78.63	78.89	78.13	76.61	84.35	77.81	1.68	**100.00**	57.81	3.30
AIDE	**93.43**	95.09	77.20	93.00	92.85	95.16	93.55	83.95	90.53	26.80	95.06	65.77	40.19
OmniDFA	85.50	**96.71**	**97.58**	**97.47**	**97.75**	**96.78**	**97.78**	**97.33**	**95.86**	77.46	88.00	**83.48**	**80.09**

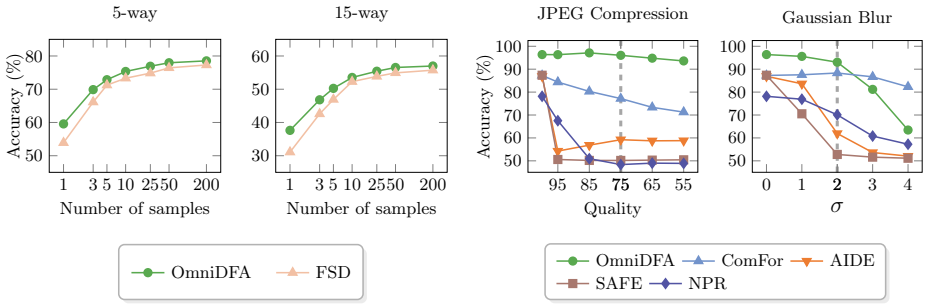
tially introducing dataset overlap, we consider this risk tolerable given the vast number of our testing categories.

The results in Tab. 3 show that our OmniDFA consistently outperforms prior methods in fake detection performance across all components of OmniFake, achieving an average improvement of 5.83%. This demonstrates the strong generalization capability of our model. We also notice that our model exhibits slightly lower accuracy in real image detection. This is attributed to our boundary update rule, which actively filters out anomalous data and may consequently exclude some marginal cases. We plan to investigate a more adaptive boundary update mechanism for better robustness in future work.

Furthermore, our results indicate that training with multiple classes yields better performance than single-generator training approaches, such as UnivFD, which shows limited effectiveness when applied to out-of-distribution data. Another observation is that FSD demonstrates good performance in fake detection but performs poorly on real images. This limitation stems from its classification mechanism that relies on the nearest prototypical centroid, which becomes less effective when dealing with diverse types of fake images. In contrast, our model explicitly addresses the distribution of real images by incorporating a center loss term, which pulls the features of real samples closer to their class center in the embedding space, thereby enhancing the discriminative power for authentic data and significantly boosting the overall detection accuracy.

5.5 Cross-Dataset Evaluation

Cross-dataset evaluation has been employed to test classifiers' generalization capability under open-set conditions [40]. To comprehensively assess the generalization capability of OmniDFA, we select two representative datasets: GenImage and Chameleon, which serve as benchmarks for standard experimental evaluation and real-world application scenarios, respectively. To rigorously ensure experimental reliability, we train an additional classifier by explicitly excluding all generator categories in both benchmarks and their related model families. For instance, we completely remove SD 1.5, SDXL, and SD3-Medium generated samples from our training data to ensure a strictly zero-shot testing condition.



(a) Influence of the number of shots **(b)** Robustness to compression and blurring

Fig. 4: Ablation studies on the number of support shots and robustness to real-world degradations. The vertical dotted line in (b) marks the augmentation threshold.

We incorporate three additional baseline methods: CNNSpot [52], DIRE [53], and PatchCraft [62]. All comparative models are developed following the experimental protocol of AIDE [58]. Since the Chameleon dataset is imbalanced, we additionally report the corresponding F1 score.

As shown in Tab. 4, OmniDFA achieves remarkably high performance in these challenging zero-shot scenarios. Our model achieves state-of-the-art classification accuracy on the GenImage benchmark, further validating the effectiveness of our multi-generator training strategy. The results reveal that the single-generator-trained generalization approach, which remains prevalent in synthetic image detection, is increasingly inadequate for addressing evolving challenges. More importantly, on the Chameleon benchmark, our model surpasses the second-best method by a significant margin of 17.71% in accuracy. It maintains balanced performance across both authentic and fake images, demonstrating superior generalization and robust practical applicability, whereas most existing models exhibit a strong systematic bias. Therefore, our work represents a significant step towards more generalizable detection models.

5.6 Ablation Studies

In this subsection, we conduct ablation studies on experimental settings and our model. We investigate the impact of sample size and data perturbations, and evaluate the effectiveness of each component.

Impact of the number of shots. To investigate the impact of the number of support samples, we evaluate two models, OmniDFA and FSD, varying the number of support samples per class from 1 to 200. For a comprehensive analysis, we report the mean accuracy across the entire OmniFake dataset. The summarized results are presented in Fig. 4a, where the horizontal axis is on a logarithmic scale. The results show that accuracy increases rapidly from 1-shot to 10-shot, after which further improvements become more gradual. This demonstrates the validity of our few-shot attribution task, in which only a small number of unseen examples are needed to achieve strong open-set identification performance.

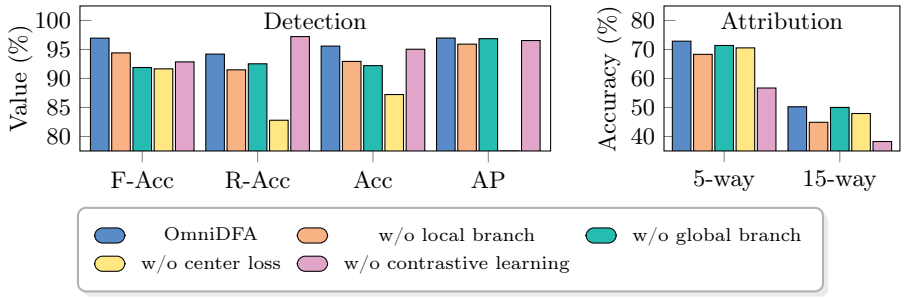


Fig. 5: Ablation study on model component. Our dual-branch architecture significantly outperforms its single-branch counterpart. Moreover, contrastive learning enables our model to capture features that are more discriminative for unseen samples.

Our results suggest that 10-shot represents a favorable trade-off between high performance and computational cost.

Robustness to real-world degradation. Image corruptions such as JPEG compression are commonly encountered during real-world image transmission and distribution. We apply JPEG compression and Gaussian blurring to examine how image quality affects model performance. Fig. 4b reports the accuracy on OmniFake part I under various perturbation strengths, and the vertical dotted line indicates the augmentation boundary adopted during our training. The results demonstrate that our model maintains robust performance within the augmentation range and exhibits strong resilience to JPEG compression even beyond the augmentation range used in training. In contrast, other methods suffer significant performance degradation from the outset. Even when trained with JPEG augmentation, ComFor remains susceptible to this influence. When it comes to Gaussian blurring, our model performs excellently within the trained augmentation range, yet its performance declines noticeably beyond it. ComFor demonstrates excellent performance in handling blur because it is trained with extensive and targeted blur augmentation. This observation underscores the importance of incorporating such perturbations during training, thus providing strong empirical support for future model development.

Effectiveness of our modules. To investigate the impact of each component, we perform a series of ablation studies. We analyze the effects of removing the local and global branches, as well as of removing the center loss and boundary threshold. Additionally, we train a binary classifier without employing contrastive learning. We train and evaluate all models under the same setup, yielding the average detection and attribution performance across all three parts of the OmniFake dataset, as shown in Fig. 5. The results demonstrate the clear advantage of our dual-path model over the single-branch variants, confirming the superiority of our multi-level feature fusion. Our OmniDFA outperforms a plain binary classifier by capturing finer inter-class distinctions, which consequently enhances fake image detection accuracy. The results demonstrate that our model

establishes a strong baseline for the novel paradigm we have introduced, thereby facilitating future research.

6 Conclusion

In this paper, we establish a new paradigm named few-shot synthetic image attribution, which aims to identify unknown image generators with only a limited number of samples. We present OmniFake, a comprehensively curated dataset that is both sufficiently large and meticulously categorized, enabling detailed investigation into model-specific artifacts. Building upon this foundation, we propose OmniDFA (Omni Detector and Few-shot Attributor), a unified framework based on few-shot learning that jointly performs authenticity detection and few-shot source attribution. Experiments show that OmniDFA not only achieves state-of-the-art generalization in synthetic image detection, but also exhibits strong open-set attribution capability with limited reference samples. The results underscore the real-world applicability of our approach, suggesting a promising path for future research.

Acknowledgements

This research is supported by Artificial Intelligence-National Science and Technology Major Project (2023ZD0121200), and the National Natural Science Foundation of China (62437001, 62436001, 62531026), and the Natural Science Foundation of Jiangsu Province under Grant BK20243051, and the Strategic Priority Research Program of Chinese Academy of Sciences under Grant XDB1350103.

References

1. Abady, L., Wang, J., Tondi, B., Barni, M.: A siamese-based verification system for open-set architecture attribution of synthetic images. Pattern Recognit. Lett. **180**, 75–81 (2024)
2. Asnani, V., Yin, X., Hassner, T., Liu, X.: Reverse engineering of generative models: Inferring model hyperparameters from generated images. IEEE Trans. Pattern Anal. Mach. Intell. **45**(12), 15477–15493 (2023)
3. Bi, X., Liu, B., Yang, F., Xiao, B., Li, W., Huang, G., Cosman, P.C.: Detecting generated images by real images only. arXiv preprint arXiv:2311.00962 (2023), `https://arxiv.org/abs/2311.00962`
4. Boychev, D., Cholakov, R.: Imaginet: A multi-content benchmark for synthetic image detection. arXiv preprint arXiv:2407.20020 (2025), `https://arxiv.org/abs/2407.20020`
5. Bui, T., Yu, N., Collomosse, J.P.: Repmix: Representation mixing for robust attribution of synthesized images. In: Computer Vision - ECCV 2022. Lecture Notes in Computer Science, vol. 13674, pp. 146–163. Springer (2022)
6. Changpinyo, S., Sharma, P., Ding, N., Soricut, R.: Conceptual 12m: Pushing web-scale image-text pre-training to recognize long-tail visual concepts. In: IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021. pp. 3558–3568. Computer Vision Foundation / IEEE (2021)

7. Chen, B., Zeng, J., Yang, J., Yang, R.: DRCT: diffusion reconstruction contrastive training towards universal detection of diffusion generated images. In: Forty-first International Conference on Machine Learning, ICML 2024. OpenReview.net (2024)
8. Chen, X., Wu, Z., Liu, X., Pan, Z., Liu, W., Xie, Z., Yu, X., Ruan, C.: Janus-pro: Unified multimodal understanding and generation with data and model scaling. arXiv preprint arXiv:2501.17811 (2025), https://arxiv.org/abs/2501.17811
9. Cioni, D., Tzelepis, C., Seidenari, L., Patras, I.: Are CLIP features all you need for universal synthetic image origin attribution? In: Computer Vision - ECCV 2024 Workshops - Milan, Italy, September 29-October 4, 2024, Proceedings, Part XXI. Lecture Notes in Computer Science, vol. 15643, pp. 363–382. Springer (2024)
10. Cocchi, F., Cornia, M., Baraldi, L., Nicolosi, A., Cucchiara, R.: Contrasting deepfakes diffusion via contrastive learning and global-local similarities. In: Computer Vision - ECCV 2024 - 18th European Conference. Lecture Notes in Computer Science, vol. 15121, pp. 199–216. Springer (2024)
11. Cubuk, E.D., Zoph, B., Shlens, J., Le, Q.: Randaugment: Practical automated data augmentation with a reduced search space. In: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020 (2020)
12. Deng, C., Zhu, D., Li, K., Gou, C., Li, F., Wang, Z., Zhong, S., Yu, W., Nie, X., Song, Z., Shi, G., Fan, H.: Emerging properties in unified multimodal pretraining. arXiv preprint arXiv:2505.14683 (2025), https://arxiv.org/abs/2505.14683
13. Epstein, D.C., Jain, I., Wang, O., Zhang, R.: Online detection of ai-generated images. In: IEEE/CVF International Conference on Computer Vision, ICCV 2023 - Workshops. pp. 382–392. IEEE (2023)
14. Esser, P., Kulal, S., Blattmann, A., Entezari, R., Müller, J., Saini, H., Levi, Y., Lorenz, D., Sauer, A., Boesel, F., Podell, D., Dockhorn, T., English, Z., Lacey, K., Goodwin, A., Marek, Y., Rombach, R.: Scaling rectified flow transformers for high-resolution image synthesis. arXiv preprint arXiv:2403.03206 (2024), https://arxiv.org/abs/2403.03206
15. Fang, S., Nguyen, T.D., Stamm, M.C.: Open set synthetic image source attribution. In: 34th British Machine Vision Conference 2023, BMVC 2023, Aberdeen, UK, November 20-24, 2023. p. 659. BMVA Press (2023)
16. Frank, J., Eisenhofer, T., Schönherr, L., Fischer, A., Kolossa, D., Holz, T.: Leveraging frequency analysis for deep fake image recognition. In: Proceedings of the 37th International Conference on Machine Learning, ICML 2020. Proceedings of Machine Learning Research, vol. 119, pp. 3247–3258. PMLR (2020)
17. Gao, Y., Chang, D., Yu, B., Qin, H., Chen, L., Liang, K., Ma, Z.: Fakereasoning: Towards generalizable forgery detection and reasoning. arXiv preprint arXiv:2503.21210 (2025), https://arxiv.org/abs/2503.21210
18. Goodfellow, I.J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A.C., Bengio, Y.: Generative adversarial nets. In: Annual Conference on Neural Information Processing Systems 2014. pp. 2672–2680 (2014)
19. Grommelt, P., Weiss, L., Pfreundt, F., Keuper, J.: Fake or jpeg? revealing common biases in generated image detection datasets. In: Computer Vision - ECCV 2024 Workshops. Lecture Notes in Computer Science, vol. 15644, pp. 80–95. Springer (2024)
20. Gu, J., Meng, X., Lu, G., Hou, L., Minzhe, N., Liang, X., Yao, L., Huang, R., Zhang, W., Jiang, X., Xu, C., Xu, H.: Wukong: A 100 million large-scale chinese cross-modal pre-training benchmark. In: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022 (2022)

21. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. In: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020 (2020)
22. Hong, Y., Feng, J., Chen, H., Lan, J., Zhu, H., Wang, W., Zhang, J.: Wildfake: A large-scale and hierarchical dataset for ai-generated images detection. In: AAAI-25, Sponsored by the Association for the Advancement of Artificial Intelligence. pp. 3500–3508. AAAI Press (2025)
23. Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W.: Lora: Low-rank adaptation of large language models. In: The Tenth International Conference on Learning Representations, ICLR 2022. OpenReview.net (2022)
24. HuggingFace: Hugging face. https://huggingface.co/ (2016)
25. IdeogramAI: Ideogram. https://ideogram.ai (2023)
26. Karras, T., Aila, T., Laine, S., Lehtinen, J.: Progressive growing of gans for improved quality, stability, and variation. In: 6th International Conference on Learning Representations, ICLR 2018. OpenReview.net (2018)
27. Khosla, P., Teterwak, P., Wang, C., Sarna, A., Tian, Y., Isola, P., Maschinot, A., Liu, C., Krishnan, D.: Supervised contrastive learning. In: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020 (2020)
28. Li, O., Cai, J., Hao, Y., Jiang, X., Hu, Y., Feng, F.: Improving synthetic image detection towards generalization: An image transformation perspective. In: Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining, V.1, KDD 2025. pp. 2405–2414. ACM (2025)
29. Li, Z., Zhang, J., Lin, Q., Xiong, J., Long, Y., Deng, X., Zhang, Y., Liu, X., Huang, M., Xiao, Z., Chen, D., He, J., Li, J., Li, W., Zhang, C., Quan, R., Lu, J., Huang, J., Yuan, X., Zheng, X., Li, Y., Zhang, J., Zhang, C., Chen, M., Liu, J., Fang, Z., Wang, W., Xue, J., Tao, Y., Zhu, J., Liu, K., Lin, S., Sun, Y., Li, Y., Wang, D., Chen, M., Hu, Z., Xiao, X., Chen, Y., Liu, Y., Liu, W., Wang, D., Yang, Y., Jiang, J., Lu, Q.: Hunyuan-dit: A powerful multi-resolution diffusion transformer with fine-grained chinese understanding. arXiv preprint arXiv:2405.08748 (2024), https://arxiv.org/abs/2405.08748
30. Liang, Z., Liu, W., Wang, R., Wu, M., Li, B., Zhang, Y., Wang, L., Yang, X.: Transfer learning of real image features with soft contrastive loss for fake image detection. In: AAAI-25, Sponsored by the Association for the Advancement of Artificial Intelligence. pp. 26281–26289. AAAI Press (2025)
31. Lin, B., Li, Z., Cheng, X., Niu, Y., Ye, Y., He, X., Yuan, S., Yu, W., Wang, S., Ge, Y., Pang, Y., Yuan, L.: Uniworld-v1: High-resolution semantic encoders for unified visual understanding and generation. arXiv preprint arXiv:2506.03147 (2025), https://arxiv.org/abs/2506.03147
32. Lin, T., Maire, M., Belongie, S.J., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft COCO: common objects in context. In: Computer Vision - ECCV 2014 - 13th European Conference. Lecture Notes in Computer Science, vol. 8693, pp. 740–755. Springer (2014)
33. Liu, F., Luo, H., Li, Y., Torr, P., Gu, J.: Which model generated this image? A model-agnostic approach for origin attribution. In: Computer Vision - ECCV 2024 - 18th European Conference. Lecture Notes in Computer Science, vol. 15120, pp. 282–301. Springer (2024)
34. Liu, Z., Mao, H., Wu, C., Feichtenhofer, C., Darrell, T., Xie, S.: A convnet for the 2020s. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022. pp. 11966–11976. IEEE (2022)
35. Lu, Z., Huang, D., Bai, L., Qu, J., Wu, C., Liu, X., Ouyang, W.: Seeing is not always believing: Benchmarking human and model perception of ai-generated images. arXiv preprint arXiv:2304.13023 (2023), https://arxiv.org/abs/2304.13023

36. Luo, Y., Du, J., Yan, K., Ding, S.: Lare^2: Latent reconstruction error based method for diffusion-generated image detection. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024. pp. 17006–17015. IEEE (2024)
37. Midjourney, I.: Midjourney. https://www.midjourney.com/home (2021)
38. Ojha, U., Li, Y., Lee, Y.J.: Towards universal fake image detectors that generalize across generative models. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023. pp. 24480–24489. IEEE (2023)
39. OpenAI: Dall-e 3. https://openai.com/index/dall-e-3 (2023)
40. Park, J., Owens, A.: Community forensics: Using thousands of generators to train fake image detectors. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2025. pp. 8245–8257. Computer Vision Foundation / IEEE (2025)
41. Rahman, M.A., Paul, B., Sarker, N.H., Hakim, Z.I.A., Fattah, S.A.: Artifact: A large-scale dataset with artificial and factual images for generalizable and robust synthetic image detection. In: IEEE International Conference on Image Processing, ICIP 2023. pp. 2200–2204. IEEE (2023)
42. Ruiz, N., Li, Y., Jampani, V., Pritch, Y., Rubinstein, M., Aberman, K.: Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023. pp. 22500–22510. IEEE (2023)
43. Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M.S., Berg, A.C., Fei-Fei, L.: Imagenet large scale visual recognition challenge. Int. J. Comput. Vis. **115**(3), 211–252 (2015)
44. Schuhmann, C., Vencu, R., Beaumont, R., Kaczmarczyk, R., Mullis, C., Katta, A., Coombes, T., Jitsev, J., Komatsuzaki, A.: Laion-400m: Open dataset of clip-filtered 400 million image-text pairs. arXiv preprint arXiv:2111.02114 (2021), https://arxiv.org/abs/2111.02114
45. Sha, Z., Li, Z., Yu, N., Zhang, Y.: 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, CCS 2023. pp. 3418–3432. ACM (2023)
46. Sun, Z., Chen, S., Yao, T., Yin, B., Yi, R., Ding, S., Ma, L.: Contrastive pseudo learning for open-world deepfake attribution. In: IEEE/CVF International Conference on Computer Vision, ICCV 2023. pp. 20825–20835. IEEE (2023)
47. Tan, C., Liu, H., Zhao, Y., Wei, S., Gu, G., Liu, P., Wei, Y.: Rethinking the upsampling operations in cnn-based generative network for generalizable deepfake detection. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024. pp. 28130–28139. IEEE (2024)
48. Tan, C., Tao, R., Liu, H., Gu, G., Wu, B., Zhao, Y., Wei, Y.: C2P-CLIP: injecting category common prompt in CLIP to enhance generalization in deepfake detection. In: AAAI-25, Sponsored by the Association for the Advancement of Artificial Intelligence. pp. 7184–7192. AAAI Press (2025)
49. Tan, C., Zhao, Y., Wei, S., Gu, G., Liu, P., Wei, Y.: Frequency-aware deepfake detection: Improving generalizability through frequency space domain learning. In: Thirty-Eighth AAAI Conference on Artificial Intelligence, AAAI 2024. pp. 5052–5060 (2024)
50. Wang, G., Zhao, S., Zhang, X., Cao, L., Zhan, P., Duan, L., Lu, S., Fu, M., Chen, X., Zhao, J., Li, Y., Chen, Q.: Ovis-u1 technical report. arXiv preprint arXiv:2506.23044 (2025), https://arxiv.org/abs/2506.23044

51. Wang, J., Tondi, B., Barni, M.: BOSC: A backdoor-based framework for open set synthetic image attribution. IEEE Trans. Inf. Forensics Secur. **20**, 8043–8058 (2025)
52. Wang, S., Wang, O., Zhang, R., Owens, A., Efros, A.A.: Cnn-generated images are surprisingly easy to spot... for now. In: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020. pp. 8692–8701. Computer Vision Foundation / IEEE (2020)
53. Wang, Z., Bao, J., Zhou, W., Wang, W., Hu, H., Chen, H., Li, H.: DIRE for diffusion-generated image detection. In: IEEE/CVF International Conference on Computer Vision, ICCV 2023. pp. 22388–22398. IEEE (2023)
54. Wen, Y., Zhang, K., Li, Z., Qiao, Y.: A discriminative feature learning approach for deep face recognition. In: European Conference on Computer Vision. pp. 499–515. Springer (2016)
55. Wu, C., Zheng, P., Yan, R., Xiao, S., Luo, X., Wang, Y., Li, W., Jiang, X., Liu, Y., Zhou, J., Liu, Z., Xia, Z., Li, C., Deng, H., Wang, J., Luo, K., Zhang, B., Lian, D., Wang, X., Wang, Z., Huang, T., Liu, Z.: Omnigen2: Exploration to advanced multimodal generation. arXiv preprint arXiv:2506.18871 (2025), https://arxiv.org/abs/2506.18871
56. Wu, S., Liu, J., Li, J., Wang, Y.: Few-shot learner generalizes across ai-generated image detection. arXiv preprint arXiv:2501.08763 (2025), https://arxiv.org/abs/2501.08763
57. Xie, J., Mao, W., Bai, Z., Zhang, D.J., Wang, W., Lin, K.Q., Gu, Y., Chen, Z., Yang, Z., Shou, M.Z.: Show-o: One single transformer to unify multimodal understanding and generation. In: The Thirteenth International Conference on Learning Representations, ICLR 2025. OpenReview.net (2025)
58. Yan, S., Li, O., Cai, J., Hao, Y., Jiang, X., Hu, Y., Xie, W.: A sanity check for ai-generated image detection. In: The Thirteenth International Conference on Learning Representations, ICLR 2025. OpenReview.net (2025)
59. Yang, T., Huang, Z., Cao, J., Li, L., Li, X.: Deepfake network architecture attribution. In: Thirty-Sixth AAAI Conference on Artificial Intelligence, AAAI 2022. pp. 4662–4670. AAAI Press (2022)
60. Yang, T., Wang, D., Tang, F., Zhao, X., Cao, J., Tang, S.: Progressive open space expansion for open-set model attribution. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023. pp. 15856–15865. IEEE (2023)
61. Yu, F., Seff, A., Zhang, Y., Song, S., Funkhouser, T., Xiao, J.: Lsun: Construction of a large-scale image dataset using deep learning with humans in the loop. arXiv preprint arXiv:1506.03365 (2016), https://arxiv.org/abs/1506.03365
62. Zhong, N., Xu, Y., Li, S., Qian, Z., Zhang, X.: Patchcraft: Exploring texture patch for efficient ai-generated image detection. arXiv preprint arXiv:2311.12397 (2024), https://arxiv.org/abs/2311.12397
63. Zhu, M., Chen, H., Yan, Q., Huang, X., Lin, G., Li, W., Tu, Z., Hu, H., Hu, J., Wang, Y.: Genimage: A million-scale benchmark for detecting ai-generated image. In: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023 (2023)