WildFake: A Large-Scale and Hierarchical Dataset for AI-Generated Images Detection

Yan Hong[1], Jianming Feng[1], Haoxing Chen[1], Jun Lan[1], Huijia Zhu[1], Weiqiang Wang[1], Jianfu Zhang[2*]
[1]Ant Group,
[2]Qing Yuan Research Institute, Shanghai Jiao Tong University
yanhong.sjtu@gmail.com, hx.chen@hotmail.com
{fengji.fjm, yelan.lj, huijia.zhj, weiqiang.wwq}@antgroup.com, c.sis@sjtu.edu.cn

Abstract

The development of text-to-image generative models has enabled the creation of images so realistic that distinguishing between AI-generated images and real photos is becoming a challenge. This progress offers new possibilities but also raises concerns over privacy, authenticity, and security. Detecting AI-generated images is crucial to prevent misuse. To assess the generalizability and robustness of AI-generated image detection, we present a large-scale dataset, referred to as WildFake. This dataset features cutting-edge image generators, a wide variety of generator categories, and generators for various applications, organized in a hierarchical framework. WildFake collects fake images from the open-source community, enriching its diversity with a broad range of image classes and image styles. Its design significantly improves the effectiveness of detection algorithms, making it a valuable resource for enhancing AI-generated image detection in practical applications. Our evaluations offer insights into the performance of generative models at various levels, showcasing WildFake's unique hierarchical structure's benefits.

1 Introduction

The development of generative models has markedly improved the creation of realistic images, simplifying the process of producing AI-generated images (i.e., fake images). This boosted accessibility has amplified concerns regarding the widespread spread of false information. Characterized by their impressive visual clarity, AI-generated images are particularly persuasive and have the potential to significantly influence public opinion in critical domains such as politics and economics. To counteract such harmful activities, the development of technologies capable of detecting altered images is essential.

Generative models typically introduce unique patterns, which don't appear in real images and vary depending on the model and its corresponding training data (Marra et al. 2019). Recent research in synthetic image detection has focused on identifying these irregular patterns through methods like color pattern analysis, light intensity evaluation, and Fourier spectrum analysis (Corvi et al. 2023; Frank et al. 2020). While traditional techniques based on manually selected features and frequency analysis show limited efficacy, deep CNN models are more effective in pattern detection (Marra et al. 2018). However, with the range of models available—from Generative Adversarial Networks (GANs) (Alanov et al. 2023; Sauer, Schwarz, and Geiger 2022; Pehlivan, Dalva, and Dundar 2023; Karras, Laine, and Aila 2019; Esser, Rombach, and Ommer 2021; Choi et al. 2020; Brock, Donahue, and Simonyan 2018; Tao et al. 2023, 2022; Kang et al. 2023) to Diffusion Models (DMs) (Saharia et al. 2022; Rombach et al. 2022), users can now easily produce high-quality and diverse images with different types of model with personalized weights. These images, often distributed across various social media platforms, present a significant challenge in terms of generalization and robustness for detection technologies. Existing detectors still face difficulties with generative models not encountered during their training phase (Aghasanli, Kangin, and Angelov 2023; Wu, Zhou, and Zhang 2023; Lorenz, Durall, and Keuper 2023; Wang et al. 2023). To aid in the development of detectors, many datasets for general AI-generated images are built (Wang et al. 2020; Verdoliva, Cozzolino, and Nagano 2022; Sha et al. 2022; Bird and Lotfi 2023; Wang et al. 2022; Rahman et al. 2023; Zhu et al. 2023; Wang et al. 2023; Lu et al. 2024). However, these existing datasets often exhibit significant limitations. They are generally restricted to one or two types of generators, constrained to producing images within fixed categories, or largely dependent on low-quality, user-generated images. These constraints hinder the effectiveness and adaptability of detectors in recognizing a broader range of AI-generated images.

In this paper, we present WildFake, a comprehensive, large-scale dataset specifically designed for the detection of AI-generated images. We summarize the comparison among fake image detection datasets in Table 1. WildFake stands out by generating a diverse array of rich, stylistically varied, and high-quality images. Within the WildFake dataset, to augment the dataset's diversity, fake images are produced either through our extensive generation pipeline or sourced from open-source communities, where users share images created with their personalized generative models. Real images are gathered from open datasets used in various tasks like image captioning, generation, and classification, ensuring a broad spectrum of styles and content. We have conducted a series of experiments on the WildFake dataset to assess the generalization capabilities of detectors trained on fake images,

*Corresponding author.
Copyright © 2025, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

Datasets	Generators			Communities	Available	Hierarchies	Image Numbers	
	GANs	DMs	Others				Fake	Real
CNNSpot (Wang et al. 2020)	✓	✗	✗	✗	✓	✗	362,000	362,000
IEEE VIP Cup (Verdoliva, Cozzolino, and Nagano 2022)	✓	✓	✗	✗	✗	✗	7,000	7,000
DE-FAKE (Sha et al. 2022)	✗	✓	✗	✗	✗	✗	20,000	60,000
CiFAKE (Bird and Lotfi 2023)	✗	✓	✗	✗	✓	✗	60,000	60,000
GenImage (Zhu et al. 2023)	✓	✓	✗	✗	✓	✗	1,331,167	1,350,000
DiffusionDB (Wang et al. 2022)	✗	✓	✗	✓	✓	✗	14,000,000	0
ArtiFact (Rahman et al. 2023)	✓	✓	✗	✗	✓	✗	1,521,900	962,200
MPBench (Lu et al. 2024)	✓	✓	✓	✗	✓	✗	2,300,000	2,250,000
DiffusionForensics (Wang et al. 2023)	✗	✓	✗	✗	✓	✗	439,020	92,000
WildFake	✓	✓	✓	✓	✓	✓	2,557,278	1,013,446

Table 1: Comparison among WildFake and existing fake image detection datasets.

demonstrating WildFake's potential to enhance the understanding of fake image detection in a multitude of real-world scenarios. Additionally, we have implemented a series of degradation tests on the WildFake testing set, illustrating the robustness of these detectors in challenging conditions. Besides, Distinct from existing datasets, WildFake categorizes generative models into three primary groups: GANs(Karras, Laine, and Aila 2019; Karras et al. 2020, 2021; Choi et al. 2018, 2020; Brock, Donahue, and Simonyan 2018; Kang et al. 2023; Tao et al. 2023), DMs (Gu et al. 2022b; Holub 2022; Ho, Jain, and Abbeel 2020; Song, Meng, and Ermon 2020; Nichol et al. 2022; Ramesh et al. 2022; Rombach et al. 2022; OpenAI 2023), and Others (Van Den Oord, Vinyals et al. 2017; Esser, Rombach, and Ommer 2021; Feichtenhofer et al. 2022; Chang et al. 2023). Based on these methods, WildFake uniquely features four levels of categorization, each based on different dimensions, as depicted in Figure 1. WildFake comprehensively incorporates multiple generator types, various architectures, different model weights, and versions of the same model series. Such a structure is conducive to a detailed analysis of various image generators, offering insights into their characteristics. WildFake dataset are available at https://github.com/hy-zpg/AIGC-Image-Detection-Dataset

2 Related Works

We offer a brief yet thorough exploration of the image generation methods, existing AI-generated image datasets, and existing AI-generated image detection approaches in Appendix Section A.

3 WildFake Dataset Construction

In this section, we present the proposed dataset, WildFake. Section 3.1 provides the objectives and an overview of the dataset. Section 3.2 details the hierarchical structure of WildFake. Finally, Section 3.3 describes the methodologies employed for data collection.

3.1 Dataset Overviews

Addressing the critical need for assessing the **generalizability** of both datasets and detectors (i.e., the ability of training detectors to accurately identify unseen images from the open world or different datasets) and **robustness** of detectors (i.e., maintaining high performance despite various corruptions to fake images) of fake image detectors, we have developed the "WildFake" dataset, which has two main characteristics:

- Diverse Content with Wild Collection: WildFake includes a wide array of high-quality fake images sourced from open-source websites, along with images produced using both user-trained and officially provided pre-trained generative models. This diverse collection ensures a comprehensive set of fake images, significantly enriching the understanding of fake image detection across numerous real-world contexts, and enhancing the generalizability and robustness of detectors.
- Hierarchical Structure: The dataset is organized hierarchically, encompassing cross-generators, cross-architectures within the same type of generator, cross-weights within identical architectures, and cross-time analysis either within the same generator type or across different versions of the same model series. This structure facilitates in-depth analysis of various image generators.

3.2 Hierarchical Organization of WildFake

As is shown in Figure 1, the proposed WildFake dataset consists of five levels consisting of *cross-generator*, *cross-architecture*, *cross-weight*, *cross-time*, and *cross-version*.

- Cross-Generator: This level encompasses DMs, GANs, and Other generators, providing a comprehensive overview of the diverse generative models in use.
- Cross-Time: Focusing on GANs and Other generators known for high-quality synthesis, we categorize them into "Early" and "Latest" groups. "Early" represents well-established, popular models, whereas "Latest" includes recent advancements. Early GANs (*resp.,* latest GANs) consists of BigGAN (Brock, Donahue, and Simonyan 2018), StyleGANs (Karras, Laine, and Aila 2019; Karras et al. 2020, 2021), and StarGANs (Choi et al. 2018, 2020), (*resp.,* GigaGAN (Kang et al. 2023), DF-GAN (Tao et al. 2022), and GALIP (Tao et al. 2023)). Similarly, for Others generators, "Early" includes VQVAE (Van Den Oord, Vinyals et al. 2017) and VQGAN (Esser, Rombach, and Ommer 2021), while "Latest" encompasses Muse (Chang et al. 2023) and MAE (Feichtenhofer et al. 2022).
- Cross-Architecture: Considering the rapid development of DMs generators, nine kinds of DMs generators comprise cross-architecture level, consisting DALLE (Ramesh et al. 2022), ADM (Dhariwal and Nichol 2021), Imagen (Saharia et al. 2022), DDPM (Ho, Jain, and Abbeel 2020), DDIM (Song, Meng, and Ermon 2020), VQDM (Gu et al. 2022b), Midjouney (Holub 2022), and SD (Rombach et al.

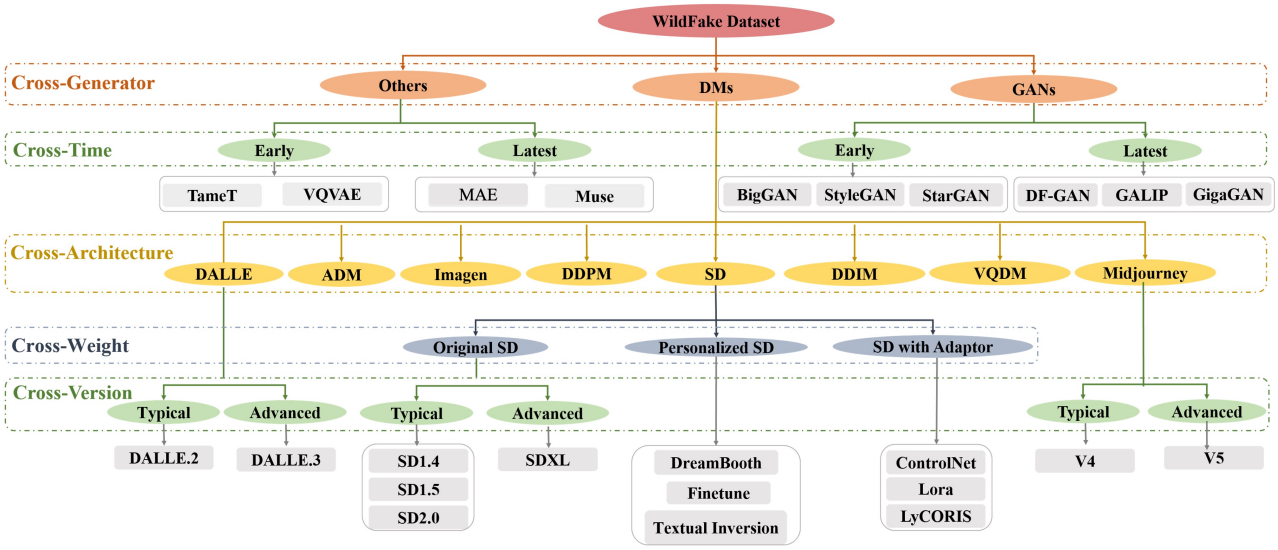

Figure 1: Overview of WildFake. At the cross-generator level, the dataset categorizes generators into DMs, GANs, and Others. The cross-architecture level distinguishes between various models/architectures within DMs, such as DALLE (Ramesh et al. 2022), Imagen (Saharia et al. 2022), Midjourney (Holub 2022), SD (Rombach et al. 2022), etc.. Furthermore, fake images from SD are divided into three subsets at the cross-weight level. A cross-version level also segments different generators into typical/advanced categories.

2022). Please note that DDPM and DDIM are commonly known as generation sampling methods. Here we use the names representing their corresponding released models in (Ho, Jain, and Abbeel 2020; Song, Meng, and Ermon 2020)

- Cross-Weight: Open-source SD (Rombach et al. 2022) has been widely spread in academia and industry, officially released pre-trained models armed with updated architecture trained on large datasets, and users also adopt different finetuning strategies such as finetuning several modules of SD (Rombach et al. 2022) or finetuning with DreamBooth (Ruiz et al. 2023) to obtain personalized models. Besides, many works focus on training different adaptors to combine with the base SD model to achieve controllable generation. ControlNet (Zhang, Rao, and Agrawala 2023) relies on paired image-prior data to control different prior information of generated images like edge, segmentation mask, style, and etc.. Lora-based methods, including Lora (Hu et al. 2021) and LyCORIS (Yeh et al. 2023) are also proposed to train extra low-rank layers to incorporate new content into the base model. Also, there are some methods (Gal et al. 2022; Voynov et al. 2023) to learn new tokens on the user-provided data for customized image generation. Thus, we classify SD-based generators into Original SD, Personalized SD, and SD with adaptors for cross-weight evaluation.
- Cross-Version: DALLE (Ramesh et al. 2022), Midjouney (Holub 2022), Imagen (Saharia et al. 2022), and SD (Rombach et al. 2022) have been widely known in academia and industry, due to the superiority of the quality of generated images. Fake images generated by DALLE (Ramesh et al. 2022) (*resp.,* Midjouney (Holub 2022)) are divided into "Typical" and "Advanced" subsets along the cross-version level.

3.3 Diverse Image Collection

Utilizing the hierarchical structure and the corresponding generators, we synthesize fake images and gather real images to construct the WildFake dataset. It's crucial to underline that the primary aim of an AI-generated image detection dataset is to achieve ***robust and generalizable detection capabilities***, rather than focusing solely on the ***image quality for quality assessment purposes***. A diverse and rich collection guarantees that the dataset encompasses a wide range of image categories, enabling detailed evaluations of AI-generated image detection algorithms and their effectiveness across various contexts.

Fake Image Collection: The guiding principle for collecting fake images is to ensure maximal diversity. We give priority to generating additional fake images using the most recent generators due to their superior quality. To collect diverse fake images from different resources, we have established a generation pipeline. This pipeline facilitates the production of images using popular generative mechanisms, including GANs, DMs, and Others generators. We strive to generate diverse content, covering people, landscapes, objects, and scenes as much as possible. Besides, we sourced user-created images from open-source platforms such as Civitai (hello@civitai.com 2022) and Midjourney (Holub 2022). On these platforms, users generate new images using either original open-source models or personalized models fine-tuned with their data. Unlike datasets primarily composed of author-generated images, such as DiffusionForensics (Wang et al. 2023), ArtiFact (Rahman et al. 2023), and GenIm-

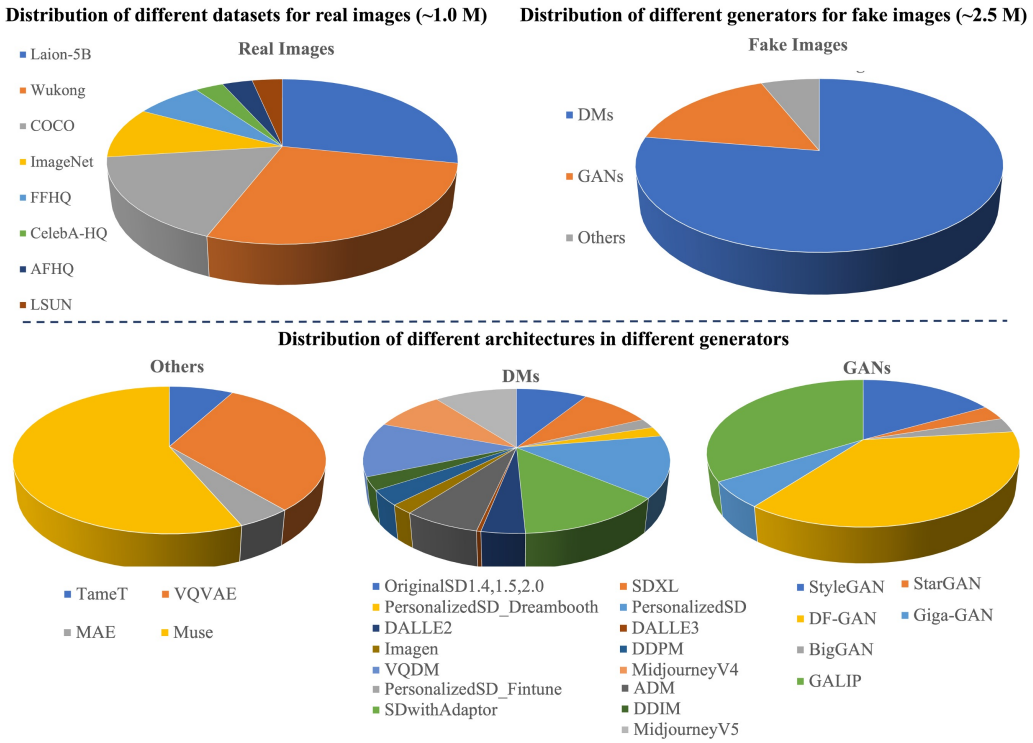

Figure 2: Overview of data distribution of WildFake. The top subfigure illustrates the distribution of real images from open-source datasets and fake images sourced from various generators, while the bottom subfigure depicts the distribution of fake images originating from different architectures within these generators.

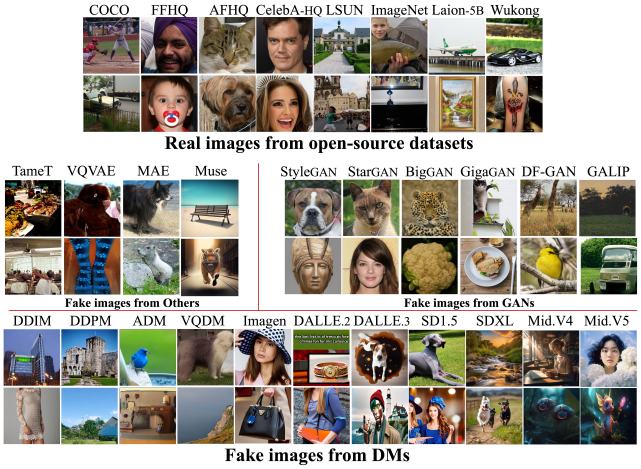

Figure 3: Samples of real and fake images from WildFake.

age (Zhu et al. 2023), our approach of collecting from open sources offers a more representative sample of the average quality of generated images. This ensures that the evaluation of detection models on our dataset is reflective of real-world applicability. For gathering images from GANs and Others, we primarily utilize official GitHub repositories and model cards from HuggingFace. When these GitHub repositories include generated samples, we directly extract fake images from there. In cases where the methods are associated with text-to-image generation, new images are produced by randomly sampling captions from their respective testing datasets. For other scenarios, we randomly generate images using the pretrained models available.

Real Image Collection: The rule of collecting real images is to similar to the training set of the generators. Considering the fact that fake images from GANs and Others are limited to specific domains determined by training datasets such as COCO (Lin et al. 2014), FFHQ (Karras, Laine, and Aila 2019), ImageNet (Deng et al. 2009), LSUN Church (Yu et al. 2015), CelebA-HQ (Karras et al. 2017), AFHQ (Choi et al. 2020) dataset, we sample parts of real images from those datasets. Besides, recent text-to-image generators mostly trained on Laion-5B (Schuhmann et al. 2022) or Chinese cross-modal Wukong (Gu et al. 2022a) datasets, we also include real image samples from these text-to-image datasets, which are commonly utilized for training DMs. This ensures a comprehensive collection of real images, facilitating a more robust and realistic evaluation of real-fake image detection.

Statistics and Settings: Examples of images from WildFake's various categories are displayed in Figure 3. To analyze WildFake, we illustrate the distribution of both real and fake images from various sources in Figure 2 for fake images and Figure 2 for real images. The WildFake dataset contains a total of 3,570,724 images, comprising 1,013,446 real images and 2,557,278 fake images. We split real images (*resp.,* fake images) into the training set and testing set as the ratio of 4 : 1. In detail, for all generators in Figure 2,

Detectors and Training Datasets	Testing Dataset					Avg
	DiffusionForensics	GenImage	DiffusionDB	ArtiFact	WildFake	
ResNet50-GenImage	84.2/95.6/77.9	**99.7/99.9/99.9**	93.0/96.3/94.3	61.3/66.1/52.8	71.6/92.1/79.0	*81.9/90.0/80.7*
ResNet50-DiffusionDB	84.2/84.1/49.4	50.1/49.3/48.5	**99.9/100/100**	61.3/61.4/50.0	81.1/82.7/72.8	75.3/75.5/64.1
ResNet50-ArtiFact	*85.4/94.9/76.7*	76.5/84.8/82.9	64.1/69.9/68.1	**97.2/99.5/99.3**	85.4/94.9/76.7	81.7/88.8/80.74
ResNet50-WildFake	**87.2/96.6/83.4**	*80.9/89.9/89.3*	*96.3/99.2/ 99.2*	*68.0/84.7/75.3*	**99.6/99.9/99.9**	**86.4/94.1/89.42**
ViT-GenImage	*84.2/97.2/85.3*	**99.6/99.9/99.9**	97.2/99.0/98.6	61.3/61.1/49.8	76.8/93.8/83.3	*83.8/90.2/83.4*
ViT-DiffusionDB	84.2/83.6/47.8	50.0/46.4/42.8	**99.9/100/100**	61.2/61.5/50.2	80.4/81.9/71.4	75.2/74.6/62.4
ViT-ArtiFact	84.2/96.1/82.5	78.5/88.1/85.0	68.4/75.3/73.2	**96.8/99.6/99.5**	*84.2/96.1/82.5*	82.4/91.0/84.4
ViT-WildFake	**95.8/99.1/97.2**	*88.6/83.6/89.7*	*99.3/99.8/99.9*	*62.2/81.9/68.8*	**99.1/99.9/99.9**	**89.0/92.84/91.1**
MultiLID-GenImage	58.8/59.9/51.5	75.8/73.9/74.9	59.1/60.4/52.0	49.9/45.3/42.9	50.5/61.5/54.3	*58.8/60.2/55.1*
MultiLID-DiffusionDB	50.0/50.8/45.6	47.0/46.5/44.9	77.2/72.4/73.4	49.9/51.9/49.0	51.0/49.4/48.9	55.0/54.2/52.3
MultiLID-ArtiFact	55.7/61.4/54.7	53.6/59.6/58.4	49.9/50.9/49.7	71.0/72.7/73.5	50.8/59.7/51.9	56.2/60.9/57.4
MultiLID-WildFake	56.7/58.8/54.4	52.1/58.2/58.1	62.1/64.5/64.6	51.2/55.1/56.0	74.3/75.9/75.9	**59.2/62.5/61.8**
DIRE-GenImage	*82.9/93.9/76.7*	**99.5/99.9/99.9**	91.3/94.5/92.5	59.1/61.8/52.7	*72.1/92.3/80.4*	*80.9/88.4/80.4*
DIRE-DiffusionDB	82.7/82.7/53.4	50.0/49.3/47.1	**99.9/99.9/99.9**	60.3/60.3/50.1	79.1/81.0/72.0	74.4/74.6/64.5
DIRE-ArtiFact	79.5/88.3/74.6	75.4/82.0/82.3	59.9/64.8/62.5	**92.4/92.6/92.4**	69.8/83.4/70.2	75.4/82.2/76.4
DIRE-WildFake	**85.5/97.5/84.9**	*77.3/85.1/84.5*	*97.2/99.3/99.3*	*67.6/84.0/74.8*	**89.3/89.6/89.7**	**83.4/91.1/86.6**
IFDL-GenImage	86.4/95.9/81.4	**99.6/99.9/99.9**	93.7/96.8/95.0	61.4/67.4/55.1	74.2/93.7/83.9	*83.0/90.7/83.0*
IFDL-DiffusionDB	88.3/88.1/53.9	52.9/51.8/51.0	**99.9/100/100**	63.4/62.4/53.5	*85.5/85.9/76.5*	78.0/77.6/66.9
IFDL-ArtiFact	*87.9/95.8/80.0*	78.3/87.1/84.9	66.2/72.1/69.1	**97.6/99.6/99.6**	71.9/90.0/75.5	80.3/88.9/81.8
IFDL-WildFake	**88.6/97.9/92.8**	*85.1/95.0/89.9*	*97.7/99.5/99.7*	*67.7/82.1/72.4*	**99.3/99.9/99.9**	**87.6/93.9/99.0**
LASTED-GenImage	87.6/96.2/83.7	**99.8/99.9/99.9**	94.3/97.9/95.7	63.4/66.7/57.0	*78.8/94.3/86.9*	*84.7/91.0/84.6*
LASTED-DiffusionDB	90.7/90.5/58.4	53.1/53.2/52.3	**99.9/100/100**	65.4/66.5/55.9	88.2/90.0/79.1	79.4/80.0/69.1
LASTED-ArtiFact	*91.3/98.6/84.1*	80.2/89.1/89.2	69.5/75.7/72.6	**98.2/99.7/99.6**	73.0/92.0/77.8	82.4/91.1/84.7
LASTED-WildFake	**94.9/98.1/96.2**	*87.0/91.4/93.8*	*98.9/99.7/99.8*	*71.4/89.0/79.1*	**99.7/99.9/99.9**	**93.0/95.6/93.7**

Table 2: Evaluating the generalized performance across different datasets and detectors. Performance metrics including (ACC(%), AP(%), and AUC(%)) are reported.

20% samples are randomly selected as the testing set from fake images generated by each generator, with the remainder forming the training set. A similar splitting strategy is applied to the real datasets shown in Figure 2.

4 Experiments

4.1 Experimental Settings

Baselines. For benchmarking purposes, we select high-quality and AI-generated image datasets (see Table 1) as baseline datasets, including *DiffusionDB* (Wang et al. 2022), *ArtiFact* (Rahman et al. 2023), *GenImage* (Zhu et al. 2023), and *DiffusionForensics* (Wang et al. 2023). Each of these datasets follows their original train-test split strategies. The relatively smaller-scale dataset, DiffusionForensics, is excluded from training considerations. DiffusionDB lacks real images, we incorporate real images from WildFake to train detectors on the DiffusionDB dataset. Additionally, WildFake's performance is compared with the recently introduced dataset, MPBench (Lu et al. 2024), in the appendix. The baseline AI-generated image detectors selected for evaluation in our study include *DIRE* (Wang et al. 2023), *IFDL* (Guo et al. 2023), *multiLID* (Lorenz, Durall, and Keuper 2023), *LASTED* (Wu, Zhou, and Zhang 2023), *ViT* (Dosovitskiy et al. 2020) and *ResNet50* (He et al. 2016). For comprehensive insights into these detectors, please refer to supplementary. In terms of baseline methodologies, our experiments conform to the configurations specified in the respective original paper. For ResNet50 and ViT, we employ pretrained models to execute binary classification distinguishing between real and fake images. All training images are resized to 224×224, with the Adam optimizer and Exponentially Decay scheduler with an initial learning rate of $1e-4$, and batch size (*resp.,* epoch) is set as 1024 (*resp.,* 15).

Evaluation Metrics. We report accuracy (ACC) and average precision (AP) in our experiments to evaluate the AI-generated image detectors. The threshold for computing accuracy is set to 0.5. Besides, we include the Area Under the ROC Curve (AUC) as another critical metric.

4.2 Comparing Generalizability of WildFake to Baseline Datasets

In our comparative analysis with WildFake, baseline datasets including DiffusionDB (Wang et al. 2022), ArtiFact (Rahman et al. 2023), and GenImage (Zhu et al. 2023), along with DiffusionForensics (Wang et al. 2023), are selected based on their volume and diversity for training detectors and evaluating their performance. We utilized foundational detectors such as ResNet50 (He et al. 2016) and ViT (Radford et al. 2021)), and advanced models like DIRE (Wang et al. 2023), IFDL (Guo et al. 2023), multiLID (Lorenz, Durall, and Keuper 2023), and LASTED (Wu, Zhou, and Zhang 2023) for this analysis. It was observed that across all detectors, each dataset demonstrated superior performance on its respective test set. However, WildFake stood out by not only excelling in its test set but also showcasing the second-best performance across the other four datasets. ***WildFake's dataset average performance is remarkable, leading by a significant margin against the second-best, GenImage.*** WildFake emerged as the dataset with the highest average test performance among

Training Dataset	Testing Dataset					Avg
	DiffusionForensics	GenImage	DiffusionDB	ArtiFact	WildFake	
WildFake(1/16)	72.9/84.8/72.6	67.4/72.8/73.4	85.2/87.4/89.4	58.6/71.8/63.0	91.9/92.1/91.7	75.2/81.8/78.1
WildFake(1/8)	76.2/87.0/75.3	71.7/77.4/78.0	88.9/91.3/92.4	60.3/74.2/66.0	94.8/94.9/95.0	78.4/84.9/81.3
WildFake(1/4)	79.1/90.1/78.7	74.3/81.7/82.2	91.8/94.9/95.0	62.1/ 78.2/70.3	96.9/96.9/97.0	80.8/88.4/84.6
WildFake(1/2)	81.8/93.3/81.2	77.8/86.9/86.4	93.9/97.8/97.9	65.2/ 81.3/72.8	98.2/98.3/98.3	83.4/91.5/87.5
GenImage	84.2/95.6/77.9	99.7/99.9/99.9	93.0/96.3/94.3	61.3/66.1/52.8	71.6/92.1/79.0	81.9/90.0/80.7
DiffusionDB	84.2/84.1/49.4	50.1/49.3/48.5	99.9/100/100	61.3/61.4/50.0	81.1/82.7/72.8	75.3/75.5/64.1
ArtiFact	85.4/94.9/76.7	76.5/84.8/82.9	64.1/69.9/68.1	97.2/99.5/99.3	85.4/94.9/76.7	81.7/88.8/87.4
WildFake	87.2/96.6/83.4	89.0/89.9/89.3	96.3/99.2/99.2	68.0/84.7/75.3	99.6/99.9/99.9	**86.4/94.1/89.4**

Table 3: Cross-dataset experiments with fractions of WildFake.

Method	DownSample		Compression		Geometric Transformation		Watemarks		Color Trans
	128	64	q=70	q=35	Flip	Crop	Text	Image	
ResNet50	91.1/95.5/93.1	71.3/65.0/39.8	84.6/95.9/92.5	85/93.1/87.7	95.1/98.4/96.4	91.3/98.7/96.9	91.0/98.8/94.1	90.8/98.8/93.9	87.9/97.3/94.8
ViT	91.8/94.6/92.4	**79.3/78.2/66.2**	**92.4/98.1/96.0**	86.6/95.1/95	**97.1/99.8/99.4**	**98.9/99.9/99.1**	93.6/99.3/96.6	92.9/99.3/96.5	**98.5/99.9/99.8**
DIRE	85.8/88.9/87.6	61.6/56.3/33.9	75.6/82.7/81.0	65.4/75.5/71.2	85.0/91.8/89.3	86.3/92.5/91.6	87.7/94.2/90.6	86.5/93.8/89.4	81.3/89.1/86.8
IFDL	91.3/95.2/93.4	73.1/69.6/49.4	87.4/97.8/95.1	82.9/94.1/89.3	94.1/99.4/98.2	94.9/99.6/99.1	93.8/99.3/96.9	91.3/99.1/94.9	96.0/98.2/97.1
Multi-LID	59.3/62.4/61.5	51.3/50.2/35.8	55.7/58.6/57.5	53.3/56.8/54.6	58.9/64.1/61.6	59.5/64.4/63.3	60.0/61.2/59.1	58.0/62.2/59.0	58.3/59.4/61.6
LASTED	**92.5/96.5/93.2**	75.8/71.3/51.8	89.6/97.7/95.0	**88.5/95.7/91.5**	96.1/99.3/99.0	98.0/99.6/99.6	**95.4/99.4/97.6**	**94.7/99.5/97.4**	92.3/99.8/99.6

Table 4: Robustness evaluation of various detectors trained on WildFake.

Training Subset	Testing Subset		
	DMs	GANs	Others
DMs	**99.7/99.9/99.9**	86.4/95.9/89.8	79.3/93.5/84.5
GANs	77.1/83.2/74.3	**98.1/99.0/99.6**	91.4/97.2/94.6
Others	76.4/77.2/70.1	82.5/96.0/91.2	**99.6/99.9/99.9**

Table 5: Cross-generator evaluation of WildFake results from DMs generators on diverse training and testing subsets via ViT detector.

all considered detectors. All these findings highlight WildFake's excellent generalizability. The comprehensive nature of WildFake, featuring a wide variety, and hierarchical quality of fake images generated by various generators, contributes to the enhanced performance of detectors trained on it. This distinct advantage over other baseline datasets highlights WildFake's significant contribution to improving the generalizability and efficacy of AI-generated image detection algorithms.

4.3 Necessity of Large Volume of WildFake For Generalizability

To elucidate the impact of data volume on detection performance, we conducted experiments utilizing ResNet50 across variously sized subsets of the WildFake dataset. These subsets were randomly selected, representing fractions from 1/2 to 1/16 of each category within the dataset. The comparative outcomes of these experiments are detailed in Table 3, with ResNet50 as the selected detector. The results distinctly highlight the correlation between dataset size and improved performance, thereby affirming the critical importance of substantial data volumes in enhancing detection capabilities. Please note ***even when scaled down to half, WildFake still outperforms GenImage, ArtiFact, and DiffusionDB*** All these datasets are larger than half of WildFake's volume. WildFake remains highly competitive even when reduced to a quarter of its size. This evidence of the high quality of WildFake, despite diminished dataset sizes, emphasizes the superior quality of data within WildFake for AI-generated image detection tasks.

4.4 Robustness Performance of Detectors on WildFake

Image degradation issues, such as low resolution, noise, and watermarks, often occur during propagation, presenting significant challenges for detectors (Schettini and Corchs 2010). Evaluating a detector's resilience to such degradation is essential for its real-world application. To this end, we applied a series of degradation techniques to the testing set images of the WildFake dataset to examine the robustness of detectors trained on it. The degradation methods include: (1) DownSample: down-sampling the high-resolution images to resolutions of 128 or 64. (2) Compression: introducing compression artifacts to the testing set by applying JPEG compression with quality ratios on the original test images. (3) Geometric Transformation: Randomly flipping or cropping the images from the testing set. (4) Watermark: randomly adding textual or visual watermarks on the random position of images from the testing set. (5) Color Transformation: we randomly change the brightness, contrast, saturation, and hue of images from the testing set. The robustness of six detectors (ResNet50, ViT, DIRE, IFDL, multiLID, and LASTED) against these degraded images was evaluated, with results detailed in Table 4. The analysis indicates that comparing ViT and ResNet-50, the ViT outperforms the ResNet-50 on degraded images, showcasing superior robustness, especially in scenarios involving geometric and color transformations. ResNet-50 shows greater sensitivity to these types of degradations, whereas ViT exhibits better resistance. Specifically, ViT achieves the best performance on geometric transformations, while LASTED shows enhanced effectiveness in handling watermarks. Both ViT and LASTED demonstrate

Training	Testing							
	ADM	DALLE	DDIM	DDPM	Imagen	VQDM	Midjourney	SD
ADM	100/100/100	93.3/98.9/97.8	78.6/90.9/84.0	80.5/92.0/86.1	96.9/99.3/99.0	90.6/97.6/96.0	84.4/92.8/89.3	87.6/97.0/94.1
DALLE	90.0/91.8/90.6	99.9/99.9/99.9	99.7/99.9/99.9	98.3/99.8/99.7	99.9/99.9/99.9	79.0/77.7/71.1	80.4/78.8/72.3	85.7/89.7/86.9
DDIM	89.9/91.0/87.1	99.7/99.9/99.9/	99.9/99.9/99.9	99.7/99.9/99.9	99.9/99.9/99.9	75.6/79.9/71.0	76.2/85.5/78.0/	88.5/90.7/86.1
DDPM	92.0/89.9/88.5	99.7/99.9/99.8	99.9/99.9/99.9	99.8/99.9/99.9	100/100/100	76.0/74.1/64.9	77.0/75.7/67.8	89.1/88.7/86.1
Imagen	80.3/92.4/85.2	98.6/99.7/99.3	95.0/98.5/97.1	94.7/98.3/96.4	100/100/100	81.8/89.1/85.4	84.0/89.8/86.9	77.3/91.3/82.7
VQDM	91.8/98.1/96.7	88.4/96.5/91.3	79.7/85.4/79.6	71.0/84.0/67.1	98.1/99.4/98.6	99.9/99.9/99.9	95.7/98.7/97.4	93.5/98.0/95.8
Midjourney	99.9/99.9/99.9	99.9/99.9/99.9	99.6/99.9/99.9	99.4/99.9/99.9	100/100/100	99.9/99.9/99.9	99.9/99.9/99.9	99.8/99.9/99.9
SD	99.9/99.0/99.9	100/100/100	99.7/99.9/99.9	99.3/99.9/99.9	100/100/100	99.9/99.9/100	99.9/100/99.9	100/99.9/100

Table 6: Cross-architecture evaluation of WildFake results from DMs generators on diverse training and testing subsets via ViT detector.

commendable robustness against DownSample and Compression effects. Furthermore, the analysis highlights that lower resolutions and lower quality significantly impair the accuracy of all trained detectors, underscoring the importance of robustness in detector design for handling various forms of image degradation in practical deployments.

4.5 Generalizability Evaluation inside WildFake Hierarchies

The unique hierarchical structure of the WildFake dataset allows for an in-depth analysis of trained detectors' generalization capabilities across different hierarchical levels, an aspect not readily feasible with other baseline datasets. As depicted in Figure 1, WildFake is systematically divided into five distinct levels: the first level is cross-generators consisting of three types of generators, the second level is cross-architecture in DMs generators consisting of eight types of DMs architectures, the third level is the cross-weight in SD consisting of three types of weights, the last two levels are cross-version and cross-time in three types of generators consisting of typical (*resp.,* early) generators and advanced (*resp.,* latest) generators. We utilize the baseline detector ViT to conduct comprehensive generalization experiments: (1) Cross-generator experimental comparison is designed to evaluate the gap among different generators in Table 5. (2) Comparison among cross-architecture is designed to evaluate the effects of DMs generators with different architectures in Table 6. (3) Cross-weight evaluation of SD, cross-time evaluation of GANs (*resp.,* Others), and cross-version evaluation of Midjourney (*resp.,* SD) are reported in Appendix.

Evaluation on Cross-Generator. We first assess the performance of the ViT detector when trained and tested on images generated by the same type of generator within WildFake. Our WildFake dataset consists of three types of generators including DMs, GANs, and Others. Accordingly, we divide the WildFake dataset into three subsets, each with its training and testing set, based on the generator type. We then train the ViT model separately on each generator type and evaluate its performance on the corresponding testing set of each type. The results in Table 5 indicate that the in-domain generalization ability is significantly superior to cross-domain generalization. Notably, models trained on DMs exhibit a lesser degree of generalization compared to those trained on GANs and Others. This suggests that the disparity between DMs and GANs (*resp*., Others) is more pronounced than that between GANs and Others. We hypothesize that this is due to the distinct image generation approaches: while Others and GANs typically employ one-step inference for image generation, DMs utilize multiple denoising steps.

Evaluation on Cross-Architecture of DMs Considering the high quality of images generated by DMs, we further classify images from DMs into 8 categories according to the difference of architectures consisting of SD, DDPM, DDIM, ADM, DALLE, Imagen, Midjourney, and VQDM. In Table 6, we can see that in-architecture testing performance reaches to high level. The performance of detectors over cross-architecture scenarios is observed to be worse than that of in-architecture validations. This suggests that different architectures within DMs might produce fake images with varying levels of sophistication. Another notable finding is that models trained on Midjourney and SD demonstrate superior generalization ability compared to other architectures. The reasons for this are threefold: (1) The volume of training images from Midjourney and SD is greater than that of other architectures, offering a more extensive learning base. (2) The content diversity of fake images from Midjourney and SD is richer, providing a broader spectrum of data for model training. (3) A portion of the fake images in Midjourney and SD are sourced from open community platforms, typically exhibiting higher quality compared to those from other architectures.

5 Conclusion

We present a large-scale dataset WildFake, to assess the generalizability and robustness of fake image detection. The dataset includes fake images generated by various types of generators, encompassing GANs, diffusion models, and other generative models. The key strengths of WildFake notably enhance the generalization and robustness of detectors trained with this dataset, showcasing its significant applicability and effectiveness in real-world scenarios for AI-generated image detection. Furthermore, our in-depth evaluation experiments are designed to provide substantial insights into the capabilities of generative models at different levels, a unique benefit derived from WildFake's distinct hierarchical structure.

Acknowledgements

This work is supported, in part, by the National Natural Science Foundation of China (Grant No. 62302295), the foundation of Key Laboratory of Artificial Intelligence, Ministry of Education, P.R. China, and the Pioneer R&D Program of Zhejiang Province (No. 2024C01024).

References

Aghasanli, A.; Kangin, D.; and Angelov, P. 2023. Interpretable-Through-Prototypes Deepfake Detection for Diffusion Models. In *ICCV*.

Alanov, A.; Titov, V.; Nakhodnov, M.; and Vetrov, D. 2023. StyleDomain: Efficient and Lightweight Parameterizations of StyleGAN for One-shot and Few-shot Domain Adaptation. In *ICCV*.

Bird, J. J.; and Lotfi, A. 2023. CIFAKE: Image Classification and Explainable Identification of AI-Generated Synthetic Images. *arXiv preprint arXiv:2303.14126*.

Brock, A.; Donahue, J.; and Simonyan, K. 2018. Large Scale GAN Training for High Fidelity Natural Image Synthesis. In *ICLR*.

Chang, H.; Zhang, H.; Barber, J.; Maschinot, A.; Lezama, J.; Jiang, L.; Yang, M.-H.; Murphy, K.; Freeman, W. T.; Rubinstein, M.; et al. 2023. Muse: Text-to-image generation via masked generative transformers. *arXiv preprint arXiv:2301.00704*.

Choi, Y.; Choi, M.; Kim, M.; Ha, J.-W.; Kim, S.; and Choo, J. 2018. StarGAN: Unified generative adversarial networks for multi-domain image-to-image translation. In *CVPR*.

Choi, Y.; Uh, Y.; Yoo, J.; and Ha, J.-W. 2020. StarGAN v2: Diverse image synthesis for multiple domains. In *CVPR*.

Corvi, R.; Cozzolino, D.; Zingarini, G.; Poggi, G.; Nagano, K.; and Verdoliva, L. 2023. On the detection of synthetic images generated by diffusion models. In *ICASSP*.

Deng, J.; Dong, W.; Socher, R.; Li, L.-J.; Li, K.; and Fei-Fei, L. 2009. Imagenet: A large-scale hierarchical image database. In *CVPR*.

Dhariwal, P.; and Nichol, A. 2021. Diffusion models beat GANs on image synthesis.

Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. 2020. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In *ICLR*.

Esser, P.; Rombach, R.; and Ommer, B. 2021. Taming transformers for high-resolution image synthesis. In *CVPR*.

Feichtenhofer, C.; Li, Y.; He, K.; et al. 2022. Masked autoencoders as spatiotemporal learners.

Frank, J.; Eisenhofer, T.; Schönherr, L.; Fischer, A.; Kolossa, D.; and Holz, T. 2020. Leveraging frequency analysis for deep fake image recognition. In *ICML*.

Gal, R.; Alaluf, Y.; Atzmon, Y.; Patashnik, O.; Bermano, A. H.; Chechik, G.; and Cohen-or, D. 2022. An Image is Worth One Word: Personalizing Text-to-Image Generation using Textual Inversion. In *ICLR*.

Gu, J.; Meng, X.; Lu, G.; Hou, L.; Minzhe, N.; Liang, X.; Yao, L.; Huang, R.; Zhang, W.; Jiang, X.; et al. 2022a. Wukong: A 100 million large-scale chinese cross-modal pre-training benchmark. In *NeurIPS*.

Gu, S.; Chen, D.; Bao, J.; Wen, F.; Zhang, B.; Chen, D.; Yuan, L.; and Guo, B. 2022b. Vector quantized diffusion model for text-to-image synthesis. In *CVPR*.

Guo, X.; Liu, X.; Ren, Z.; Grosz, S.; Masi, I.; and Liu, X. 2023. Hierarchical fine-grained image forgery detection and localization. In *CVPR*.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *CVPR*.

hello@civitai.com. 2022. civitai.

Ho, J.; Jain, A.; and Abbeel, P. 2020. Denoising diffusion probabilistic models. In *NeurIPS*.

Holub, O. 2022. Midjourney.

Hu, E. J.; Shen, Y.; Wallis, P.; Allen-Zhu, Z.; Li, Y.; Wang, S.; Wang, L.; and Chen, W. 2021. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*.

Kang, M.; Zhu, J.-Y.; Zhang, R.; Park, J.; Shechtman, E.; Paris, S.; and Park, T. 2023. Scaling up GANs for text-to-image synthesis. In *CVPR*.

Karras, T.; Aila, T.; Laine, S.; and Lehtinen, J. 2017. Progressive growing of GANs for improved quality, stability, and variation. *arXiv preprint arXiv:1710.10196*.

Karras, T.; Aittala, M.; Laine, S.; Härkönen, E.; Hellsten, J.; Lehtinen, J.; and Aila, T. 2021. Alias-free generative adversarial networks. In *NeurIPS*.

Karras, T.; Laine, S.; and Aila, T. 2019. A style-based generator architecture for generative adversarial networks. In *CVPR*.

Karras, T.; Laine, S.; Aittala, M.; Hellsten, J.; Lehtinen, J.; and Aila, T. 2020. Analyzing and improving the image quality of StyleGAN. In *CVPR*.

Lin, T.-Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; and Zitnick, C. L. 2014. Microsoft coco: Common objects in context. In *ECCV*.

Lorenz, P.; Durall, R. L.; and Keuper, J. 2023. Detecting Images Generated by Deep Diffusion Models using their Local Intrinsic Dimensionality. In *ICCV*.

Lu, Z.; Huang, D.; Bai, L.; Qu, J.; Wu, C.; Liu, X.; and Ouyang, W. 2024. Seeing is not always believing: Benchmarking human and model perception of ai-generated images. In *NeurIPS*.

Marra, F.; Gragnaniello, D.; Cozzolino, D.; and Verdoliva, L. 2018. Detection of GAN-generated fake images over social networks. In *MIPR*.

Marra, F.; Gragnaniello, D.; Verdoliva, L.; and Poggi, G. 2019. Do GANs leave artificial fingerprints? In *MIPR*.

Nichol, A. Q.; Dhariwal, P.; Ramesh, A.; Shyam, P.; Mishkin, P.; Mcgrew, B.; Sutskever, I.; and Chen, M. 2022. GLIDE: Towards Photorealistic Image Generation and Editing with Text-Guided Diffusion Models. In *ICML*.

OpenAI. 2023. DALL·E 3 System Card.

Pehlivan, H.; Dalva, Y.; and Dundar, A. 2023. Styleres: Transforming the residuals for real image editing with stylegan. In *CVPR*.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *ICML*.

Rahman, M. A.; Paul, B.; Sarker, N. H.; Hakim, Z. I. A.; and Fattah, S. A. 2023. ArtiFact: A Large-Scale Dataset with Artificial and Factual Images for Generalizable and Robust Synthetic Image Detection. *arXiv preprint arXiv:2302.11970*.

Ramesh, A.; Dhariwal, P.; Nichol, A.; Chu, C.; and Chen, M. 2022. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*.

Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; and Ommer, B. 2022. High-resolution image synthesis with latent diffusion models. In *CVPR*.

Ruiz, N.; Li, Y.; Jampani, V.; Pritch, Y.; Rubinstein, M.; and Aberman, K. 2023. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *CVPR*.

Saharia, C.; Chan, W.; Saxena, S.; Li, L.; Whang, J.; Denton, E. L.; Ghasemipour, K.; Gontijo Lopes, R.; Karagol Ayan, B.; Salimans, T.; et al. 2022. Photorealistic text-to-image diffusion models with deep language understanding. In *NeurIPS*.

Sauer, A.; Schwarz, K.; and Geiger, A. 2022. Stylegan-xl: Scaling stylegan to large diverse datasets. In *ACM SIGGRAPH*.

Schettini, R.; and Corchs, S. 2010. Underwater image processing: state of the art of restoration and image enhancement methods. *EURASIP journal on advances in signal processing*, 2010: 1–14.

Schuhmann, C.; Beaumont, R.; Vencu, R.; Gordon, C.; Wightman, R.; Cherti, M.; Coombes, T.; Katta, A.; Mullis, C.; Wortsman, M.; et al. 2022. Laion-5b: An open large-scale dataset for training next generation image-text models. In *NeurIPS*.

Sha, Z.; Li, Z.; Yu, N.; and Zhang, Y. 2022. De-fake: Detection and attribution of fake images generated by text-to-image diffusion models. *arXiv preprint arXiv:2210.06998*.

Song, J.; Meng, C.; and Ermon, S. 2020. Denoising Diffusion Implicit Models. In *ICLR*.

Tao, M.; Bao, B.-K.; Tang, H.; and Xu, C. 2023. GALIP: Generative Adversarial CLIPs for Text-to-Image Synthesis. In *CVPR*.

Tao, M.; Tang, H.; Wu, F.; Jing, X.-Y.; Bao, B.-K.; and Xu, C. 2022. Df-GAN: A simple and effective baseline for text-to-image synthesis. In *CVPR*.

Van Den Oord, A.; Vinyals, O.; et al. 2017. Neural discrete representation learning. In *NeurIPS*.

Verdoliva, L.; Cozzolino, D.; and Nagano, K. 2022. Image and Video Processing Cup Synthetic Image Detection.

Voynov, A.; Chu, Q.; Cohen-Or, D.; and Aberman, K. 2023. *P+*: Extended Textual Conditioning in Text-to-Image Generation. *arXiv preprint arXiv:2303.09522*.

Wang, S.-Y.; Wang, O.; Zhang, R.; Owens, A.; and Efros, A. A. 2020. CNN-generated images are surprisingly easy to spot... for now. In *CVPR*.

Wang, Z.; Bao, J.; Zhou, W.; Wang, W.; Hu, H.; Chen, H.; and Li, H. 2023. DIRE for Diffusion-Generated Image Detection. *arXiv preprint arXiv:2303.09295*.

Wang, Z. J.; Montoya, E.; Munechika, D.; Yang, H.; Hoover, B.; and Chau, D. H. 2022. Diffusiondb: A large-scale prompt gallery dataset for text-to-image generative models. *arXiv preprint arXiv:2210.14896*.

Wu, H.; Zhou, J.; and Zhang, S. 2023. Generalizable Synthetic Image Detection via Language-guided Contrastive Learning. *arXiv preprint arXiv:2305.13800*.

Yeh, S.-Y.; Hsieh, Y.-G.; Gao, Z.; Yang, B. B.; Oh, G.; and Gong, Y. 2023. Navigating Text-To-Image Customization: From LyCORIS Fine-Tuning to Model Evaluation. *arXiv preprint arXiv:2309.14859*.

Yu, F.; Seff, A.; Zhang, Y.; Song, S.; Funkhouser, T.; and Xiao, J. 2015. Lsun: Construction of a large-scale image dataset using deep learning with humans in the loop. *arXiv preprint arXiv:1506.03365*.

Zhang, L.; Rao, A.; and Agrawala, M. 2023. Adding conditional control to text-to-image diffusion models. In *ICCV*.

Zhu, M.; Chen, H.; Yan, Q.; Huang, X.; Lin, G.; Li, W.; Tu, Z.; Hu, H.; Hu, J.; and Wang, Y. 2023. GenImage: A Million-Scale Benchmark for Detecting AI-Generated Image. *arXiv preprint arXiv:2306.08571*.