Where's the liability in the Generative Era? Recovery-based Black-Box Detection of AI-Generated Content

Haoyue Bai
University of Wisconsin-Madison
haoyue.bai@wisc.edu

Yiyou Sun
University of California, Berkeley
sunyiyou@berkeley.edu

Wei Cheng
NEC Laboratories America
weicheng@nec-labs.com

Haifeng Chen
NEC Laboratories America
haifeng@nec-labs.com

Abstract

The recent proliferation of photorealistic images created by generative models has sparked both excitement and concern, as these images are increasingly indistinguishable from real ones to the human eye. While offering new creative and commercial possibilities, the potential for misuse, such as in misinformation and fraud, highlights the need for effective detection methods. Current detection approaches often rely on access to model weights or require extensive collections of real image datasets, limiting their scalability and practical application in real-world scenarios. In this work, we introduce a novel black-box detection framework that requires only API access, sidestepping the need for model weights or large auxiliary datasets. Our approach leverages a corrupt-and-recover strategy: by masking part of an image and assessing the model's ability to reconstruct it, we measure the likelihood that the image was generated by the model itself. For black-box models that do not support masked-image inputs, we incorporate a cost-efficient surrogate model trained to align with the target model's distribution, enhancing detection capability. Our framework demonstrates strong performance, outperforming baseline methods by 4.31% in mean average precision across eight diffusion model variant datasets.

1. Introduction

The rapid advancement of generative models [4, 30, 35] has driven remarkable progress in synthesizing photorealistic images, offering numerous benefits yet also raising concerns about potential misuse. For instance, the creation of fake images, such as the widely circulated "Trump getting arrested" photo [20], can escalate public confusion and fuel misinformation. Similarly, a Hong Kong employee was deceived into transferring money to criminals through an AI-generated video call [41].

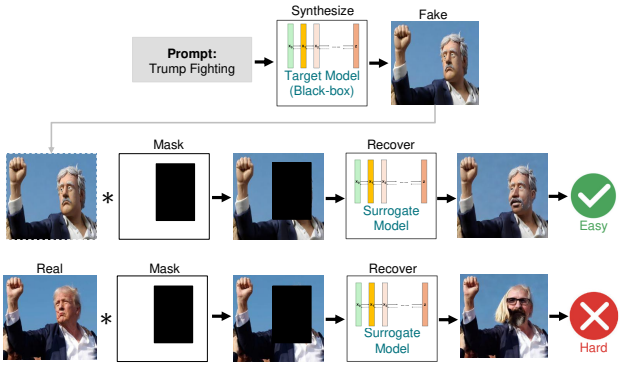


Figure 1. Visual quality comparison of surrogate model recovery output between real images and images generated by the target model from masked examples. We observe that images generated by target models with masks are more likely to be accurately recovered compared to real images. These results are based on Stable Diffusion as the surrogate model.

This underscores the urgent need for robust methods to distinguish real images from AI-generated ones. However, developing such methods is challenging given that current generative models can produce images with a photorealistic quality. For instance, Appendix E showcases examples where humans struggle to accurately differentiate real from fake, with accuracy often as low as chance level (50-50). In a curated dataset of challenging cases, humans could only identify 70% of the fake images correctly as we show in Section 4.4.

Considerable effort has gone into developing detection methods [6, 16, 21, 44]. However, many of these are "white-box" methods that require access to model weights or token information. The real challenge, however, lies in detecting

fake images generated through widely accessible, low-cost "black-box" APIs [3, 40]. In these model-agnostic or black-box settings, many other detection approaches [28, 42] typically rely on binary classifiers trained on limited datasets of real versus fake images. However, given the vast diversity of real-world images, such classifiers are prone to overfitting and will exhibit overconfidence or vulnerability when exposed to previously unseen data [1] in open-world environments. As diffusion-based generative models continue to evolve, there remains a critical need for practical, generalizable, interpretable, and robust methodologies to reliably detect AI-generated images.

To address these limitations, our work introduces a novel black-box fake detection method that requires only API access, without needing direct access to model weights or auxiliary datasets. Our framework is grounded in a straightforward intuition:

A generative model should more easily recover its own generated images when corrupted than it would with real images.

This intuition, validated in previous work [46], is based on the premise that the distribution of AI-generated content inside the network significantly differs from that of real-world images. Using this insight, we developed a detection algorithm that follows a pipeline illustrated in Figure 1: (1) corrupt the target image by applying a mask, (2) use the generative model to recover the masked content, and (3) compare the quality of the recovery to the uncorrupted version. Ideally, if the target model generated the image, it should easily recover the masked areas, given that the image aligns with its own distribution. For example, a model that generates an image of a "clay" Trump can likely recover clay-specific features, such as a clay mouth, but would struggle to accurately reconstruct details in a realistic photo of Trump. Thus, we can discern real versus fake by calculating a measurable score that compares the original image to the corrupt-recover version. We expect this score to be higher for real images and lower for fake ones.

For public models where the API only supports generation from scratch (without a masked-image input), we employ a surrogate model. This surrogate is trained in a cost-effective manner to closely align its probability distribution with that of the target black-box model, effectively synchronizing their distributions for accurate detection. Unlike previous methods [28], which require a large corpus of real images (approximately 400k) and significant computational resources, our approach requires fewer than 1,000 samples from the target model's API and less than 2 GPU hours of compute time.

Our main contributions are summarized as follows:

- **Competitive Black-Box Detection Framework.** We introduce a novel black-box detection framework that identifies AI-generated images using only API access, eliminating the need for access to model weights or large auxiliary datasets.
- **Novel Corrupt-and-Recover Detection Paradigm.** Our approach leverages a unique corrupt-and-recover pipeline, where a model's ability to reconstruct its own generated content provides an effective measure to differentiate between real and fake images.
- **Efficient Resource Requirements.** Unlike prior methods that require extensive datasets and computational resources, our approach achieves high detection performance with fewer than 1,000 API samples and minimal GPU time.
- **Improved Generalizability and Practicality**. Our method demonstrates enhanced generalizability across diverse generative models, making it a practical and robust solution as generative models continue to improve in quality and accessibility.

2. Related Works

2.1. Synthetic Image Generation

The field of generative models has been significantly advanced with the introduction of Generative Adversarial Networks (GANs) [5, 8, 50]. Some works have exploited Transformers to enhance generated image quality [7, 29, 48]. The advent of diffusion models has led to significantly improved cutting-edge generation models, including Stable Diffusion [35], DALL-E [31], DALL-E 2 [31], DALL-E 3 [4], GLIDE [27], and others [12, 52].

2.2. Detection of AI-Generated Content

With the proliferation of synthetic image generators, designing methods to detect generated content has attracted much attention [51]. Some earlier works focus on detecting fake faces. The work in [25] leverages visual features such as eyes, teeth, and facial contours. The work of [26] examines color information related to the synthesis of RGB color channels. Most recent detectors rely on traces that are invisible to the human eye, inherent to the generation process, and based on semantic, physical, or statistical inconsistencies. One direction identifies feature frequency artifacts for GAN-generated images [24, 49]. The work in [47] studies learning GAN fingerprints for image attribution. Patchfor [6] uses classifiers with limited receptive fields to focus on local artifacts instead of global semantics of the images.

The work in [9] shows that GAN detectors perform poorly on diffusion model-generated images. Recent detection techniques have begun studying diffusion model-based images. Synthbuster [2] investigates the inherent frequency artifacts during the diffusion process and leverages spectral analysis to highlight the artifacts in the Fourier transform of a residual image for fake detection. Other

works exploit lighting [13] and perspective [14] inconsistencies of DALL-E 2 generated images. DE-FAKE [39] focuses on advanced text-to-image generation models including DALL-E 2 and Stable Diffusion, and observes that incorporating prompts or generated captions into the detector improves classification. DIRE [43] observes that diffusion-generated images can be approximately reconstructed by a diffusion model while real images cannot. The work in [32] observes that diffusion models produce fewer detectable artifacts and are more difficult to detect compared to GANs, and explores retraining GAN detectors on diffusion model-generated images to show improved detection. The work in [44] defines a reverse-engineering task for generative models and analyzes the disparities in reconstruction loss between the generated samples of the specific model and others. The work in [16] exploits denoising diffusion probabilistic models as denoising autoencoders and uses the resulting multi-dimensional reconstruction error to classify out-of-distribution inputs. Universal fake detector [28] proposes using a pre-trained vision transformer with a final classification layer for fake detection of both GAN and diffusion model-generated images. However, if a network is trained on a specific model, its performance degrades when used to detect images generated by another architecture [10]. This suggests that each generation architecture contains its own peculiar traces.

2.3. Black-box Detection

The specific model used by the attacker is often unavailable. One approach is to train a classifier with fewer or even no fake images from the pre-trained generative model. AutoGAN [49] investigates the artifacts induced by the upsampler of GAN pipelines in the frequency domain to develop robust spectrum-based fake image classifiers. Some recent works [15, 23, 42] claim to perform well on images from unseen generative models. However, it remains unclear whether this performance holds for images generated by diffusion models.

3. Methodology

Motivation. The core idea behind our approach is: a generative model should more easily recover its own generated images when corrupted than it would with real images. Technically, we hypothesize that a generative model $\mathbb{G}_t$ can inherently "recognize" its own outputs, allowing it to reconstruct masked regions of its own generated images more effectively than it can for other images, such as those from the real world. This occurs because the model's learned distribution aligns closely with the statistical properties of its own outputs, making these images easier to "fill in" when corrupted. In contrast, real images or images generated by other models follow distributions that $\mathbb{G}_t$ hasn't explicitly learned, so it struggles to accurately restore missing content in these cases.

Problem Setup. Our task is to detect whether a given image $\mathbf{x}$ is produced by a target generative model $\mathbb{G}_t$ (where we only have API access), which can be framed as a binary classification problem where we want to determine if $\mathbf{x}$ is generated by $\mathbb{G}_t$ (i.e., $y = 1$) or not (i.e., $y = 0$):

$$\hat{y} = \begin{cases} 1, & \text{if } \delta(\mathbf{x}) < \tau \\ 0, & \text{otherwise} \end{cases} \tag{1}$$

Here, τ is a predefined threshold. If the discrepancy score $\delta(\mathbf{x})$ is below this threshold, we classify the image as being generated by $\mathbb{G}_t$ ($\hat{y} = 1$); otherwise, we classify it as not generated by $\mathbb{G}_t$ ($\hat{y} = 0$).

There are typically two different task settings for AI-generated content detection: black-box detection (with access only to input and output) and white-box detection (with additional information about the model internals). We focus on the black-box setting in this work, as large tech companies and AI research organizations often keep their most advanced models closed-source, such as DALL-E 3 and SORA by OpenAI. In these cases, we can only access the model through their API, while the underlying code and model weights remain unavailable to the public. Therefore, our goal is to improve black-box detection without any access to the target model weights.

Preliminary on Diffusion Models. Diffusion Models are a group of probabilistic generative models. Since the milestone work DDPM [18], there are numerious improvements with higher fidelity and diversity [27, 31, 35]. A diffusion probabilistic model is a parameterized Markov chain trained using variational inference to produce samples matching the data after finite time, which gradually diffuse a sample from this distribution and then learn to reverse this diffusion process.

In the diffusion (or forward) process for DDPMs, a sample $\mathbf{x}_0$ (e.g., an image) is repeatedly corrupted by Gaussian noise in sequential steps $t = 1, \ldots, T$ in dependence of a monotonically increasing noise schedule $\{\beta_t\}_{t=1}^T$:

$$q(\mathbf{x}_t|\mathbf{x}_{t-1}) = \mathcal{N}(\sqrt{1-\beta_t}\mathbf{x}_{t-1}, \beta_t\mathbf{I}) \ . \tag{2}$$

With $\alpha_t = 1 - \beta_t$ and $\bar{\alpha}_t = \prod_{s=1}^t \alpha_t$, we can directly sample from the forward process at arbitrary times:

$$q(\mathbf{x}_t|\mathbf{x}_0) = \mathcal{N}(\sqrt{\bar{\alpha}_t}\mathbf{x}_0, (1-\bar{\alpha}_t)\mathbf{I}) \ . \tag{3}$$

The noise schedule is typically designed to satisfy $q(\mathbf{x}_T|\mathbf{x}_0) \approx \mathcal{N}(\mathbf{0}, \mathbf{I})$. During the denoising process, we aim to iteratively sample from $q(\mathbf{x}_{t-1}|\mathbf{x}_t)$ to ultimately obtain a clean image from $\mathbf{x}_T \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$. However, since $q(\mathbf{x}_{t-1}|\mathbf{x}_t)$ is intractable as it depends on the entire underlying data distribution, it is approximated by a deep neural network. More formally, $q(\mathbf{x}_{t-1}|\mathbf{x}_t)$ is approximated by

$$p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t) = \mathcal{N}(\mu_\theta(\mathbf{x}_t, t), \Sigma_\theta(\mathbf{x}_t, t)) \ , \tag{4}$$

where mean μ_θ and covariance Σ_θ are given by the output of model (or the latter is set to a constant as shown in [18]).

3.1. Recovery-based Detection Methods

Given an input image $\mathbf{x}$, our goal is to determine whether it is synthesized by a generative model or if it is a real image. We define a mask m to divide the image into two parts: the known pixels $(1-m)\odot\mathbf{x}$ and the unknown pixels $m\odot\mathbf{x}$. As illustrated in Figure 2, we apply a generative model to recover the unknown pixels $m \odot \mathbf{x}$ conditioned on the known pixels $(1 - m) \odot \mathbf{x}$. The surrogate generative model $\mathbb{G}_s$ is usually shared with an inpainting model, which provides both the ability to generate from scratch and to generate from partially masked inputs [34]. The difference between the input $m \odot \mathbf{x}$ and the recovered $m \odot \mathbf{x}$ helps distinguish between real and generated images. We compute a metric δ of this discrepancy gap and use it as a scoring function to classify the source image as either real or generated. In practice, we sample the recovery results K times to account for the stochastic nature of the process and to obtain a more robust evaluation.

Details of Scoring Function δ. Our learning framework is orthogonal, thus compatible with various metrics δ used for measuring discrepancy as a scoring function. We evaluate four different types of scoring functions in this work: *i*) Peak Signal-to-Noise Ratio (PSNR), which measures the ratio of signal to noise; *ii*) Structural Similarity Index (SSIM), which quantifies structural similarity; *iii*) L1 distance, which measures absolute pixel-wise differences; and *iv*) L2 distance, which measures squared pixel-wise differences. More details of δ can be found in Appendix B.

As shown in Section 4.3, PSNR achieves better performance in fake image detection compared to other metrics. This may be due to the fact that fake images often contain subtle alterations, and PSNR excels at detecting small pixel-wise differences, making it highly sensitive to fine-grained changes. In contrast, L1 and L2 distances do not normalize these differences relative to the image's dynamic range, while PSNR normalizes the error against the maximum possible pixel value, making it more interpretable and robust to variations in intensity. Additionally, SSIM focuses more on structural similarity and perceptual quality, which may cause it to overlook subtle pixel-level deviations.

3.2. Distribution-Aligned Black-Box Detection

For public models where the API only supports generation from scratch, we utilize a distribution-aligned surrogate model $\mathbb{G}_s$ to recover masked images generated by a target model $\mathbb{G}_t$. We observe that images generated by target models with masks can be accurately recovered by a distribution-aligned model, as shown in Figure 1 where a "clay trump"'s mouth can be recovered in similar style.

In the black-box detection setting, selecting an appropriate surrogate model is crucial for achieving accurate and reliable results. There exists a distribution gap between the given surrogate model and the target model. We aim to obtain a surrogate model that approaches the distribution of the target model by utilizing images generated by the target model. We propose a novel and efficient framework to train a distribution-aligned surrogate model that achieves good performance for black-box detection with a small-sized dataset. As shown in Figure 2, we first collect a small set of training data generated by the target model from the publicly shared API. Then, we perform parameter-efficient fine-tuning of the surrogate model using this training dataset to align its distribution with the source model.

Alignment Data Collection To align the distribution of the surrogate model $\mathbb{G}_s$ with the target model $\mathbb{G}_t$, we collect a small-sized dataset $S = \{x_i\}_{i=1}^N$ for a specific target model, referred to as the alignment dataset. Here, N denotes the number of collected images, and x_i refers to an image generated by the target model through publicly shared APIs. We then utilize the collected dataset S to fine-tune the surrogate model $\mathbb{G}_s$, aligning its distribution with that of the target model $\mathbb{G}_t$.

Distribution-Aligned Surrogate Model Fine-Tuning As shown in Figure 2, we implement low-rank adaptation (LoRA) [19] for the surrogate model $\mathbb{G}s$ to enable parameter-efficient fine-tuning. The LoRA model $\mathbb{G}s + \theta$ is trained with the collected dataset S, while the parameters of the original surrogate model $\mathbb{G}_s$ remain frozen. After training, the previously misaligned model generates a distribution similar to that of the target model $\mathbb{G}_t$. Consequently, this distribution-aligned surrogate model can be utilized to perform recovery evaluation for downstream fake detection.

3.3. Theoretical Insights

Formally, given an input image $\mathbf{x}$, we define a mask m for dividing the image into two parts: $X = (1 - m) \odot x$, and $Y_0 = m \odot x$. Next, we ask the generative model to continue generating the remaining pixels purely based on X, and the generated results are denoted by $Y' \sim G(\cdot|X)$. In practice, we sample the new results for K times (refer to a principled choice of $K = \Omega\big(\sigma \log(1/\delta)/\Delta^2\big)$ in Appendix A.2) to get a set of sequences $\Omega = \{Y_1, ..., Y_k, ..., Y_K\}$. Our method is based on the hypothesis that the generation process G of the machine typically maximizes the log probability function throughout the generation process, while real image creation process is different. In other words, the thought process of real images does not simply follow the likelihood maximization criterion. We find that this discrepancy between machine and real is especially enormous when conditioned on the input pixels X, and we state this hypothesis formally as:

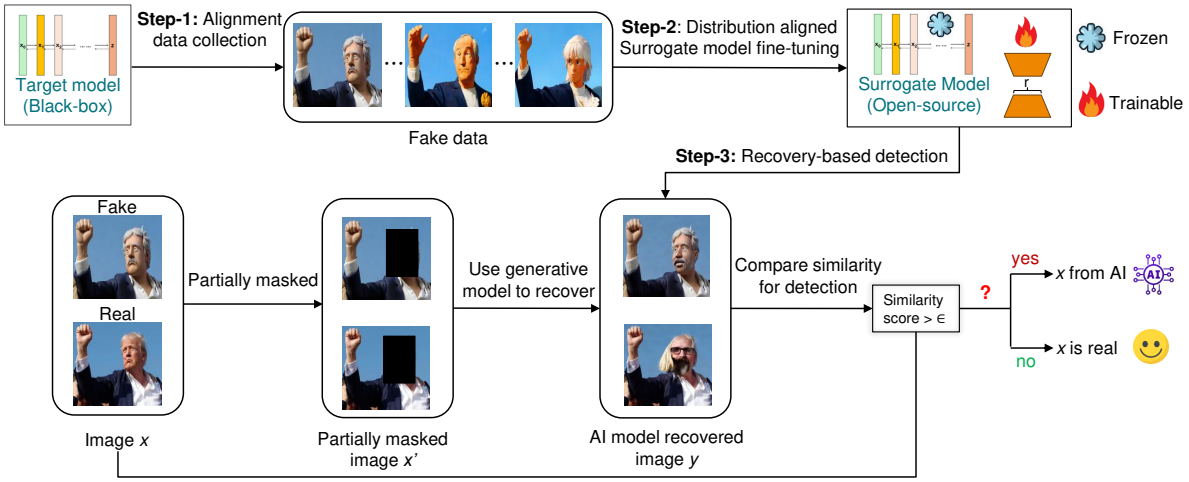


Figure 2. An overview of our proposed black-box content detection framework. Given a candidate input image, we aim to determine whether it is generated by a target model or if it is a real image. Our method first aligns the distribution of the surrogate model with that of the target model via alignment data collection and parameter-efficient surrogate model fine-tuning. We then perform recovery-based detection to calculate the scoring function for classification.

Likelihood-Gap Hypothesis. The expected log-likelihood of the machine generation process G has a positive gap $\Delta > 0$ over that of real generation process H :

$$\mathbb{E}_{Y\sim G(\cdot|X)}[\log p(Y|X)] - \mathbb{E}_{Y\sim H(\cdot|X)}[\log p(Y|X)] > \Delta.$$

This hypothesis states that, conditioned on the input parts of image, the log-likelihood value of the machine-generated remaining parts of image is significantly higher than the human-generated remaining pixels. An implication is that

$$\begin{aligned}\Delta &\leq \mathbb{E}_{Y\sim G(\cdot|X)}[\log p(Y|X)] - \mathbb{E}_{Y\sim H(\cdot|X)}[\log p(Y|X)] \\ &\leq \|\log p(\cdot|X)\|_\infty \cdot d_{\mathrm{TV}}(G,H) \\ &\leq \|\log p(\cdot|X)\|_\infty \cdot \sqrt{\frac{1}{2} d_{\mathrm{KL}}(G,H)}.\end{aligned}$$

The second inequality holds due to the definition of the total-variation distance; the third inequality holds due to Pinsker's inequality. When there is no ambiguity, we omit the parenthesis and condition, denote $G(\cdot|X)$ as G and the same for H.

4. Experiments

4.1. Experiments Setup

Datasets. We consider a variety of generative models, including Guided diffusion [11], the Latent Diffusion Model (LDM)[35], Glide[27], DALL-E [30], and DALL-E 3. For these methods, we use the LAION [37] dataset as the real class, while fake images are generated based on the corresponding text descriptions from LAION.

Following the data setup in [28], LDMs can generate images in various ways. The standard practice involves using a text prompt as input and performing 200 steps of noise refinement (LDM 200). Additionally, images can be generated with guidance (LDM 200 w/CFG) or using fewer steps for faster sampling (LDM 100).

Similarly, we test on different variants of a pre-trained Glide model, which consists of two stages of noise refinement. The standard approach uses 100 steps to create a low-resolution image at 64 x 64 pixels, followed by 27 steps to upscale the image to 256 x 256 pixels (Glide 100-27). We also consider two other variants: Glide 50-27 and Glide 100-10, which differ in the number of refinement steps used in the two stages.

Baselines. We consider several strong baseline methods: *i*) Trained Deep Network [42]: This method uses a ResNet-50 [17] pre-trained on ImageNet, fine-tuned on ProGAN's real and fake images to make real/fake decisions using binary cross-entropy loss; *ii*) Patch Classifier [6]: This approach trains a similar classification network, but operates at the patch level instead; *iii*) Freq-Spec [49]: This technique trains a classification network on the frequency spectrum of real and fake images.

Metrics. We use the metrics of Average Precision (AP), Area Under The ROC Curve (AUROC) score, and FPR95 to evaluate the detection quality. The threshold for the detector is selected based on the fake data when 95% of fake test data points are declared as fake.

Experimental details. In detecting AI-generated content, we consider two realistic scenarios: the white-box scenario, where we have access to the target generative model, and the black-box scenario, where we do not. In the white-box

Detection method	Variant	Guided	LDM 200 steps	LDM 200 w/ CFG	Glide 100 steps	Glide 100 27	Glide 50 27	Glide 100 10	DALL-E	Average mAP
Trained deep network [42]	Blur+JPEG (0.1)	73.72	70.62	71.0	70.54	80.65	84.91	82.07	70.59	75.51
	Blur+JPEG (0.5)	68.57	66.0	66.68	65.39	73.29	78.02	76.23	65.93	70.01
	ViT:CLIP (B+J 0.5)	55.74	52.52	54.51	52.2	56.64	61.13	56.64	62.74	56.52
Patch classifier [6]	ResNet50-Layer1	70.05	87.84	84.94	88.1	74.54	76.28	75.84	**77.07**	79.33
	Xception-Block2	75.03	87.1	**86.72**	86.4	85.37	83.73	78.38	75.67	82.30
Freq-spec [49]	CycleGAN	57.72	77.72	77.25	76.47	68.58	64.58	61.92	67.77	69.00
Ours	Stable Diffusion	**92.97**	**89.40**	82.84	**90.41**	**87.75**	**86.78**	**86.75**	75.98	**86.61**

Table 1. Fake image detection results, evaluated with the Average Precision (AP) metric. Results are presented for different generative models (Guided Diffusion, LDM, Glide, and DALL-E) under varying configurations, such as sampling steps and guidance levels. The "Average mAP" column represents the mean AP across all generative model variants.

scenario, we directly perform recovery-based fake detection using the generative model that produced the fake images. In the black-box scenario, we employ a strong stable diffusion model as a surrogate in our experiments. Additional experimental details can be found in the Appendix C.

4.2. Main Results and Analysis

We begin by comparing our approach against baseline methods for identifying fake images generated by various models. In addition, we perform ablation studies to analyze the impact of different components of our approach.

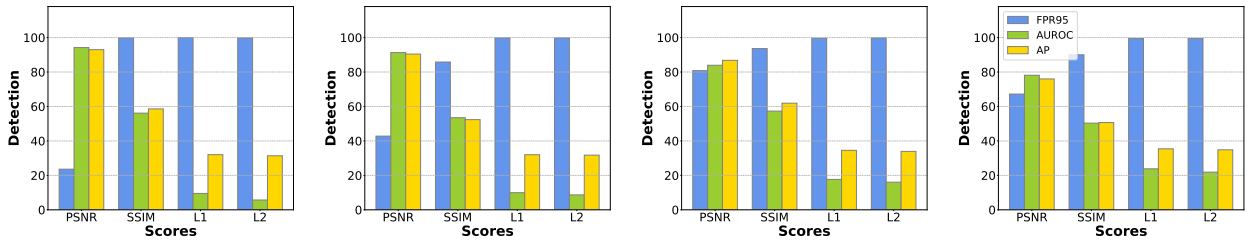
Compared against baseline methods for fake image detection. Table 1 presents the average precision (AP) of various methods for detecting AI-generated content across different generative models. While the trained classifier baseline achieves high accuracy for GAN variants [42], its performance significantly drops for modern diffusion-based generative models. This trend remains consistent even when switching the backbone from standard deep neural networks to the CLIP:ViT model, which performs slightly worse. These results suggest that detection accuracy may suffer from overfitting when using models with larger capacities. Performing classification at the patch level [6] or utilizing the frequency domain [49], does not achieve consistent detection performance. This indicates that learning patterns from small image patches alone is insufficient to address the problem. Current fake detection baselines struggle to reliably identify complex generative content.

On the other hand, our approach demonstrates significantly higher average precision in identifying fake images. By using Stable Diffusion as a surrogate model, our method outperforms baseline approaches—such as trained deep networks [42], patch classifiers [6], and frequency-based methods [49]—across a diverse range of advanced diffusion-based generative model datasets. Our method maintains a high mean Average Precision (mAP) of approximately 86.81% for fake content detection, clearly surpassing the performance of these baselines. This result highlights the effectiveness of our approach in recovering from masked images, particularly for the challenging task of black-box AI-generated content detection.

New dataset for advanced generative models DALL-E 3. Recent commercial tools (e.g., DALL-E 3) have made remarkable strides in synthesizing photorealistic images. However, DALL-E 3 is currently only accessible via APIs, and there are very few benchmarks focused on detecting fake content from this new model. Additionally, most existing AI-generated content detection benchmarks lack paired images representing real and fake distributions. Typically, these benchmarks consist of randomly selected images from different generative models, which may introduce confounding factors, leading to shortcut learning in fake detection and hindering their generalization capabilities.

In this work, we design a new benchmark specifically for DALL-E 3-based content detection. We begin by collecting a set of high-resolution, high-quality real images covering a wide range of real-world categories, referencing the LAION dataset. Next, we generate detailed captions for these real images using the LLaVA model. Based on these captions, we then generate corresponding images using DALL-E 3. This process allows us to create paired data for evaluating fake image detection, effectively eliminating confounding factors. This paired dataset provides a more challenging and robust evaluation setting, offering a better measure of detection methods' performance.

Our proposed dataset is highly challenging, with existing methods showing a significant drop in performance compared to traditional benchmarks. In summary, the newly proposed DALL-E 3 fake detection dataset is more challenging due to the model's photorealistic generation capabilities and the inclusion of paired real-fake images, which reduces unrelated factors in detection tasks. This benchmark provides a valuable resource for advancing future research in AI content detection.



(a) Guided (b) LDM (c) Glide (d) DALL-E

Figure 3. Evaluation on differen scores and evaluation metrics. The PSNR scores here show better detection accuracy in terms of FPR95, AUROC, and AUPR. This trend is consistent across different datasets, including (a) Guided, (b) LDM, (c) Glide, (d) DALL-E.

Method	Scores	FPR↓	AUROC↑	AP↑
Ours w/o fine-tuning	PSNR	47.90	87.84	86.74
	SSIM	100	45.28	44.36
Ours with fine-tuning	PSNR	23.60	94.19	92.97
	SSIM	99.80	56.13	58.60

Table 2. Experiments study on comparing the performance of the model after applying fine-tuning versus without parameter-efficient fine-tuning. The study utilized datasets tailored for guided diffusion models, with Stable Diffusion serving as the base model and testing on guided diffusion datasets.

Method	Scores	FPR↓	AUROC↑	AP↑
Guided	PSNR	10.80	97.18	96.69
	SSIM	76.70	67.25	63.81
DALL-E	PSNR	67.80	76.52	73.18
	SSIM	93.40	60.01	60.95
Glide	PSNR	77.70	85.67	87.75
	SSIM	95.50	48.34	48.25

Table 3. Evaluation was conducted on different datasets using guided diffusion as the base model. The experiments were performed without parameter-efficient fine-tuning.

4.3. Ablation Studies

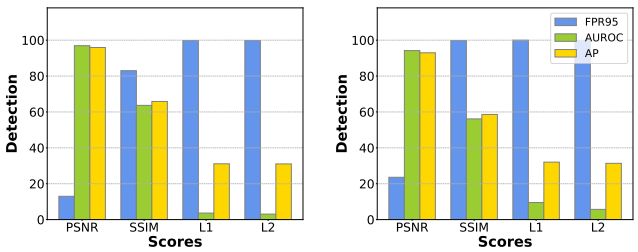
Ablations on different components. As shown in Table 2, we experiment with a variant of our recovery-based black-box detection method. We observe that the inpainting-based recovery approach, without parameter-efficient fine-tuning, results in a 47.90% FPR and an 86.74% average precision on the guided diffusion datasets. However, our approach, which leverages a surrogate model (e.g., stable diffusion) with parameter-efficient fine-tuning on a small set of examples from the target generative model (e.g., guided diffusion), significantly enhances fake detection accuracy for the unknown target model. This method achieves a 23.60% FPR and a 92.97% average precision, representing a 24.3% reduction in FPR. These results highlight the importance of parameter-efficient fine-tuning for effective black-box detection when only API access to the target model is available, without knowledge of its internal weights.

We further conduct an ablation study using guided diffusion as surrogate model, testing it on various diffusion datasets, including guided diffusion, DALL-E, and GLIDE. With appropriate scoring measures, we observe that when the guided diffusion model is tested on guided datasets, it demonstrates significantly better detection performance, achieving an FPR of 10.80%, an AUROC of 97.18%, and an average precision of 96.69%. In contrast, the detection performance on DALL-E yields an FPR of 67.80% and an AUROC of 76.75%, while on GLIDE, it shows an FPR of 77.70% and an AUROC of 85.67%. These results are notably lower compared to detecting guided diffusion images.

This indicates that diffusion models are better at distinguishing images that are closer to the distribution they are trained on, making recovery easier compared to real images. As a result, the discrepancy between fake and real images is more pronounced, leading to improved detection accuracy. This also supports the effectiveness of our recovery-based fake detection method in a white-box scenario. Specifically, if we have access to the target diffusion model, we can use recovery scores for fake detection without the need for fine-tuning, demonstrating the flexibility of our method for both white-box and black-box detection scenarios.

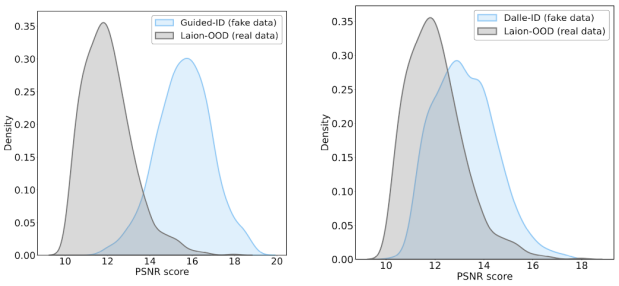
Ablations on different scores. Figure 3 and Figure 7 present ablations using different metrics—PSNR, SSIM, L1, and L2—for measuring the discrepancy between real and fake images across various datasets. We observe that PSNR demonstrates significantly better performance in detecting fake images compared to the other three metrics. For instance, when using stable diffusion as the surrogate model and guided diffusion as the target model, the AUROC achieved with the PSNR is 94.19%, whereas SSIM yields only 56.13% AUROC under the same setting—a direct improvement of 38.06%. This trend consistently holds across different surrogate models, target models, and masks.

These highlight that selecting an appropriate metric, such as PSNR, is crucial for achieving high detection per-



(a) Thick Mask (b) Genhalf Mask

Figure 4. Different evaluation scores demonstrate significant differences in fake detection performance. The PSNR scores here show better detection accuracy in terms of FPR95, AUROC, and AUPR. This trend is consistent across different types of masks used for recovery, such as (a) thick masks and (b) genhalf masks.



(a) Guided vs Laion (b) DALL-E vs Laion

Figure 5. KDE visualization without fine-tuning steps. The surrogate model is a guided diffusion model. The images generated by the guided diffusion model are better identified, as they are more aligned with the distribution of the base model.

formance, while the use of less suitable metrics can severely hinder the model's ability to identify fake images. This observation also suggests that designing more tailored metrics could further enhance recovery-based detection methods, offering promising directions for future research.

Ablations on different masks. We evaluate different variants of mask types, as shown in Figure 7, with visualizations provided in Appendix D. The choice of mask type influences the performance of recovery-based black-box detection. Notably, the genhalf mask demonstrates slightly better fake detection accuracy across all four metrics compared to thick-type masks. This improvement may be attributed to the larger masked region in the genhalf mask, which increases the recovery area used to compute the discrepancy.

Visualization and qualitative analysis. We visualize the score distributions in Figure 5 (a) and (b) for the Guided vs. Laion and DALL-E vs. Laion settings. There are two key obervations: first, the PSNR scores for fake data are consistently higher than those for real data (Laion), indicating that fake images are better recovered with higher quality using our recovery-based black-box fake detection model. Additionally, the score distributions for Guided vs. Laion show better separation compared to DALL-E vs. Laion. This suggests that our fake detection method is more effective when the generative model's distribution aligns closely with the target test model. This finding highlights the importance of parameter-efficient fine-tuning steps for detecting fake images from closed-source, advanced generative models.

4.4. Human preference evaluation.

We conducted a human preference evaluation for AI-generated content detection, focusing on comparing images generated by advanced AI models, such as DALL-E 3, to real or human-created images. Respondents were asked to carefully observe the provided images and identify which ones they believed were generated by DALL-E 3. The evaluation consisted of 100 questions, with images randomly selected from our curated DALL-E 3 fake detection dataset.

Our observations are as follows: *i*) The average accuracy of human respondents was 72.33%, indicating that DALL-E 3-generated images are challenging to distinguish even for human eyes. *ii*) Distinguishing between DALL-E 3-generated and human-created content was notably harder for categories like art paintings and cartoons compared to realistic photographs such as landscapes and human figures. This could be because people are more familiar with real-world scenes, making it easier to identify subtle inconsistencies in those contexts. *iii*) The accuracy of individual respondents ranged from 59% to 89%, highlighting that people's ability to differentiate between AI-generated and real images varies based on their backgrounds, areas of expertise, and familiarity with modern commercial AI tools.

5. Conclusion

In this work, we introduce a novel recovery-based black-box detection framework for distinguishing AI-generated content from real images. Our method leverages the differences in the recovery quality between real and synthetic images from the masked regions. By aligning the distribution of a surrogate model with that of a black-box target model through parameter-efficient fine-tuning, we achieved significant improvements in detection performance. Extensive experiments across various diffusion models demonstrate the effectiveness of our method. Notably, we observe a specific scoring measure proved superior in detecting fake images compared to other metrics. This highlights the critical role of selecting appropriate discrepancy measures in enhancing detection accuracy. Our findings underscore the need for robust, cost-efficient detection methods, particularly in scenarios where access to model internals is restricted. The success of our recovery-based strategy also opens up new avenues for developing more tailored scoring metrics and recovery techniques to address evolving capabilities of advanced generative models.

References

[1] Haoyue Bai, Gregory Canal, Xuefeng Du, Jeongyeol Kwon, Robert D Nowak, and Yixuan Li. Feed two birds with one scone: Exploiting wild data for both out-of-distribution generalization and detection. In *International Conference on Machine Learning*, pages 1454–1471. PMLR, 2023. 2

[2] Quentin Bammey. Synthbuster: Towards detection of diffusion model generated images. *IEEE Open Journal of Signal Processing*, 2023. 2

[3] James Betker, Gabriel Goh, Li Jing, Tim Brooks, Jianfeng Wang, Linjie Li, Long Ouyang, Juntang Zhuang, Joyce Lee, Yufei Guo, Wesam Manassra, Prafulla Dhariwal, Casey Chu, and Yunxin Jiao. https://openai.com/dall-e-3, 2023. 2, 12

[4] James Betker, Gabriel Goh, Li Jing, Tim Brooks, Jianfeng Wang, Linjie Li, Long Ouyang, Juntang Zhuang, Joyce Lee, Yufei Guo, et al. Improving image generation with better captions. *Computer Science. https://cdn. openai. com/papers/dall-e-3. pdf*, 2(3):8, 2023. 1, 2

[5] Andrew Brock, Jeff Donahue, and Karen Simonyan. Large scale gan training for high fidelity natural image synthesis. *arXiv preprint arXiv:1809.11096*, 2018. 2

[6] Lucy Chai, David Bau, Ser-Nam Lim, and Phillip Isola. What makes fake images detectable? understanding properties that generalize. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXVI 16*, pages 103–120. Springer, 2020. 1, 2, 5, 6

[7] Huiwen Chang, Han Zhang, Jarred Barber, AJ Maschinot, Jose Lezama, Lu Jiang, Ming-Hsuan Yang, Kevin Murphy, William T Freeman, Michael Rubinstein, et al. Muse: Text-to-image generation via masked generative transformers. *arXiv preprint arXiv:2301.00704*, 2023. 2

[8] Yunjey Choi, Minje Choi, Munyoung Kim, Jung-Woo Ha, Sunghun Kim, and Jaegul Choo. Stargan: Unified generative adversarial networks for multi-domain image-to-image translation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8789–8797, 2018. 2

[9] Riccardo Corvi, Davide Cozzolino, Giada Zingarini, Giovanni Poggi, Koki Nagano, and Luisa Verdoliva. On the detection of synthetic images generated by diffusion models. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2023. 2

[10] Davide Cozzolino, Justus Thies, Andreas Rössler, Christian Riess, Matthias Nießner, and Luisa Verdoliva. Forensictransfer: Weakly-supervised domain adaptation for forgery detection. *arXiv preprint arXiv:1812.02510*, 2018. 3

[11] Prafulla Dhariwal and Alex Nichol. Diffusion models beat gans on image synthesis, 2021. 5

[12] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in neural information processing systems*, 34:8780–8794, 2021. 2, 12

[13] Hany Farid. Lighting (in) consistency of paint by text. *arXiv preprint arXiv:2207.13744*, 2022. 3

[14] Hany Farid. Perspective (in) consistency of paint by text. *arXiv preprint arXiv:2206.14617*, 2022. 3

[15] Diego Gragnaniello, Davide Cozzolino, Francesco Marra, Giovanni Poggi, and Luisa Verdoliva. Are gan generated images easy to detect? a critical analysis of the state-of-the-art. In *2021 IEEE international conference on multimedia and expo (ICME)*, pages 1–6. IEEE, 2021. 3

[16] Mark S Graham, Walter HL Pinaya, Petru-Daniel Tudosiu, Parashkev Nachev, Sebastien Ourselin, and Jorge Cardoso. Denoising diffusion models for out-of-distribution detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2948–2957, 2023. 1, 3

[17] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 5

[18] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020. 3, 4

[19] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021. 4

[20] Kayleen Devlin and Joshua Cheetham. Fake trump arrest photos: How to spot an ai-generated image. *BBC News*, 2024. 1

[21] Mike Laszkiewicz, Jonas Ricker, Johannes Lederer, and Asja Fischer. Single-model attribution of generative models through final-layer inversion. *arXiv preprint arXiv:2306.06210*, 2023. 1

[22] Lucien Le Cam. *Asymptotic methods in statistical decision theory*. Springer Science & Business Media, 2012. 11

[23] Sara Mandelli, Nicolò Bonettini, Paolo Bestagini, and Stefano Tubaro. Detecting gan-generated images by orthogonal training of multiple cnns. In *2022 IEEE International Conference on Image Processing (ICIP)*, pages 3091–3095. IEEE, 2022. 3

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

[25] Falko Matern, Christian Riess, and Marc Stamminger. Exploiting visual artifacts to expose deepfakes and face manipulations. In *2019 IEEE Winter Applications of Computer Vision Workshops (WACVW)*, pages 83–92. IEEE, 2019. 2

[26] Scott McCloskey and Michael Albright. Detecting gan-generated imagery using color cues. *arXiv preprint arXiv:1812.08247*, 2018. 2

[27] Alex Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and Mark Chen. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. *arXiv preprint arXiv:2112.10741*, 2021. 2, 3, 5, 12

[28] Utkarsh Ojha, Yuheng Li, and Yong Jae Lee. Towards universal fake image detectors that generalize across generative models. In *Proceedings of the IEEE/CVF Conference*

on Computer Vision and Pattern Recognition, pages 24480–24489, 2023. 2, 3, 5

[29] Niki Parmar, Ashish Vaswani, Jakob Uszkoreit, Lukasz Kaiser, Noam Shazeer, Alexander Ku, and Dustin Tran. Image transformer. In *International conference on machine learning*, pages 4055–4064. PMLR, 2018. 2

[30] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. In *International conference on machine learning*, pages 8821–8831. Pmlr, 2021. 1, 5, 12

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

[32] Jonas Ricker, Simon Damm, Thorsten Holz, and Asja Fischer. Towards the detection of diffusion model deepfakes. *arXiv preprint arXiv:2210.14571*, 2022. 3

[33] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models, 2021. 12

[34] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10684–10695, 2022. 4

[35] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022. 1, 2, 3, 5

[36] Vinu Sankar Sadasivan, Aounon Kumar, Sriram Balasubramanian, Wenxiao Wang, and Soheil Feizi. Can ai-generated text be reliably detected? *arXiv preprint arXiv:2303.11156*, 2023. 11

[37] Christoph Schuhmann, Richard Vencu, Romain Beaumont, Robert Kaczmarczyk, Clayton Mullis, Aarush Katta, Theo Coombes, Jenia Jitsev, and Aran Komatsuzaki. Laion-400m: Open dataset of clip-filtered 400 million image-text pairs. *arXiv preprint arXiv:2111.02114*, 2021. 5

[38] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. Laion-5b: An open large-scale dataset for training next generation image-text models. *Advances in Neural Information Processing Systems*, 35:25278–25294, 2022. 12

[39] Zeyang Sha, Zheng Li, Ning Yu, and Yang Zhang. De-fake: Detection and attribution of fake images generated by text-to-image generation models. In *Proceedings of the 2023 ACM SIGSAC Conference on Computer and Communications Security*, pages 3418–3432, 2023. 3

[40] Gemini Team, Petko Georgiev, Ving Ian Lei, Ryan Burnell, Libin Bai, Anmol Gulati, Garrett Tanzer, Damien Vincent, Zhufeng Pan, Shibo Wang, et al. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*, 2024. 2

[41] The Guardian. Uk engineering giant arup falls victim to deepfake scam in hong kong. *The Guardian*, 2024. 1

[42] Sheng-Yu Wang, Oliver Wang, Richard Zhang, Andrew Owens, and Alexei A Efros. Cnn-generated images are surprisingly easy to spot... for now. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8695–8704, 2020. 2, 3, 5, 6

[43] Zhendong Wang, Jianmin Bao, Wengang Zhou, Weilun Wang, Hezhen Hu, Hong Chen, and Houqiang Li. Dire for diffusion-generated image detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 22445–22455, 2023. 3

[44] Zhenting Wang, Chen Chen, Yi Zeng, Lingjuan Lyu, and Shiqing Ma. Where did i come from? origin attribution of ai-generated images. *Advances in neural information processing systems*, 36, 2024. 1, 3

[45] Larry Wasserman. Lecture notes for stat 705: Advanced data analysis, 2023. 11

[46] Xianjun Yang, Wei Cheng, Yue Wu, Linda Ruth Petzold, William Yang Wang, and Haifeng Chen. Dna-gpt: Divergent n-gram analysis for training-free detection of gpt-generated text. In *The Twelfth International Conference on Learning Representations*. 2

[47] Ning Yu, Larry S Davis, and Mario Fritz. Attributing fake images to gans: Learning and analyzing gan fingerprints. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 7556–7566, 2019. 2

[48] Bowen Zhang, Shuyang Gu, Bo Zhang, Jianmin Bao, Dong Chen, Fang Wen, Yong Wang, and Baining Guo. Styleswin: Transformer-based gan for high-resolution image generation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11304–11314, 2022. 2

[49] Xu Zhang, Svebor Karaman, and Shih-Fu Chang. Detecting and simulating artifacts in gan fake images. In *2019 IEEE international workshop on information forensics and security (WIFS)*, pages 1–6. IEEE, 2019. 2, 3, 5, 6

[50] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pages 2223–2232, 2017. 2

[51] Mingjian Zhu, Hanting Chen, Qiangyu Yan, Xudong Huang, Guanyu Lin, Wei Li, Zhijun Tu, Hailin Hu, Jie Hu, and Yunhe Wang. Genimage: A million-scale benchmark for detecting ai-generated image. *Advances in Neural Information Processing Systems*, 36, 2024. 2

[52] Yuanzhi Zhu, Zhaohai Li, Tianwei Wang, Mengchao He, and Cong Yao. Conditional text image generation with diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14235–14245, 2023. 2