FakeInversion: Learning to Detect Images from Unseen Text-to-Image Models by Inverting Stable Diffusion

George Cazenavette[1]* Avneesh Sud[2] Thomas Leung[2] Ben Usman[2]

[1]Massachusetts Institute of Technology [2]Google Research

fake-inversion.github.io

Figure 1. **(left)** We propose a new synthetic image detector that uses two additional input signals derived from a fixed pre-trained Stable Diffusion [45]: an inverted latent noise map and the reconstructed input image. **(middle)** Our detector is trained using fake images generated using Stable Diffusion and real LAION images. It achieves state-of-the-art generalization performance in detecting unseen text-to-image generators. **(right)** To ensure that the performance evaluation does not favor detectors that are biased towards particular themes or styles, we introduce a new thematically and stylistically aligned evaluation benchmark – we measure detector's ability to discriminate fake images (*e.g*. DALL·E 3, Imagen) from real images with matching content and style found on the Internet using reverse image search (RIS).

Abstract

Due to the high potential for abuse of GenAI systems, the task of detecting synthetic images has recently become of great interest to the research community. Unfortunately, existing image-space detectors quickly become obsolete as new high-fidelity text-to-image models are developed at blinding speed. In this work, we propose a new synthetic image detector that uses features obtained by inverting an open-source pre-trained Stable Diffusion model. We show that these inversion features enable our detector to generalize well to unseen generators of high visual fidelity (e.g., DALL·E 3) even when the detector is trained only on lower fidelity fake images generated via Stable Diffusion. This detector achieves new state-of-the-art across multiple training and evaluation setups. Moreover, we introduce a new challenging evaluation protocol that uses reverse image search to mitigate stylistic and thematic biases in the detector evaluation. We show that the resulting evaluation scores align well with detectors' in-the-wild performance, and release these datasets as public benchmarks for future research.

*Work done during an internship at Google Research.

1. Introduction

Recent advances in text-to-image modeling have made it easier than ever to generate harmful or misrepresentative content at scale. Moreover, new versions of most photorealistic commercial models are being continuously updated and released behind closed APIs, making it harder to keep fake image detectors up to date. In this work, we make significant strides towards building a GenAI detector that can reliably identify images from unseen photorealistic text-to-image models. Specifically, we propose a model that can be trained using fake images *only* from Stable Diffusion (SD) [45] and reliably detect images generated by recent open (Kandinsky [51], Wüerstchen [39], PixArt-α [16], etc.) and closed-source text-to-image models (Imagen [46], Midjourney [2], DALL·E 3 [12], etc.) of *significantly higher* visual fidelity.

Existing methods [17, 37, 54] focus primarily on detecting traces left by convolutional generators in a way that is robust to re-compression, resizing and other in-the-wild transformations. While these methods worked well for GANs and early diffusion models, we show that they, unfortunately, fail to generalize well to current photorealistic generative models, even when re-trained using better data. Recent diffusion detectors that rely on CLIP embeddings [37] or inversions [55] fail to generalize to challenging benchmarks. Drawing inspiration from recent works that showed that GANs tend to "omit hard objects" [11] and that text-to-image models lean towards "easily captionable" images [50], in this paper, we focus on detecting fake images by analyzing internal representations of an existing off-the-shelf text-to-image model.

In this paper, we introduce a new synthetic image detection method: **FakeInversion**. Our method uses features extracted from a lower-fidelity open-source text-to-image model (Stable Diffusion [45]) to detect images generated by unseen text-to-image generators. Specifically, our model takes as input 1) the original image, 2) the approximate noise map recovered via text-conditioned DDIM [52] inversion with Stable Diffusion (SD), and 3) the reconstruction obtained by "denoising" the approximate noise map (Figure 1). We show that these additional signals significantly improve the performance of the resulting detector on unseen new proprietary and open-source photorealistic text-to-image models, attaining a new state-of-the-art. We also provide an intuitive justification for why a diffusion detector needs such features to generalize well to unseen diffusion generators.

To deploy a synthetic image detector at scale, we need to make sure that it is not relying on content signals such as the presence of specific objects or styles in the image. If left unmitigated, such bias towards particular themes or styles would disproportionately marginalize particular groups when applied to detecting healthcare misinfo [1] or forged art [3] at scale. Unfortunately, existing evaluation protocols that measure a detector's ability to differentiate between real and fake images drawn from very visually and thematically different distributions can not be used to test for the presence of such bias in the detector. For example, evaluating a fake detector using real COCO [33] images and fake images generated by DALL·E 3 [12] could favour a detector that assigns higher fakeness score to digital art, since COCO contains mostly natural images. To circumvent these issues, we propose a new evaluation protocol: **SynRIS**. For each set of synthetic images, we evaluate a given detector against a set of real images obtained by applying reverse image search (Figure 1) to given fake images – the resulting evaluation does not favor models biased towards any particular topic, theme, or style. We show that the proposed evaluation protocol is more reliable at evaluating the quality of the synthetic image detector, especially when applied to closed-source text-to-image models. We will release our evaluation benchmark (including datasets) for future research.

To summarize our contributions: 1) we introduce a new synthetic image detector that uses text-conditioned inversion maps extracted from Stable Diffusion; 2) we show that this additional feature improves the detector's ability to detect images generated by unseen text-to-image models, achieving new state-of-the-art; 3) we propose a new challenging evaluation protocol that uses reverse image search to ensure that the classifier is not biased towards any particular theme or style; 4) we verify that this evaluation protocol reliably measures detector generalization to closed text-to-image models; and 5) we release our challenging benchmark for future research.

2. Related Work

In this section we first give a brief overview of the state-of-art in image-space detectors, then we discuss how recent works attempt to detect semantic inconsistencies in generated images, and finally discuss how our evaluation protocol compares to evaluation protocols used in prior work.

Artifact Detectors. Wang et al. [54] were among the first to show that a CNN detector (CNNDet) generalized well from more powerful GANs (*e.g.* ProGAN) to less powerful ones. Soon after, Chai et al. [14] extended this idea with a fully-convolutional network that classified individual image patches, Ju et al. [27] explored fusing global and local image features, Corvi et al. [17] explored better augmentation and downsampling strategies, Zhang et al. [61] and Frank et al. [20] explored artifacts in the spectrum of GAN-generated images, and Marra et al. [35] explored GAN fingerprinting.

Generation Inconsistencies. Several works have focused on understanding the semantic properties of generated images. For example, Bau et al. [11] showed that GANs avoid generating "hard objects" such as mirrors and TVs – that both humans and discriminators fail to notice missing. Recently, Ojha et al. [37] showed that image CLIP [42] embeddings are highly predictive of whether an image is fake, and Sha et al. [50] showed that images generated using text-to-image models tend to have higher CLIP similarity to their automatically inferred captions, suggesting that images generated by text-to-image models can often be described more fully by short text captions compared to images naturally occurring on the web. Inspired by these works, we also focus on properties of images beyond their low-level convolutional traces by examining internal representations of diffusion models obtained using DDIM inversion [52]. A concurrent work [55] found that DDIM image reconstruction residuals are predictive of whether an image is fake. In this paper we justify why image-space residuals are insufficient, and empirically verify that a detector that uses internal representations of a diffusion model generalizes better. We evaluate our model against the official DIRE checkpoint and perform an ablation using only reconstruction residuals.

Evaluation Protocols. Given that internal representations of diffusion models lack the low-level features necessary to perform generator trace detection, we need a way to ensure that the learned classifier does not overfit to particular objects or styles. Unfortunately, prior works focus either on open-source models with known training sets but lower visual fidelity or use dataset pairs of real and in-the-wild fake images that are too different both in style and content

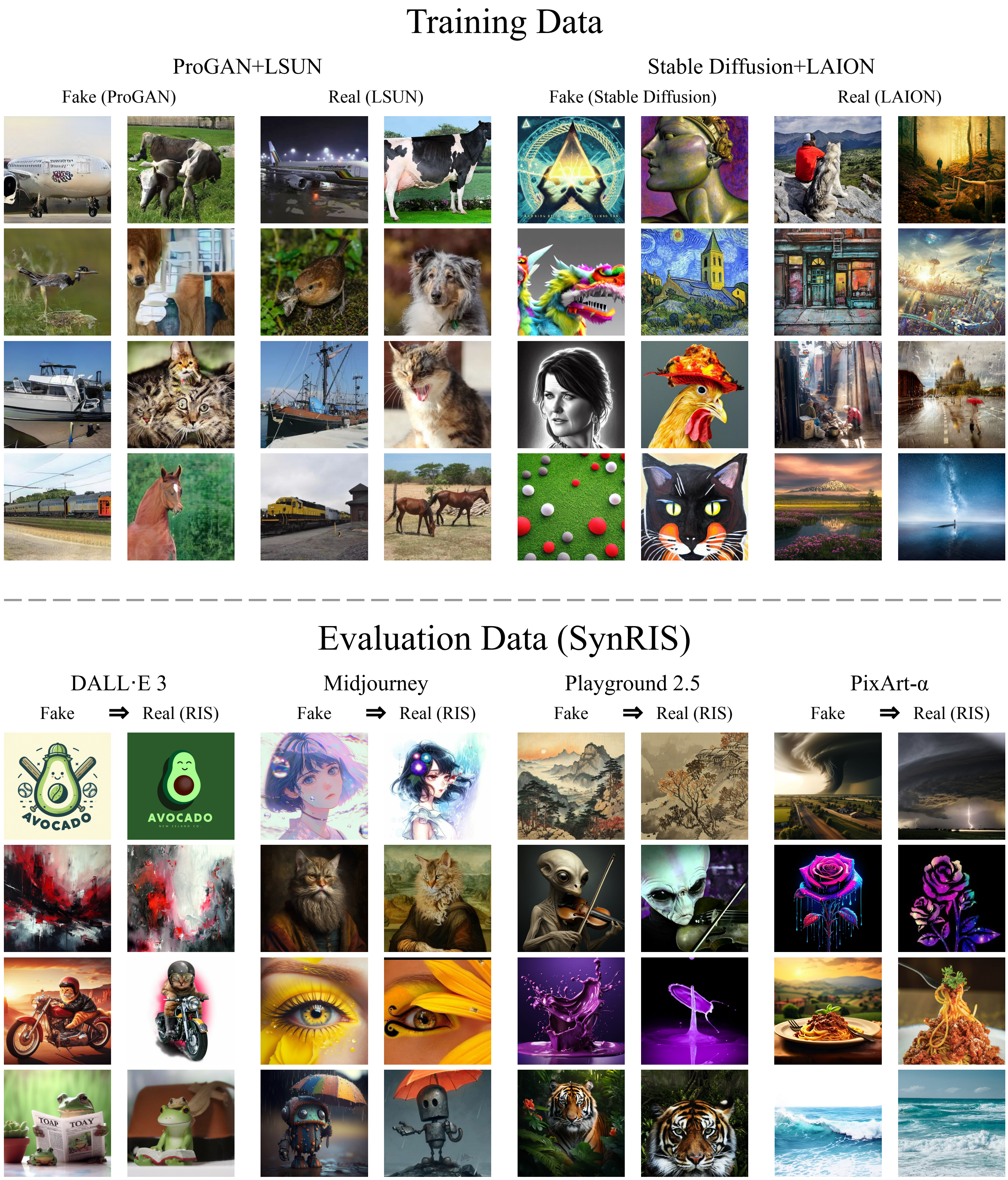

Figure 2. **Training and evaluation datasets.** We train all methods on **two training sets (top)**: ProGAN+LSUN from [54] and Stable Diffusion+LAION with fake images taken from DiffusionDB [56]. We construct a new **evaluation benchmark SynRIS (bottom)** using reverse image search (RIS) on fake images generated by both proprietary (*e.g.*, DALL·E 3 [12], Midjourney [2]) and open-source (*e.g.*, Playground [31], PixArt-α) models. Note that generators used for evaluation are of *significantly higher visual fidelity* than those used for training.

to ensure the lack of such bias. For example, recent works of Corvi et al. [17] and Ojha et al. [37] measure detectors' ability to discriminate between DALL·E 2 images and a mix of Imagenet, COCO and UCID [48] or LAION respectively, and DIRE [55] focuses only on open-source lower fidelity generators such as SD and older generators trained on ImageNet and LSUN-Bedrooms [59]. In a concurrent work, Epstein et al. [19] evaluate how adding training data from older models affects the performance of the classifier on newer fakes, which is an important problem but different from the one we address in this paper.

To summarize, we are the first to show that text-conditioned DDIM inversion feature maps extracted from one diffusion model improve the ability of a detector to identify images generated by other higher-fidelity diffusion models. Moreover, we are the first to propose an evaluation procedure for GenAI detectors that ensures that the learned detector is not biased towards any style or theme, and to quantitatively verify that the resulting evaluation is more reliable.

3. Method

In this section, we first provide a background on diffusion models and DDIM inversion. Then, we introduce our detection method that makes use of text-conditioned DDIM inversion and give an intuitive justification for why having this signal is helpful for generalization.

Latent Diffusion Models. LDMs [45] first map high-resolution (in our case, 512×512×3) RGB images x into low-resolution (64×64×4) latent images z using a pre-trained encoder $E : \mathcal{X} \to \mathcal{Z}$. The original image can be recovered almost perfectly via a pre-trained decoder $D : \mathcal{Z} \to \mathcal{X}$. In the derivation below z_* correspond to such *latent images*, rather than RGB images.

Conditional Diffusion Models and DDIM Inversion. To generate a new latent image z conditioned on some vector c, a conditional denoising diffusion model [25] starts from a random noise map z_T of the same shape and iteratively stochastically denoises it using a learned denoising network ϵ_θ for a fixed number of steps, until a clean latent image z_0 is obtained. The process of sampling from a pre-trained diffusion model can be discretized into fewer steps and made deterministic through the use of DDIM sampling [52]. Notably, this sampling procedure enables "inverting" a clean image z_0 into a corresponding noise map z_T, such that when z_T is denoised via DDIM sampling, we obtain a new latent $\hat{z}_0$ that is very close to the original z_0. Formally, to invert an image z_0, *i.e.*, to obtain a corresponding noise map z_T, we iteratively add noise to its current estimate z_t via the following *conditional forward process* starting from a clean latent image z_0:

$$z_{t+1} = \sqrt{\bar{\alpha}_{t+1}} f_\theta(z_t, t, c) + \sqrt{1 - \bar{\alpha}_{t+1}} \epsilon_\theta(z_t, t, c) \quad (1)$$

where z_t is the noisy latent at time t, vector c is the conditioner, value $\bar{\alpha}$ is the DDIM noise scaling factor [52], noise $\epsilon_\theta(z_t, t, c)$ is the prediction of the learned denoising function ϵ_θ at time t, and $f_\theta(z_t, t, c)$ is the best current estimate of the clean latent z_0:

$$f_\theta(z_t, t, c) = \frac{z_t - \sqrt{1 - \bar{\alpha}_t} \epsilon_\theta(z_t, t, c)}{\sqrt{\bar{\alpha}_t}} \quad (2)$$

A imperfect reconstruction $\hat{z}_0$ can then be obtained via the deterministic *conditional reverse process* starting from z_T:

$$\hat{z}_{t-1} = \sqrt{\bar{\alpha}_{t-1}} f_\theta(\hat{z}_t, t, c) + \sqrt{1 - \bar{\alpha}_{t-1}} \epsilon_\theta(\hat{z}_t, t, c). \quad (3)$$

We will refer to such full forward and reverse mapping as:

$$\hat{z}_T = F_\theta(z_0, c), \quad \hat{z}_0 = R_\theta(\hat{z}_T, c). \quad (4)$$

Text Conditioning. In our case, the conditioning vector c used to modulate the forward and reverse sampling processes is the embedding of the text prompt describing an image. In this work, we use an off-the-shelf captioner (BLIP 2 [32]) to obtain a text prompt describing an input image, and CLIP [42] to embed this text. Prior work showed that the realism of generated images can be improved through the use of classifier-free guidance [24]. Later, Mokady et al. [36] showed that classifier-free guidance leads to instability in DDIM inversion and proposed a mitigation strategy through fine-tuning parts of the model. Since we cannot afford fine-tuning on each incoming image, in this paper we do not use classifier-free guidance during inversion and sampling and use the original conditional update rules described above.

GenAI Detector. As shown in Figure 3, given an input image x, we first caption that image using BLIP [32] and embed that caption into a vector c using CLIP [42]. Then we compute the corresponding latent image $z_0 = E(x)$ using a pre-trained encoder and obtain a latent DDIM noise map $\hat{z}_T$ using text-conditioned DDIM inversion with a pre-trained diffusion model. Then, we obtain a reconstructed latent image $\hat{z}_0$ using text-conditioned DDIM sampling and decode both the latent noise map $\hat{z}_T$ and the reconstruction $\hat{z}_0$ to image space using the decoder D. Finally, we apply a learned mapping $h_\phi : \mathcal{X}^3 \to \mathbb{R}$ to these "images" to get a prediction logit. We learn the parameters of this function ϕ via backpropagation of the binary cross-entropy loss $\ell(\hat{y}, y)$ on the training set of known fake and real images $\{(x, y)\}$:

$$c = \text{CLIP}(\text{BLIP}(x)) \quad (5)$$

$$\hat{z}_T = F_\theta(E(x), c), \quad \hat{z}_0 = R_\theta(\hat{z}_T, c) \quad (6)$$

$$\phi^* = \underset{\phi}{\operatorname{argmin}}\, \mathbb{E}_{x,y}[\, \ell\left(h_\phi\left(x, D(\hat{z}_T), D(\hat{z}_0)\right), y\right)] \quad (7)$$

Intuition. But why would a diffusion detector benefit from having access to DDIM inversion of an image if it already has access to the image itself? Recent works [34, 47] showed that DDIM can be viewed as a first-order discretization of a neural probability-flow ODE. The bijection between observations z_0 and noise maps z_t induced by this ODE can be used to evaluate the likelihood of the data via the change of variable. If we view the forward DDIM mapping F_θ as an approximation of that true bijective mapping between

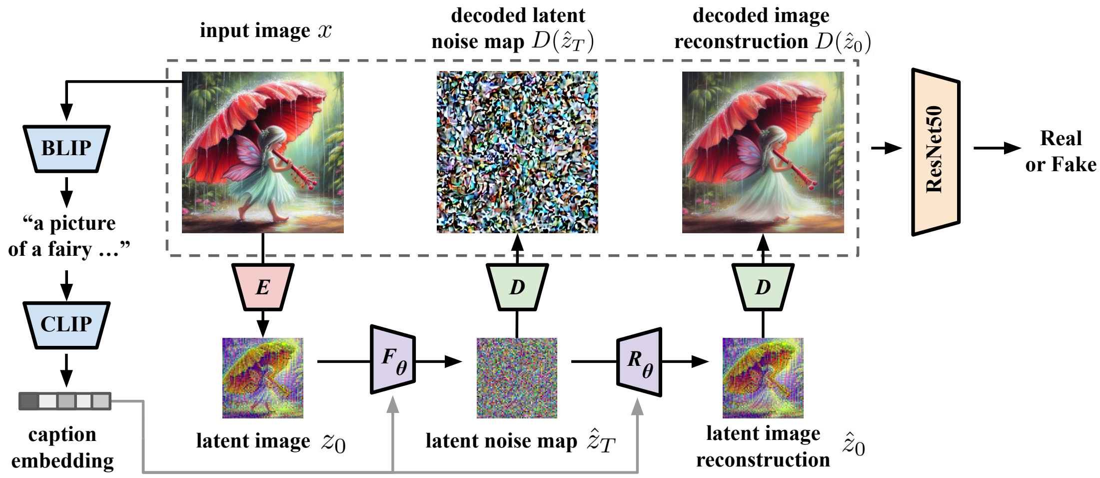

Figure 3. **Proposed method.** In addition to the original image itself (x), we also train our detector using the (decoded) noise map $D(\hat{z}_T)$ and reconstruction $D(\hat{z}_0)$ obtained by inverting the image through Stable Diffusion using DDIM. The original image is first mapped to the latent space with Stable Diffusion's **VAE Encoder**. The latent image is then inverted and reconstructed through Stable Diffusion's **U-Net** using DDIM while conditioned on the **CLIP** embedding of the image's predicted **BLIP** caption. The latent noise map and reconstruction are mapped back to image space using Stable Diffusion's **VAE Decoder**. The original image, decoded noise map, and decoded reconstruction are then concatenated and used as input for our **ResNet Classifier**.

the z_0 and z_T that introduces a discretization error δ into the inverted noise maps $\hat{z}_T$, causing the resampled image $\hat{z}_0$ to deviate from the original image z_0, it can be shown (see App. B.1) that, in the first-order approximation, the log-likelihood of the data given that underlying model can be estimated from the input image z_0, its imperfect reconstruction $\hat{z}_0$, and the noise map z_T alone:

$$\log p(z_0) \propto \log p_z(z_T) - \langle \delta, \hat{z}_0 - z_0 \rangle / \|\delta\|^2 \qquad (8)$$

Notably, the expression above does not explicitly depend on the parameters θ other than through these three signals. Given that, under model misspecification, likelihood-based models tend to overgeneralize [26], producing samples that are unlikely under the true data distributions, and assuming that the class of diffusion-based models is similar enough to overgeneralize in similar ways, we propose that a model that has access to all signals required to internally perform some form of likelihood testing on input data against a particular text-to-image model (the image, its imperfect reconstruction, and the intermediate noise map) would generalize better to detect images generated via other diffusion models. Text conditioning c further amplifies differences between log-likelihoods of distributions of fake and real images, making the corresponding test more powerful and consequently making inversions even more useful for detecting fake images. To sum up, GenAI detectors find discrepancies between the real data distribution and the approximation learned by the GenAI model. The equation above shows that proposed features enable a detector to get a rough estimate of whether a given image is of high probability under the approximate distribution learned by Stable Diffusion. We empirically verify that this "SD-likelihood" signal helps in detecting other generators.

4. Experiments

In this section, we discuss our training and evaluation sets, which baseline methods we use to compare, and the details of how we train our classifier.

Baselines. We compare our model to a representative set of the most recent state-of-art baselines (all published in 2023) that open-sourced their training or inference code. DMDet [17] is a state-of-art RGB-only method that achieved significant generalization performance through the use of augmentations and a modified down-sampling strategy; authors released only the inference checkpoint. UFD [37] is another recent state-of-the-art method that trains a linear classification head on top of the CLIP [42] embeddings of real and fake images. We use the official checkpoint and also retrain it from scratch on each of our training sets using the official code. DIRE [55] is a concurrent work that showed that using image-space DDIM reconstruction residuals helps detection. The official checkpoint open-sourced by the authors has an issue causing its performance to be much lower than the performance reported in the paper; we discuss this in more detail in App. E.1. We also provide an ablation of our method that uses only DDIM residuals. This serves as a close approximation of what a DIRE-like method *could* achieve. We also include an older convolutional baseline CNNDet [54] using its official checkpoint and code to retrain on our data.

Training data – ProGAN+LSUN. Most prior works use the ProGAN training set introduced in CNNDet [54]. This training set consists of 350k images from class-conditioned pre-trained ProGAN [30] combined with a set of real images

Dataset	Data Size	Real Data	Real/Fake Source
DALL·E 2 [44]	700/700	RIS (ours)	fakes from [17]
DALL·E 3 [12]	3.3k/3.3k	RIS (ours)	fakes from [8]
Midjourney [2]	4.4k/4.4k	RIS (ours)	fakes from [9]
Imagen [46]	700/700	RIS (ours)	our fakes (see App. A)
Open-Source[1]	3.5k/3.5k (x11)	RIS (ours)	our fakes (see App. A)
DALL·E 2 (A)	1k/5k	Imagenet, COCO, UCID [48]	both reals/fakes from [17]
Craiyon (A)	1k/1k	LAION	both from [37]
LDM (A)	1k/1k	LAION	both from [37]

Table 1. **Evaluation Datasets.** We evaluate using 15 new RIS-based evaluation benchmarks (**SynRIS (ours)** – top) as well as existing academic (A) text-to-image evaluation dataset (bottom).

Eval Data		Training Data		FID	KID $\times 10^{-2}$
Fake	Real	ProGAN+LSUN	SD+LAION		
DALL·E 2	LAION	0.233	0.043	163.5	2.7
	RIS (ours)	**0.457**	**0.406**	**88.5**	**0.3**
DALL·E 3	LAION	0.794	0.360	126.1	2.6
	RIS (ours)	**0.920**	**0.795**	**93.6**	**0.4**
Imagen	LAION	0.406	0.360	127.1	2.1
	WebLI	0.559	0.664	101.9	1.2
	RIS (ours)	**0.620**	**0.720**	**83.6**	**0.2**
Kandinsky 2	LAION	0.655	0.189	118.0	2.1
	RIS (ours)	**0.857**	**0.686**	**88.9**	**0.4**
SDXL	LAION	0.689	0.106	108.7	1.6
	RIS (ours)	**0.874**	**0.551**	**93.0**	**0.4**

Table 2. **Difficulty of RIS vs LAION eval.** FPR@0.8 recall for the state-of-the-art detector (UFD [37]) evaluated using fakes from respective generators and real images from LAION, reverse image search (RIS), and WebLI (Imagen's training set [60]) - higher FPR is harder; FID and KID between reals and fakes - lower is closer.

from LSUN [59]. Training on these images has yielded good results in the detection of GAN-generated images [17, 37, 54], and we find that this set continues to show promise when applied to images from newer diffusion models.

Training data – Stable Diffusion+LAION. Similar to the concurrent work of Epstein et al. [19], we first train detectors using 300k fake Stable Diffusion v1 images from DiffusionDB [56] and 300k random real LAION [49] images. While state-of-the-art at the time of its release, Stable Diffusion (v1) has since been eclipsed in quality by many new text-to-image models. We find that training on fake images from this relatively old diffusion model still yields models capable of identifying fakes from much newer and more powerful generators.

Evaluation data (fakes). We obtain several thousand images generated by closed-source photorealistic text-to-image models using APIs (Imagen [46]), using existing databases of fakes on HuggingFace (Midjourney [9], DALL·E 3 [8]) and by taking fake images from prior academic benchmarks (DALL·E 2 from [17]). We also generate several thousand images using high-fidelity open-source text-to-image models[1] conditioned on Midjourney prompts [57].

Evaluation data (reals) – Reverse Image Search (RIS). To ensure that detectors are not biased toward any particular theme or style, we need sets of real and fake images that are themselves stylistically and thematically aligned. We address this issue using a reverse image search API to find a visually and thematically similar real image for each fake image from the eval fake set defined above. Examples of images found using this procedure can be found in Figure 2. We define real images as images not generated using a text-to-image model, even if other tools (such as Photoshop) were used. To ensure that our real images are not contaminated with similar images generated by text-to-image models, we include only matches found on pages created before January 1, 2021. As a result, our real sets include only images published before the original DALL·E [43] was announced. The exact sizes of all evaluation and training sets can be found in Table 1.

Evaluation data – prior academic benchmarks. We evaluate competing methods on academic text-to-image benchmarks from published prior work [17, 37] that evaluate methods' abilities to differentiate between a set of fakes from a text-to-image model (*e.g.*, DALL·E 2) and an unrelated set of real images (*e.g.*, Imagenet, COCO). Consequently, these benchmarks can not be used to test whether a detector focuses on the styles and themes of a particular generator.

Detector architecture. We use ResNet50 trained from scratch as a detector backbone. In each experiment, we select the best checkpoint via validation on the held-out set sampled from the same source as the training set. We augment each image via a suite of random transforms *before* performing DDIM inversion: flip, crop, color jitter, grayscale, cutout, noise, blur, jpeg, and rotate. We use BLIP [32] to compute image captions. See App. C for more details.

Metrics. We report detection AUCROC as the main metric. We also provide tables with average precision and accuracy, along with PR, ROC and DET curves in the appendix. To ensure that trained and evaluated detectors can not exploit differences in image resolutions and aspect ratios, each image is resized to 256px along the shortest side and saved losslessly.

[1] Open-Source dataset includes fake images from Kandinsky 2 [51], Kandinsky 3 [10], PixArt-α [16], SDXL-DPO [53], SDXL [41], SegMoE [58], SSD-1B [21], Stable-Cascade [39], Segmind-Vega [21], and Würstchen 2 [39].

Train Data	ProGAN + LSUN					Stable Diffusion + LAION			
Eval Set | Model	DIRE	CNNDet	DMDet	UFD	Ours	CNNDet†	DMDet*	UFD†	Ours
DALL·E 2 [44]	0.561	0.455	0.656	0.728	**0.854**	0.680	0.672	**0.776**	0.747
DALL·E 3 [12]	0.524	0.378	0.409	0.323	**0.642**	0.716	0.415	0.480	**0.759**
Midjourney v5/6 [2]	0.538	0.473	0.544	0.397	**0.750**	0.630	0.484	0.592	**0.664**
Imagen [46]	0.562	0.452	0.502	0.637	**0.776**	0.714	0.573	0.575	**0.807**
Kandinsky 2 [51]	0.463	0.492	0.468	0.474	**0.758**	0.600	0.478	0.562	**0.699**
Kandinsky 3 [10]	0.491	0.480	0.593	0.469	**0.845**	0.659	0.614	0.637	**0.743**
PixArt-α [16]	0.478	0.487	0.599	0.506	**0.854**	0.627	0.580	0.647	**0.730**
Playground 2.5 [31]	0.453	0.528	0.661	0.466	**0.778**	0.582	0.517	0.587	**0.625**
SDXL-DPO [53]	0.458	0.486	0.603	0.464	**0.841**	0.843	0.563	0.702	**0.881**
SDXL [41]	0.459	0.525	0.667	0.464	**0.764**	**0.814**	0.568	0.663	0.807
Seg-MOE [58]	0.459	0.429	0.467	0.401	**0.796**	0.663	0.476	0.620	**0.713**
SSD-1B [21]	0.449	0.589	0.689	0.515	**0.827**	0.726	0.556	0.628	**0.794**
Stable-Cascade [39]	0.465	0.447	0.603	0.341	**0.882**	0.705	0.565	0.682	**0.749**
Segmind Vega [21]	0.471	0.556	0.645	0.468	**0.823**	0.742	0.540	0.623	**0.811**
Würstchen 2 [39]	0.456	0.510	0.671	0.616	**0.792**	0.610	0.675	0.697	**0.705**
DALL·E 2 [44] (A)	0.554	0.466	0.646	**0.662**	0.623	0.566	**0.727**	0.590	0.571
Craiyon [18] (A)	0.523	0.660	0.941	**0.974**	0.874	0.763	**0.988**	0.918	0.886
LDM [45] (A)	0.512	0.653	0.854	**0.924**	0.878	0.913	**1.000**	0.919	0.979
Average	0.493	0.504	0.623	0.546	**0.798**	0.697	0.611	0.661	**0.759**

Table 3. **Main Results – Detector AUCROC.** Detectors trained on ProGAN+LSUN and SD+LAION are evaluated using proprietary (first panel) and open-source (second panel) generators, and academic (A) benchmarks from prior work (last panel). *Note: This DMDet classifier was trained with fakes from an LDM checkpoint rather than Stable Diffusion. †These models were re-trained by us.

5. Results

In this section we discuss following key findings: 1) the proposed thematically and stylistically aligned RIS-based evaluation protocol is harder and is more reliable then protocols used in prior work; 2) the proposed detector outperforms prior work on both prior academic and this new RIS-based evaluation; 3) the DDIM inversion features were crucial in achieving high generalization in all cases.

RIS-based evaluation is harder and more reliable. Table 2 compares False Positive Rate (FPR) of the state-of-the-art detector [37] on different evaluation sets at the threshold that attains 80% recall (fake images are the same, so the threshold at given recall is the same as well). Results show that LAION-based evaluation significantly underestimates the false positive rate of the detector when evaluating its ability to discriminate fakes from closed-source text-to-image models (Imagen, DALL·E 2/3) across both training sets. We also obtained real examples from the multimodal dataset used to train Imagen (WebLI [60]), and evaluated the detector against these real examples and these results closely align with our RIS-based eval (see Fig. A.1 for PR curves). The FID [23] and KID [13] between real and fake images is also lower for RIS eval, and matches FID/KID between WebLI and Imagen fakes, suggesting better stylistic and thematic alignment. Similar trends can be seen on open-source models (Kadnisky, SDXL) and across both training sets. These results suggest that our RIS-based eval is a more reliable way to estimate a model's ability to detect images from closed-sourced text-to-image models trained on unknown data.

FakeInversion achieves state-of-the-art performance. Table 3 shows that our method consistently scores best at detecting both closed and open-source methods across various training sets. It also matches the performance of prior work on academic benchmarks. On average, our method outperforms prior work by **at least 4pp** on both training sets.

Inversions are crucial for generalization. To ensure that the observed gains are not coming from a particular choice of hyperparameters in our detector, we performed an ablation training the exact same network using only RGB images and only absolute DDIM image reconstruction residuals $\mathrm{Res} = |x - D(\hat{z}_0)|$ (similar to DIRE [55]). Table 4 shows that both RGB and reconstruction residual-based models perform significantly worse than the proposed method that uses both the input image, its reconstruction, and the inversion map, confirming all three are essential to achieve state-of-art generalization to unseen detectors. In the appendix we show

Train Data	ProGAN + LSUN			SD + LAION		
Eval Set | Model	RGB	Res	Ours	RGB	Res	Ours
DALL·E 2 [43]	0.410	0.650	**0.854**	0.592	0.650	**0.747**
DALL·E 3 [12]	0.399	**0.672**	0.642	0.676	0.672	**0.759**
Midjourney v5 [2]	0.434	0.590	**0.750**	0.530	0.590	**0.664**
Imagen [46]	0.530	0.670	**0.776**	0.729	0.670	**0.807**
Kandinsky 2 [51]	0.462	0.600	**0.716**	0.614	0.607	**0.714**
Kandinsky 3 [10]	0.434	0.617	**0.824**	0.606	0.679	**0.774**
PixArt-α [16]	0.470	0.604	**0.647**	0.594	0.570	**0.707**
Playground 2.5 [31]	0.439	0.604	**0.726**	0.510	0.533	**0.660**
SDXL-DPO [53]	0.338	0.643	**0.704**	0.738	0.711	**0.837**
SDXL [41]	0.410	0.612	**0.691**	0.784	0.709	**0.884**
Seg-MOE [58]	0.416	0.585	**0.799**	0.607	0.611	**0.781**
SSD-1B [21]	0.494	0.672	**0.775**	0.690	0.648	**0.813**
Stable-Cascade [39]	0.448	0.674	**0.743**	0.557	0.686	**0.766**
Segmind Vega [21]	0.465	0.677	**0.781**	0.683	0.631	**0.829**
Würstchen 2 [39]	0.563	0.624	**0.664**	0.588	0.605	**0.702**

Table 4. **Input Signal Ablation – AUCROC**. Detectors trained on ProGAN+LSUN and Stable Diffusion+LAION and evaluated on proprietary and open generators. Using the original images, inversion maps, and reconstructions (**Ours**) yields better performance than RGB or DDIM Residuals (as in DIRE [55]) alone.

		ProGAN + LSUN			SD + LAION		
		CNNDet	UFD	Ours	CNNDet	UFD	Ours
noise	Imagen	0.477	0.579	**0.758**	0.730	0.529	**0.822**
	MJ	0.481	0.383	**0.665**	0.600	0.533	**0.624**
	DALL·E 3	0.390	0.315	**0.598**	0.700	0.449	**0.750**
blur	Imagen	0.447	0.595	**0.793**	0.730	0.570	**0.812**
	MJ	0.463	0.379	**0.747**	0.639	0.583	**0.658**
	DALL·E 3	0.375	0.315	**0.639**	0.738	0.501	**0.756**
JPEG	Imagen	0.463	0.651	**0.769**	0.715	0.555	**0.804**
	MJ	0.466	0.383	**0.743**	0.624	0.610	**0.654**
	DALL·E 3	0.372	0.327	**0.632**	0.713	0.477	**0.754**
crop	Imagen	0.436	0.561	**0.781**	0.704	0.546	**0.797**
	MJ	0.471	0.383	**0.742**	0.623	0.597	**0.680**
	DALL·E 3	0.375	0.298	**0.642**	0.702	0.457	**0.769**

Table 5. **Performance on Corrupted Images – AUCROC**. Our method remains robust to common image degredations [20].

that text conditioning also helps generalization.

Robustness and Interpretability. Table 5 shows that our method is sufficiently robust to in-the-wild transformations [20] such as JPEG re-compression and blur. Figure 4 shows that a model that uses inversion maps not only generalizes better but also focuses more on features that humans recognize as GenAI artifacts (*e.g.* malformed hands).

Discussion. Our results suggest that, while both our method and recent methods (UFD [37], DMDet [17]) consistently outperform older prior methods (CNNDet [54]) on prior academic benchmarks, recent methods struggle to maintain

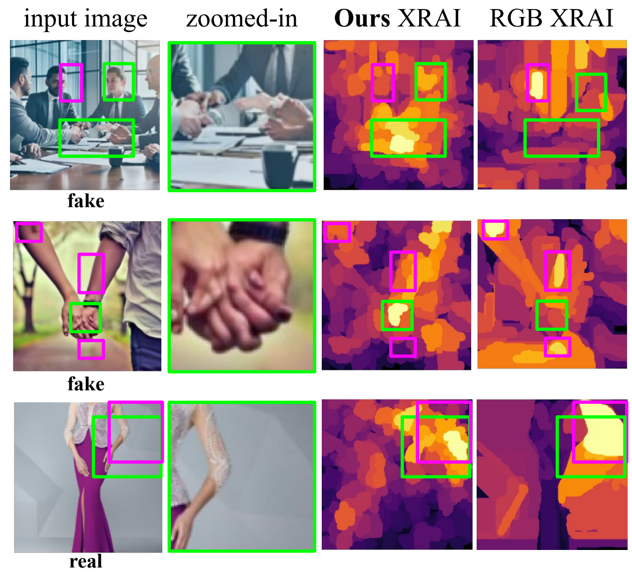

Figure 4. **Saliency Analysis**. Green boxes highlight the most salient regions according to our model and purple boxes for an equivalent RGB-only model. We use a post-hoc explainability technique, XRAI [29]. The regions of anatomical inconsistencies in fakes are most salient in our model.

the same level of exceptional performance when evaluated against our new RIS-based evaluation benchmark, even when retrained on better data. Our method and some of the older baselines, on the other hand, perform well on both. We attribute this discrepancy to the drastic shift between real and fake images used in prior evaluation – suggesting that some of the recent methods were in part overfitting to the distributions of styles and content of natural images, which appears to be less of an issue for our method.

6. Conclusion

In this paper, we introduce **FakeInversion**: a GenAI detection method that uses text-conditioned inversion maps extracted from a pre-trained Stable Diffusion to achieve a new state-of-the-art at detecting images generated via unseen text-to-image diffusion models. We also propose **SynRIS**: a new challenging evaluation protocol that uses reverse image search to ensure that the evaluation is not biased towards any styles and themes. We show that the new protocol is also more reliable at evaluating detectors on images generated using proprietary models trained on unknown data. While FakeInversion improves upon the state-of-the-art on this challenging benchmark, there clearly remains **much work to be done**; the detection performance on the new evaluation benchmark is far from saturated. We invite future researchers to use these new datasets to explore, build, and deploy better GenAI detectors at scale, with confidence that their solutions will not favor any content and style.

References

[1] Deepfakes could supercharge health care's misinformation problem. https://www.axios.com/2023/11/14/ai-deepfake-health-misinformation-fake-pictures-videos. 2

[2] Midjourney. https://www.midjourney.com/. 1, 3, 6, 7, 8, 15, 16, 25

[3] An a.i.-generated picture won an art prize. artists aren't happy. https://www.nytimes.com/2022/09/02/technology/ai-artificial-intelligence-artists.html. 2

[4] Sd pokémon diffusers. https://huggingface.co/lambdalabs/sd-pokemon-diffusers. 17, 39

[5] Pokémon. pokemon.com. 39

[6] Danbooru 2022 dataset. https://huggingface.co/datasets/animelover/danbooru2022, 2022. 17, 40

[7] Animagine XL 2.0. https://huggingface.co/Linaqruf/animagine-xl-2.0, 2023. 17, 40

[8] laion/dalle-3-dataset · datasets at hugging face, 2023. 6

[9] wanng/midjourney-v5-202304-clean · datasets at hugging face, 2023. 6

[10] Vladimir Arkhipkin, Andrei Filatov, Viacheslav Vasilev, Anastasia Maltseva, Said Azizov, Igor Pavlov, Julia Agafonova, Andrey Kuznetsov, and Denis Dimitrov. Kandinsky 3.0 technical report, 2023. 6, 7, 8, 13, 15, 16, 29

[11] David Bau, Jun-Yan Zhu, Jonas Wulff, William Peebles, Hendrik Strobelt, Bolei Zhou, and Antonio Torralba. Seeing what a gan cannot generate. In *ICCV*, 2019. 2

[12] 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. Improving image generation with better captions. https://cdn.openai.com/papers/dall-e-3.pdf. 1, 2, 3, 6, 7, 8, 15, 16, 25, 27

[13] Mikołaj Bińkowski, Danica J Sutherland, Michael Arbel, and Arthur Gretton. Demystifying mmd gans. *arXiv preprint arXiv:1801.01401*, 2018. 7

[14] Lucy Chai, David Bau, Ser-Nam Lim, and Phillip Isola. What makes fake images detectable? understanding properties that generalize. In *ECCV*, 2020. 2

[15] 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. 12

[16] Junsong Chen, Jincheng Yu, Chongjian Ge, Lewei Yao, Enze Xie, Yue Wu, Zhongdao Wang, James Kwok, Ping Luo, Huchuan Lu, et al. Pixart-alpha: Fast training of diffusion transformer for photorealistic text-to-image synthesis. *arXiv preprint arXiv:2310.00426*, 2023. 1, 6, 7, 8, 13, 15, 16, 30

[17] 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. 1, 2, 4, 5, 6, 8, 12, 26

[18] Boris Dayma, Suraj Patil, Pedro Cuenca, Khalid Saifullah, Tanishq Abraham, Phúc Lê Khac, Luke Melas, and Ritobrata Ghosh. Dall-e mini, 2021. 7, 15, 16

[19] David C. Epstein, Ishan Jain, Oliver Wang, and Richard Zhang. Online detection of ai-generated images. In *ICCV Workshop*, 2023. 4, 6

[20] Joel Frank, Thorsten Eisenhofer, Lea Schönherr, Asja Fischer, Dorothea Kolossa, and Thorsten Holz. Leveraging frequency analysis for deep fake image recognition. In *ICML*, 2020. 2, 8

[21] Yatharth Gupta, Vishnu V. Jaddipal, Harish Prabhala, Sayak Paul, and Patrick Von Platen. Progressive knowledge distillation of stable diffusion xl using layer level loss, 2024. 6, 7, 8, 13, 15, 16, 35, 37

[22] 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. 14

[23] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *NeurIPS*, 2017. 7

[24] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022. 4

[25] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *NeurIPS*, 2020. 4

[26] Ferenc Huszár. How (not) to train your generative model: Scheduled sampling, likelihood, adversary? *arXiv preprint arXiv:1511.05101*, 2015. 5

[27] Yan Ju, Shan Jia, Lipeng Ke, Hongfei Xue, Koki Nagano, and Siwei Lyu. Fusing global and local features for generalized ai-synthesized image detection. In *ICIP*, 2022. 2

[28] Minguk Kang, Jun-Yan Zhu, Richard Zhang, Jaesik Park, Eli Shechtman, Sylvain Paris, and Taesung Park. Scaling up gans for text-to-image synthesis. In *CVPR*, 2023. 12

[29] Andrei Kapishnikov, Tolga Bolukbasi, Fernanda Viégas, and Michael Terry. Xrai: Better attributions through regions. In *ICCV*, 2019. 8

[30] Tero Karras, Timo Aila, Samuli Laine, and Jaakko

Lehtinen. Progressive growing of gans for improved quality, stability, and variation. *ICML*, abs/1710.10196, 2017. 5, 12

[31] Daiqing Li, Aleks Kamko, Ehsan Akhgari, Ali Sabet, Linmiao Xu, and Suhail Doshi. Playground v2.5: Three insights towards enhancing aesthetic quality in text-to-image generation, 2024. 3, 7, 8, 13, 15, 16, 31

[32] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *arXiv preprint arXiv:2301.12597*, 2023. 4, 6, 14, 17, 39

[33] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *ECCV*, 2014. 2

[34] Cheng Lu, Yuhao Zhou, Fan Bao, Jianfei Chen, Chongxuan Li, and Jun Zhu. Dpm-solver: A fast ode solver for diffusion probabilistic model sampling in around 10 steps. *NeurIPS*, 2022. 4

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

[36] Ron Mokady, Amir Hertz, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Null-text inversion for editing real images using guided diffusion models. In *CVPR*, 2023. 4

[37] Utkarsh Ojha, Yuheng Li, and Yong Jae Lee. Towards universal fake image detectors that generalize across generative models. In *CVPR*, 2023. 1, 2, 4, 5, 6, 7, 8, 15, 16

[38] Gaurav Parmar, Krishna Kumar Singh, Richard Zhang, Yijun Li, Jingwan Lu, and Jun-Yan Zhu. Zero-shot image-to-image translation. In *ACM TOG*, 2023. 14

[39] Pablo Pernias, Dominic Rampas, Mats L. Richter, Christopher Pal, and Marc Aubreville. Wuerstchen: Efficient pretraining of text-to-image models, 2023. 1, 6, 7, 8, 13, 15, 16, 36, 38

[40] Justin N. M. Pinkney. Pokemon blip captions. `https://huggingface.co/datasets/lambdalabs/pokemon-blip-captions/`, 2022. 17, 39

[41] Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. Sdxl: Improving latent diffusion models for high-resolution image synthesis. *arXiv preprint arXiv:2307.01952*, 2023. 6, 7, 8, 13, 15, 16, 33

[42] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *ICML*, 2021. 2, 4, 5

[43] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. In *ICML*, 2021. 6, 8

[44] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 2022. 6, 7, 15, 16, 26

[45] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, 2022. 1, 2, 4, 7, 14, 15, 16

[46] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. *NeurIPS*, 2022. 1, 6, 7, 8, 15, 16, 24

[47] Tim Salimans and Jonathan Ho. Progressive distillation for fast sampling of diffusion models. In *ICLR*, 2022. 4

[48] Gerald Schaefer and Michal Stich. Ucid: An uncompressed color image database. In *Storage and retrieval methods and applications for multimedia 2004*. SPIE, 2003. 4, 6

[49] 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. *NeurIPS*, 2022. 6, 12, 14

[50] Zeyang Sha, Zheng Li, Ning Yu, and Yang Zhang. Defake: Detection and attribution of fake images generated by text-to-image generation models, 2023. 2

[51] Arseniy Shakhmatov, Anton Razzhigaev, Aleksandr Nikolich, Vladimir Arkhipkin, Igor Pavlov, Andrey Kuznetsov, and Denis Dimitrov. kandinsky 2.2, 2023. 1, 6, 7, 8, 12, 15, 16, 28

[52] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *ICLR*, 2020. 2, 4

[53] Bram Wallace, Meihua Dang, Rafael Rafailov, Linqi Zhou, Aaron Lou, Senthil Purushwalkam, Stefano Ermon, Caiming Xiong, Shafiq Joty, and Nikhil Naik. Diffusion model alignment using direct preference optimization. *arXiv preprint arXiv:2311.12908*, 2023. 6, 7, 8, 13, 15, 16, 32

[54] Sheng-Yu Wang, Oliver Wang, Richard Zhang, Andrew Owens, and Alexei A Efros. Cnn-generated images are surprisingly easy to spot...for now. In *CVPR*, 2020. 1, 2, 3, 5, 6, 8, 12, 14, 15, 16

[55] Zhendong Wang, Jianmin Bao, Wengang Zhou, Weilun

Wang, Hezhen Hu, Hong Chen, and Houqiang Li. Dire for diffusion-generated image detection. *ICCV*, 2023. 1, 2, 4, 5, 7, 8, 15

[56] Zijie J Wang, Evan Montoya, David Munechika, Haoyang Yang, Benjamin Hoover, and Duen Horng Chau. Diffusiondb: A large-scale prompt gallery dataset for text-to-image generative models. *arXiv preprint arXiv:2210.14896*, 2022. 3, 6, 12

[57] wangjunjie. midjourney-v5-202304-clean. https://huggingface.co/datasets/wanng/midjourney-v5-202304-clean, 2023. 6

[58] Harish Prabhala Yatharth Gupta, Vishnu V Jaddipal. Segmoe: Segmind mixture of diffusion experts. https://github.com/segmind/segmoe, 2024. 6, 7, 8, 13, 15, 16, 34

[59] Fisher Yu, Ari Seff, Yinda Zhang, Shuran Song, Thomas Funkhouser, and Jianxiong Xiao. Lsun: Construction of a large-scale image dataset using deep learning with humans in the loop. *arXiv preprint arXiv:1506.03365*, 2015. 4, 6, 12

[60] Jiahui Yu, Yuanzhong Xu, Jing Yu Koh, Thang Luong, Gunjan Baid, Zirui Wang, Vijay Vasudevan, Alexander Ku, Yinfei Yang, Burcu Karagol Ayan, Ben Hutchinson, Wei Han, Zarana Parekh, Xin Li, Han Zhang, Jason Baldridge, and Yonghui Wu. Scaling autoregressive models for content-rich text-to-image generation. *TMLR*, 2022. 6, 7, 14

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