RAID: A Dataset for Testing the Adversarial Robustness of AI-Generated Image Detectors

Hicham Eddoubi[1,5] **Jonas Ricker**[2] **Federico Cocchi**[3,4] **Lorenzo Baraldi**[4]
Angelo Sotgiu[1,6] **Maura Pintor**[1,6] **Marcella Cornia**[3] **Lorenzo Baraldi**[3]
Asja Fischer[2] **Rita Cucchiara**[3] **Battista Biggio**[1,6]
[1]University of Cagliari, Italy [2]Ruhr University Bochum, Germany
[3]University of Modena and Reggio Emilia, Italy [4]University of Pisa, Italy
[5]Sapienza University of Rome, Italy [6]CINI, Italy
{hicham.eddoubi, angelo.sotgiu, maura.pintor, battista.biggio}@unica.it
{jonas.ricker, asja.fischer}@rub.de {federico.cocchi, marcella.cornia,
lorenzo.baraldi, rita.cucchiara}@unimore.it lorenzo.baraldi@phd.unipi.it

Abstract

AI-generated images have reached a quality level at which humans are incapable of reliably distinguishing them from real images. To counteract the inherent risk of fraud and disinformation, the detection of AI-generated images is a pressing challenge and an active research topic. While many of the presented methods claim to achieve high detection accuracy, they are usually evaluated under idealized conditions. In particular, the *adversarial robustness* is often neglected, potentially due to a lack of awareness or the substantial effort required to conduct a comprehensive robustness analysis. In this work, we tackle this problem by providing a simpler means to assess the robustness of AI-generated image detectors. We present RAID (Robust evaluation of AI-generated image Detectors), a dataset of 72k diverse and highly transferable adversarial examples. The dataset is created by running attacks against an ensemble of seven state-of-the-art detectors and images generated by four different text-to-image models. Extensive experiments show that our methodology generates adversarial images that transfer with a high success rate to unseen detectors, which can be used to quickly provide an approximate yet still reliable estimate of a detector's adversarial robustness. Our findings indicate that current state-of-the-art AI-generated image detectors can be easily deceived by adversarial examples, highlighting the critical need for the development of more robust methods. We release our dataset at `https://huggingface.co/datasets/aimagelab/RAID` and evaluation code at `https://github.com/pralab/RAID`.

1 Introduction

Over the last years, generative artificial intelligence (GenAI) has evolved from a mere research topic to a vast collection of commonly available tools. While the inception of large language models (LLMs), most notably ChatGPT, has been most transformative for our everyday life, the evolution of generative image modeling has drastically shifted our understanding of visual media. This development was initiated by the discovery of diffusion models [71], which utilize an iterative noising and denoising process [34] to learn the distribution of natural images. Later work [66] improved this process by performing the generation process in the compressed latent space of a pre-trained variational autoencoder [42] that essentially reduces computational overhead and preserves semantic information while discarding high-frequency noise, in addition to introducing flexible conditional generation with the use of cross-attention layers. This rapid development, while improving computer vision tasks

Submitted to 39th Conference on Neural Information Processing Systems (NeurIPS 2025). Do not distribute.

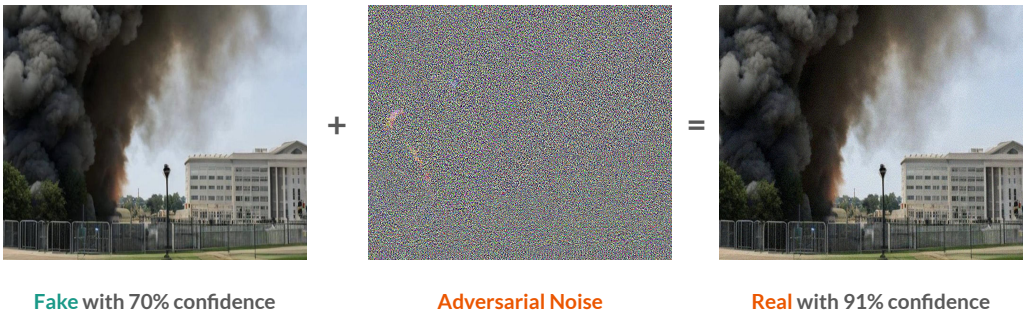

Figure 1: Adversarial attack on the commercially available detector provided by *sightengine*. Successful evasion is shown by adding adversarial noise computed on the Corvi et al. [16] detector.

such as image upsampling [78] and dataset augmentation [3], poses a considerable risk of nefarious misuse leading to the spread of misinformation, privacy violation, and identity theft [63, 82]. This underscores the urgent need for detection methods that generalize and keep up with the ever-evolving image generation technology while maintaining robustness to adversarial attempts to evade detection.

To mitigate the harmful consequences of AI-generated images, a variety of detection approaches have been proposed in the literature [4, 12, 32, 50, 57, 75, 76]. Based on the reported results, according to which many detectors can distinguish real from generated images with almost perfect accuracy, one could get the impression that the problem of detecting synthetic images is already solved. However, evaluations are typically conducted within an idealized lab setting that does not consider real-world risks. One major factor is the effect of common processing operations, such as resizing or compression, which have already been shown to have drastic effects on the performance of AI-generated image detectors [15, 31, 80].

An important factor, which the majority of existing work has neglected, is the *adversarial robustness* of detectors. Take, for instance Figure 1, a synthetic image generated and spread across social media outlets [58] that despite the relatively quick intervention to debunk it as a *fake image*, still had an impact on the stock market; we can detect such an image as being AI-generated with an off-the-shelf detector, but when we modify the image using carefully designed adversarial perturbations, it can evade the detection of said detector and others, with the effectiveness increasing for those that share similar architecture [52]. The adversarial robustness is often not investigated in works proposing synthetic image detectors, partially due to the significant effort required to generate adversarial examples. Due to many different attacks and hyperparameters and the required technical knowledge, conducting a comprehensive robustness analysis is not straightforward.

Existing work [52] unveils this failure to show robustness in white-box scenarios where the malicious actor has access to the architecture and training parameters of the detector, and also in black-box scenarios where the attacker's knowledge is limited. However, we note that the attacks used for the evaluation remain restricted in using techniques that increase their success and transferability. In this work, we extend this concept to bridge this evaluation gap by providing a standard and effective means to assess the adversarial robustness of AI-generated image detectors. In particular, we propose RAID (Robust evaluation of AI-generated image Detectors), a large-scale dataset of *diverse* and *transferable* adversarial examples created using an ensemble of state-of-the-art detectors that employ different architectures. As we experimentally demonstrate, testing the detection performance on RAID provides a solid estimate of the adversarial robustness of a detector. Our benchmark on seven recently proposed detectors shows that the current landscape of AI-generated image detection is not yet expansive nor reliable for widespread adoption in the real world, without properly ensuring adversarial robustness to evasion attacks.

Contributions. In summary, we make the following contributions:

- We create RAID, the first dataset of transferable adversarial synthetic images, constructed using highly transferable attacks, to standardize testing the adversarial robustness of SoTA synthetic image detectors.

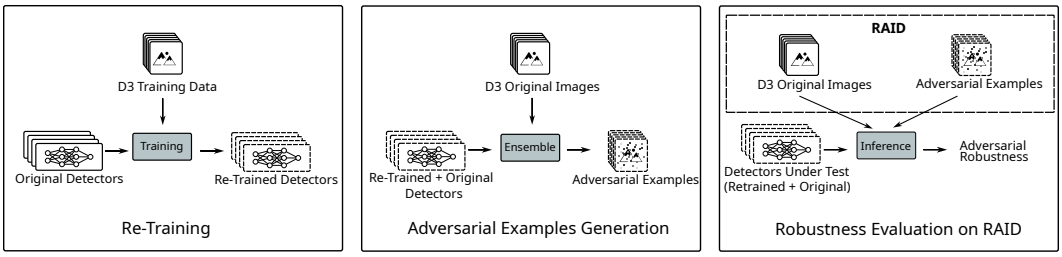

Figure 2: A diagram of the three experimental pipelines used in the paper for generating the RAID dataset. Left: We keep the top three performing detectors intact while we re-train the rest. Middle: We generate adversarial examples from adversarial attacks on an ensemble of detectors. Right: We evaluate the baseline detectors on the RAID dataset.

- We conduct a large-scale study showing that adversarial perturbations transfer across several SoTA synthetic image detectors.
- We show that the transferability of adversarial perturbations to AI-generated detectors increases when we use an ensemble adversarial attack with comparable results to a white-box attack.

2 The RAID Dataset

This section describes how we constructed our dataset of transferable adversarial examples. An overview of how we created RAID is given in Figure 2.

2.1 Source Dataset

RAID is built upon the D^3 dataset [4]. In total, D^3 consists of 11.5M images. It is constructed from 2.3M real images taken from the LAION-400M [70] dataset. Using the corresponding caption as prompts, synthetic images are generated from four open-source text-to-image models: Stable Diffusion v1.4 [66], Stable Diffusion v2.1 [66], Stable Diffusion XL [59], and DeepFloyd IF [2]. For details on data selection and prompt engineering, we refer to the original publication [4]. It should be noted that each generated image is post-processed such that the image format and compression strength match that of the real distribution present in the corresponding real image. This not only reduces the risk of unwanted biases between real and generated images but also makes the dataset significantly more challenging for the detection task.

We built our dataset using D^3 in two phases. First, to re-train the detectors used to compute the adversarial examples, we take the training subset of D^3 comprising 2,311,429 real and 9,245,716 generated images. As we show in Section 3.2, re-raining helps to ensure a sufficient detection performance on the original images and, subsequently, the generation of effective adversarial examples. Second, we use the same procedure as in [4] to construct the actual RAID dataset to generate synthetic images based on 4,800 new real images. For each of the resulting 24,000 images (i.e., the real images and the synthetic images from four generators), we create matching adversarial examples using the attack presented in Section 2.2 for each ϵ attack parameter. Thus, our proposed dataset consists of 72,000 adversarial examples — 24,000 adversarial examples for each attack parameter ϵ, of which we consider $\frac{8}{255}$, $\frac{16}{255}$ and $\frac{32}{255}$ — in addition to original images, for a total of 96,000 images.

2.2 Crafting Adversarial Perturbations

Adversarial Examples Optimization Given an input sample $\boldsymbol{x} \in [0,1]^d$ and a victim model with parameters $\boldsymbol{\theta}$, adversarial examples [5, 72] can be crafted with evasion attacks, which aim to solve the following optimization problem to compute the adversarial perturbation $\boldsymbol{\delta} \in \mathbb{R}^d$:

$$\underset{\boldsymbol{\delta}:\|\boldsymbol{\delta}\|_\infty \leq \epsilon}{\arg\min}\ \mathcal{L}(\boldsymbol{x}+\boldsymbol{\delta}, \boldsymbol{\theta})\ , \tag{1}$$

where ϵ is the applied bound on the perturbation size, and $\mathcal{L}$ is a loss function encoding the attacker's objective (i.e., a misclassification).

The above optimization problem is commonly solved with gradient-based techniques, which require access to the model's parameters and gradients. Despite this approach working well in this white-box setting, it is often not possible to get full access to the model's internals. Nevertheless, it has been shown that the adversarial examples produced against a model can still be effective on different models, although this phenomenon, called *transferability*, is weaker across different model architectures. To increase transferability, previous works proposed to simultaneously attack an ensemble of models [23, 46]. We thus leverage an ensemble attack by extending Eq. 1 to a set of M models:

$$\underset{\boldsymbol{\delta}:\|\boldsymbol{\delta}\|_\infty \leq \epsilon}{\arg\min}\, \mathcal{L}(\boldsymbol{x}+\boldsymbol{\delta}, \boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \ldots, \boldsymbol{\theta}_M) \ . \tag{2}$$

Among several possible choices, we define the loss function as:

$$\mathcal{L}(\boldsymbol{x}+\boldsymbol{\delta}, \boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \ldots, \boldsymbol{\theta}_M) = \frac{1}{M}\sum_{m=1}^{M} CE(l_{\boldsymbol{\theta}_m}(\boldsymbol{x}), y_t) \ , \tag{3}$$

where CE is the Cross-Entropy loss, and $l_{\boldsymbol{\theta}_m}$ is the output logit of the m-th model and y_t is the target class, i.e., the output label desired by the attacker.

Detection Ensemble. The key requirement for our dataset is that the adversarial examples are *transferable*, i.e., they are effective against unseen detectors. To achieve high transferability, we use an ensemble of a diverse set of seven different detectors:

Wang et al. [75] In this seminal work, the authors show that a ResNet-50 [33] trained on real and generated images from a single generator (ProGAN [38]) is sufficient to successfully detect images from a variety of other GANs. During training, extensive data augmentation is used to account for different image processing and increase generalization.

Corvi et al. [16] This architecture is adapted from [31] and is a modification of the detector proposed by Wang et al. [75]. It uses the same backbone, but avoids down-sampling in the first layer to preserve high-frequency features and uses stronger augmentation.

Ojha et al. [57] Unlike previous work, this detector uses a pre-trained vision foundation model (CLIP [60]) as a feature extractor to avoid overfitting on a particular class of generated images and, thus, improve generalization. To obtain a final classification, a single linear layer is added and trained on top of the 768-dimensional feature vector.

Koutlis and Papadopoulos [43] The approach of *RINE* is similar to that presented by Ojha et al. [57], but additionally uses intermediate encoder-blocks of CLIP [60]. The resulting features are weighted using a learnable projection network, followed by a classification head.

Cavia et al. [8] Instead of classifying the entire image at once, *LaDeDa* operates on 9×9 patches. The architecture is based on ResNet-50 [33] but uses 1×1 convolutions to limit the receptive field. The scores of all patches are combined using average pooling to obtain a final score.

Chen et al. [12] *DRCT* is a training paradigm that can increase the generalizability of AI-generated image detectors. It uses the reconstruction capabilities of DMs to generate images that are semantically similar to real ones, but contain the artifacts used for detection. The authors provide pre-trained detectors based on two backbones, ConvNeXt [47] and CLIP [60].

As we show in Section 3.2, the published version of most detectors does not perform well on the original images in our dataset. We hypothesize that this may be attributed to the applied post-processing and the specific test images, which differ greatly from the training data of the various detectors. Since creating adversarial examples based on detectors that are already ineffective against clean samples reduces the impact of the work, we re-train detectors on the training subset described in Section 2.1, following the original authors' training instructions.

3 Experimental Analysis

In this section, we evaluate how well RAID can be used to estimate the adversarial robustness of AI-generated image detectors. To this end, we initially test the performance on unperturbed images and conduct classical white-box attacks on each detector, demonstrating their susceptibility to evasion attacks. We subsequently analyze the transferability of white-box attacks and compare them to our ensemble attack, showing the effectiveness of RAID.

3.1 Experimental Setup

Detectors. In addition to the seven detectors described in 2.2, we use four additional detectors whose architecture is a pre-trained visual backbone with a linear layer added on top. During training, a Binary Cross Entropy Loss is considered to discern between real and fake images. All model weights are frozen except for the linear layer trained using the D³ train set. In particular, we use two different versions of DINO [7, 19] to highlight the behavior of self-supervised models. At the same time, to explore the impact of model size, we also consider a ViT-Tiny. We follow the transformation pipeline introduced by Ojha et al. [57] for training and evaluation. This setup enables a consistent comparison across different architectures and training paradigms applied to deepfake detection. Finally, we adapt the CoDE model [4] for this analysis. Specifically, we used the feature extractor trained with a contrastive loss tailored for deepfake detection and trained a linear classification layer on top. In this case, we follow the transformation strategy described in the original CoDE implementation.

Dataset. As noted in 2.2, we use the images of the D³ test data set and the adversarial images generated by the adversarial attacks. We run the evaluations for our experiments on 1k clean images and 1k adversarial images. *All images are center-cropped* to ensure consistent input dimensions for efficient batch processing, before applying the detector-specific preprocessing and evaluation. Moreover, in the creation of the RAID dataset, following the approach used for the D³ dataset, only images labeled as *safe* in the LAION metadata were considered as real images. This ensured that the generated images also adhere to this safeguard. To this end, samples depicting NSFW images has been excluded.

Experimental environment. During the re-training phase of the different detectors, 4x A100 GPUs are used in a distributed data parallel setup. Each experiment runs for a maximum of 18 hours until convergence is reached. In contrast, the ensemble attack is conducted using a single A100 GPU, which takes a total of 8 hours. The evaluation script is lightweight and takes less than 1 hour for each detector in a non-distributed setting.

Attack setting. The attack optimization is performed with a PGD algorithm, with 10 iterations and a step size of 0.05. We select two perturbations for the attacks: $\epsilon = \frac{16}{255}$ and $\epsilon = \frac{32}{255}$.

Evaluation Metrics. To evaluate the performance of the detectors, we make use of the following metrics:

- *F1-score.* The F1 score measures the harmonic mean of the precision and true positive rate (TPR), which provides a metric capable of reliably computing the model's performance in the presence of unbalanced class distributions. It is defined as: $F1 = 2 \times \frac{\text{Precision} \times \text{TPR}}{\text{Precision} + \text{TPR}}$.
- *Accuracy.* Accuracy is the ratio of correctly predicted samples over the total number of samples. It can be misleading in unbalanced datasets as it does not consider class distributions. We take the classification threshold = 0.5, as was done for all the detectors considered.
- *AUROC.* The Area Under the Receiver Operating Characteristic Curve metric summarizes the ROC curve, which plots the True Positive Rate against the False Positive Rate, by correctly measuring the model's capability to identify samples across all classification thresholds. An AUROC of 1 corresponds to a perfect classifier with a 0 False Positive Rate across all thresholds, and an AUROC of 0.5 corresponds to a random chance classifier.

3.2 Experimental Results

Initial Evaluation. Prior to evaluating the adversarial robustness of the considered detectors, we first evaluate their performance on the D³ test set. The initial performance evaluation of the detectors with the provided weights in Table 1 shows mixed results, particularly regarding F1-score and Accuracy measures. For instance, Cavia et al. [8] and Wang et al. [75] show very low F1-score of 20 and 20.6, along with an AUROC of 44.6 and 45.8, respectively. Additionally, Ojha et al. [57] shares a similar trend with F1-score = 32.8, while other detectors show a decent performance. We hypothesize that the poor adaptability of the first three detectors is due to the data drift between the D³ test set and the datasets used in their respective papers, in addition to the post-processing applied to the images, especially the compression of images when using lossy formats.

Table 1: Evaluation of each model on a subset of the clean D³ test set (1000 samples). † refers to detectors trained on the D³ training set.

	F1	Accuracy	AUROC
Cavia *et al.* [8] (Ca24)	0.2	0.07	0.45
Corvi *et al.* [16] (Co23)	0.75	0.81	0.91
Ojha *et al.* [57] (O23)	0.33	0.29	0.68
Wang *et al.* [75] (W20)	0.21	0.02	0.46
Cavia *et al.* [8]† (Ca24†)	0.55	0.63	0.84
Chen *et al.*(CLIP) [12] (Ch24C)	0.81	0.89	0.73
Chen *et al.*(ConvNext) [12] (Ch24CN)	0.86	0.91	0.94
Corvi *et al.* [16]† (Co23†)	0.98	0.99	0.99
Koutlis and Papadopoulos [43] (K24)	0.74	0.81	0.88
Ojha *et al.* [57]† (O23†)	0.92	0.95	0.95
Wang *et al.* [75]† (W20†)	0.99	0.99	0.99

In our work, we focus on robustness to adversarial attacks, and as such, we re-train on the D³ training set the top four detectors that perform the worst: Cavia et al. [8], Corvi et al. [16], Ojha et al. [57], and Wang et al. [75]. We retain the same weights for the rest of the detectors (Koutlis and Papadopoulos [43], Chen et al. [12] (ConvNext) and Chen et al. [12] (CLIP)), both for the acceptable performance and to ensure the generalization of attacks without the induced risk of a dataset-bias if all detectors are trained on the D³ dataset. After re-training, we note an improvement across metrics for the four detectors; however, Cavia et al. [8]'s metrics do not improve similarly to the others.

Table 2: White-box and Black-box Adversarial Robustness Evaluation (F1/AUROC) on PGD with 10 steps, step size = 0.05 and ϵ = 16/255 - Adversarial attacks are ran against the model in each row and evaluated on models in each column. † refers to detectors trained on the D³ training set. Final row N - *model* refers to ensemble attacks targeting all models, excluding the evaluated against model. **bold** values corresponding to white-box evaluation in single detector attacks

	Ca24†	Ch24C	Ch24CN	Co23†	K24	O23†	W20†
Ca24†	**0.0/0.5**	0.9/0.63	0.9/0.85	0.92/0.91	0.85/0.65	0.78/0.8	0.8/0.83
Ch24C	0.6/0.56	**0.0/0.5**	0.28/0.58	0.87/0.87	0.07/0.51	0.23/0.56	0.76/0.81
Ch24CN	0.69/0.54	0.86/0.61	**0.29/0.48**	0.87/0.88	0.85/0.59	0.79/0.8	0.75/0.8
Co23†	0.74/0.48	0.86/0.64	0.48/0.64	**0.0/0.5**	0.78/0.6	0.67/0.73	0.05/0.51
K24	0.55/0.56	0.31/0.51	0.3/0.59	0.83/0.85	**0.0/0.5**	0.41/0.62	0.8/0.83
O23†	0.59/0.58	0.46/0.52	0.41/0.63	0.84/0.85	0.31/0.54	**0.0/0.5**	0.74/0.79
W20†	0.7/0.54	0.88/0.66	0.25/0.56	0.09/0.52	0.84/0.68	0.69/0.75	**0.0/0.5**
N - *model*	0.66/0.43	0.34/0.51	0.2/0.56	0.53/0.67	0.08/0.51	0.18/0.54	0.4/0.62

Adversarial Robustness Evaluation. We evaluate the adversarial robustness of the seven detectors, using the Projected Gradient Descent (PGD) adversarial attack [48], a well-established iterative approach to generate adversarial examples. We employ PGD with the following configuration: *10 steps*, *step size = 0.05* and two perturbation budgets $\epsilon = \frac{16}{255}$ and $\epsilon = \frac{32}{255}$. and we report the results in Tables 2 and 3.

First, we assess the adversarial robustness of detectors in a white-box setting where each detector is attacked using adversarial perturbations crafted against it. This scenario represents the worst-case scenario in which the attacker has full knowledge of the target detector's architecture and parameters. The results for both $\epsilon = \frac{16}{255}$ and $\epsilon = \frac{32}{255}$ reveal a general lack of adversarial robustness. In particular, looking at Table 2, the F1-score drops to 0 for all detectors except for Chen et al. [12] (ConvNext), which exhibits inherent robustness with F1-score = 28.9. This tendency is carried over to the results in Table 3, where we report the attack against the detectors with $\epsilon = \frac{32}{255}$, as even under such large adversarial perturbations, the F1-score (25.2) does not drop to 0 similarly to the rest of the detectors. However, we underline that the drop in performance is still steep, from the initial F1 score = 86.5 on the clean examples 1. Additionally, an evaluation of the AUROC reveals

Table 3: White-box and Black-box Adversarial Robustness Evaluation (F1/AUROC) on PGD with 10 steps, step size = 0.05 and ϵ = 32/255 - Adversarial attacks are ran against the model in each row and evaluated on models in each column. † refers to detectors trained on the D³ training set. Final row N - *model* refers to ensemble attacks targeting all models, excluding the evaluated against model. **bold** values corresponding to white-box evaluation in single detector attacks

	Ca24†	Ch24C	Ch24CN	Co23†	K24	O23†	W20†
Ca24†	**0.0/0.5**	0.87/0.61	0.82/0.81	0.93/0.89	0.84/0.69	0.80/0.80	0.69/0.76
Ch24C	0.12/0.49	**0.0/0.5**	0.03/0.51	0.86/0.87	0.0/0.5	0.06/0.51	0.61/0.72
Ch24CN	0.33/0.5	0.79/0.61	**0.25/0.51**	0.81/0.83	0.79/0.69	0.76/0.76	0.58/0.70
Co23†	0.67/0.48	0.76/0.59	0.17/0.53	**0.0/0.5**	0.65/0.6	0.69/0.73	0.03/0.51
K24	0.23/0.54	0.20/0.51	0.13/0.53	0.87/0.87	**0.0/0.5**	0.36/0.6	0.74/0.79
O23†	0.13/0.51	0.19/0.51	0.10/0.52	0.85/0.86	0.08/0.51	**0.0/0.5**	0.63/0.73
W20†	0.59/0.49	0.82/0.60	0.06/0.5	0.04/0.51	0.68/0.6	0.67/0.72	**0.0/0.5**
N - *model*	0.28/0.33	0.08/0.49	0.01/0.5	0.32/0.59	0.01/0.5	0.06/0.52	0.17/0.54

similar results. Furthermore, we investigate the transferability of adversarial examples, referring to whether adversarial perturbations designed to fool one detector can also fool the others. Many adversarial examples generated for one detector do indeed reduce the performance substantially among the other detectors. For example, when we look at adversarial examples generated for Koutlis and Papadopoulos [43] in Table 2, an AUROC performance drop can be seen across the following detectors: Cavia et al. [8] to 0.56 (from 0.84), Chen et al. [12] (CLIP) to 0.51 (from 0.73), Chen et al. [12] (ConvNext) to 0.59 (from 0.94), and Ojha et al. [57] to 0.5 (from 0.95). The other two detectors Corvi et al. [16] and Wang et al. [75] seem resilient, although this trend continues for perturbations generated for other detectors.

Next, we evaluate the transferability of ensemble attacks, in a *leave-one-out manner*, in which the assessed detector is excluded from the attacked ensemble. This ensemble approach described in Section 2.2 significantly boosts transferability. It allows us to evaluate the detector's performance in a transferable black-box scenario where access to the targeted detector's architecture or parameters is unavailable. Therefore, the attack (in a white-box manner) is done against an ensemble of detectors to increase its effectiveness further. We find that the ensemble attacks can drop the performance of detectors to metrics similar to a white-box attack. For example, looking at Table 2 we note that even for the two detectors Corvi et al. [16] and Wang et al. [75] that showed a decent robustness to perturbations generated for other detectors, the AUROC metric drops to 0.67 and 0.62 respectively (from 0.99 and 0.99), and the F1-score metric drops to 0.53 and 0.4 (from 0.98 and 0.99). This drop in performance is further highlighted in Table 3, where we use a higher perturbation budget in which the AUROC metric drops to 0.59 and 0.54 for the two detectors, respectively.

These results motivate the creation of our RAID dataset, which is composed of adversarial examples crafted using attacks on an ensemble of SoTA detectors. Our dataset enables a fast and standardized benchmark of new detectors against strong transferable perturbations, facilitating the assessment of their robustness to adversarial attacks.

RAID — Tensors vs Images. We construct the RAID dataset as shown in Figure 2 by running the adversarial attack on the ensemble of detectors using the entire D³ dataset. We generate the dataset and save the adversarial examples as PNG images, avoiding the use of the lossy format (JPEG, JPG). The reason for this is that we only consider the worst-case adversarial scenario in which no post-processing operations are done, which could reduce the transferability of the attack. While the previous evaluation provides a good assessment of RAID, as only one detector is missing from the ensemble, we perform one additional evaluation on the full RAID dataset, to ensure that the effectiveness of the adversarial perturbations is not reduced with their quantization when we save them as images. We use four additional baseline detectors detailed in 3.1 tested against the adversarial examples saved as tensors with float values, and the adversarial images. We find no significant drop in effectiveness except for a few fluctuations as reported in Table 4. We release our full dataset with a total of 96,000 images: 24,000 adversarial examples for each attack parameter ϵ considering $\frac{8}{255}$, $\frac{16}{255}$ and $\frac{32}{255}$, in addition to the original images.

Table 4: Adversarial Robustness Evaluation of the 4 trained baseline classifiers. All detectors were trained on the D^3 training set. `RAID` refers to the dataset of the generated adversarial examples saved as raw tensors, while `RAID_IMG` refers to the released dataset of adversarial images

Metric	Dataset	DINOv2	DINOv2-Reg	ViT-T	ViT-T CoDE[4]
Accuracy	D^3	83.1	81.5	83.7	91.9
	RAID(ϵ=16/255)	54.6	55.4	68.2	75.6
	RAID(ϵ=32/255)	43.6	46.8	63.4	76.6
	RAID_IMG(ϵ=16/255)	54.8	53.3	69.5	72.2
	RAID_IMG(ϵ=32/255)	43.6	46.8	63.5	76.6
F1	D^3	88.8	87.6	89.4	94.9
	RAID(ϵ=16/255)	60.9	62	76.2	82.8
	RAID(ϵ=32/255)	46.0	50.7	71.3	84.1
	RAID_IMG(ϵ=16/255)	61.4	59.4	77.7	80
	RAID_IMG(ϵ=32/255)	46.1	50.7	71.4	84.1
AUROC	D^3	81.2	81.6	79.9	87.5
	RAID(ϵ=16/255)	69.9	70.3	75.1	78.9
	RAID(ϵ=32/255)	63.8	65.6	73.2	75.4
	RAID_IMG(ϵ=16/255)	69.9	69.2	74.2	76.3
	RAID_IMG(ϵ=32/255)	63.8	65.6	73.2	75.7

4 Related Work

Generative Image Modeling. Generating images requires learning the underlying data distribution, which is non-trivial for high-dimensional distributions such as natural images. Several approaches have been proposed, such as autoregressive models [73, 74], VAEs [42, 62], and GANs [30]. The latter architecture has been continuously improved [14, 38, 85] to improve quality. Especially StyleGAN [39] and its successors [40, 41] achieved unprecedented visual quality, making generated images impossible to distinguish from real ones [28, 56]. While it has been shown that DMs [21, 34, 71] can surpass GANs with respect to quality, the costly iterative denoising process prevented the generation of high-resolution images. As a remedy, LDMs [66] perform the diffusion and denoising process in a smaller latent space instead of the high-dimensional pixel space, using a pre-trained VAE [42] as a translation layer across both domains. Moreover, the addition of cross-attention layers based on U-Net [67] allows controlling the generative process based on, for instance, a textual prompt, laying the foundation for powerful text-to-image models [55, 61, 69]. Recent models have significantly advanced the resolution of generated images (up to 4k) [13, 84], improved prompt following and human preference [26, 79].

AIGI Detection Methods. Due to the continuously increasing capabilities of generative models, the detection of AI-generated images is an active area of research with a plethora of proposed methods. Early approaches, mostly targeting images generated by GANs, exploit visible flaws like differently colored irises [51] or irregular pupil shapes [32]. Since the occurrence of such imperfections is becoming less likely as generative models become more advanced, several methods rely on imperceptible artifacts. Such features include model-specific fingerprints [50, 83] or unnatural patterns in the frequency domain [24, 25, 27]. Instead of identifying generated images based on selected features, the majority of methods are data-driven. In their seminal work, Wang et al. [75] demonstrate that training a ResNet-50 [33] on real and generated images from ProGAN [38], paired with strong data augmentation, suffices to detect images generated by several other GANs. Subsequent works propose improved model architectures [31] or learning paradigms [12, 17, 49], with a particular focus on generalization [10, 44]. A promising direction is the use of foundation models as feature extractors to avoid overfitting on images generated by a single class of models [18, 43, 57]. While most works attempt to detect images from all kinds of generative models, several methods explore unique characteristics of DMs for detection, like frequency artifacts [16, 64] or features obtained by inverting the diffusion process [9, 65, 76].

Datasets for Evaluating the Robustness of AIGI Detectors. To the best of our knowledge, we are the first to propose a dataset for evaluating the *adversarial* robustness of AIGI detectors. However, several datasets exist to test their generalizability and robustness to common image degradations.

GenImage [86] is a large-scale dataset comprising 1.35 million generated images based on the 1 000 class labels of ImageNet [68]. Within their benchmark, they evaluate how well seven deepfake detectors generalize to images from unseen generators as well as their robustness to downsampling, JPEG compression, and blurring. *WildFake* [35] is a hierarchically-structured dataset featuring images generated by GANs, DMs, and other generative models, which are partly sourced from platforms such as Civitai to cover a broad range of content and styles. Yan et al. [81] also collect images from popular image-sharing websites. However, their *Chameleon* dataset features images that were misclassified by human annotators, allowing for the evaluation of deepfake detectors on particularly challenging images. Furthermore, *Deepfake-Eval-2024* [11] contains deepfake videos, audio, and images that circulated on social media and deepfake detection platforms in 2024. Through their evaluation, the authors find that detectors trained on academic datasets fail to generalize to real-world deepfakes.

Adversarial Robustness of AIGI Detectors. The vulnerability of deepfake detectors to adversarial examples was first explored by Carlini and Farid [6]. They demonstrate that by adding visually imperceptible perturbations to an image, the AUC of a forensic classifier can be reduced from 0.95 down to 0.0005 in the white-box and 0.22 in the black-box setting. Subsequent work explores the applicability of attacks in practical scenarios [37, 52, 54] as well as possible defenses [29]. De Rosa et al. [20] study the robustness of CLIP-based detectors, finding that adversarial examples computed for CNN-based classifiers are not easily transferable and vice versa. Besides the addition of adversarial noise, it has been shown that deepfake detectors can also be attacked by applying natural degradations (e.g., local brightness changes) [36] or by removing generator-specific artifacts [22, 77]. Other attacks leverage image generators themselves to perform semantic adversarial attacks, which adversarially manipulate a particular attribute of an image [53] that can even be controlled through a text prompt [1, 45].

5 Discussion

Adversarial robustness should always be evaluated when proposing new AI-generated image detection methods, as in current SoTA detectors, it is vastly neglected in favor of an evaluation against *naturally occurring* post-processing operations, such as resizing, cropping, blurring, jpeg compression or noise. While the robustness to these operations is indeed important, introducing a *malicious actor* that utilizes carefully crafted adversarial noise can lead to the evasion of detection by most methods, as highlighted in our work. Nonetheless, the lack of a standard benchmark that serves as a comparability reference for detectors contributes further to this lack of evaluation against adversarial attacks in AI-generated image detection. As such, we introduce RAID to address this gap in the current literature and provide a more comprehensive solution for evaluating generative models. However, it is important to acknowledge that our method is not without its limitations. One key challenge is that RAID requires frequent updates to stay relevant as new generative models emerge, and these models would need to be incorporated into our proposed ensemble attacks for continued effectiveness. This dynamic nature of generative models demands a proactive approach to maintain the robustness of RAID over time. Additionally, our perturbations are not designed to be robust to post-processing operations, which should be considered in future work. Finally, due to the inherent restrictions of input sizes of the architecture of some detectors, when considering attacks on ensemble models, we are restricted to the center region of the image to be perturbed.

6 Conclusions

We introduce RAID, the first dataset of transferable adversarial examples for robustness evaluation of AI-generated image detection. We employ an ensemble attack that demonstrates strong transferability against seven diverse detectors and cover images generated from four text-to-image generative models. Our results further highlight the existing gap in current evaluations of SoTA detectors, as more often than not, they are tested on naturally occurring post-processing as images are disseminated and shared, but remain highly vulnerable to adversarial attacks. RAID addresses this gap by providing a simple and reliable benchmark for adversarial robustness evaluation, ensuring that detection models can be tested under more realistic and challenging adversarial conditions.

References

[1] Sifat Muhammad Abdullah, Aravind Cheruvu, Shravya Kanchi, Taejoong Chung, Peng Gao, Murtuza Jadliwala, and Bimal Viswanath. An analysis of recent advances in deepfake image detection in an evolving threat landscape. In *SP*, 2024. 9

[2] DeepFloyd Lab at StabilityAI. DeepFloyd IF: a novel state-of-the-art open-source text-to-image model with a high degree of photorealism and language understanding. `https://huggingface.co/DeepFloyd/IF-I-XL-v1.0`, 2023. 3

[3] Shekoofeh Azizi, Simon Kornblith, Chitwan Saharia, Mohammad Norouzi, and David J Fleet. Synthetic data from diffusion models improves imagenet classification. *arXiv preprint arXiv:2304.08466*, 2023. 2

[4] Lorenzo Baraldi, Federico Cocchi, Marcella Cornia, Lorenzo Baraldi, Alessandro Nicolosi, and Rita Cucchiara. Contrasting Deepfakes Diffusion via Contrastive Learning and Global-Local Similarities. In *ECCV*, 2024. 2, 3, 5, 8

[5] B. Biggio, I. Corona, D. Maiorca, B. Nelson, N. Šrndić, P. Laskov, G. Giacinto, and F. Roli. Evasion attacks against machine learning at test time. In Hendrik Blockeel, Kristian Kersting, Siegfried Nijssen, and Filip Železný, editors, *Machine Learning and Knowledge Discovery in Databases (ECML PKDD), Part III*, volume 8190 of *LNCS*, pages 387–402. Springer Berlin Heidelberg, 2013. 3

[6] Nicholas Carlini and Hany Farid. Evading deepfake-image detectors with white- and black-box attacks. In *CVPRW*, 2020. 9

[7] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *ICCV*, 2021. 5

[8] Bar Cavia, Eliahu Horwitz, Tal Reiss, and Yedid Hoshen. Real-time deepfake detection in the real-world. *arXiv preprint arXiv:2406.09398*, 2024. 4, 5, 6, 7

[9] George Cazenavette, Avneesh Sud, Thomas Leung, and Ben Usman. FakeInversion: Learning to detect images from unseen text-to-image models by inverting Stable Diffusion. In *CVPR*, 2024. 8

[10] Lucy Chai, David Bau, Ser-Nam Lim, and Phillip Isola. What makes fake images detectable? Understanding properties that generalize. In *ECCV*, 2020. 8

[11] Nuria Alina Chandra, Ryan Murtfeldt, Lin Qiu, Arnab Karmakar, Hannah Lee, Emmanuel Tanumihardja, Kevin Farhat, Ben Caffee, Sejin Paik, Changyeon Lee, Jongwook Choi, Aerin Kim, and Oren Etzioni. Deepfake-eval-2024: A multi-modal in-the-wild benchmark of deepfakes circulated in 2024. *arXiv Preprint*, 2025. 9

[12] Baoying Chen, Jishen Zeng, Jianquan Yang, and Rui Yang. DRCT: Diffusion reconstruction contrastive training towards universal detection of diffusion generated images. In *ICML*, 2024. 2, 4, 6, 7, 8

[13] Junsong Chen, Chongjian Ge, Enze Xie, Yue Wu, Lewei Yao, Xiaozhe Ren, Zhongdao Wang, Ping Luo, Huchuan Lu, and Zhenguo Li. PIXART-Σ: Weak-to-strong training of diffusion transformer for 4K text-to-image generation. In *ECCV*, 2024. 8

[14] 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 *CVPR*, 2018. 8

[15] Federico Cocchi, Lorenzo Baraldi, Samuele Poppi, Marcella Cornia, Lorenzo Baraldi, and Rita Cucchiara. Unveiling the impact of image transformations on deepfake detection: An experimental analysis. In *ICIAP*, 2023. 2

[16] 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, 4, 6, 7, 8

[17] Davide Cozzolino, Diego Gragnaniello, Giovanni Poggi, and Luisa Verdoliva. Towards universal GAN image detection. In *VCIP*, 2021. 8

[18] Davide Cozzolino, Giovanni Poggi, Riccardo Corvi, Matthias Nießner, and Luisa Verdoliva. Raising the Bar of AI-generated Image Detection with CLIP. In *CVPRW*, 2024. 8

[19] Timothée Darcet, Maxime Oquab, Julien Mairal, and Piotr Bojanowski. Vision transformers need registers. In *ICLR*, 2024. 5

[20] Vincenzo De Rosa, Fabrizio Guillaro, Giovanni Poggi, Davide Cozzolino, and Luisa Verdoliva. Exploring the adversarial robustness of CLIP for AI-generated image detection. In *WIFS*, 2024. 9

[21] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat GANs on image synthesis. In *NeurIPS*, 2021. 8

[22] Chengdong Dong, Ajay Kumar, and Eryun Liu. Think twice before detecting GAN-generated fake images from their spectral domain imprints. In *CVPR*, 2022. 9

[23] Yinpeng Dong, Fangzhou Liao, Tianyu Pang, Hang Su, Jun Zhu, Xiaolin Hu, and Jianguo Li. Boosting Adversarial Attacks with Momentum . In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9185–9193, Los Alamitos, CA, USA, June 2018. IEEE Computer Society. doi: 10.1109/CVPR.2018.00957. URL `https://doi.ieeecomputersociety.org/10.1109/CVPR.2018.00957`. 4

[24] Ricard Durall, Margret Keuper, and Janis Keuper. Watch your up-convolution: CNN based generative deep neural networks are failing to reproduce spectral distributions. In *CVPR*, 2020. 8

[25] Tarik Dzanic, Karan Shah, and Freddie Witherden. Fourier spectrum discrepancies in deep network generated images. In *NeurIPS*, 2020. 8

[26] Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, Dustin Podell, Tim Dockhorn, Zion English, and Robin Rombach. Scaling rectified flow transformers for high-resolution image synthesis. In *ICML*, 2024. 8

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

[28] Joel Frank, Franziska Herbert, Jonas Ricker, Lea Schönherr, Thorsten Eisenhofer, Asja Fischer, Markus Dürmuth, and Thorsten Holz. A Representative Study on Human Detection of Artificially Generated Media Across Countries . In *SP*, 2024. 8

[29] Apurva Gandhi and Shomik Jain. Adversarial perturbations fool deepfake detectors. In *IJCNN*, 2020. 9

[30] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *NeurIPS*, 2014. 8

[31] D. Gragnaniello, D. Cozzolino, F. Marra, G. Poggi, and L. Verdoliva. Are GAN generated images easy to detect? A critical analysis of the state-of-the-art. In *ICME*, 2021. 2, 4, 8

[32] Hui Guo, Shu Hu, Xin Wang, Ming-Ching Chang, and Siwei Lyu. Eyes tell all: Irregular pupil shapes reveal GAN-generated faces. In *ICASSP*, 2022. 2, 8

[33] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 4, 8

[34] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *NeurIPS*, 33:6840–6851, 2020. 1, 8

[35] Yan Hong and Jianfu Zhang. WildFake: A large-scale challenging dataset for AI-generated images detection. *arXiv Preprint*, 2024. 9

[36] Yang Hou, Qing Guo, Yihao Huang, Xiaofei Xie, Lei Ma, and Jianjun Zhao. Evading DeepFake detectors via adversarial statistical consistency. In *CVPR*, 2023. 9

[37] Shehzeen Hussain, Paarth Neekhara, Malhar Jere, Farinaz Koushanfar, and Julian McAuley. Adversarial deepfakes: Evaluating vulnerability of deepfake detectors to adversarial examples. In *WACV*, 2021. 9

[38] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of GANs for improved quality, stability, and variation. In *ICLR*, 2018. 4, 8

[39] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *CVPR*, 2019. 8

[40] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of StyleGAN. In *CVPR*, 2020. 8

[41] Tero Karras, Miika Aittala, Samuli Laine, Erik Härkönen, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Alias-Free Generative Adversarial Networks. In *NeurIPS*, 2021. 8

[42] Diederik P. Kingma and Max Welling. Auto-encoding variational bayes. In *ICLR*, 2014. 1, 8

[43] Christos Koutlis and Symeon Papadopoulos. Leveraging representations from intermediate encoder-blocks for synthetic image detection. In *European Conference on Computer Vision*, pages 394–411. Springer, 2024. 4, 6, 7, 8

[44] Huan Liu, Zichang Tan, Chuangchuang Tan, Yunchao Wei, Jingdong Wang, and Yao Zhao. Forgery-aware adaptive transformer for generalizable synthetic image detection. In *CVPR*, 2024. 8

[45] Jiang Liu, Chen Wei, Yuxiang Guo, Heng Yu, Alan Yuille, Soheil Feizi, Chun Pong Lau, and Rama Chellappa. Instruct2Attack: Language-guided semantic adversarial attacks. *arXiv Preprint*, 2023. 9

[46] Yanpei Liu, Xinyun Chen, Chang Liu, and Dawn Song. Delving into transferable adversarial examples and black-box attacks. In *International Conference on Learning Representations*, 2017. URL `https://openreview.net/forum?id=Sys6GJqxl`. 4

[47] Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A ConvNet for the 2020s. In *CVPR*, 2022. 4

[48] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. In *International Conference on Learning Representations (ICLR)*, 2018. 6

[49] Sara Mandelli, Nicolò Bonettini, Paolo Bestagini, and Stefano Tubaro. Detecting GAN-generated images by orthogonal training of multiple CNNs. In *ICIP*, 2022. 8

[50] Francesco Marra, Diego Gragnaniello, Luisa Verdoliva, and Giovanni Poggi. Do GANs leave artificial fingerprints? In *MIPR*, 2019. 2, 8

[51] Falko Matern, Christian Riess, and Marc Stamminger. Exploiting visual artifacts to expose deepfakes and face manipulations. In *WACVW*, 2019. 8

[52] Sina Mavali, Jonas Ricker, David Pape, Yash Sharma, Asja Fischer, and Lea Schönherr. Fake it until you break it: On the adversarial robustness of ai-generated image detectors. *arXiv preprint arXiv:2410.01574*, 2024. 2, 9

[53] Xiangtao Meng, Li Wang, Shanqing Guo, Lei Ju, and Qingchuan Zhao. AVA: Inconspicuous attribute variation-based adversarial attack bypassing DeepFake detection. In *SP*, 2024. 9

[54] Paarth Neekhara, Brian Dolhansky, Joanna Bitton, and Cristian Canton Ferrer. Adversarial threats to DeepFake detection: A practical perspective. In *CVPRW*, 2021. 9

[55] Alexander Quinn 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. In *ICML*, 2022. 8

[56] Sophie J. Nightingale and Hany Farid. AI-synthesized faces are indistinguishable from real faces and more trustworthy. *Proceedings of the National Academy of Sciences*, 2022. 8

[57] Utkarsh Ojha, Yuheng Li, and Yong Jae Lee. Towards universal fake image detectors that generalize across generative models. In *CVPR*, pages 24480–24489, 2023. 2, 4, 5, 6, 7, 8

[58] Will Oremus, Drew Harwell, and Teo Armus. A fake image of an explosion at the pentagon caused a stir online. it was likely created by ai, May 2023. URL `https://www.washingtonpost.com/technology/2023/05/22/pentagon-explosion-ai-image-hoax/`. The Washington Post. 2

[59] 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. 3

[60] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *ICML*, 2021. 4

[61] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with CLIP latents. *arXiv Preprint*, 2022. 8

[62] Danilo Jimenez Rezende, Shakir Mohamed, and Daan Wierstra. Stochastic backpropagation and approximate inference in deep generative models. In *ICML*, 2014. 8

[63] Jonas Ricker, Dennis Assenmacher, Thorsten Holz, Asja Fischer, and Erwin Quiring. AI-generated faces in the real world: A large-scale case study of Twitter profile images. In *RAID*, 2024. 2

[64] Jonas Ricker, Simon Damm, Thorsten Holz, and Asja Fischer. Towards the detection of diffusion model deepfakes. In *VISIGRAPP*, 2024. 8

[65] Jonas Ricker, Denis Lukovnikov, and Asja Fischer. AEROBLADE: Training-free detection of latent diffusion images using autoencoder reconstruction error. In *CVPR*, 2024. 8

[66] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, pages 10684–10695, 2022. 1, 3, 8

[67] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *MICCAI*, pages 234–241, 2015. 8

[68] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. ImageNet large scale visual recognition challenge. *IJCV*, 2015. 9

[69] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L. Denton, Seyed Kamyar Seyed Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, Jonathan Ho, David J. Fleet, and Mohammad Norouzi. Photorealistic text-to-image diffusion models with deep language understanding. In *NeurIPS*, 2022. 8

[70] 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. In *Adv. Neural Inform. Process. Syst. Workshops*, 2021. 3

[71] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *ICML*, 2015. 1, 8

[72] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. In *International Conference on Learning Representations*, 2014. 3

[73] Aaron van den Oord, Nal Kalchbrenner, Lasse Espeholt, koray kavukcuoglu, Oriol Vinyals, and Alex Graves. Conditional image generation with PixelCNN decoders. In *NeurIPS*, 2016. 8

[74] Aaron van den Oord, Nal Kalchbrenner, and Koray Kavukcuoglu. Pixel recurrent neural networks. In *ICML*, 2016. 8

[75] 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*, pages 8695–8704, 2020. 2, 4, 5, 6, 7, 8

[76] Zhendong Wang, Jianmin Bao, Wengang Zhou, Weilun Wang, Hezhen Hu, Hong Chen, and Houqiang Li. DIRE for diffusion-generated image detection. In *ICCV*, 2023. 2, 8

[77] Vera Wesselkamp, Konrad Rieck, Daniel Arp, and Erwin Quiring. Misleading deep-fake detection with GAN fingerprints. In *IEEE Deep Learning and Security Workshop (DLS)*, 2022. 9

[78] Rongyuan Wu, Lingchen Sun, Zhiyuan Ma, and Lei Zhang. One-step effective diffusion network for real-world image super-resolution. *NeurIPS*, 37:92529–92553, 2024. 2

[79] Enze Xie, Junsong Chen, Junyu Chen, Han Cai, Haotian Tang, Yujun Lin, Zhekai Zhang, Muyang Li, Ligeng Zhu, Yao Lu, and Song Han. SANA: Efficient high-resolution image synthesis with linear diffusion transformers. *arXiv Preprint*, 2024. 8

[80] Huiyu Xu, Yaopeng Wang, Zhibo Wang, Zhongjie Ba, Wenxin Liu, Lu Jin, Haiqin Weng, Tao Wei, and Kui Ren. ProFake: Detecting deepfakes in the wild against quality degradation with progressive quality-adaptive learning. In *CCS*, 2024. 2

[81] Shilin Yan, Ouxiang Li, Jiayin Cai, Yanbin Hao, Xiaolong Jiang, Yao Hu, and Weidi Xie. A sanity check for AI-generated image detection. In *ICLR*, 2025. 9

[82] Kai-Cheng Yang, Danishjeet Singh, and Filippo Menczer. Characteristics and prevalence of fake social media profiles with ai-generated faces. *arXiv preprint arXiv:2401.02627*, 2024. 2

[83] Ning Yu, Larry S. Davis, and Mario Fritz. Attributing fake images to GANs: Learning and analyzing GAN fingerprints. In *ICCV*, 2019. 8

[84] Jinjin Zhang, Qiuyu Huang, Junjie Liu, Xiefan Guo, and Di Huang. Diffusion-4K: Ultra-high-resolution image synthesis with latent diffusion models. In *CVPR*, 2025. 8

[85] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *ICCV*, 2017. 8

[86] 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. *NeurIPS*, 2023. 9

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The claims made in the paper match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings. Refer to section 3 for these results.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: Refer to section 5, in which we discussed the limitations of this work.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [NA]

   Justification: We don't present theoretical results in this work.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: We discussed the steps to prepare and eventually reproduce the dataset in Section 2. Moreover, we release the code publicly to run the experiments and create the dataset in the URL included in the abstract. The evaluation is conducted using public models, thus we provide the pointers to access these models and their checkpoints.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: We provide the dataset and code in the URLs included in the abstract. The code includes a README file with instructions to reproduce the experiments.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: Refer to section 3.1, in which we describe experimental setting.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: Refer to the supplementary material.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

Answer: [Yes]

Justification: Refer to section 3.1 in which we describe the computer resources used to conduct the experiments.

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: Refer to 1, where we discuss the negative societal impacts addressed by this work.

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [Yes]

    Justification: Our work provides measures to safeguard synthetic image detectors and restrict the misuse of AI-generated images. In section 3.1 we provide more details.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: Refer to the dataset and code in the URLs included in the abstract.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: Refer to dataset and evaluation code released in addition to sections 2 and 3.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: Paper does not involve crowsourcing nor research with human subjects.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: Paper does not involve crowdsourcing nor research with human subjects.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

Answer: [NA]

Justification: LLMs are used only for editing and formatting purposes.