Computer Vision and Image Understanding 274 (2027) 105002

Contents lists available at ScienceDirect

Computer Vision and Image Understanding

journal homepage: www.elsevier.com/locate/cviu

RAID: A benchmark dataset for testing the adversarial robustness of AI-generated image detectors☆

Hicham Eddoubi [a,*], Jonas Ricker [b], Federico Cocchi [c], Angelo Sotgiu [a], Maura Pintor [a], Marcella Cornia [c], Lorenzo Baraldi [c], Asja Fischer [b], Battista Biggio [a]

[a] University of Cagliari, Italy
[b] Ruhr University Bochum, Germany
[c] University of Modena and Reggio Emilia, Italy

ARTICLE INFO

Dataset link: https://github.com/pralab/RAID, https://huggingface.co/datasets/aimagelab/RAID

Keywords:
AI-generated images
Adversarial robustness
Benchmark dataset
Adversarial attacks

ABSTRACT

AI-generated images (AIGIs) have reached a level of quality at which humans are unable to reliably distinguish them from real images. To counter the inherent risks of fraud and disinformation, their detection is a pressing challenge and an active research topic. While many of the presented methods claim high detection accuracy, they are often evaluated without considering their *adversarial robustness*. In real-world scenarios, malicious actors can modify AIGIs, e.g., through imperceptible perturbations, to evade detection systems. In this work, we provide a dataset to assess the robustness of AI-generated image detectors. We present **RAID** (**R**obust evaluation of **AI**-generated image **D**etectors), a benchmark dataset of 216k diverse and transferable adversarial images. The proposed dataset is created starting from images generated by four different text-to-image models (to promote diversity) and optimizing attacks against an ensemble of seven state-of-the-art detectors (to promote transferability). Extensive experiments show that our methodology generates adversarial images that transfer to unseen detectors, degrading the performance of every detector outside the attack ensemble that we evaluate, and provides a practical, worst-case robustness screen: failure on RAID is a strong indicator that a detector is not adversarially robust.

1. Introduction

In recent years, generative artificial intelligence has evolved from a mere research topic to a vast collection of commonly available tools. While improving tasks such as image super-resolution (Wang et al., 2024) and dataset augmentation (Rahat et al., 2025), this rapid development poses considerable risks of misuse leading to misinformation spread, privacy violation, and identity theft (Ricker et al., 2024a). This underscores the urgent need for detection methods that generalize and keep up with ever-evolving image generation technology. To mitigate the harmful consequences of AI-generated images (AIGIs), a variety of detection approaches have been proposed in the literature (Wang et al., 2020; Guo et al., 2022; Ojha et al., 2023; Chen et al., 2024; Baraldi et al., 2024; Wang et al., 2023). Based on the reported near-perfect accuracy, the problem of detecting AIGIs appears to have already been solved. However, evaluations are typically conducted without considering 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 AIGI detectors' performance (Gragnaniello et al., 2021; Xu et al., 2024; Cocchi et al., 2023).

Another important factor that the majority of existing work has neglected is the *adversarial robustness* of detectors. In fact, attackers can introduce carefully crafted, often imperceptible modifications to AIGIs and systematically bypass detection without degrading image quality. Take, for instance, Fig. 1, an AIGI that was spread across social media outlets in 2023 (Oremus et al., 2023) that, despite being quickly debunked as fake, still had an impact on the stock market. We can detect such an image as AI-generated with an off-the-shelf detector, but when we modify the image with adversarial perturbations, it can evade detection by that detector and others, with increased effectiveness against those with a similar architecture (Mavali et al., 2024). The adversarial robustness of AIGI detectors is often not investigated, partly because generating adversarial images requires significant effort. Due to the many attack algorithms, hyperparameters, and the required technical knowledge, conducting a comprehensive robustness analysis is not straightforward.

☆ This article is part of a Special issue entitled: 'YCVIU_MF-DISM' published in Computer Vision and Image Understanding.

* Corresponding author.
E-mail address: hicham.eddoubi@unica.it (H. Eddoubi).

https://doi.org/10.1016/j.cviu.2026.105002
Received 15 May 2026; Received in revised form 30 September 2026; Accepted 7 November 2026
Available online 12 November 2026
1077-3142/© 2026 The Author(s). Published by Elsevier Inc. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).

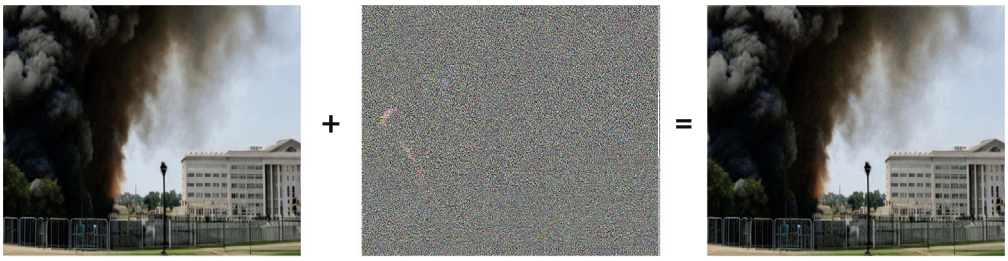


Fig. 1. Adversarial attack on the AIGI detector provided by *sightengine*. Successful evasion is shown by adding adversarial noise computed on the detector by Corvi et al. (2023). We perform an extended evaluation on commercial detectors in Appendix D.

Existing work (Mavali et al., 2024) reveals the lack of robustness of AIGI detectors in white-box scenarios, where the malicious actor has access to the model's architecture and training parameters, and in black-box scenarios, where the attacker's knowledge is limited. However, we note that the attacks used for the evaluation remain restricted to techniques that increase their success and transferability. In this work, we extend this concept to bridge this evaluation gap by providing a dataset to assess the adversarial robustness of AIGI detectors. In particular, we propose **RAID** (**R**obust evaluation of **AI**-generated image **D**etectors), a large-scale benchmark dataset of *diverse* and *transferable* adversarial images. To promote diversity and transferability, we use images generated by four different text-to-image models and optimize attacks against an ensemble of state-of-the-art AIGI detectors with different architectures. As we demonstrate experimentally, testing detection performance on RAID provides a fast and necessary robustness screen: a detector that fails on RAID is very unlikely to be robust to stronger, detector-specific attacks. Our benchmark across seven recently proposed detectors shows that the current landscape of AIGI detection is neither expansive nor reliable enough for widespread real-world adoption without proper adversarial robustness.

Contributions. In summary, we make the following contributions:

- We create RAID, the first benchmark dataset of transferable adversarial images to test the adversarial robustness of state-of-the-art AIGI detectors, released in three splits generated with different ensemble attacks (PGD, APGD, and CWA), enabling reproducible and comparable assessment.
- We conduct a large-scale study across diverse AIGI detectors, showing that black-box adversarial images evade detection in both open-source and commercial closed-source detectors, and that no current detector is reliably robust to transferable adversarial attacks.
- We show that the RAID's adversarial images provide a fast robustness screen without white-box access to the target model, making it a practical drop-in robustness test for new detectors: failure on RAID implies a lack of adversarial robustness.

2. Adversarial problem definition

AIGI Detection. We consider the problem of detecting AIGIs as a binary classification task. Let $\boldsymbol{x} \in [0,1]^d$ denote an input image and $y \in \{0,1\}$ be the corresponding ground-truth label, *i.e.*, 1 for AIGIs and 0 for real images. An AIGI detector can be modeled as a function $f_\theta : [0,1]^d \to [0,1]$ parametrized by $\boldsymbol{\theta}$, whose output represent the probability that $\boldsymbol{x}$ is AI-generated. The predicted label $\hat{y}$ can then be obtained by applying a fixed threshold τ, such that a sample is considered an AIGI if $f_\theta(\boldsymbol{x}) \geq \tau$, and a real image otherwise. This formulation is shared by most state-of-the-art AIGI detectors, independently of their underlying architecture.

Crafting Adversarial Examples. Given an image with ground-truth label $y \in \{0,1\}$, the goal of an adversarial attack is to produce a perturbation that makes the detector misclassify it, while ensuring that the perturbation does not significantly corrupt the image. We therefore cast evasion as an *untargeted* attack: the attacker seeks a perturbation $\boldsymbol{\delta} \in \mathbb{R}^d$ maximizing a loss representing the attacker's objective, solving the following optimization problem (Biggio et al., 2013; Szegedy et al., 2014):

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

where ϵ bounds the perturbation size in the ℓ_∞ norm, and $\mathcal{L}$, in our case, is the cross-entropy $\mathcal{L}_{CE}$ with respect to the true class y. The optimization problem in Eq. (1) is commonly solved using gradient-based techniques, which assume full access to the detector's parameters and gradients. This corresponds to a *white-box* attack setting. In practice, however, such access is rarely available, as deployed detection systems are typically proprietary or accessible only through black-box interfaces. Despite this limitation, adversarial perturbations crafted against a specific model can still induce misclassification on other, unseen detectors. This phenomenon, known as *transferability*, enables *black-box* attacks in which the adversary optimizes perturbations on surrogate models and deploys them against a target detector. While transferability is generally weaker across heterogeneous architectures, it can be substantially improved by jointly attacking multiple models (Liu et al., 2017; Dong et al., 2018).

Promoting Transferability. To increase transferability, we extend Eq. (1) to an ensemble of M detectors with parameters $\{\boldsymbol{\theta}_m\}_{m=1}^M$ with an ensemble loss computed as the average of the individual detectors' losses:

$$\arg\max_{\boldsymbol{\delta}:\|\boldsymbol{\delta}\|_\infty \leq \epsilon} \frac{1}{M}\sum_{m=1}^{M} \mathcal{L}_{CE}(\boldsymbol{x}+\boldsymbol{\delta}, y, \boldsymbol{\theta}_m)\ . \tag{2}$$

We solve Eq. (2) with (untargeted) Projected Gradient Descent (PGD) (Madry et al., 2018). Algorithm 1 describes the complete procedure in detail. By simultaneously optimizing against a diverse set of detectors, ensemble-based attacks encourage perturbations that exploit weaknesses shared across models, rather than detector-specific artifacts. As a result, such perturbations exhibit significantly higher transferability to unseen detectors. From an evaluation perspective, *if a detector fails under strong transferable adversarial attacks, it is in practice unlikely to be robust against more powerful attacks*, including fully white-box ones. This observation motivates the construction of transferable adversarial benchmarks for robustness assessment.

Threat Model. To avoid ambiguity between the experiments we run in the paper and the released dataset, we make the access assumptions explicit. We distinguish two settings: *white-box*, where the attacker has full access to the target detector's architecture, parameters, and gradients; and *black-box* (*transfer*), where the target detector is excluded from the surrogate ensemble and is unknown to the attacker. Table 1 maps every experiment in the paper to its setting, the surrogate ensemble used, and whether the corresponding images are part of the public RAID release. The first six listed experiments are target-specific diagnostics and are

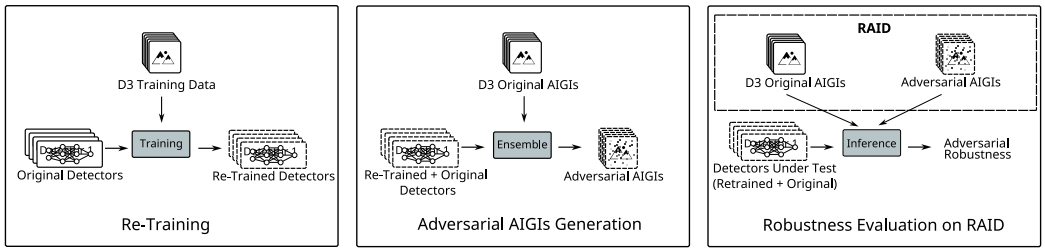


Fig. 2. A diagram of the three experimental pipelines used in the paper for generating the RAID dataset. Left: We keep the top three-performing detectors intact while retraining the rest. Middle: We generate adversarial images optimizing on an ensemble of detectors. Right: We evaluate the baseline detectors on the RAID images.

Table 1
Threat model for each experiment: access setting, surrogate ensemble, and whether the images belong to the public RAID release. N is the seven-detector pool of Section 3.2.

Experiment	Setting	Surrogate ensemble	In release?
PGD Hyperparameters Ablation (Fig. 3)	black-box	$N \setminus \{\text{target}\}$	no
Single-detector attack (Tables 4, 5, C.11, diag.)	white-box	target only	no
Cross-model transfer (Tables 4, 5, C.11, off-diag.)	black-box	one source detector	no
Leave-one-out ensemble (Tables 4, 5, C.11, last row)	black-box	$N \setminus \{\text{target}\}$	no
Additional generators (Table 6)	black-box	$N \setminus \{\text{target}\}$	no
Other ensemble attacks (Table B.9)	black-box	$N \setminus \{\text{target}\}$	no
RAID – extra detectors (Tables 7–8)	black-box	full N	yes
Commercial detectors Appendix D	black-box	full N	yes

not part of the release, while the released RAID dataset is generated with the full seven-detector ensemble and is evaluated only on detectors outside that ensemble.

3. The RAID benchmark

This section describes how we constructed our dataset of transferable adversarial images[1]. First, we detail the design choices that led to the creation of RAID, namely the choice of the initial dataset and the pool of AIGI detectors involved in the study. We then illustrate our three-step pipeline for creating and using the dataset, shown in Fig. 2.

3.1. Source dataset

RAID is built upon the D^3 dataset (Baraldi et al., 2024). In total, D^3 consists of 11.5M images. It is constructed from 2.3M real images taken from the LAION-400M (Schuhmann et al., 2021) dataset. Using the corresponding caption as prompts, AIGIs are generated from four open-source text-to-image models: Stable Diffusion v1.4 (Rombach et al., 2022), Stable Diffusion v2.1 (Rombach et al., 2022), Stable Diffusion XL (Podell et al., 2024), and DeepFloyd IF (DeepFloyd Lab at StabilityAI, 2023). For details on data selection and prompt engineering, we refer to the original publication (Baraldi et al., 2024). It should be noted that each AIGI is post-processed such that the image format and compression strength match that of the corresponding image from the distribution of real images. 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 focus on diffusion-based text-to-image models, as opposed to GAN-based models, because of their higher image quality and wider use in practice (Dhariwal and Nichol, 2021; Bie et al., 2025; Lin et al., 2024).

3.2. Detection ensemble

The key requirement for our benchmark is that adversarial perturbations are *transferable* (*i.e.*, they are effective against unseen detectors). To achieve that, we use an ensemble of a diverse set of seven different detectors:

Wang et al. (2020). It employs a ResNet-50 (He et al., 2016) trained on real and generated images from ProGAN (Karras et al., 2018). During training, extensive data augmentation is used to account for different image processing and increase generalization.

Corvi et al. (2023). This architecture is adapted from Gragnaniello et al. (2021) and is a modification of the detector by Wang et al. (2020), that avoids downsampling in the first layer to preserve high-frequency features and applies stronger augmentation.

Ojha et al. (2023). It uses a pre-trained vision foundation model (CLIP Radford et al., 2021) as a feature extractor to avoid overfitting on a particular class of generated images and, thus, improve generalization. A linear classifier is added and trained on top of the 768-dimensional feature vector.

Koutlis and Papadopoulos (2024). It extends Ojha et al. (2023) by also using intermediate encoder blocks from CLIP. The resulting features are weighted using a learnable projection network, followed by a classification head.

Cavia et al. (2024). Instead of classifying the entire image at once, this detector operates on 9×9 patches. The architecture is based on ResNet-50 (He et al., 2016) but uses 1×1 convolutions to limit the receptive field.

Chen et al. (2024). It is a training paradigm that can increase the generalizability of AIGI detectors by leveraging diffusion models to reconstruct semantically similar images containing detectable artifacts. The authors provide pre-trained detectors based on two backbones, ConvNeXt (Liu et al., 2022) and CLIP (Radford et al., 2021).

3.3. Creating RAID

Fig. 2 illustrates the overall workflow adopted in this paper for constructing and using RAID. The pipeline consists of three main stages: (i) re-training a subset of detectors to ensure reliable clean performance, (ii) generating transferable adversarial images via ensemble-based attacks, and (iii) evaluating the adversarial robustness of detectors on the resulting RAID benchmark. In the following, we describe each step in detail and provide practical guidance on assessing the robustness of AIGI detectors using RAID.

[1] Dataset: https://huggingface.co/datasets/aimagelab/RAID; code: https://github.com/pralab/RAID.

Table 2
Evaluation of each model on a subset of the clean D^3 test set. † refers to detectors trained on D^3. F1-score computed with the AI-generated class as positive.

	F1	Acc	AUROC
Corvi et al. (2023) - Co23	0.75	0.81	0.91
Ojha et al. (2023) - O23	0.33	0.29	0.68
Wang et al. (2020) - W20	0.21	0.02	0.46
Cavia et al. (2024) - Ca24	0.20	0.07	0.45
Wang et al. (2020)† - W20†	0.99	0.99	0.99
Corvi et al. (2023)† - Co23†	0.98	0.99	0.99
Ojha et al. (2023)† - O23†	0.92	0.95	0.95
Chen et al. (2024) (ConvNext) - Ch24CN	0.86	0.91	0.94
Koutlis and Papadopoulos (2024) - K24	0.74	0.81	0.88
Cavia et al. (2024)† - Ca24†)	0.55	0.63	0.84
Chen et al. (2024) (CLIP) - Ch24C	0.81	0.89	0.73

Re-training. The first step of the pipeline (Fig. 2, *left*) aims to ensure that the detectors used to generate adversarial perturbations exhibit adequate performance on clean images. As shown in Table 2, the published versions of most detectors do 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 images 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 3.1, following the original authors' training instructions. We note that the subset of re-trained detectors was chosen to maintain a practical balance between the diversity of the training data and performance on clean samples, thereby avoiding dataset bias. As shown in Section 4.4, retraining ensures sufficient performance and, subsequently, the generation of effective adversarial images.

Adversarial Images Generation. In the second step (Fig. 2, *middle*), we generate adversarial images by optimizing perturbations against an ensemble of detectors. To create the initial dataset, we use the same procedure as in Baraldi et al. (2024) to generate AIGIs from 4800 new real images, obtaining $4800 \times 4 = 19{,}200$ AIGIs (four generators per real image) alongside the 4800 real images, for 24,000 unique images in total. Following the untargeted formulation of Section 2, we perturb images of *both* classes: each of the 24,000 images is attacked at each budget ϵ (real images are themselves perturbed, and no image is duplicated across splits). The released benchmark comprises three attack splits, generated with PGD, APGD (Croce and Hein, 2020), and CWA (Chen et al., 2023) using the same ensemble, budgets, and hyperparameters Appendix B, each covering the three budgets $\epsilon \in \{\frac{8}{255}, \frac{16}{255}, \frac{32}{255}\}$. Each split contains $24{,}000 \times 3 = 72{,}000$ adversarial images – of which $19{,}200 \times 3 = 57{,}600$ are adversarial AI-generated images and $4800 \times 3 = 14{,}400$ are adversarial real images – for 216,000 adversarial images in total, released together with the 24,000 original images. The PGD split is the main split of RAID, on which all experiments in this paper are based. The dataset documentation reports the exact class distribution of every split.

Robustness Evaluation on RAID. The final step of the pipeline (Fig. 2, *right*) consists of evaluating AIGI detectors on the RAID benchmark. This evaluation additionally serves as a practical tutorial on how RAID should be used to evaluate the robustness of new AIGI detectors. Given a detector under test, robustness is assessed by measuring its detection performance on adversarial images. Since adversarial perturbations are generated using an ensemble that excludes the detector under evaluation, this setting naturally corresponds to a transferable black-box attack. A significant performance drop on RAID indicates high vulnerability. *Importantly, failure under this evaluation suggests that the detector is unlikely to remain reliable in real-world adversarial settings, where attackers may have access to even stronger, detector-specific attacks.* We emphasize that RAID provides a *one-directional* (necessary, not sufficient) screen: failure on RAID implies a lack of adversarial robustness, whereas maintaining high performance on RAID is a favorable but not conclusive signal, since a detector could still be vulnerable to a detector-specific white-box attack or to a different transfer distribution. RAID should therefore be reported as a strong lower bound on adversarial vulnerability rather than as a complete characterization of robustness.

4. Experimental analysis

In this section, we show how RAID can be used to estimate the adversarial robustness of AIGI 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 adversarial images. We subsequently analyze the transferability of white-box attacks and compare them to our ensemble attack, showing the effectiveness of RAID as an easy-to-use tool for assessing adversarial robustness.

4.1. Experimental setup

Detectors. In addition to the seven detectors described in Section 3.2, we build four additional detectors by training a linear classifier on top of four pre-trained visual backbones, using the D^3 train set. In particular, we use two different versions of DINO (Caron et al., 2021; Darcet et al., 2024) 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. Additionally, we adapt CoDE (Baraldi et al., 2024) for this analysis. Specifically, we use the feature extractor trained with a contrastive loss tailored for AIGI detection and train a linear classification layer on top. For DINO and ViT-Tiny, we use the transformation pipeline introduced by Ojha et al. (2023) for training and evaluation, whereas for CoDE, we use the transformation strategy described in its original implementation. This setup enables consistent comparisons across different architectures and training paradigms.

Dataset. As noted in Section 3, we consider the images of the D^3 test data set. Following the approach used for D^3, only images labeled *safe* in the LAION metadata are considered real. This ensures that the generated images also adhere to this safeguard. To this end, samples depicting NSFW images have been excluded. *All images are center-cropped* to 256×256 rather than resized, before applying the detector-specific pre-processing and evaluation: down-sampling smooths local pixel correlations and can distort the low-level artifacts that detectors rely on, whereas cropping preserves them (O. Li et al., 2025; Z. Li et al., 2025). A fixed crop size also provides consistent input dimensions for efficient batch processing, and the perturbation is optimized over the entire 256×256 crop, which constitutes the released image. During dataset construction, we save the adversarial images as PNGs, avoiding lossy formats (*e.g.*, JPEG), since we only consider the worst-case adversarial scenario in which no post-processing operations, which could reduce the attack's transferability (Guo et al., 2018), are performed.

Evaluation Protocol. Initially, we use all the 24,000 unperturbed images to evaluate the detectors' performance in the absence of adversarial attacks. We then select a subset of 1000 test images per class (5000 in total) and use them to conduct a thorough evaluation of the detector's performance with and without adversarial attacks across several settings. 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. Furthermore, we investigate the cross-model transferability of adversarial images computed against each detector to the other six detectors, examining whether adversarial perturbations designed to fool one detector can also fool the others.

Next, we apply the ensemble attack approach described in Section 2 in a *leave-one-out manner*, *i.e.*, excluding the assessed detector from the attacked ensemble. This ensemble approach enables us to evaluate both the transferability of ensemble attacks and the detector's performance in a black-box scenario, where access to the targeted detector architecture or parameters is unavailable. Finally, to assess transfer to genuinely unseen detectors whose architectures and training data are independent of our ensemble, we evaluate five recent state-of-the-art detectors: Effort (Z. Yan et al., 2025), AIDE (S. Yan et al., 2025), OmniAID (Guo et al., 2026), DDA (Chen et al., 2025), and DDD (Yang et al., 2025), directly on the pre-computed RAID dataset.

Additionally, to demonstrate the adaptability of our method and broaden the evaluation scope, we apply our attack methodology to a set of 11 additional text-to-image models: Kandinsky 2.1, Kandinsky 2.2, Kandinsky 3 (Razzhigaev et al., 2023), aMUSEd (Patil et al., 2024), PixArt-α (Chen et al., 2024b), SD-XL-Turbo (Podell et al., 2024), SAG (Hong et al., 2023), unCLIP (Ramesh et al., 2022), MultiDiffusion (Bar-Tal et al., 2023), and the latent consistency models LCM (Luo et al., 2023a) and LCM-XL (Luo et al., 2023b). The selected models span a diverse range of architectures to test the generalizability of our approach under different conditions. We generate synthetic counterparts to the 4800 real images in our benchmark, yielding a total of 52,800 AIGIs, and apply our leave-one-out ensemble attack to a subset of 200 images per generator, together with 200 real images, varying the perturbation budget.

Finally, we construct the RAID dataset, as shown in Fig. 2, by running the adversarial attack on the ensemble of seven detectors using the entire D^3 test set, and then evaluate its effectiveness on nine additional detectors.

Attack Setting. The attack optimization is performed with the Projected Gradient Descent (PGD) adversarial attack (Madry et al., 2018)[2] (Algorithm 1), a well-established iterative approach to generate adversarial examples. We employ PGD with iterations fixed at 10, as a longer optimization tends to overfit the surrogate models and reduce transferability (Demontis et al., 2019; Xiong et al., 2022; Chen et al., 2023), zero initialization, and a budget-dependent step size, selected with the ablation of Section 4.2: $\alpha = 0.01$ for $\epsilon = \frac{8}{255}$, and $\alpha = 0.03$ for $\epsilon = \frac{16}{255}$ and $\epsilon = \frac{32}{255}$.[3]

Experimental Environment. During the detectors' retraining phase, 4 A100 GPUs are used in a distributed data-parallel setup, requiring each up to 18 h until convergence. In contrast, the ensemble attack is run on a single A100 GPU and takes 8 h. The evaluation script is lightweight and takes less than 1 h for each detector in a non-distributed setting.

Metrics. We restrict the adversarial analysis to two threshold-free measures: the AUROC and the *attack success rate* (ASR), defined as the relative AUROC drop with respect to the clean images, $\text{ASR} = 100 \cdot \frac{\text{AUROC}_{\text{clean}} - \text{AUROC}_{\text{adv}}}{\text{AUROC}_{\text{clean}}}$. Accuracy and F1-score depend on the decision threshold, which is not calibrated for our data. We therefore report them only for the clean evaluation. As untargeted attacks perturb both classes, an AUROC of 0.5 corresponds to random ranking, and 0.0 corresponds to inverted ranking. For commercial detectors (Appendix D), we additionally report accuracy with threshold fixed at 0.5.

4.2. Ablation on PGD hyperparameters

We select the attack configuration with an ablation of the leave-one-out ensemble attack on a subset of 1000 test images (200 real images and 200 AIGIs from each of the four generators). For every budget, we vary the number of iterations $K \in \{10, 20, 30\}$ and the step size $\alpha \in \{0.01, 0.03, 0.05\}$, with three seeded random starts per configuration, and report the ASR averaged over the random starts and the seven held-out detectors in Fig. 3. As a benign-distortion control, we add Gaussian noise with standard deviation ϵ, clipped to $[-\epsilon, \epsilon]$. With $K = 10$, the best step size depends on the budget: $\alpha = 0.01$ at $\epsilon = \frac{8}{255}$ (ASR 64.8%, against 57.6% and 40.1% for $\alpha = 0.03$ and 0.05, whose steps are comparable to or larger than the budget itself), and $\alpha = 0.03$ at $\frac{16}{255}$ (78.7%, against 69.8% and 73.2%). At $\frac{32}{255}$, $\alpha = 0.03$ and 0.05 are within one point (86.8% and 87.5%), and we keep $\alpha = 0.03$. Increasing K from 10 to 30 changes the ASR of the selected configurations by at most 4.6 points for three times the computation, so we use $K = 10$ for all budgets. Gaussian noise of the same ℓ_∞ norm yields an ASR of only 9.8%–15.9%, against 64.8%–86.8% for the attack, confirming that the degradation stems from the adversarial optimization rather than from generic image corruption.

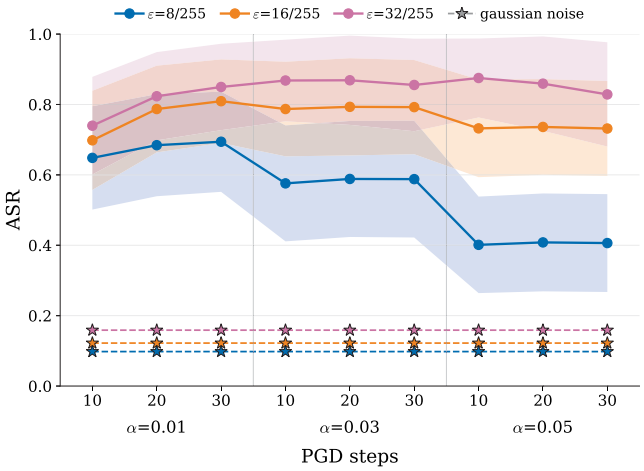


Fig. 3. ASR of the leave-one-out ensemble PGD attack for every number of iterations K (ticks) and step size α (groups), for each budget ϵ (colors). Dashed lines: Gaussian noise clipped to the same ℓ_∞ budget. Mean over 3 random starts and the 7 held-out detectors, on 1000 images; shaded areas: ±1 standard deviation.

Table 3
LPIPS and SSIM between the adversarial images of the released RAID dataset and their clean counterparts, per perturbation budget ϵ. Mean ± standard deviation over all images.

LPIPS ↓			SSIM ↑		
$\epsilon = \frac{8}{255}$	$\epsilon = \frac{16}{255}$	$\epsilon = \frac{32}{255}$	$\epsilon = \frac{8}{255}$	$\epsilon = \frac{16}{255}$	$\epsilon = \frac{32}{255}$
$0.09_{\pm 0.07}$	$0.24_{\pm 0.13}$	$0.40_{\pm 0.16}$	$0.85_{\pm 0.07}$	$0.63_{\pm 0.12}$	$0.47_{\pm 0.13}$

4.3. Perceptibility of adversarial perturbations

An ℓ_∞ bound alone does not quantify how much a perturbation distorts an image. We therefore measure LPIPS (Zhang et al., 2018) and SSIM (Wang et al., 2004) between each adversarial image and its clean counterpart, for the released RAID dataset (Table 3). Both are image-distance measures, LPIPS being learned: a lower value supports a claim of lower measured distortion, not of human imperceptibility or of a perturbation below a visibility threshold. Accordingly, we do not describe the RAID perturbations as imperceptible: $\epsilon = \frac{8}{255}$ yields the lowest measured distortion, whereas $\frac{16}{255}$ and $\frac{32}{255}$ introduce clearly measurable structural changes and should be read as bounded, worst-case perturbations.

[2] The released RAID additionally includes splits generated with APGD (Croce and Hein, 2020) and CWA (Chen et al., 2023), using the same ensemble, budgets, and hyperparameters; all experiments in this paper use the PGD split. Appendix B compares the three attacks.

[3] The results for the perturbation $\epsilon = \frac{8}{255}$ are reported in Appendix C.

Table 4
White-box and black-box adversarial robustness evaluation (AUROC/ASR) on PGD with $\epsilon = 16/255$ perturbation. Adversarial attacks are ran against the model in each row and evaluated on models in each column. † refers to detectors trained on the D^3 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	**0.00/100.0**	0.70/0.6	0.85/5.9	0.90/7.6	0.68/18.5	0.87/4.3	0.94/4.2
Ch24C	0.58/27.8	**0.00/100.0**	0.49/46.3	0.93/3.6	0.04/95.4	0.08/91.6	0.94/4.3
Ch24CN	0.52/34.5	0.59/17.0	**0.03/96.9**	0.89/7.7	0.61/26.9	0.88/3.4	0.95/3.6
Co23	0.32/59.4	0.65/7.5	0.08/90.8	**0.00/100.0**	0.49/41.4	0.82/9.9	0.01/99.3
K24	0.54/31.8	0.08/89.3	0.43/52.4	0.93/4.4	**0.05/93.7**	0.20/77.6	0.94/3.8
O23	0.57/28.6	0.03/95.3	0.71/21.6	0.93/4.2	0.06/92.4	**0.00/100.0**	0.93/5.5
W20	0.35/56.5	0.69/2.2	0.31/65.9	0.06/93.3	0.54/35.2	0.84/7.3	**0.00/100.0**
N− model	0.36/54.7	0.02/97.6	0.20/78.2	0.13/87.0	0.05/94.3	0.12/86.5	0.26/73.5

Table 5
White-box and black-box adversarial robustness evaluation (AUROC/ASR) on PGD with $\epsilon = 32/255$ perturbation. Adversarial attacks are ran against the model in each row and evaluated on models in each column. † refers to detectors trained on the D^3 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	**0.00/100.0**	0.74/−4.5	0.76/15.4	0.88/9.3	0.72/13.4	0.87/4.0	0.92/6.0
Ch24C	0.54/32.3	**0.00/100.0**	0.35/61.1	0.92/4.8	0.01/98.7	0.03/97.1	0.92/6.3
Ch24CN	0.48/39.7	0.55/21.9	**0.01/99.2**	0.87/9.7	0.57/31.8	0.86/4.9	0.94/4.3
Co23	0.26/67.2	0.60/15.5	0.01/98.4	**0.00/100.0**	0.44/47.6	0.70/23.4	0.00/100.0
K24	0.51/35.5	0.08/88.2	0.38/57.8	0.92/5.2	**0.07/91.6**	0.22/75.4	0.93/5.2
O23	0.54/32.0	0.01/98.7	0.51/43.5	0.91/5.9	0.03/96.4	**0.00/100.0**	0.90/8.5
W20	0.28/64.4	0.67/5.8	0.19/79.3	0.01/98.5	0.45/46.1	0.77/14.7	**0.00/100.0**
N− model	0.31/61.5	0.01/98.6	0.15/83.0	0.05/94.8	0.01/98.5	0.10/88.9	0.04/95.8

4.4. Experimental results

Initial Evaluation. Prior to evaluating the adversarial robustness of the considered detectors, we first evaluate their performance on the D^3 test set. As already noted in Section 3.3, the initial performance evaluation of the detectors with the provided weights in Table 2 shows mixed results, particularly regarding F1-score and accuracy measures. We thus retrain a selected subset of detectors on the D^3 training set: Cavia et al. (2024), Corvi et al. (2023), Ojha et al. (2023), and Wang et al. (2020), while we retain the same weights for the rest of the detectors (Koutlis and Papadopoulos (2024), Chen et al. (2024) (ConvNext), and Chen et al. (2024) (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 D^3 dataset. After retraining, we observe improvements across all metrics for the four detectors, except for Cavia et al. (2024), which shows limited gains. As the empirical results reported in the following show, however, high performance on unperturbed images does not imply greater robustness to adversarial attacks.

White-box Attacks. White-box attacks on individual detectors reveal a general lack of adversarial robustness. At $\epsilon = \frac{16}{255}$ (Table 4), the AUROC drops to 0.00 for five of the seven detectors and to 0.03 and 0.05 for Chen et al. (2024) (ConvNeXt) and Koutlis and Papadopoulos (2024), *i.e.*, an ASR of at least 93.7%; at $\epsilon = \frac{32}{255}$ (Table 5), the ASR is at least 91.6%. Since both classes are perturbed, AUROC values close to 0 indicate a fully inverted ranking rather than chance-level behavior.

Cross-model Transferability. Adversarial images generated via white-box attacks against each detector were applied to the other six detectors, revealing that transfer from a single detector largely follows architecture families (Table 4). Perturbations crafted on a CLIP-based detector (Ch24C, K24, O23) transfer strongly to the other CLIP-based detectors, with the lowest ASR being 77.6% for perturbations crafted on K24 and evaluated against O23. Conversely, they barely affect the CNN-based Co23 and W20 (ASR $\leq 5.5\%$, AUROC ≥ 0.93). Perturbations crafted on Co23 or W20, in turn, transfer strongly to each other (ASR 99.3% and 93.3%) and to Ch24CN (90.8% and 65.9%), moderately to K24 (41.4% and 35.2%), and only weakly to Ch24C and O23 (ASR $\leq$ 9.9%). Perturbations crafted on Ch24CN and on the patch-based Ca24 transfer weakly to all other detectors (ASR $\leq 34.5\%$ and $\leq 18.5\%$, respectively). Consequently, every single-source attack leaves at least one target almost unaffected (lowest off-diagonal ASR $\leq 7.5\%$ in every row), so that a single surrogate is effective only if it shares the architecture family of the target.

Leave-one-out Ensemble Attack. In this setting, we find that the ensemble attack transfers strongly to every held-out detector, regardless of its architecture family, providing generalizability with respect to unseen detectors. Looking at Table 4 (last row), the AUROC of the held-out detectors drops to between 0.02 and 0.36 (ASR 54.7%–97.6%). Even the two detectors Corvi et al. (2023) and Wang et al. (2020), which were barely affected by perturbations crafted on CLIP-based detectors, drop to an AUROC of 0.13 and 0.26, respectively (from 0.99 and 0.99). This drop in performance is further highlighted in the last row of Table 5, where we use a higher perturbation budget, with the AUROC dropping to between 0.01 and 0.31 (ASR 61.5%–98.6%), and to 0.05 and 0.04 for the two detectors, respectively. The ensemble does not always outperform the best single surrogate of the same family, but it does not require knowing the architecture family of the target, which is the property RAID relies on. Cavia et al. (2024) is consistently the most resistant held-out detector (ASR 54.7% and 61.5%).

These results motivate the creation of our RAID dataset, which comprises adversarial images generated via attacks on an ensemble of state-of-the-art detectors. Our benchmark dataset enables fast, standardized evaluation of new detectors against strong, transferable perturbations, facilitating assessment of their robustness to adversarial attacks.

Generalizability to AIGIs from Other Generators. We apply the leave-one-out ensemble attack to the images of the 11 additional generators and report AUROC and ASR averaged over all of them in Table 6. While most detectors achieve high clean AUROC (0.74–0.99), the ASR

Table 6
Evaluation of our attack methodology on AIGIs from additional generators. For each detector, we report AUROC and ASR on images from 11 generative models, under the *N − model* ensemble attack (all models excluding the evaluated one).

Metric	Perturbation	Ca24†	Ch24C	Ch24CN	Co23†	K24	O23†	W20†
AUROC	Clean	0.77	0.74	0.97	0.99	0.89	0.96	0.99
	8/255	0.46	0.06	0.43	0.32	0.09	0.24	0.46
	16/255	0.46	0.02	0.23	0.20	0.04	0.15	0.29
	32/255	0.40	0.01	0.19	0.08	0.02	0.12	0.06
ASR (%)	8/255	39.5	91.9	55.5	67.4	89.5	75.4	53.5
	16/255	39.8	97.1	75.8	79.9	95.5	84.6	70.8
	32/255	48.2	98.9	79.9	91.7	97.4	87.1	94.4

Table 7
Adversarial robustness evaluation of the four trained baseline classifiers. All detectors are trained on the D^3 training set. RAID refers to the dataset of the adversarial images saved as raw tensors, while RAID Img refers to the released dataset of adversarial images saved as PNG images. ASR (%) is the relative AUROC drop with respect to the clean D^3 test set.

		DINOv2		DINOv2-Reg		ViT-T		CoDE	
Dataset	Perturbation	AUROC	ASR	AUROC	ASR	AUROC	ASR	AUROC	ASR
D^3	Clean	0.85	–	0.85	–	0.82	–	0.89	–
RAID	8/255	0.76	10.1	0.78	8.9	0.74	9.0	0.72	19.2
RAID	16/255	0.73	14.7	0.75	12.6	0.70	14.2	0.65	27.7
RAID	32/255	0.67	21.8	0.70	17.8	0.66	19.7	0.59	34.2
RAID Img	8/255	0.76	10.2	0.78	9.0	0.74	9.2	0.72	19.2
RAID Img	16/255	0.73	14.7	0.74	12.7	0.70	14.3	0.65	27.8
RAID Img	32/255	0.66	21.9	0.70	17.8	0.66	19.7	0.59	34.4

Table 8
Adversarial robustness evaluation of five independently trained detectors, not part of the attack ensemble. AIDE, Effort, DDD (Yang et al., 2025), and OmniAID use their released checkpoints trained on GenImage (SD1.4); DDA uses its released checkpoint trained on DDA-aligned MSCOCO. RAID refers to the adversarial images saved as raw tensors, while RAID Img refers to the released dataset of adversarial images saved as PNGs. ASR (%) is the relative AUROC drop with respect to the clean D^3 test set.

		AIDE		Effort		DDD		OmniAID		DDA	
Dataset	Perturbation	AUROC	ASR	AUROC	ASR	AUROC	ASR	AUROC	ASR	AUROC	ASR
D^3	Clean	0.78	–	0.70	–	0.89	–	0.78	–	0.93	–
RAID	8/255	0.66	15.5	0.55	20.5	0.71	19.5	0.29	63.1	0.52	44.2
RAID	16/255	0.60	22.1	0.56	20.0	0.67	24.9	0.15	81.1	0.34	62.9
RAID	32/255	0.50	35.1	0.48	30.7	0.57	35.1	0.09	87.9	0.16	82.5
RAID Img	8/255	0.66	15.6	0.56	20.0	0.71	19.5	0.29	63.4	0.52	44.2
RAID Img	16/255	0.60	22.3	0.56	20.2	0.66	25.0	0.15	81.4	0.34	63.1
RAID Img	32/255	0.50	35.2	0.48	30.8	0.57	35.2	0.09	88.1	0.16	82.6

reaches 39.8%–97.1% at $\epsilon = \frac{16}{255}$ and 48.2%–98.9% at $\frac{32}{255}$, with Cavia et al. (2024) again the most resistant. This further demonstrates that our proposed approach is effective across multiple generator families, thereby extending its applicability.

RAID Evaluation. While the previous evaluation provides a good assessment of RAID, as only one detector is missing from the ensemble, we additionally evaluate the released RAID dataset, generated with the full seven-detector ensemble, on detectors that took no part in its construction. We use the four additional baseline detectors detailed in Section 4.1, tested against the adversarial examples saved as tensors (RAID) and as PNG images (RAID Img). As reported in Table 7, RAID reduces their AUROC at every budget; at $\epsilon = \frac{16}{255}$, the ASR is 12.6%–27.7% and the AUROC remains between 0.65 and 0.75. Saving the perturbations as PNG images does not reduce their effectiveness: across all nine detectors and budgets, the AUROC on RAID Img differs from RAID by at most 0.01 and the ASR by at most 0.5 points (Tables 7 and 8). We further evaluate five independently trained detectors with their released checkpoints (Table 8): AIDE (S. Yan et al., 2025), Effort (Z. Yan et al., 2025), DDD (Yang et al., 2025), OmniAID (Guo et al., 2026), and DDA (Chen et al., 2025). At $\epsilon = \frac{16}{255}$, AIDE, Effort, and DDD are moderately affected (ASR 20.0%–24.9%), while OmniAID and DDA are strongly affected, with their AUROC dropping from 0.78 and 0.93 to 0.15 and 0.34. Overall, even recent state-of-the-art detectors, trained independently of RAID and never seen by the attack, remain vulnerable to our black-box ensemble attack, with an ASR of 20.0%–81.1% at $\epsilon = \frac{16}{255}$, which shows that RAID provides a meaningful robustness test beyond the detectors used to construct it.

5. Related work

Generative Image Modeling. Learning to generate new samples from a high-dimensional data distribution, such as natural images, is not trivial. Several approaches, such as autoregressive models (van den Oord et al., 2016b,a), VAEs (Kingma and Welling, 2014; Rezende et al., 2014), and GANs (Goodfellow et al., 2014; Zhu et al., 2017; Choi et al., 2018; Karras et al., 2018, 2019, 2020, 2021), have been proposed. While it has been shown that DMs (Sohl-Dickstein et al., 2015; Ho et al., 2020; Dhariwal and Nichol, 2021) can surpass GANs in terms of quality, the costly iterative denoising process has prevented the generation of high-resolution images. As a remedy, LDMs (Rombach et al., 2022) perform the diffusion and denoising process in the latent space of a pre-trained VAE (Kingma and Welling, 2014). Moreover, with the addition of cross-attention layers based on U-Net (Ronneberger et al., 2015), the generation can be controlled by textual prompts (Nichol et al., 2022; Ramesh et al., 2022; Saharia et al., 2022). Recent models have significantly advanced the resolution of generated images (up to 4k) (Chen et al., 2024a; Zhang et al., 2025), improved prompt following, and human preference (Esser et al., 2024; Xie et al., 2025).

AIGI Detection Methods. Most early approaches for detecting AIGIs exploit visible flaws like differently colored irises (Matern et al., 2019)

or irregular pupil shapes (Guo et al., 2022). Since such imperfections are becoming less likely, several methods rely on imperceptible artifacts instead. Such features include model-specific fingerprints (Yu et al., 2019) or unnatural frequency patterns (Frank et al., 2020; Durall et al., 2020; Dzanic et al., 2020). Instead of using handcrafted features, Wang et al. (2020) demonstrate that training a ResNet-50 (He et al., 2016) on real and generated images from ProGAN (Karras et al., 2018), paired with strong data augmentation, suffices to detect images generated by several other GANs. Subsequent works propose improved model architectures (Gragnaniello et al., 2021) or learning paradigms (Cozzolino et al., 2021; Mandelli et al., 2022; Chen et al., 2024), with a particular focus on generalization (Chai et al., 2020; Liu et al., 2024). A promising direction is the use of foundation models as feature extractors to avoid overfitting on images generated by a single class of models (Ojha et al., 2023; Cozzolino et al., 2024; Koutlis and Papadopoulos, 2024). Recently, some works have explored unique characteristics of DMs for detection, like frequency artifacts (Corvi et al., 2023) or features obtained by inverting the diffusion process (Wang et al., 2023; Ricker et al., 2024b; Cazenavette et al., 2024).

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* (Zhu et al., 2023) is a large-scale dataset comprising 1.35 million generated images based on ImageNet (Russakovsky et al., 2015). *WildFake* (Hong and Zhang, 2024) features images generated by GANs, DMs, and other generative models, sourced from platforms such as Civitai to cover a broad range of content and styles. S. Yan et al. (2025) also collects images from popular image-sharing websites. However, their *Chameleon* dataset consists of images misclassified by human annotators, making it particularly challenging. Furthermore, *Deepfake-Eval-2024* (Chandra et al., 2025) contains deepfake videos, audio, and images that circulated on social media and deepfake detection platforms in 2024.

Adversarial Robustness of AIGI Detectors. The vulnerability of AIGI detectors to adversarial images was first explored by Carlini and Farid (2020). 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 (Neekhara et al., 2021; Hussain et al., 2021; Mavali et al., 2024) as well as possible defenses (Gandhi and Jain, 2020). De Rosa et al. (2024) study the robustness of CLIP-based detectors, finding that adversarial images computed for CNN-based classifiers are not easily transferable and vice versa. Besides the addition of adversarial noise, it has been shown that AIGI detectors can also be attacked by applying natural degradations (e.g., local brightness changes) (Hou et al., 2023) or by removing generator-specific artifacts (Dong et al., 2022; Wesselkamp et al., 2022). Other attacks leverage image generators themselves to perform semantic adversarial attacks, which adversarially manipulate a particular attribute of an image (Meng et al., 2024) that can even be controlled through a text prompt (Liu et al., 2023; Abdullah et al., 2024). A further line of work designs *learned* anti-forensic attacks specifically against generated-image detectors, either by training an auxiliary network to remove GAN fingerprints via a contrastive objective (Lou et al., 2023) or a deep-dithering generator (Xie et al., 2022), or by exploiting spatial–frequency feature importance (Zhu et al., 2026). These attacks are typically tailored to a single detector and to GAN artifacts, and are therefore complementary to, but not directly usable within, the differentiable ensemble-transfer setting we adopt (see Appendix B). Most closely related to our imperceptibility analysis, Musso et al. (2026) craft transferable adversarial perturbations against frozen foundation-model (CLIP/Swin/DINOv2) synthetic-image detectors and report perceptual-quality metrics; unlike that work, which attacks individual frozen backbones, RAID releases a reusable, multi-generator, multi-detector transferable benchmark.

Benchmarking Transfer Attacks and Real-World Robustness. Orthogonal to adversarial robustness, a growing body of work benchmarks the effect of real-world dissemination and processing on AIGI detection. TrueFake (Dell'Anna et al., 2025) evaluates detectors on last-generation fakes shared through social platforms such as Facebook, X, and Telegram, while QuAD (Guillaro et al., 2026) studies repeated recompression, resizing, cropping, and online near-duplicates. These works target *natural* degradations and are complementary to RAID, which isolates the *adversarial* worst case. On the attack side, Brau et al. (2025) systematically benchmarks ensemble-based black-box transfer attacks across many surrogate–target configurations, and finds that a simple ensemble-averaged objective performs on par with or better than more elaborate ensemble attacks; our own comparison (Appendix B) shows that this holds only partly in our setting.

6. Discussion

Adversarial robustness should always be evaluated when proposing new AIGI detection methods, as in current state-of-the-art 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 robustness to these operations is indeed important, introducing a *malicious actor* that leverages carefully crafted adversarial noise can evade detection by most methods, as highlighted in our work. Nonetheless, the absence of a standard benchmark for comparison with detectors further contributes to the lack of evaluation against adversarial attacks in AIGI detection. As such, we introduce RAID to address this gap in the current literature and provide a more comprehensive solution for evaluating the adversarial robustness of AIGI detectors. Through the direct evaluation on RAID, first on the original AIGIs and subsequently on their adversarial counterparts with the chosen perturbation budget ϵ, the authors can obtain metrics reflecting the detector's performance, in addition to its adversarial robustness. We recommend authors evaluating on RAID to report results on both the clean and adversarial splits, using $\epsilon = 16/255$ and $\epsilon = 32/255$.

As a control, we compare the adversarial degradation against a benign-noise baseline (Gaussian noise matched by ℓ_∞ norm; Section 4.2), which yields an ASR of at most 15.9%, against 64.8%–86.8% for the attack, confirming that the observed drops stem from adversarial optimization rather than generic image corruption.

However, our method is not without its limitations. One key challenge is that a dataset like RAID requires frequent updates to remain relevant as new generative models emerge, and these models must be incorporated into our proposed ensemble attacks to maintain effectiveness. The dynamic nature of generative models demands a proactive approach to maintaining RAID's robustness over time, and we plan to address this by periodically extending the generator ensemble and releasing updated dataset versions.

Furthermore, using PGD in our attack algorithm requires that all ensemble detectors be end-to-end differentiable, which constrains our detector selection. Although several attacks do not have this constraint, their inclusion in the model ensemble requires a significant modification to the method.

Additionally, our perturbations are not designed to be robust to post-processing operations, and due to the inherent restrictions on the input sizes of some detector architectures, all images are center-cropped to 256×256 before the attack; the perturbation covers this entire crop, which constitutes the released image. Four of the seven ensemble detectors further apply a 224×224 center crop as their first preprocessing step, meaning the image border is perturbed only through the gradients of the remaining three detectors, which process the full crop.

Several directions remain for future work: extending the attack to detectors with non-differentiable operations, including additional attack methods, and designing perturbations robust to post-processing transformations.

7. Conclusions

We introduce RAID, the first benchmark dataset of transferable adversarial images for evaluating the robustness of AIGI detectors. 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 state-of-the-art detectors, which are typically tested against natural post-processing but not against adversarial perturbations, to which they remain highly vulnerable. RAID addresses this gap by providing a simple, reproducible worst-case benchmark for evaluating adversarial robustness, enabling detection models to be screened under strong, transferable adversarial conditions.

CRediT authorship contribution statement

Hicham Eddoubi: Writing – review & editing, Writing – original draft, Software, Methodology, Formal analysis, Data curation. **Jonas Ricker:** Writing – review & editing, Writing – original draft, Software, Methodology, Formal analysis, Data curation. **Federico Cocchi:** Writing – review & editing, Software, Resources. **Angelo Sotgiu:** Writing – review & editing, Writing – original draft, Validation, Software, Methodology, Formal analysis, Data curation. **Maura Pintor:** Writing – review & editing, Supervision, Methodology. **Marcella Cornia:** Writing – review & editing, Validation, Supervision, Resources. **Lorenzo Baraldi:** Writing – review & editing, Software, Resources. **Asja Fischer:** Writing – review & editing, Validation, Supervision. **Battista Biggio:** Writing – review & editing, Validation, Supervision.

Ethics statement

Research on the adversarial robustness of classifiers always poses a dual-use risk. Our proposed attack methodology could be misused by an adversary to craft AIGIs that bypass detectors. However, our work does not introduce any new attack methodologies that were not previously available to adversaries. Therefore, we argue that our benchmark dataset primarily benefits defenders, enabling developers to evaluate and improve detectors' robustness.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This work was carried out while H. Eddoubi was enrolled in the Italian National Doctorate on Artificial Intelligence run by Sapienza University of Rome in collaboration with the University of Cagliari. We acknowledge the CINECA award under the ISCRA initiative, for the availability of high-performance computing resources and support. This research has been also partially supported by the Horizon Europe projects ELSA (GA no. 101070617), Sec4AI4Sec (GA no. 101120393), and CoEvolution (GA no. 101168560); and by SERICS (PE00000014) and FAIR (PE00000013) under the MUR NRRP funded by the EU-NGEU.

Algorithm 1 Ensemble PGD for crafting transferable adversarial examples.

Require: Image $\boldsymbol{x} \in [0,1]^d$; detectors $\{f_{\theta_m}\}_{m=1}^{M}$ with preprocessing $\{\pi_m\}_{m=1}^{M}$; budget ϵ; step size α; iterations K
Ensure: adversarial image $\boldsymbol{x}+\boldsymbol{\delta}$ with $\|\boldsymbol{\delta}\|_\infty \le \epsilon$

$\boldsymbol{x} \leftarrow \text{center_crop}(\boldsymbol{x}, 256\times256)$ ▷ center crop for batch processing
$\boldsymbol{\delta} \leftarrow \mathbf{0}$ ▷ zero initialization
for $k = 1$ **to** K **do**
  $\boldsymbol{g} \leftarrow \frac{1}{M}\sum_{m=1}^{M} \nabla_{\boldsymbol{\delta}} \mathcal{L}_{CE}\left(\pi_m(\boldsymbol{x}+\boldsymbol{\delta}),\, y, \theta_m\right)$ ▷ aggregate gradients
  $\boldsymbol{\delta} \leftarrow \boldsymbol{\delta} + \alpha\ \text{sign}(\boldsymbol{g})$ ▷ gradient ascent step
  $\boldsymbol{\delta} \leftarrow \text{clip}_{[-\epsilon,\epsilon]}(\boldsymbol{\delta})$ ▷ project onto ℓ_∞ ball
  $\boldsymbol{\delta} \leftarrow \text{clip}_{[0,1]}(\boldsymbol{x}+\boldsymbol{\delta}) - \boldsymbol{x}$ ▷ keep image in valid range
end for
return $\boldsymbol{x}+\boldsymbol{\delta}$

Appendix A. Detailed ensemble attack algorithm

In Alg. 1, we report the complete procedure used to craft the adversarial images, making explicit the per-detector preprocessing, the aggregation of per-model gradients, the signed update, the projection onto the ℓ_∞ ball, and the clipping that keeps the image in the valid range $[0,1]^d$. The perturbation is initialized to zero ($\boldsymbol{\delta} \leftarrow \mathbf{0}$); we report a random-restart ablation in Section 4.2. Because the ensemble detectors require different input resolutions, each detector m applies its own preprocessing π_m (*e.g.*, resizing and normalization); the perturbation is optimized over the entire 256×256 image: Ojha et al. (2023), Koutlis and Papadopoulos (2024), and both variants of Chen et al. (2024) center-crop it to 224×224, whereas Cavia et al. (2024), Corvi et al. (2023), and Wang et al. (2020) process the full image.

Appendix B. Other adversarial attacks: Discussion and results

To complement our PGD-based ensemble attack, we compare it with two widely used alternatives in the same leave-one-out setting: APGD (Croce and Hein, 2020), extended to the ensemble by averaging the per-model cross-entropy losses, and the Common Weakness Attack (CWA) (Chen et al., 2023), which is explicitly designed for transferability from model ensembles. All three attacks are untargeted, perturb both classes, start from zero, and use the budgets, number of iterations, and step sizes of RAID (Section 4.1). Table B.9 reports the ASR on each held-out detector.

APGD performs on par with PGD, with a mean ASR 1.2–1.5 points higher at every budget. CWA transfers more strongly at the smaller budgets, with a mean ASR of 85.7% and 89.6% at $\epsilon = 8/255$ and $16/255$, against 70.6% and 81.7% for PGD. At $\epsilon = 32/255$, the three attacks are within 1.2 points on average (88.7%–89.9%), and CWA is weaker than PGD on Ojha et al. (2023) (70.9% vs. 88.9%).

Brau et al. (2025) benchmark 13 ensemble-based black-box transfer attacks (including momentum-, variance-reduction-, weight-averaging-, and query-refinement-based methods such as CWA, SVRE, LGV, MBA, and SASD-WS) across 17 surrogate–target settings, and find that a simple query-free averaging of per-model losses (exactly the objective of Eq. (2)) performs on par with or better than the more elaborate ensemble attacks. Our comparison shows that, in our setting, the ranking depends on the budget: CWA transfers more strongly at $\epsilon \le 16/255$, where it outperforms PGD on all seven held-out detectors, whereas at $\epsilon = 32/255$ the three attacks perform similarly on average (88.7%–89.9%) and PGD is stronger on three detectors, most notably (Ojha et al., 2023) (88.9% vs. 70.9%). We therefore release, alongside the main PGD split on which all experiments in this paper are based, two additional splits computed with APGD and CWA.

We further note that the specialized anti-forensic attacks against AIGI detectors suggested in the literature are, by design, not well suited

Table B.9
Comparison with other ensemble attacks in the leave-one-out setting: ASR (%) on each held-out detector. Attacks are computed on 1000 images per class.

ϵ	Attack	Ca24	Ch24C	Ch24CN	Co23	K24	O23	W20	Mean
8/255	PGD (Ours)	48.3	93.6	61.0	73.4	89.1	75.3	53.8	70.6
	APGD	47.6	94.6	62.5	74.1	89.6	75.8	60.0	72.0
	CWA	51.7	96.5	82.4	85.9	98.8	90.6	93.7	85.7
16/255	PGD (Ours)	54.7	97.6	78.2	87.0	94.3	86.5	73.5	81.7
	APGD	53.7	97.8	81.5	89.1	95.6	87.0	77.7	83.2
	CWA	59.9	98.1	92.5	94.1	99.6	87.5	95.0	89.6
32/255	PGD (Ours)	61.5	98.6	83.0	94.8	98.5	88.9	95.8	88.7
	APGD	64.3	98.8	86.1	94.4	98.7	89.7	97.0	89.9
	CWA	74.1	98.4	93.9	92.9	99.8	70.9	96.3	89.5

Table C.10
Leave-one-out ensemble PGD attack with a post-processing operation applied to the adversarial images before classification: ASR (%) on each held-out detector, on 1000 images per class. ASR is computed with respect to the clean AUROC without post-processing. Resize and random crop are followed by upsampling back to 256 × 256; random flip is horizontal with probability 0.5.

ϵ	Post-processing	Ca24†	Ch24C	Ch24CN	Co23†	K24	O23†	W20†
8/255	No transform	48.3	93.6	61.0	73.4	89.1	75.3	53.8
	Resize 64 × 64	6.2	9.3	24.5	56.2	38.0	11.4	26.5
	Random crop 64 × 64	9.6	23.9	26.5	40.7	44.2	27.7	31.4
	Random flip	46.9	87.2	61.2	72.9	82.3	66.8	53.4
	JPEG ($q = 75$)	50.6	29.1	54.6	55.2	49.3	21.4	47.5
16/255	No transform	54.7	97.6	78.2	87.0	94.3	86.5	73.5
	Resize 64 × 64	3.2	4.4	34.2	60.9	52.5	13.0	27.4
	Random crop 64 × 64	7.7	22.7	34.7	43.7	42.6	31.9	35.2
	Random flip	54.2	94.1	78.0	86.4	91.2	78.6	72.6
	JPEG ($q = 75$)	55.6	55.3	73.9	83.2	67.2	45.2	72.8
32/255	No transform	61.5	98.6	83.0	94.8	98.5	88.9	95.8
	Resize 64 × 64	4.1	7.5	45.8	69.3	59.5	14.9	39.7
	Random crop 64 × 64	15.3	25.8	40.7	43.6	41.1	40.5	35.3
	Random flip	60.6	97.2	82.6	94.4	96.8	82.7	95.3
	JPEG ($q = 75$)	64.2	69.1	59.0	93.1	84.7	65.2	92.2

Table C.11
White-box and black-box adversarial robustness evaluation (AUROC/ASR) on PGD with $\epsilon = 8/255$ perturbation. Adversarial attacks are ran against the model in each row and evaluated on models in each column. † refers to detectors trained on the D^3 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	**0.00/100.0**	0.68/3.5	0.89/1.7	0.92/4.9	0.65/22.0	0.88/2.8	0.96/2.4
Ch24C	0.62/22.5	**0.00/100.0**	0.66/27.5	0.94/2.9	0.04/95.2	0.09/90.5	0.96/2.4
Ch24CN	0.57/28.0	0.60/15.0	**0.01/99.2**	0.91/6.3	0.71/14.2	0.88/2.9	0.95/2.7
Co23	0.36/55.1	0.72/−2.2	0.12/86.7	**0.00/100.0**	0.56/33.2	0.86/4.9	0.01/99.1
K24	0.58/27.8	0.10/85.2	0.61/32.9	0.93/4.4	**0.03/96.1**	0.24/73.3	0.95/3.3
O23	0.60/24.3	0.03/96.2	0.85/6.3	0.94/3.2	0.05/93.7	**0.00/100.0**	0.95/3.1
W20	0.41/48.8	0.72/−1.9	0.39/56.6	0.09/90.7	0.68/18.0	0.88/3.0	**0.00/100.0**
$N-$ *model*	0.41/48.3	0.05/93.6	0.35/61.0	0.26/73.4	0.09/89.1	0.22/75.3	0.45/53.8

to constructing a transferable ensemble benchmark. The contrastive-perturbation attack (Lou et al., 2023) and the deep-dithering approach (Xie et al., 2022) are *learned* anti-forensic methods that train an auxiliary network (an encoder–decoder and a VQ-VAE-based simulator, respectively) against a *single* target detector, and are oriented toward removing *GAN* fingerprints rather than the diffusion-model traces relevant to our data; the dual-domain feature-importance attack (Zhu et al., 2026) is likewise a single-detector spatial–frequency attack. None of these admit the differentiable, ensemble-averaged objective of Eq. (2) that produces a fixed, transferable dataset across heterogeneous generators and detectors, which is the same differentiability constraint we discuss in Section 6.

Appendix C. Additional adversarial robustness evaluations

Sensitivity to Image Transformations. In our experiments, the attack aims to provide an approximate estimation of the adversarial robustness in worst-case settings. For this reason, we do not consider optimizing the adversarial perturbation to make it robust against any transformations. Nevertheless, this could be achieved by introducing data augmentations during attack optimization and leveraging the Expectation Over Transformation (EOT) algorithm (Athalye et al., 2018), although we expect that optimizing against transformations would decrease the adversarial images' transferability to other never-before-seen detectors. Likely, in real scenarios, an attacker would perform a stronger, yet more costly, attack tailored to the target detector that may resist image transformations better. In this context, our dataset can quickly and costless provide valuable insights: if a model is not robust with respect to it, it will almost certainly be vulnerable to more realistic threat models (Carlini et al., 2019).

Nonetheless, to provide additional insights, we apply several post-processing operations to the adversarial images of the leave-one-out setting (1000 images per class) before classification, and compute the ASR with respect to the clean AUROC without post-processing (Table C.10). At $\epsilon = \frac{16}{255}$, a random horizontal flip barely affects the attack (ASR change of at most 7.9 points). JPEG compression ($q = 75$) mostly preserves the attack on the CNN-based detectors (ASR

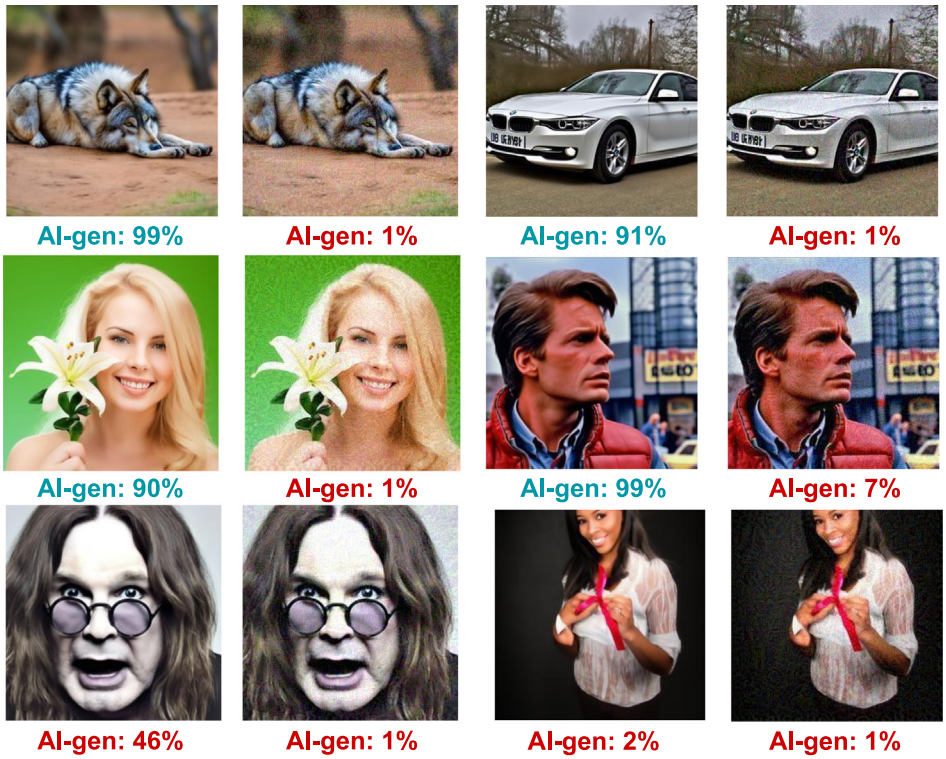


Fig. D.4. Detection scores returned by Sightengine detector on a subset of clean and adversarial images. Higher scores indicate higher confidence that the image is AI-generated.

Table C.12
Leave-one-out ensemble PGD attack on each held-out detector at $\epsilon = 16/255$ ($K = 10$, $\alpha = 0.03$): mean ± standard deviation over three seeded random starts, on 1000 test images. ASR (%) is the relative AUROC drop with respect to the clean images.

Metric	Ca24†	Ch24C	Ch24CN	Co23†	K24	O23†	W20†
AUROC	$0.35_{\pm 0.00}$	$0.02_{\pm 0.00}$	$0.24_{\pm 0.00}$	$0.16_{\pm 0.00}$	$0.06_{\pm 0.01}$	$0.16_{\pm 0.01}$	$0.33_{\pm 0.01}$
ASR	$56.1_{\pm 0.6}$	$96.8_{\pm 0.1}$	$72.9_{\pm 0.4}$	$83.8_{\pm 0.2}$	$92.7_{\pm 1.0}$	$82.5_{\pm 1.4}$	$66.2_{\pm 0.5}$

Table D.13
Performance on commercial AIGI detectors. We report mean detector scores on AIGIs, with adversarial variants evaluated on RAID examples. Metrics on real images are also included.

Detector	Score	Score (Real)	Adversarial Score	Acc	Acc (Real)	Adversarial Acc
Sightengine	0.65	0.99	0.01	0.66	1.00	0.00
HIVE	0.59	0.99	0.45	0.60	1.00	0.44

55.6%–83.2%) but weakens it on the CLIP-based ones (45.2%–67.2%, against 86.5%–97.6% without post-processing). Downscaling to 64×64 and random 64×64 crops, which also remove most of the image content, reduce the ASR to 3.2%–60.9% and 7.7%–43.7%, respectively. Larger budgets are more robust to compression (JPEG ASR 59.0%–93.1% at $\epsilon = \frac{32}{255}$).

Results with $\epsilon = 8/255$. We evaluate the detectors under a smaller perturbation budget of $\epsilon = 8/255$, commonly used in white-box evaluations but less relevant in the transfer-based scenario, where larger perturbations (*e.g.*, $\epsilon = 16/255$ or $\epsilon = 32/255$) are typically required to achieve effective cross-model fooling. The results reported in Table C.11 confirm that. In the white-box setting, the AUROC drops to at most 0.03 (ASR $\geq$ 96.1%). In the leave-one-out setting, transfer is weaker than at larger budgets, with an ASR of 48.3%–93.6%, against 54.7%–97.6% and 61.5%–98.6% at $\epsilon = \frac{16}{255}$ and $\frac{32}{255}$ (Table 4 and Table 5).

Variance Across Random Starts. We assess the variability of the leave-one-out ensemble attack with respect to its initialization by repeating it with three seeded random starts at $\epsilon = \frac{16}{255}$ with the selected configuration in Section 4.2 ($K = 10$, $\alpha = 0.03$), on the 1000 images of the ablation. Table C.12 reports the mean and standard deviation of the AUROC and ASR on each held-out detector: the standard deviation is at most 0.01 in AUROC and 1.4 points in ASR, so the results do not depend on the random start.

Appendix D. Evaluation on commercial detectors

We use two commercial AIGI detectors provided by Sightengine[4] and HIVE[5] to evaluate RAID in real-world settings. We report the results of a set of 50 real images and 50 pairs of clean and adversarial images in Table D.13. While these results are preliminary due to the dataset size, they still suggest that RAID remains effective in assessing the adversarial robustness of detectors deployed in commercial detection APIs. We compute Accuracy using a 0.5 threshold on the detectors'

[4] https://dashboard.sightengine.com/ai-image-detection
[5] https://hivemoderation.com/ai-generated-content-detection

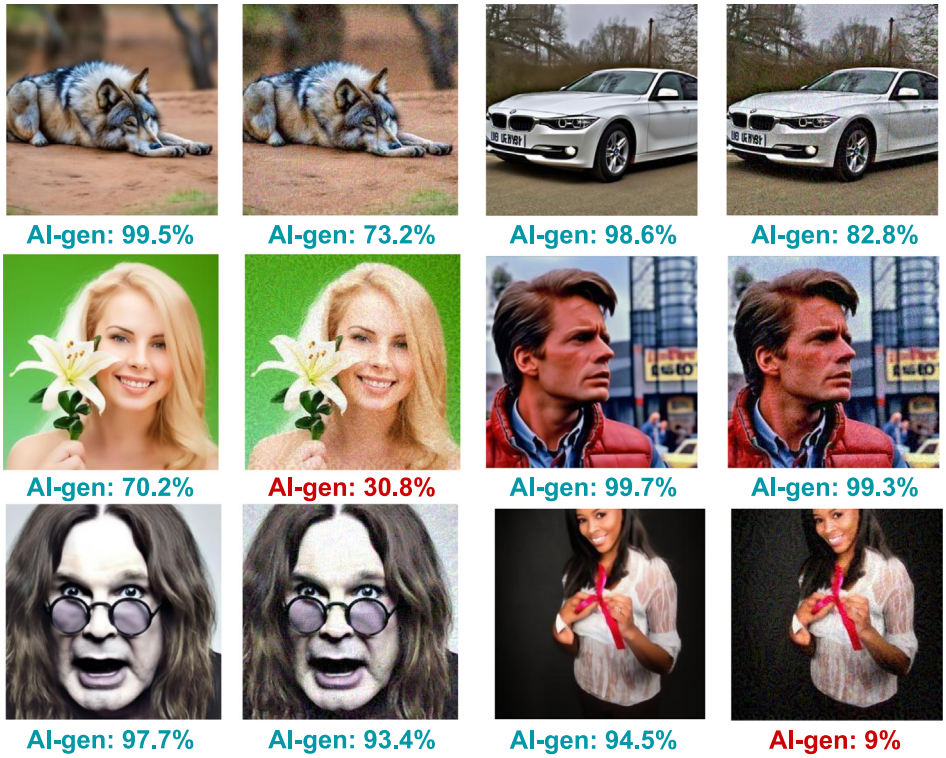


Fig. D.5. Detection scores returned by HIVE Moderation detector on a subset of clean and adversarial images. Higher scores indicate higher confidence that the image is AI-generated.

confidence scores. We also provide a few examples of the detection scores returned by commercial detectors on randomly selected images in Figs. D.4 and D.5.

Data availability

We release our evaluation code at https://github.com/pralab/RAID and dataset at https://huggingface.co/datasets/aimagelab/RAID.

References

Abdullah, S.M., Cheruvu, A., Kanchi, S., Chung, T., Gao, P., Jadliwala, M., Viswanath, B., 2024. An Analysis of Recent Advances in Deepfake Image Detection in an Evolving Threat Landscape. In: SP.

Athalye, A., Engstrom, L., Ilyas, A., Kwok, K., 2018. Synthesizing robust adversarial examples. In: ICML.

Bar-Tal, O., Yariv, L., Lipman, Y., Dekel, T., 2023. MultiDiffusion: Fusing diffusion paths for controlled image generation. In: International Conference on Machine Learning. ICML.

Baraldi, L., Cocchi, F., Cornia, M., Baraldi, L., Nicolosi, A., Cucchiara, R., 2024. Contrasting Deepfakes Diffusion via Contrastive Learning and Global-Local Similarities. In: ECCV.

Bie, F., Yang, Y., Zhou, Z., Ghanem, A., Zhang, M., Yao, Z., Wu, X., Holmes, C., Golnari, P., Clifton, D.A., He, Y., Tao, D., Song, S.L., 2025. RenAIssance: A survey into AI text-to-image generation in the era of large model. IEEE TPAMI.

Biggio, B., Corona, I., Maiorca, D., Nelson, B., Šrndić, N., Laskov, P., Giacinto, G., Roli, F., 2013. Machine Learning and Knowledge Discovery in Databases. Springer.

Brau, F., Pintor, M., Cinà, A., Mura, R., Scionis, L., Oneto, L., Roli, F., Biggio, B., 2025. TransferBench: Benchmarking ensemble-based black-box transfer attacks. In: Belgrave, D., Zhang, C., Lin, H., Pascanu, R., Koniusz, P., Ghassemi, M., Chen, N. (Eds.), Advances in Neural Information Processing Systems. 38, Main Conference, Curran Associates, Inc., http://dx.doi.org/10.52202/085713-4665, URL: https://proceedings.neurips.cc/paper_files/paper/2025/file/cc41a858306c9bbb72127c90c1b9f5d6-Paper-Datasets_and_Benchmarks_Track.pdf.

Carlini, N., Athalye, A., Papernot, N., Brendel, W., Rauber, J., Tsipras, D., Goodfellow, I., Madry, A., Kurakin, A., 2019. On evaluating adversarial robustness. arXiv preprint arXiv:1902.06705.

Carlini, N., Farid, H., 2020. Evading Deepfake-Image Detectors with White- and Black-Box Attacks. In: CVPR Workshops.

Caron, M., Touvron, H., Misra, I., Jégou, H., Mairal, J., Bojanowski, P., Joulin, A., 2021. Emerging properties in self-supervised vision transformers. In: ICCV.

Cavia, B., Horwitz, E., Reiss, T., Hoshen, Y., 2024. Real-Time Deepfake Detection in the Real-World. arXiv preprint arXiv:2406.09398.

Cazenavette, G., Sud, A., Leung, T., Usman, B., 2024. FakeInversion: Learning to detect images from unseen text-to-image models by inverting Stable Diffusion. In: CVPR.

Chai, L., Bau, D., Lim, S.-N., Isola, P., 2020. What makes fake images detectable? Understanding properties that generalize. In: ECCV.

Chandra, N.A., Murtfeldt, R., Qiu, L., Karmakar, A., Lee, H., Tanumihardja, E., Farhat, K., Caffee, B., Paik, S., Lee, C., Choi, J., Kim, A., Etzioni, O., 2025. Deepfake-eval-2024: A multi-modal in-the-wild benchmark of deepfakes circulated in 2024. arXiv preprint arXiv:2503.02857.

Chen, J., Ge, C., Xie, E., Wu, Y., Yao, L., Ren, X., Wang, Z., Luo, P., Lu, H., Li, Z., 2024a. PIXART-Σ: Weak-to-strong training of diffusion transformer for 4K text-to-image generation. In: ECCV.

Chen, R., Xi, J., Yan, Z., Zhang, K.-Y., Wu, S., Xie, J., Chen, X., Xu, L., Guan, I., Yao, T., Ding, S., 2025. Dual data alignment makes AI-generated image detector easier generalizable. In: Advances in Neural Information Processing Systems (NeurIPS).

Chen, J., Yu, J., Ge, C., Yao, L., Xie, E., Wang, Z., Kwok, J.T., Luo, P., Lu, H., Li, Z., 2024b. PixArt-(α): Fast training of diffusion transformer for photorealistic text-to-image synthesis. In: ICLR. OpenReview.net.

Chen, B., Zeng, J., Yang, J., Yang, R., 2024. DRCT: Diffusion reconstruction contrastive training towards universal detection of diffusion generated images. In: ICML.

Chen, H., Zhang, Y., Dong, Y., Yang, X., Su, H., Zhu, J., 2023. Rethinking model ensemble in transfer-based adversarial attacks. In: ICML.

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

Cocchi, F., Baraldi, L., Poppi, S., Cornia, M., Baraldi, L., Cucchiara, R., 2023. Unveiling the impact of image transformations on deepfake detection: An experimental analysis. In: ICIAP.

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

Cozzolino, D., Gragnaniello, D., Poggi, G., Verdoliva, L., 2021. Towards universal GAN image detection. In: VCIP.

Cozzolino, D., Poggi, G., Corvi, R., Nießner, M., Verdoliva, L., 2024. Raising the Bar of AI-generated Image Detection with CLIP. In: CVPR Workshops.

Croce, F., Hein, M., 2020. Reliable evaluation of adversarial robustness with an ensemble of diverse parameter-free attacks. In: ICML.

Darcet, T., Oquab, M., Mairal, J., Bojanowski, P., 2024. Vision transformers need registers. In: ICLR.

De Rosa, V., Guillaro, F., Poggi, G., Cozzolino, D., Verdoliva, L., 2024. Exploring the adversarial robustness of CLIP for AI-generated image detection. In: WIFS.

DeepFloyd Lab at StabilityAI, 2023. 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.

Dell'Anna, S., Montibeller, A., Boato, G., 2025. TrueFake: A real world case dataset of last generation fake images also shared on social networks. In: 2025 International Joint Conference on Neural Networks. IJCNN, pp. 1–8. http://dx.doi.org/10.1109/IJCNN64981.2025.11228911.

Demontis, A., Melis, M., Pintor, M., Jagielski, M., Biggio, B., Oprea, A., Nita-Rotaru, C., Roli, F., 2019. Why do adversarial attacks transfer? Explaining transferability of evasion and poisoning attacks. In: 28th USENIX Security Symposium (USENIX Security 19). USENIX Association, Santa Clara, CA, ISBN: 978-1-939133-06-9, pp. 321–338, URL: https://www.usenix.org/conference/usenixsecurity19/presentation/demontis.

Dhariwal, P., Nichol, A., 2021. Diffusion models beat GANs on image synthesis. In: NeurIPS.

Dong, C., Kumar, A., Liu, E., 2022. Think twice before detecting GAN-generated fake images from their spectral domain imprints. In: CVPR.

Dong, Y., Liao, F., Pang, T., Su, H., Zhu, J., Hu, X., Li, J., 2018. Boosting adversarial attacks with momentum. In: CVPR.

Durall, R., Keuper, M., Keuper, J., 2020. Watch your up-convolution: CNN based generative deep neural networks are failing to reproduce spectral distributions. In: CVPR.

Dzanic, T., Shah, K., Witherden, F., 2020. Fourier Spectrum Discrepancies in Deep Network Generated Images. In: NeurIPS.

Esser, P., Kulal, S., Blattmann, A., Entezari, R., Müller, J., Saini, H., Levi, Y., Lorenz, D., Sauer, A., Boesel, F., Podell, D., Dockhorn, T., English, Z., Rombach, R., 2024. Scaling Rectified Flow Transformers for High-Resolution Image Synthesis. In: ICML.

Frank, J., Eisenhofer, T., Schönherr, L., Fischer, A., Kolossa, D., Holz, T., 2020. Leveraging Frequency Analysis for Deep Fake Image Recognition. In: ICML.

Gandhi, A., Jain, S., 2020. Adversarial Perturbations Fool Deepfake Detectors. In: IJCNN.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y., 2014. Generative adversarial nets. In: NeurIPS.

Gragnaniello, D., Cozzolino, D., Marra, F., Poggi, G., Verdoliva, L., 2021. Are GAN generated images easy to detect? A critical analysis of the state-of-the-art. In: ICME.

Guillaro, F., De Rosa, V., Cozzolino, D., Verdoliva, L., 2026. Quality-aware calibration for AI-generated image detection in the wild. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops.

Guo, H., Hu, S., Wang, X., Chang, M.-C., Lyu, S., 2022. Eyes Tell All: Irregular Pupil Shapes Reveal GAN-generated Faces. In: ICASSP.

Guo, C., Rana, M., Cisse, M., van der Maaten, L., 2018. Countering adversarial images using input transformations. In: International Conference on Learning Representations. ICLR.

Guo, Y., Ye, J., Zhang, C., Kang, H., Fu, H., He, C., Li, W., 2026. OmniAID: Decoupling semantics and artifacts for universal AI-generated image detection in the wild. In: International Conference on Machine Learning. ICML.

He, K., Zhang, X., Ren, S., Sun, J., 2016. Deep Residual Learning for Image Recognition. In: CVPR.

Ho, J., Jain, A., Abbeel, P., 2020. Denoising diffusion probabilistic models. In: NeurIPS.

Hong, S., Lee, G., Jang, W., Kim, S., 2023. Improving Sample Quality of Diffusion Models Using Self-Attention Guidance. In: ICCV.

Hong, Y., Zhang, J., 2024. WildFake: A large-scale challenging dataset for AI-generated images detection. arXiv preprint arXiv:2402.11843.

Hou, Y., Guo, Q., Huang, Y., Xie, X., Ma, L., Zhao, J., 2023. Evading DeepFake Detectors via Adversarial Statistical Consistency. In: CVPR.

Hussain, S., Neekhara, P., Jere, M., Koushanfar, F., McAuley, J., 2021. Adversarial deepfakes: Evaluating vulnerability of deepfake detectors to adversarial examples. In: WACV.

Karras, T., Aila, T., Laine, S., Lehtinen, J., 2018. Progressive growing of GANs for improved quality, stability, and variation. In: ICLR.

Karras, T., Aittala, M., Laine, S., Härkönen, E., Hellsten, J., Lehtinen, J., Aila, T., 2021. Alias-Free Generative Adversarial Networks. In: NeurIPS.

Karras, T., Laine, S., Aila, T., 2019. A Style-Based Generator Architecture for Generative Adversarial Networks. In: CVPR.

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

Kingma, D.P., Welling, M., 2014. Auto-Encoding Variational Bayes. In: ICLR.

Koutlis, C., Papadopoulos, S., 2024. Leveraging representations from intermediate encoder-blocks for synthetic image detection. In: ECCV.

Li, O., Cai, J., Hao, Y., Jiang, X., Hu, Y., Feng, F., 2025. Improving synthetic image detection towards generalization: An image transformation perspective. In: Proceedings of the ACM SIGKDD Conference on Knowledge Discovery and Data Mining. KDD.

Li, Z., Yan, J., He, Z., Zeng, K., Jiang, W., Xiong, L., Fu, Z., 2025. Is artificial intelligence generated image detection a solved problem?. arXiv preprint arXiv:2505.12335.

Lin, Z., Pathak, D., Li, B., Li, J., Xia, X., Neubig, G., Zhang, P., Ramanan, D., 2024. Evaluating text-to-visual generation with image-to-text generation. In: ECCV.

Liu, Y., Chen, X., Liu, C., Song, D., 2017. Delving into Transferable Adversarial Examples and Black-box Attacks. In: ICLR.

Liu, Z., Mao, H., Wu, C.-Y., Feichtenhofer, C., Darrell, T., Xie, S., 2022. A ConvNet for the 2020s. In: CVPR.

Liu, H., Tan, Z., Tan, C., Wei, Y., Wang, J., Zhao, Y., 2024. Forgery-Aware Adaptive Transformer for Generalizable Synthetic Image Detection. In: CVPR.

Liu, J., Wei, C., Guo, Y., Yu, H., Yuille, A., Feizi, S., Lau, C.P., Chellappa, R., 2023. Instruct2Attack: Language-Guided Semantic Adversarial Attacks. arXiv preprint arXiv:2311.15551.

Lou, Z., Cao, G., Lin, M., 2023. Black-box attack against GAN-generated image detector with contrastive perturbation. Eng. Appl. Artif. Intell. (ISSN: 0952-1976) 124, 106594. http://dx.doi.org/10.1016/j.engappai.2023.106594, URL: https://www.sciencedirect.com/science/article/pii/S0952197623007789.

Luo, S., Tan, Y., Huang, L., Li, J., Zhao, H., 2023a. Latent consistency models: Synthesizing high-resolution images with few-step inference. arXiv preprint arXiv:2310.04378.

Luo, S., Tan, Y., Patil, S., Gu, D., von Platen, P., Passos, A., Huang, L., Li, J., Zhao, H., 2023b. LCM-LoRA: A universal stable-diffusion acceleration module. arXiv preprint arXiv:2311.05556.

Madry, A., Makelov, A., Schmidt, L., Tsipras, D., Vladu, A., 2018. Towards deep learning models resistant to adversarial attacks. In: ICLR.

Mandelli, S., Bonettini, N., Bestagini, P., Tubaro, S., 2022. Detecting GAN-generated images by orthogonal training of multiple CNNs. In: ICIP.

Matern, F., Riess, C., Stamminger, M., 2019. Exploiting Visual Artifacts to Expose Deepfakes and Face Manipulations. In: WACV Workshops.

Mavali, S., Ricker, J., Pape, D., Sharma, Y., Fischer, A., Schönherr, L., 2024. Fake It Until You Break It: On the Adversarial Robustness of AI-generated Image Detectors. arXiv preprint arXiv:2410.01574.

Meng, X., Wang, L., Guo, S., Ju, L., Zhao, Q., 2024. AVA: Inconspicuous attribute variation-based adversarial attack bypassing DeepFake detection. In: SP.

Musso, C., Battocchio, J., Montibeller, A., Boato, G., 2026. Backbone is all you need: Assessing vulnerabilities of frozen foundation models in synthetic image forensics. In: Proceedings of the 2026 ACM Workshop on Information Hiding and Multimedia Security. In: IH&MMSec '26, Association for Computing Machinery, New York, NY, USA, ISBN: 9798400723766, pp. 223–228. http://dx.doi.org/10.1145/3785353.3815075.

Neekhara, P., Dolhansky, B., Bitton, J., Ferrer, C.C., 2021. Adversarial threats to DeepFake detection: A practical perspective. In: CVPR Workshops.

Nichol, A.Q., Dhariwal, P., Ramesh, A., Shyam, P., Mishkin, P., McGrew, B., Sutskever, I., Chen, M., 2022. GLIDE: towards photorealistic image generation and editing with text-guided diffusion models. In: ICML.

Ojha, U., Li, Y., Lee, Y.J., 2023. Towards universal fake image detectors that generalize across generative models. In: CVPR.

Oremus, W., Harwell, D., Armus, T., 2023. A fake image of an explosion at the Pentagon caused a stir online. URL: https://www.washingtonpost.com, The Washington Post.

Patil, S., Berman, W., Rombach, R., von Platen, P., 2024. aMUSEd: An Open MUSE Reproduction. arXiv preprint arXiv:2401.01808.

Podell, D., English, Z., Lacey, K., Blattmann, A., Dockhorn, T., Müller, J., Penna, J., Rombach, R., 2024. SDXL: Improving Latent Diffusion Models for High-Resolution Image Synthesis. In: ICLR.

Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., Sutskever, I., 2021. Learning Transferable Visual Models from Natural Language Supervision. In: ICML.

Rahat, F., Hossain, M.S., Ahmed, M.R., Jha, S.K., Ewetz, R., 2025. Data augmentation for image classification using generative AI. In: WACV.

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

Razzhigaev, A., Shakhmatov, A., Maltseva, A., Arkhipkin, V., Pavlov, I., Ryabov, I., Kuts, A., Panchenko, A., Kuznetsov, A., Dimitrov, D., 2023. Kandinsky: an Improved Text-to-Image Synthesis with Image Prior and Latent Diffusion. arXiv preprint arXiv:2310.03502.

Rezende, D.J., Mohamed, S., Wierstra, D., 2014. Stochastic Backpropagation and Approximate Inference in Deep Generative Models. In: ICML.

Ricker, J., Assenmacher, D., Holz, T., Fischer, A., Quiring, E., 2024a. AI-generated faces in the real world: A large-scale case study of Twitter profile images. In: RAID.

Ricker, J., Lukovnikov, D., Fischer, A., 2024b. AEROBLADE: Training-free detection of latent diffusion images using autoencoder reconstruction error. In: CVPR.

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

Ronneberger, O., Fischer, P., Brox, T., 2015. U-Net: Convolutional Networks for Biomedical Image Segmentation. In: MICCAI.

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A.C., Fei-Fei, L., 2015. ImageNet Large Scale Visual Recognition Challenge. IJCV 115 (3), 211–252.

Saharia, C., Chan, W., Saxena, S., Li, L., Whang, J., Denton, E.L., Ghasemipour, S.K.S., Lopes, R.G., Ayan, B.K., Salimans, T., Ho, J., Fleet, D.J., Norouzi, M., 2022. Photorealistic text-to-image diffusion models with deep language understanding. In: NeurIPS.

Schuhmann, C., Vencu, R., Beaumont, R., Kaczmarczyk, R., Mullis, C., Katta, A., Coombes, T., Jitsev, J., Komatsuzaki, A., 2021. LAION-400M: Open dataset of CLIP-filtered 400 million image-text pairs. In: NeurIPS Workshops.

Sohl-Dickstein, J., Weiss, E., Maheswaranathan, N., Ganguli, S., 2015. Deep unsupervised learning using nonequilibrium thermodynamics. In: ICML.

Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I., Fergus, R., 2014. Intriguing properties of neural networks. In: ICLR.

van den Oord, A., Kalchbrenner, N., Espeholt, L., kavukcuoglu, k., Vinyals, O., Graves, A., 2016a. Conditional image generation with PixelCNN decoders. In: NeurIPS.

van den Oord, A., Kalchbrenner, N., Kavukcuoglu, K., 2016b. Pixel Recurrent Neural Networks. In: ICML.

Wang, Z., Bao, J., Zhou, W., Wang, W., Hu, H., Chen, H., Li, H., 2023. DIRE for diffusion-generated image detection. In: ICCV.

Wang, Z., Bovik, A.C., Sheikh, H.R., Simoncelli, E.P., 2004. Image quality assessment: from error visibility to structural similarity. IEEE Trans. Image Process. 13 (4), 600–612.

Wang, S.-Y., Wang, O., Zhang, R., Owens, A., Efros, A.A., 2020. CNN-Generated Images Are Surprisingly Easy to Spot... for Now. In: CVPR.

Wang, Y., Yang, W., Chen, X., Wang, Y., Guo, L., Chau, L.-P., Liu, Z., Qiao, Y., Kot, A.C., Wen, B., 2024. SinSR: Diffusion-based image super-resolution in a single step. In: CVPR.

Wesselkamp, V., Rieck, K., Arp, D., Quiring, E., 2022. Misleading Deep-Fake Detection with GAN Fingerprints. In: IEEE Deep Learning and Security Workshop.

Xie, E., Chen, J., Chen, J., Cai, H., Tang, H., Lin, Y., Zhang, Z., Li, M., Zhu, L., Lu, Y., Han, S., 2025. SANA: Efficient High-Resolution Text-to-Image Synthesis with Linear Diffusion Transformers. In: ICLR.

Xie, H., Ni, J., Zhang, J., Zhang, W., Huang, J., 2022. Evading generated-image detectors: A deep dithering approach. Signal Process. (ISSN: 0165-1684) 197, 108558. http://dx.doi.org/10.1016/j.sigpro.2022.108558, URL: https://www.sciencedirect.com/science/article/pii/S0165168422001050.

Xiong, Y., Lin, J., Zhang, M., Hopcroft, J.E., He, K., 2022. Stochastic variance reduced ensemble adversarial attack for boosting the adversarial transferability. In: 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. CVPR, pp. 14963–14972. http://dx.doi.org/10.1109/CVPR52688.2022.01456.

Xu, H., Wang, Y., Wang, Z., Ba, Z., Liu, W., Jin, L., Weng, H., Wei, T., Ren, K., 2024. ProFake: Detecting deepfakes in the wild against quality degradation with progressive quality-adaptive learning. In: CCS.

Yan, S., Li, O., Cai, J., Hao, Y., Jiang, X., Hu, Y., Xie, W., 2025. A sanity check for AI-generated image detection. In: The Thirteenth International Conference on Learning Representations. URL: https://openreview.net/forum?id=ODRHZrkOQM.

Yan, Z., Wang, J., Jin, P., Zhang, K., Liu, C., Chen, S., Yao, T., Ding, S., Wu, B., Yuan, L., 2025. Orthogonal subspace decomposition for generalizable AI-generated image detection. In: Singh, A., Fazel, M., Hsu, D., Lacoste-Julien, S., Berkenkamp, F., Maharaj, T., Wagstaff, K., Zhu, J. (Eds.), Forty-Second International Conference on Machine Learning, ICML 2025, Vancouver, BC, Canada, July 13-19, 2025. In: Proceedings of Machine Learning Research, vol. 267, PMLR / OpenReview.net, URL: https://proceedings.mlr.press/v267/yan25b.html.

Yang, Y., Qian, Z., Zhu, Y., Russakovsky, O., Wu, Y., 2025. D3: Scaling up deepfake detection by learning from discrepancy. In: 2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition. CVPR, pp. 23850–23859. http://dx.doi.org/10.1109/CVPR52734.2025.02221.

Yu, N., Davis, L.S., Fritz, M., 2019. Attributing fake images to GANs: Learning and analyzing GAN fingerprints. In: ICCV.

Zhang, J., Huang, Q., Liu, J., Guo, X., Huang, D., 2025. Diffusion-4K: Ultra-High-Resolution Image Synthesis with Latent Diffusion Models. In: CVPR.

Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O., 2018. The unreasonable effectiveness of deep features as a perceptual metric. In: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. IEEE, pp. 586–595.

Zhu, W., Cao, G., Liu, J., Yu, L., Weng, S., 2026. Transferable dual-domain feature importance attack against AI-generated image detector. IEEE Signal Process. Lett. 33, 818–822. http://dx.doi.org/10.1109/LSP.2026.3658434.

Zhu, M., Chen, H., Yan, Q., Huang, X., Lin, G., Li, W., Tu, Z., Hu, H., Hu, J., Wang, Y., 2023. GenImage: A million-scale benchmark for detecting AI-generated image. In: NeurIPS.

Zhu, J.-Y., Park, T., Isola, P., Efros, A.A., 2017. Unpaired Image-to-Image Translation Using Cycle-Consistent Adversarial Networks. In: ICCV.