Organic or Diffused: Can We Distinguish Human Art from AI-generated Images?

Anna Yoo Jeong Ha*
University of Chicago
Chicago, IL, USA

Josephine Passananti*
University of Chicago
Chicago, IL, USA

Ronik Bhaskar
University of Chicago
Chicago, IL, USA

Shawn Shan
University of Chicago
Chicago, IL, USA

Reid Southen
Concept Artist
Detroit, MI, USA

Haitao Zheng
University of Chicago
Chicago, IL, USA

Ben Y. Zhao
University of Chicago
Chicago, IL, USA

Abstract

The advent of generative AI images has completely disrupted the art world. Distinguishing AI generated images from human art is a challenging problem whose impact is growing over time. A failure to address this problem allows bad actors to defraud individuals paying a premium for human art and companies whose stated policies forbid AI imagery. It is also critical for content owners to establish copyright, and for model trainers interested in curating training data in order to avoid potential model collapse.

There are several different approaches to distinguishing human art from AI images, including classifiers trained by supervised learning, research tools targeting diffusion models, and identification by professional artists using their knowledge of artistic techniques. In this paper, we seek to understand how well these approaches can perform against today's modern generative models in both benign and adversarial settings. We curate real human art across 7 styles, generate matching images from 5 generative models, and apply 8 detectors (5 automated detectors and 3 different human groups including 180 crowdworkers, 3800+ professional artists, and 13 expert artists experienced at detecting AI). Both Hive and expert artists do very well, but make mistakes in different ways (Hive is weaker against adversarial perturbations while Expert artists produce higher false positives). We believe these weaknesses will persist, and argue that a combination of human and automated detectors provides the best combination of accuracy and robustness.

CCS Concepts

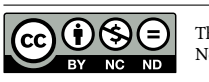
• **Security and privacy → Human and societal aspects of security and privacy**.

Keywords

AI-generated Art Detection, Automated and Human Detectors

ACM Reference Format:
Anna Yoo Jeong Ha, Josephine Passananti, Ronik Bhaskar, Shawn Shan, Reid Southen, Haitao Zheng, and Ben Y. Zhao. 2024. Organic or Diffused: Can We Distinguish Human Art from AI-generated Images?. In *Proceedings of the 2024 ACM SIGSAC Conference on Computer and Communications Security (CCS '24), October 14–18, 2024, Salt Lake City, UT, USA.* ACM, New York, NY, USA, 15 pages. https://doi.org/10.1145/3658644.3670306

*Both authors contributed equally to the paper



CCS '24, October 14–18, 2024, Salt Lake City, UT, USA.

ACM ISBN 979-8-4007-0636-3/24/10
https://doi.org/10.1145/3658644.3670306

1 Introduction

Creative expression through artwork is intrinsic to the human experience. From cave paintings by Neanderthals and Homo sapiens, to modern abstract masters like Kandinsky, Pollock and Mitchell, human art connects us by evoking shared experiences, trauma, and hopes and dreams.

Yet this might all be changing with the proliferation of output from generative image models like Midjourney, DALL-E 3, Stable Diffusion XL (SDXL) and Adobe Firefly. Given prompts as short as a single word, these models can generate glossy images that at a glance, resemble the work of a professional artist or fine photographer. As they continue to evolve, it is becoming increasingly difficult to distinguish art produced by human creatives and images produced by generative AI.

Identifying if a piece of art is human-made or AI-generated is critical for a number of reasons. First, individuals and companies are often willing to pay a premium for human art over AI content. Also, companies or creative groups have policies restricting the use of AI-generated imagery in competitions, work product, or ad campaigns. Yet recent news is littered with examples of fraud, where AI-generated images are sold as human art to individuals [48, 77] and publishers [15, 62], and used in ad campaigns or submitted to creative competitions against AI policies [51, 58]. This has resulted in numerous controversies [7, 22, 47], retracted ads and publications [14, 70], and public apologies [45, 62].

Second, identification of AI images is also a legal and regulatory issue. Commercial companies want to copyright their creative content, but the US Copyright Office has ruled that only human created artwork (or human contributions to hybrid artwork) can be copyrighted [64]. Thus businesses using generative AI might try to pass off AI images as human art to obtain copyright. Finally, multiple projects have shown evidence that both text and image AI models will degrade if only trained on output of AI models [2, 10, 63]. Thus AI model trainers also need to distinguish AI-generated images from human art for training purposes.

All of this begs the question, do we have the tools today to reliably and consistently distinguish AI-generated imagery from human-created art? There are multiple potential solutions. First, human artists are often quite good at recognizing human art, and

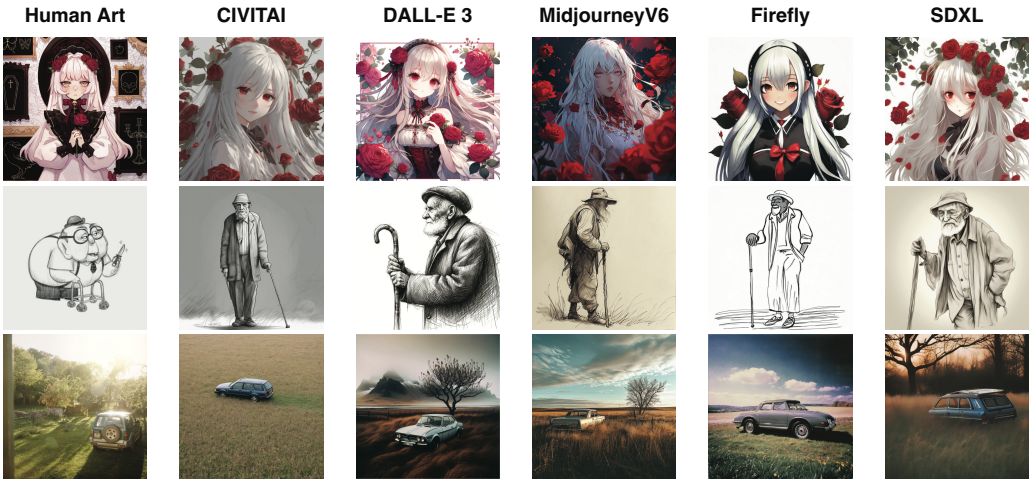


Figure 1: Samples of human art and matching images produced by generative AI models. Copyright held by respective artists, ©Kirsty (@kirue_t), ©Nguyen Viet, ©Liam Collod

some experts have demonstrated a consistent ability to detect AI images trying to pass as human art [25, 48, 58]. Alternatively, specific companies like Hive, Optic and Illuminarty have trained supervised classifiers to distinguish AI imagery from human art. Online media has raised questions on the accuracy of these detectors amid their growing impact on news media [44]. Finally, recent research results like DIRE and DE-FAKE [60, 75] suggest specific techniques to recognize images produced by diffusion models, including all of the major generative models today.

The goal of our work is to systematically and comprehensively explore the effectiveness of these detectors at distinguishing AI-generated images and human art[1]. We consider a wide range of imagery, including 7 distinct art styles, each represented by samples of human art, images generated by each of 5 generative AI models, AI images painted over by humans (hybrid), and human photography enhanced by AI. We also consider a range of "detection methods," including 5 automated tools (3 deployed classifiers Hive, Optic, Illuminarty and 2 research detectors DIRE and DE-FAKE) and 3 different populations of "human detectors" (crowdsourced non-artists, professional artists, and expert artists). Finally, we also consider adversarial scenarios, where AI-generated images are augmented with noise and adversarial perturbations with the intention of bypassing detection.

In total, we curated a dataset of 280 real human art images across 7 different styles, and 350 generated AI-based images from generative models using prompts automatically extracted from each of the human art images. One component of our study tests the efficacy of automated detectors on these images and their perturbed variants; and the other part evaluates human-based detection. The latter involves 3 separate user studies on: a) 180 crowdworkers on the Prolific platform, b) 3800+ professional artist volunteers recruited from social media artist groups, and c) 13 expert professional artists who have experience identifying generative AI images.

Our study produces a number of significant findings:

- We find that normal, non-artist users are generally unable to tell the difference between human art and AI-generated images. Professional artists are more confident and perform much better on average, and expert artists more so.
- Supervised classification does surprisingly well, and Hive outperforms all detectors (human and ML), and produces zero false positives. Unsurprisingly, accuracy seems to correlate with expected training data availability: biggest classifier (Hive) performs best; all classifiers perform the worst on Firefly, the newest of the major generative models.
- Fine tuned models that mimic human artists, and real images modified with AI upscaling posed no significant challenges to human or ML detectors.
- Adversarial perturbations did have significant impact on ML detectors such as Hive, with feature space perturbations being the most effective.
- Expert human artists perform very well in detecting AI-generated images, but often attribute mistakes and lower skill by human artists as evidence of AI images, thus producing false positives.
- A combined team of human and automated detectors provides the best combination of accuracy and robustness.

2 Background and Related Work

We begin by providing background on generative AI image models using the diffusion model architecture, and on currently available automated detectors of generative AI images. We then discuss existing work related to our study.

2.1 Generative Image Diffusion models

First introduced in 2020, diffusion models quickly replaced GANs as the state-of-the-art models for image synthesis [21]. This was quickly followed by extension to text-to-image generation [46, 54]. Later instances included multiple open-source models from Stable Diffusion [52, 56, 67–69], and commercial models from Midjourney, DALL-E, and Adobe Firefly.

Diffusion models face a growing number of social and ethical concerns. Base models require enormous amounts of training data,

[1]Unlike most prior studies, our focus is not identifying deepfake images from real photographs. Instead our focus is on determining the provenance of creative art imagery, and our inclusion of photography focuses on photographs as art, not as records of real events.

often obtained without consent through web-scraping. Midjourney trained their model using data from over 16,000 artists, the vast majority without consent [31]. Stability AI trained on datasets from LAION [56, 59], containing millions of copyrighted works. These copyright infringement issues have led to multiple class-action lawsuits [17, 35]. Even in smaller volumes, artists are finding their works being used without consent in finetuned models using techniques such as LoRA [3].

As these models continue to improve in quality, many of their users have attempted to pass AI-generated images as human art. AI-generated images have been used to win art competitions, fooling judges of digital art, photography, and book covers [27, 57, 58]. Companies that promote human art have found themselves using AI content provided to them by third-party vendors [50].

2.2 Automated AI-Image Detectors

A number of software and web services offer the ability to detect if an image is generated by generative AI image models. We group these detectors into two categories: deployed commercial detectors and research-based detectors.

Research-based detectors. These come from published research papers that often offer source code and training/testing datasets, sometimes also with pretrained models. While they lack the same public reach and influence as commercial detectors, they offer the benefit of transparency in methodology. Two such detectors are the DIRE detector [75] and DE-FAKE [60]. DIRE, or Diffusion Reconstruction Error, aims to exploit the forwards and backwards diffusion processes to identify images generated by that model. DE-FAKE uses both image-only and multimodal embeddings [53] to create a model-agnostic detector [60].

Commercial detectors. These detectors are deployed online, generally as web services with a tiered pricing model and a web-based non-transparent (black-box) detection API. They provide easy access to image classification with minimal computational requirements, Among the most popular are Hive AI Detector (Hive), Optic AI or Not (Optic), and Illuminarty [30, 33, 49]. All three services advertise free demo plans via a web UI with high accuracy, and are well covered by popular media [23, 71, 72].

Beyond DIRE and DE-FAKE, other techniques have been proposed to detect AI-generated images. Diffusion models can embed invisible watermarks into images during generation [8]. Diffusion watermarks involve manipulating the initial noise vector, creating watermarks robust to some perturbations [76, 79]. However, this approach requires modifying the diffusion model. Finally, other detection methods make use of frequency domain analysis to detect AI-generated images as outliers [5].

2.3 Related work

Detecting Deepfake Photographs. While our study focuses on distinguishing human art from AI-generated images, several prior studies have focused on human detection of deepfake photos generated by machine learning models. [9] evaluates users' ability to detect deepfakes of human faces using StyleGAN2 [36], and finds that human participants have below 65% accuracy in all experiments, even when taught how to recognize deepfakes. Similarly, [42] evaluates human detection on a set of real photos and photorealistic images from Midjourney V5. They also create a dataset of roughly two million fake images to train ML detectors. While humans misclassify 37% of images, the best-performing ML detector misclassifies 13% of the same evaluation set.

Several projects explore robust evaluation and robust training techniques to improve detection accuracy. [73] proposes training data augmentation using flipping, blurring, and JPEG compression; [65] evaluates detection under perturbations of color contrast, color saturation, blurring, and pixelation; [4] performs data augmentation with JPEG compression; and [32] uses an ensemble of detectors over the frequency domain to improve detection robustness.

Explainability in Image Identification. Some have explored explainability in detecting AI images. [55] studies distributions of GAN, diffusion, and real images, showing greater overlap between diffusion and real distributions than between GANs and real distributions. [6] creates a counterfactual, generated dataset to CIFAR-10 [37] and uses gradient heatmaps to visualize important features for detection. [16] performs forensic analysis on the frequency domain distributions of various diffusion and GAN models.

AI Images and Art. The abundance of prior work has almost entirely focused on detecting deepfakes and photorealistic images, including some very large fake image benchmarks [42, 80]. DE-FAKE briefly mentions detecting art but only tests on 50 pieces of human art and 50 AI-generated images. Deepart [74] is an art-based dataset composed of a random selection of images selected from LAION-5B [59], designed as a training dataset for a classifier to detect AI-generated art.

The most related work on this topic was presented recently at IEEE S&P 2024 [24]. Where our work focuses entirely on creative visual art, this prior study covered generative AI detection broadly across images, audio and text across Internet users in multiple countries.

3 Methodology

AI-generated images have already become exceptionally good at mimicking human art. Distinguishing these generated images from human art is critical for individual and institutional consumers, for copyright reasons, and for AI models seeking to curate their training datasets. The goal of this study is to understand how feasible this task is today given recent advances in these generative models, how and why current detectors make mistakes, and what that portends for the future.

3.1 Overview: Goals and Challenges

In tackling this multifaceted problem, our goal is to try to explore several broad questions on this topic:

(1) Are there detection methods today, human or automated, that can accurately distinguish between human art and AI-generated images? How do artists using their knowledge of art fundamentals fare against semantically-agnostic supervised classification and research tools designed specifically to detect diffusion model output?
(2) What are limitations of current detectors, and why do they make mistakes?
(3) How well do detectors perform under adversarial conditions, i.e. against images altered to avoid identification?

(4) Are there fundamental trends in performance of detection approaches, and what are implications as models continue to evolve?

As the first research study to perform a comprehensive analysis of classifying human art and AI-generated images, our most significant challenge is how to capture the numerous dimensions of this problem. Most specifically, we consider and incorporate five different dimensions into our study. We summarize these here and present further details as we describe our experimental methodology in the remainder of this section.

- **Art Styles.** Generative AI image models have a wide range of success when mimicking different styles of art. Therefore, our evaluation must cover a wide range of art styles, from anime to sketches to fine photography.
- **Sources/Types of AI-Generated Images.** Different AI models vary in their ability to mimic human art. Thus we must consider a representative set of current diffusion models, as well as more unorthodox image types such as hybrid (AI-generated images painted/altered by humans) and upscaled (human-generated photography expanded in resolution using AI models).
- **Range of Automated AI Image Detectors.** We include results of the most popular available automated detectors (Hive, Optic, Illuminarty) as well as research prototypes (DIRE, DE-FAKE).
- **Range of Human AI Image Detectors.** Humans will vary significantly in their ability to identify human art vs AI images, depending on their knowledge and experience in producing art. We consider three user groups: regular users (non-artists), professional artists, and expert artists experienced in identifying AI images.
- **Range of Adversarial Countermeasures.** Multiple factors incentivize AI model users to alter their images to escape identification as AI images. Thus our study also considers multiple types of adversarial perturbations and explore their ability to confuse different detectors (both automated and human).

3.2 Evaluating Automated Detectors

For automated software-based detectors, we consider both deployed commercial systems, as well as research-based systems. There are three well-known deployed commercial systems:

- Hive: AI content detection using supervised classification provided by thehive.ai.
- Optic: "AI or Not" is a free service (for limited queries) running a proprietary algorithm to detect AI images and audio.
- Illuminarty: an AI detection service running a proprietary algorithm including an implementation of DIRE.

For research-based systems, we selected two recent systems that had code (and models) available for testing.

- DIRE [75]: DIRE (Diffusion Reconstruction Error) pushes a test sample forwards and backwards through a diffusion pipeline and measures its changes to detect if the image came from that pipeline. DIRE has pretrained models with a public implementation.
- DE-FAKE [60]: DE-FAKE uses both image-only and multimodal embeddings to create a model-agnostic detector. We trained a model based on techniques from the paper.

We evaluate both automated detectors and human detection on the same core test dataset of images (280 human art pieces, 350 AI images, 40 hybrid images), described in more detail in Section 4. However, we also test automated detectors against a variety of adversarial perturbations including Gaussian noise, JPEG compression, adversarial perturbations, and the Glaze style mimicry protection tool [61].

3.3 Evaluating Human Detection: User Studies

Recent events have shown human artists to be exceptionally successful at identifying AI-generated imagery masquerading as human art [25, 48, 58]. Instead of looking for statistical properties of images, human artists look for inconsistencies in artistic technique, flaws in logic/composition, and other domain-specific properties that diffusion models do not understand.

Our study evaluates how well skilled artists can use their understanding of art to detect AI-generated images, by performing separate user studies for 3 separate user populations.

- Baseline Participants. We recruited 180 crowdworkers through the Prolific online crowdsourcing platform (177 completed and passed attention checks). Participants were compensated $2/10min and this group took on average 8 minutes to complete. This group included no full-time professional artists and 7 part-time artists.
- Professional Artist Volunteers. We asked for artist volunteers on social media to participate. Of more than 4000 who responded, 3803 completed the survey and passed all attention checks.
- Expert Participants. We recruited 13 high-profile professional artists known by members of the research team to have experience identifying AI imagery. These expert artists are compensated $25 for completing the initial survey and detailed feedback, and $25 more for participating in the Glaze perturbation user study.

Procedure. The basic user survey included a randomized sample set of real human artwork, hybrid images, and generative AI images. We ask participants to classify each image as human-generated, unsure, or AI-generated. We also ask if they have seen the image displayed before, and answers to previously seen images are discarded. We ask questions about their artistic expertise, what styles of art they found easier to classify than others, and factors that influenced their classification.

We also presented the expert team with a small fixed sample set of AI generated images that produced the most misclassifications in the other user studies. In an interactive chat setting, we asked the experts for detailed feedback on techniques and specific examples applicable to each of these difficult images.

3.4 Data Collection

We curated our own dataset of real human-created artwork, AI-generated images, and hybrid images. We define real images as original artwork drawn or created by human artists. AI-generated images are images that are generated using AI models like Midjourney, Stable Diffusion and DALL-E 3 from text prompts. *Hybrid images* are images that are AI-generated, retouched, and partially drawn over by humans. One of the coauthors is a professional artist with over 30 years of experience. He scanned numerous social media sites and art platforms and collected a set of 40 images whose

creators admitted they were AI-generated images altered by human artists later. We describe our data collection process in detail next.

4 Constructing the Dataset

We consider images of diverse art styles and sources. We curate a dataset consisting of four different groups of images: artworks handcrafted by human artists (§4.1), AI-generated images (§4.2), perturbed versions of human artworks and AI images (§4.3), and unusual images created by combining human and AI efforts (§4.4).

4.1 Human Artworks

Human artworks are novel creations by artists that capture their personal touch and emotions. They showcase the unique techniques, styles and perspectives of individual artists that only come from years of training and experiences. With help from the artist community, we collected artworks across 7 major art styles, including anime, cartoon, fantasy, oil/acrylic, photography, sketch, and watercolor. We recruited artist volunteers from Cara [11], a major portfolio platform dedicated to human-created art which uses filters to detect AI images and peer-based validation between artists. When recruiting volunteers, we provided artists with a detailed explanation of the study's scope and operations. We sought their consent to use their artworks in the study and offered them the option to opt out if they were not comfortable with their works being included.

Overall, we recruited 53 artists and received 280 distinct artworks, mapping to 40 images per style. For each style, we recruited 5 artists specialized for this style and each artist sent us 8 digital images of their own artworks. The only exception is the watercolor style, where we recruited 7 artists and the number of images sent per artist varies between 4 and 11.

Many artists choose to protect their intellectual property by adding digital signatures or watermarks onto the images of their original artworks. However, many AI-generated images do not have these distinctive marks. Therefore, the presence of signatures or watermarks can potentially influence the perception of human art, introducing unwanted bias and susceptibility to manipulation. To address this issue, we obtained consent from the artists to crop out any signatures or watermarks from their submitted images. In cases where the mark was too adjacent to the art subject, we communicated with the artists to request the original artwork image free of such markings. Finally, all images were cropped to achieve a square shape, with efforts made to minimize any potential loss of content. This is to maintain consistency between human artworks and AI-generated images, since the latter is of a square shape.

Ethics. Aside from obtaining consent from artists, we take great efforts to minimize exposure of human artworks to external sources. We provide details in §9.

4.2 AI-Generated Images

For AI-generated images, we take effort to cover the 7 art styles (listed above) and different AI generators. We consider the five most popular AI generators: CIVITAI [13], DALL-E 3, Adobe Firefly, MidjourneyV6, and Stable Diffusion XL (SDXL). All were the latest release at the time of submission. For each art style, we prompt each AI generator to produce 10 images, for a total of 50 images across all five generators.

Configurating Prompts for Each Art Style. We create prompts for AI generators by running BLIP [41] on human artworks submitted by artists, generating captions that effectively capture both the artwork's style and content. BLIP stands out as the state-of-the-art model for image captioning. We apply this method to improve consistency, because artworks of the same style often display large variation in content type and scene. For each art style, we randomly select 10 human artworks, making sure to include at least one piece per contributing artist. The chosen images are input to BLIP to extract the captions.

Here we encounter an issue where, for some artworks, BLIP struggles to extract the correct art style or any style at all. For example, for some *anime* artworks, BLIP generates captions accurately describing the content but fails to include any style. When prompted by this caption, the AI generators consistently produce images in the *photohumanistic* style instead of the intended *anime* style, despite the substantial difference between the two. Similarly, BLIP also fails to extract the watercolor style. In our study, we address this issue by adjusting the BLIP-generated captions to include the style of the artwork, for which we have ground truth. Table 14 in Appendix summarizes the modification made for each art style.

Customizing Prompts per AI Generator. We also make customized adjustments on BLIP-generated captions to address unique restrictions and configurations that each AI generator impose on prompts. Specifically, Adobe's Firefly and OpenAI's DALL-E 3 models do not respond to prompts that contain certain content. For instance, Firefly does not generate any image when prompted with "a fantasy style image of a woman in black holding a knife in the snow," but responds properly when the word "knife" is replaced with "sword." Similarly, DALL-E 3 does not react to prompts containing copyrighted materials such as Marvel character names (e.g., Spider-Man) and Nintendo game names (e.g., The Legend of Zelda). To address this, we manually substitute such content with more generic terms like " superhero-themed action figures" or "a fantasy-themed action figure on a horse." We verify that the modified prompts do produce images that aligned with the intended description.

Another issue is the inconsistent aspect ratio of generated images. Four out of the five generators consistently produce square images. DALL-E 3, on the other hand, generates images with random aspect ratios (e.g. 1024×1792). DALL-E 3 also tends to self-elaborate on the input prompt, producing extraneous intricacies. To address these artifacts, we include, in each input prompt to DALL-E 3, the additional phrase of "square image prompt the text to Dall-e exactly, with no modifications." Doing so effectively restricts its operation to adhere to the original prompt and return a square image.

Selecting Art Style from CIVITAI. Unlike other models, CIVITAI hosts instances of SDXL fine-tuned on specific art styles. For each art style, we locate the most frequently downloaded model from CIVITAI with that style. For instance, we use "Anime Art Diffusion XL" to generate *anime* style images.

4.3 Perturbed Images

Users of AI-generated images can intentionally add perturbations to images to deter their identification as AI images. We consider

four representative types of perturbations and describe each below. Appendix (Fig 9) provides visual samples of perturbed images.

Perturbation #1: JPEG Compression. Existing work has shown that compression artifacts can reduce the accuracy of image classifiers [29, 39, 40]. To study its impact on AI image detectors, we follow prior work to apply JPEG compression of a quality factor 15 [66] to AI-generated images before querying these classifiers.

Perturbation #2: Gaussian Noise. Similarly, digital noises can be introduced to disrupt classification-based detectors. For our study, we apply zero-mean Gaussian noise to each pixel value, with a standard deviation limited to 0.025, a parameter sweetspot with maximum impact and minimal visual disturbance.

Perturbation #3: CLIP-based Adversarial Perturbation. A more advanced (and costly) approach is to apply adversarial perturbations on AI images. Adversarial perturbations [28, 43] are carefully crafted pixel-level perturbations that can confuse ML classifiers. Automated AI image detectors are known to rely on the public CLIP model [18, 60] for detection, and thus, we leverage the CLIP model to craft our adversarial perturbations to maximize their transferability to AI detectors [20]. Specifically, we compute LPIPS-based adversarial perturbation [26] on each AI-generated image. We ensure that the perturbation is sufficient to confuse the CLIP model (i.e., LPIPS budget = 0.03).

Perturbation #4: Glaze. Glaze [61] is a tool for protecting human artists from unauthorized style mimicry. It introduces imperceivable perturbations on each artwork, which transforms the image's art style to a completely different one in the feature space. The widespread use of Glaze by artists has sparked extensive online discussions focused on instances where the use of Glaze on human art results in detection as AI images, while applying Glaze on AI-generated images can evade detection. To understand its impact, we use the public WebGlaze [38] tool to perturb both human art and AI images. We choose both the default medium intensity and also the high intensity, as artists often employ strongest protection to safeguard their online images.

4.4 Unusual Images

Hybrid Images. Users can create "hybrid" images by painting over AI-generated images. When posting them online, many include in the caption the generative models used. One of the coauthors, a professional artist over 30 years of experience, collected 40 hybrid images to include in our dataset and verified their sources. The images cover a variety of styles and subjects, including anime, cartoon, industrial design, and photography.

Human Artworks with Upscaling. Some artists use tools like image upscalers to enhance the quality of photography images, e.g., reducing blur or noise introduced during image capturing. We have 70 images in this group, upscaled using the baseline function of MagnificAI [1], a web upscaling tool endorsed by artists.

5 Accuracy of Automated Detectors

Using the dataset outlined in §4, we examine the efficacy of automated detectors in detecting AI-generated images. We first report the results on unperturbed imagery and the impact of AI generator choice. We then consider advanced scenarios where the detectors face different types of perturbed images, benign or malicious.

5.1 Experiment Setup

As discussed in §3, we consider five classification-based detectors, including three commercial detectors (Hive, Optic, Illuminarty), and two detectors built by academic researchers (DIRE, DE-FAKE). Our experiments use both original, unperturbed imagery (280 human artworks and 350 AI-generated images) and their perturbed versions. We delay the study of unusual images to §7.3 due to their decision complexity.

Detector Decisions. We study the ability of automated detectors to identify a human artwork as human and an AI-generated image as AI, mapping to a binary decision. However, today's automated detectors all output a probabilistic score indicating the likelihood or confidence of the input being AI-generated. A score of 100% implies absolute certainty that the input is AI-generated, while 0% indicates it is definitely human art. To convert this score into a binary decision, we need to establish a boundary that distinguishes between the two classes. Relying on a single threshold, such as x% (e.g., 50%), is obviously too fragile for this purpose.

Instead, we leverage the widely used 5-point Likert scale in user studies [34], designed to obtain quantitative measures of perception/decision. Specifically, scores ranging from 0-20% are associated with category human artwork (very confident), 20-40% with human artwork (somewhat confident), 40-60% with "not sure," 60-80% with AI-generated (somewhat confident), and 80-100% with AI-generated (very confident). It is important to note that our user studies maintain consistency by adopting the same rating scale. Next, to produce binary decisions, we designate any score below 40% as a decision of human art and any score above 60% as a decision of AI-generated. Those inputs yielding scores between 40-60%, indicating uncertainty ("not sure"), are excluded from the experiment. This exclusion is based on two practical considerations. First, there is no equitable method for comparing a "not sure" decision against the definitive ground truth [12]. Second, the occurrence of "not sure" is minimal across the machine detectors, e.g., 0.54% for Hive and 4.29% for Optic, thus their removal has a negligible impact on overall performance.

Evaluation Metrics. We report detector performance using four metrics. For easy notation, let [H→H] represent the # of human artworks detected as human-made, and [AI→AI] represent the # of AI-generated images detected as AI-generated. Let [H] and [AI] represent the total # of human artworks and AI-generated images included in this test, respectively.

- **Overall accuracy (ACC)** measures the accuracy of the detector regardless of the data origin. ACC= ([H→H]+[AI→AI])/([H]+[AI]).
- **False positive rate (FPR)** represents the ratio of human artworks misdetected as AI-generated. FPR=1- [H→H]/[H].
- **False negative rate (FNR)** measures the ratio of AI-generated images misdetected as human artworks. FNR=1 - [AI→AI]/[AI].
- **AI Detection success rate (ADSR)** captures the detector accuracy on AI-generated images. ADSR= [AI→AI]/[AI]. We use this metric to examine how generation-related factors would affect the detection outcome of AI-generated images.

Tested on Human Artworks + AI-generated Images			
Detector	ACC (%) ↑	FPR (%) ↓	FNR (%) ↓
Hive	**98.03**	**0.00**	**3.17**
Optic	90.67	24.47	1.15
Illuminarty	72.65	67.40	4.69
DE-FAKE	50.32	41.79	56.00
DIRE (a)	55.40	99.29	0.86
DIRE (b)	51.59	25.36	66.86
Ensemble	**98.75**	**0.48**	**1.71**

Table 1: Performance of automated detectors tested on unperturbed human artworks and AI-generated images. The Ensemble detector (Hive+Optic+Illuminarty) takes scores from Hive, Optic, and Illuminarty, using the highest confidence value as the determining score.

5.2 Results of Unperturbed Imagery

We start from unperturbed human artworks (280 images) and AI-generated images (350 images). Table 1 summarizes the detection performance in terms of ACC, FPR and FNR.

Hive. Hive is the clear winner among all five detectors, with a 98.03% accuracy, 0% FPR (i.e., it never misclassifies human artworks), and 3.17% FNR (i.e., it rarely misclassifies AI-generated images).

Optic and Illuminarty. Both perform worse than Hive, except that Optic has a lower FNR (1.15%) and thus is more effective at flagging AI-generated images. However, this comes at the expense of a high 24.47% FPR, where human art is misclassified as AI-generated. Illuminarty demonstrates even harsher treatment towards human art, with a very high FPR of 67.4%.

DE-FAKE and DIRE. We experiment with all 6 versions of DIRE, representing its 6 checkpoints published online. The detailed performance is listed in Table 12 in Appendix. The overall accuracy is consistently low for all 6 models (< 55.5%). The top-2 models' performances are listed in Table 1 as DIRE (a) and DIRE (b), one with nearly 100% FPR, and another with 66% FNR. DE-FAKE shows a similar pattern, with a low 50% accuracy and high FPR and FNR values. Given their poor performance, we do not conduct any additional experiments with both detectors.

Variation across Automated Detectors. The large performance variations across detectors could be attributed to the diversity and coverage of their training data. According to its website, Hive utilizes a rich collection of generative AI datasets and can identify the generative model used for the current input from a pool of nine models [2]. Similarly, Optic can pinpoint an input image to one of the 4 generative models. In contrast, Illuminarty's training data coverage is limited, particularly since it does not support image files exceeding 3MB[3]. Illuminarty and Optic are smaller companies with less training data [4].

DE-FAKE and DIRE's training data are constrained and lack representation from art. DIRE models are trained on interior design images (LSUN-Bedroom), human faces (CelebA-HQ) and ImageNet. The detection accuracy on art images is around 50%. For DE-FAKE, we follow the methodology described by [60] to train the classifier using images generated by SDXL and MSCOCO captions. When tested on SDXL images produced from MSCOCO prompts, the classifier reproduces a high ACC of 92.44% similar to [60]. However, when tested on images generated using artwork prompts, the accuracy drops to 50.3% for SDXL images and 46.43% for those produced by other generators. We attribute performance discrepancy to the issue of transferability. Training images generated from MSCOCO prompts do not follow the art dataset distribution, so these open-source detectors did not transfer well to the art dataset.

[2]As of April when conducting additional tests on Hive, it now detects 27 models.
[3]We downsample images generated by Firefly from 2048x2048 to 1024x1024 to meet this restriction.
[4]Illuminarty's creator informed us the model has not been updated for 6 months in Novemeber 2023.

	ADSR (%) ↑				
	CIVITAI	DALL-E 3	Firefly	MJv6	SDXL
Hive	100.00	98.57	91.04	94.29	100.00
Optic	100.00	97.14	97.06	100.00	100.00
Illuminarty	94.03	100.00	92.42	91.67	98.41

Table 2: The impact of AI-generator choice on detector performance, represented by ADSR, the % of AI-generated images correctly detected as AI-generated.

Combining Detectors (Hive+Optic+Illuminarty). We also study an Ensemble detector that leverages decisions from Hive, Optic, and Illuminarty. In case of a disagreement between the detectors, it opts for the decision marked with highest confidence. The confidence is calculated relative to the classification, by computing | detector score - 0.5 |. The result in Table 1 shows that the improvement over Hive is minor: 0.6% increase in ACC while FNR reduces from 3.17% to 1.71% and FPR increases from 0% to 0.48%.

5.3 Impact of AI Generator Choice

We also investigate the factors that may affect detection performance. While one might anticipate that the art style could have an impact, we do not observe any notable effect, at least for the seven major styles considered by our study. Instead, we observe a considerable impact by the choice of AI-generator. This can be seen from both Table 2 and Figure 2. Table 2 lists ADSR, the % of AI-generated images correctly detected as AI, while Figure 2 displays the raw confidence score produced by the detectors.

Across the five AI generators, images by CIVITAI and SDXL are the "easiest" to detect by Hive and Optic, i.e., 100% ADSR and not a single "not sure" decision. Since CIVITAI models are fine-tuned versions of SDXL, this suggests fine-tuning has minimal impact on AI image detection.

On the other hand, Firefly images are the least detectable – Hive marks 6 out of 70 Firefly images as human art and 3 as "not sure" while Optic marks 2 as human art and 2 as "not sure" (Fig. 2).

We hypothesize this is due to lack of training data. Firefly is a relatively new model, so Hive and Optic are likely to have much less training data relative to other models

5.4 Impact of Perturbations

Our goal is to understand whether adding perturbations to images, whether benign or malicious, could change detection outcomes. This triggers two questions below.

Question 1: How do perturbations affect the detection of AI images? We explore four perturbations: JPEG compression, Gaussian noise, adversarial perturbations on CLIP, and Glaze at medium and high intensity. The details of each perturbation are discussed at length in Section §4.3.

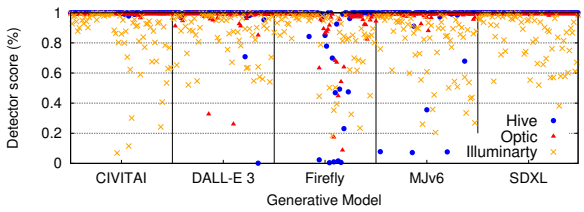


Figure 2: The confidence score produced by automated detectors on images generated by 5 generators. Detecting images generated by Firefly is the hardest.

	ADSR (%) ↑					
	JPEG comp.	Gaus. noise	Adver.	Glaze (med.)	Glaze (high)	Unperturbed
Hive	91.88	88.73	93.00	69.73	67.56	96.83
Optic	97.98	52.63	80.42	62.62	58.00	98.85
Illuminarty	93.19	61.43	94.34	89.80	89.22	95.31
Ensemble	94.29	88.86	94.29	73.71	71.43	98.29

Table 3: The impact of perturbations on AI-generated images, represented by ADSR, the % of AI-generated images correctly detected as AI-generated.

Table 3 reports ADSR, the success rate of detecting AI-generated images as AI when they include one of the four perturbations. As a reference, we include the ADSR for unperturbed images. We discuss each of the perturbations individually below, analyzing their impact on each detector.

Figure 3 plots a detailed view of the impacts of perturbations on the scores assigned by Hive to each image. While we examined these plots for each detector, we observed relatively similar trends across all, except that Optic and Illuminarty exhibited higher levels of noise. Hive's plots are easier to interpret, both for the perturbed and unperturbed data, thus we only present them for clarity. Additionally, we identify a set of images that demonstrate 'robustness' to any perturbation against Hive and provide a more in-depth discussion of these images in Table 13 in Appendix.

JPEG Compression. JPEG compression shows minimal impact on performance across all detectors, as they all remain above 91%. The lossy compression artifacts don't hinder the detectors' ability to detect AI images.

Gaussian Noise. Gaussian noise has little impact on Hive, it does drop both Optic and Illuminarty in ADSR. Optic's ADSR decreases to 52.63%, and Illuminarty's ADSR decreases to 61.43%, both of which are near random guessing. On the other hand, Hive's performance remains relatively high at 88.73% ADSR. These classifiers may have already adapted to the presence of mild noise in training images and learned to suppress the effects of noise.

Adversarial Perturbation. Of all the perturbations, the adversarial perturbations on CLIP have the least impact on performance for Hive and Illuminarty. Optic's ADSR drops to 80.42%, but Hive and Illuminarty both remain above 93%. This may indicate relatively low transferability of targeted attacks in the CLIP space in black-box settings.

Glaze. Across detectors, Glaze at both intensities consistently has a significant impact on ADSR. Additionally Glaze affects each detector similarly between medium and high intensity, with the increase in intensity only resulting in a drop in ADSR ranging from 0.5% to 4%. Glaze has the least effect on Illuminarty, dropping the ADSR from 95.31% to 89.22% for high-intensity Glaze. In comparison, Glaze reduces the ADSR for both Hive and Optic to below 70% while the Ensemble detector is only able to achieve 71.43%. We explore the effects of Glaze further below.

	ACC (%) ↑	FPR (%) ↓	FNR (%) ↓
Hive	80.81 / 98.03	3.23 / 0	32.44 / 3.17
Optic	61.92 / 90.67	33.59 / 24.47	42.00 / 1.15
Illuminarty	68.66 / 72.65	56.91 / 67.40	10.78 / 4.69
Ensemble	82.70 / 98.75	3.21 / 0.48	28.57 / 1.71

Table 4: Detection performance on human art and AI-generated images with and without Glaze (at high intensity), shown as: Glazed / Unperturbed.

Question 2: Does Glaze Affect Accuracy Similarly on Human Artwork vs AI-generated Images? To investigate if the detection of human artwork is impacted similarly to AI-generated images when Glazed, we applied high-intensity Glaze to both. The detection outcomes, including ACC, FPR and FNR are presented in Table 4 comparing results with and without Glaze. Our findings reveal a large reduction in ACC with the use of Glaze, as expected. However, the FNR increases across all detectors, while the FPR remains relatively consistent. On Hive, ACC is decreased around 20% and yet FPR increases only 3.23% while FNR increases almost 30%. Our Ensemble detector is able to achieve the highest accuracy on Glazed images, but maintains a FNR of 28.57%. This suggests that glazing human artwork typically does not affect classification success, but glazing AI-generated images often leads to misclassification as human artwork. We attribute this to the scarcity of Glazed images online and thus the lack of Glazed images in the detector's training datasets. Additionally Glaze was created to protect human artwork and has since been adopted by many artists, therefore the distribution of Glazed human art vs. Glazed AI-generated images online is likely skewed. Once again this demonstrates the impact of insufficient training data on image-based classifiers.

5.5 Summary of Findings

Our study leads to three key findings.

- Commercial detectors perform surprisingly well, and Hive performs the best. It is highly accurate (98.03% accuracy) and never misclassifies human artwork in our test. The other two detectors (Optic and Illuminarty) tend to misclassify human art as AI.
- Commercial detectors are heavily affected by feature space perturbations (i.e., Glaze) added to AI-generated images. On the other hand, human artworks with Glaze are mostly unaffected, as they are still largely detected as human.
- Poor performance on Firefly and Glaze indicates the results are correlated with training data. Performance of supervised classification depends heavily on the availability of training data. Detectors are more vulnerable to newer models with less available training data (Firefly) and adversarial inputs they have not seen before.

5.6 Followup Robustness Tests on Hive

After the paper was accepted, we discovered a Reddit post that claimed overlaying a real image of a white wall onto an AI-generated

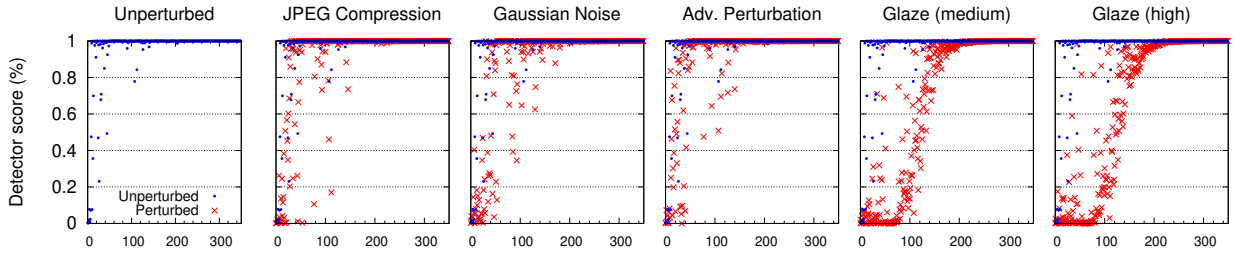


Figure 3: Impact of five different perturbations on the Hive confidence score, for 350 AI-generated images. In each figure, the images are indexed by the increasing Hive score of unperturbed versions.

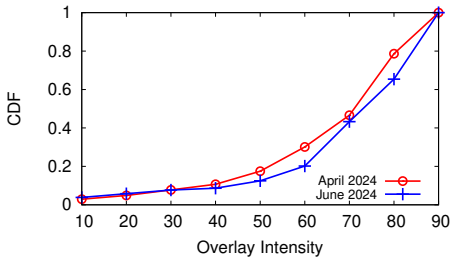


Figure 4: CDF of overlay intensity required to change Hive's decision over the period of 3 months.

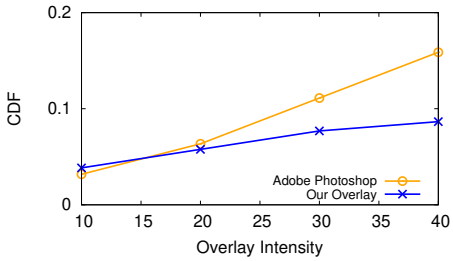


Figure 5: CDF of overlay intensity required to change Hive's decision using different overlay methods in June 2024.

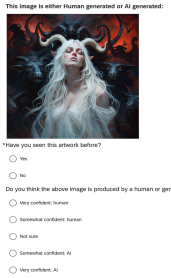


Figure 6: User study interface.

image could bypass Hive's detection [78]. We investigate this as a new adversarial approach to bypassing detection, and conduct a large-scale test on 105 AI-generated images, randomly sampled from our dataset to include 3 images from each of the 7 styles and 5 AI-generators. We implement a python script to overlay each image with a white wall image from Adobe Shutterstock, at intensity levels varying from 10% to 90% in increments of 10. Hive confirms the Adobe's white wall image as not AI-generated with 100% confidence.

Figure 4 illustrates the CDF graph of the distribution of overlay intensities required to flip Hive's decision from AI-generated to not AI-generated. The red curve from April 2024 shows there is a 17.5% chance that Hive's decision will change at or below intensity levels of 50%. Yet, there are corner cases where images (2.91% in April and 3.85% in June) were able to fool Hive with the white wall overlaid at only 10% intensity. Most images require around overlay intensity of 60-80% to change Hive's classification.

We also note that Hive's performance improved over time. Figure 4 displays Hive results collected in April 2024 and June 2024. The shift in the CDF shows that Hive's accuracy and robustness increased in the 2 month period, possibly due to adversarial training on inputs or updates to its supervised classification algorithm.

We conduct another smaller-scale test to see the effect of different overlay methods on Hive's output. We apply and compare Adobe Photoshop's "multiply blend" overlay option (the method used in the Reddit post) against our scripted blending algorithm, varying intensities from 10% to 40% in increments of 10 across 63 AI-generated images[5]. Figure 5 shows Adobe's Photoshop method has a significantly stronger effect against Hive at higher intensities. This gap is somewhat unexpected, and just confirms that even small variations in adversarial attacks can have large unpredictable impact on Hive's detection robustness.

6 User Studies on Human Detection

Next, we measure the ability of human users and artists in identifying AI-generated images. As discussed in §3.3, we perform user studies with 3 separate user populations, including crowdworkers, professional artists and expert artists.

To ensure uniformity, each group is provided with the same user study, but the expert group received extended followup questions asking for detailed examples (Section 6.4). Our study is approved by our local Institutional Review Board (IRB). We omit the IRB number for anonymity.

6.1 Study Setup

Participants. As discussed in §3.3, we recruited three participant groups, and a total of **3993** participants completed our study and passed all attention checks. These include 177 baseline participants recruited from Prolific, referred to as general users, 3803 professional artist volunteers, and 13 high profile experts.

Task. Our study takes the form of a user survey. We added attention-check questions in both the middle and concluding sections of the survey to filter out low attention participants.

After a brief introduction on generative models and the current issue of distinguishing between human artworks and AI-generated images, we present the participant with a sequence of 20 images, shown one at a time, and ask them to decide whether each image

[5]Dataset is narrowed down to 3 images from MJv6, Firefly, and SDXL across 7 styles.

is human-made or AI-generated. For each image, we ask two questions. The first question is "Have you seen this image before?" If yes, we disregard the response for this image, since the participant has likely seen this image and is possibly aware of the image's true source. The second question asks them to rate the current image with one of the five choices: "human art (very confident)," "human art (somewhat confident)," "not sure," "AI-generated (somewhat confident)," "AI-generated (very confident)." For each participant, we randomly sample 20 images from a collection of 670 images, consisting of 210 human artworks, 70 human artworks after upscaling, 350 AI-generated images, and 40 hybrid images. Every image is seen by five participants.

Next, we ask the participant, for each of the 7 art styles, whether they are confident at distinguishing human art from AI-generated images; the user study interface is presented in Figure 6. If confident, we ask them to describe the properties of the art that contributed to their decisions. Here we present seven options: "content," "complexity," "technical skill," "perspective," "lighting," "consistency," and an additional text response section for additional details. Finally, we ask them if they self-identify as full-time or part time artists.

Performance Metrics. Same as the evaluation of automated detectors in §5.1, we convert each 5-point Likert scale rating into a binary decision. That is, both "human art (very confident)" and "human art (somewhat confident)" map to a decision of human art, and "AI-generated (somewhat confident)" and "AI-generated (very confident)" map to AI-generated. Not decision is produced for the "not sure" responses, and we ignore them when computing ACC, FPR, FNR and ADSR (same as §5.1).

6.2 Detection Accuracy for General Users

Table 5 shows detection performance of general users, which is only slightly better than random coin-flip. This shows that general, non-artist users are unable to tell the difference between human art and AI-generated images.

Table 6 examines the impact of art style on detection accuracy (ACC). For general users, accuracy is slightly higher on cartoon, fantasy and photography styles. These three represent a collection of "digital" artworks more frequently accessible to general users, compared to "physical" art styles like oil/acrylic, watercolor, and sketch, and other "digital" artworks like anime. We attribute the slightly higher accuracy to this increased familiarity. Finally, Table 7 reports detection success rate on AI-generated images, which is around 60% and varies slightly across the five AI generators. Images generated by Firefly and MidjourneyV6 are harder for non-artist users to recognize.

6.3 Detection Accuracy for Professional Artists

Compared to general users, professional artists take more time to inspect images and are more effective at distinguishing between human art and AI images. They produce a detection accuracy of 75.32% (Table 5). Their 23.53% FPR and 25.37% FNR indicate that their detection performance is not skewed toward either human art or AI-generated images.

Decision Factors. To understand why professional artists are better at evaluating art images, we study the key factors that influenced their decisions. The first is the image's art style. Table 6 shows

	ACC (%) ↑	FPR (%) ↓	FNR (%) ↓
General user	59.23	40.81	40.75
Professional artist	75.32	23.53	25.37
Expert artist	83.00	20.78	14.63

Table 5: Performance of human detection on both human art and AI-generated images.

	Anime	Cartoon	Fantasy	Photo.	Oil/Acrylic	Sketch	Watercolor
Gen.	57.43	63.58	62.80	63.23	56.81	54.99	56.16
Pro.	79.16	82.42	81.72	76.05	66.01	70.27	70.51

Table 6: Impact of art style on detection accuracy (ACC) for general users (Gen.) and professional artists (Pro.). We exclude the result of expert artists due to insufficient coverage.

	ADSR (%) ↑				
	CIVITAI	DALL-E 3	Firefly	MJv6	SDXL
General user	66.77	60.63	51.18	50.00	67.56
Professional artist	83.56	67.56	75.40	61.53	84.43
Expert artist	90.32	86.96	65.22	86.96	95.65

Table 7: ADSR of human detection on AI-generated images by different AI generators. ADSR is the success rate of detection AI-generated images as AI-generated.

that the image's artistic style has a clear impact on the performance of professional artists. Among the 7 styles, the top-3 "easier-to-detect" styles are anime, cartoon and fantasy, and the bottom one is oil/acrylic for which the detection accuracy drops nearly 20%. This aligns with the participants' feedback on the styles that they feel most confident on detection, where 82.42% selected cartoon, followed by fantasy and anime.

For these top-3 styles, the artists select "consistency" as the most dominant decision factor. Specifically, in the text response section on anime images, the most frequently entered words are hands, hair, details, eyes and lines. Similarly, in the case of fantasy images, artists observe specific details such as asymmetric armor or symbols that should be symmetric. This shows that professional artists can apply their knowledge and experiences of art creation to identify inconsistencies in AI-generated images, with particular focus on fine-grained details.

Impact of AI generator. Table 7 lists the detection success rate of AI-generated images broken down by source generator model. For professional artists, accuracy clearly varies across generator models. The top-2 "easiest-to-detect" are CivitAI and SDXL, with detection success rate above 83%. Interesting to note that these two are also the most easy-to-identify generators for automatic detectors. Next, detection accuracy reduces to below 70% for images produced by MidjourneyV6 and DALL-E 3, suggesting that MidjourneyV6 and DALL-E 3 produce better copies of human art styles, making it harder for artists to spot inconsistencies.

6.4 Detection Accuracy for Expert Artists

Our 13 expert artists show greater proficiency in the detection task, raising overall detection accuracy (ACC) to 83% (see last row in Table 5). They also produce slightly imbalanced FPR (20.78%) and FNR (14.63%), indicating that they are better at spotting AI-generated images than human art. When we gave feedback to the experts on their results, many were frustrated that they committed errors by marking human art as AI (false positive). In retrospect, they

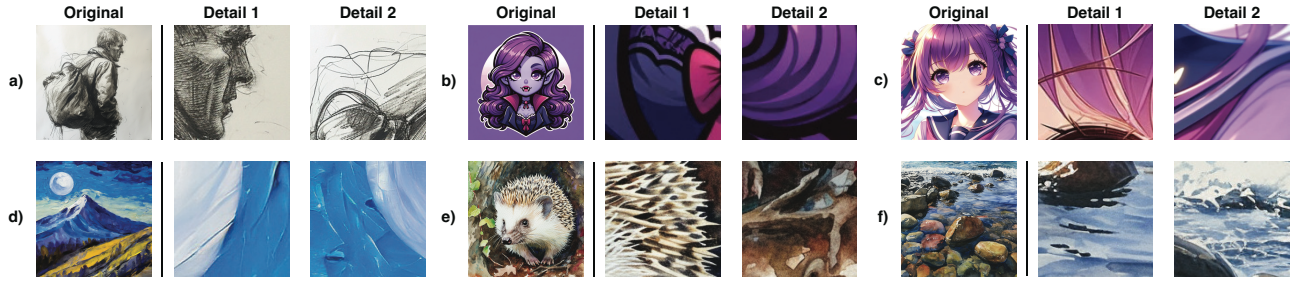


Figure 7: Six hard-to-detect AI-generated images, and their artistic errors/inconsistencies discovered by our expert artists. In each 3-image group, the left one is the full image, and the right two are zoomed-in view of discovered artifacts.

explained that they identified detailed mistakes and inconsistencies which were likely due to inexperience and human error by the human artists. For example, in a painting of a bedroom bathed in moonlight, the shadow of a window pane had slightly offset position of the latch compared to the window itself. This was seen by expert artists as an inconsistency, but later attributed to lack of attention to detail by a human artist. This attests for the drop in ADSR for Firefly images. Since expert artists look at fine-grained detail in AI-generated images, they are overfitted to spot irregularities from popular generators and have yet been accustomed to the newer style of images.

Decision factors. Our expert artists are generally very confident at judging images from more than 2 art styles. The most frequently selected one is fantasy art. The primary decision factors are intrinsic artistic details, which often go beyond the "(in)consistency" element used by professional artists in their detection efforts (as discussed in §6.3). Specifically, our expert artists point out that AI-generated images generally "look too clean, rendered, and detailed" and "have no variety in composition, edges, distribution of detail," and had "design elements are nonsensical or blend into each other in telltale ways," while human-made fantasy art "contains components that are novel such as armor or jewelry."

Focus Study on "Hard-to-Detect" AI Images. For a comprehensive view of how experts identify AI-generated images, we presented our expert group with a fixed set of the six AI-generated images that produced the most false negative errors by the professional (non-expert) artist group. They cover 5 styles: sketch, oil/acrylic, cartoon, watercolor, and anime. For each of these difficult images, we asked the experts for detailed feedback on what exactly they would use to identify these images as AI. Figure 7 shows all six images, each followed by two regions of the image zoomed in to show details of artifacts identified by our expert artists.

Next we summarize general techniques identified by experts, and use specific images to illustrate these methods.

- *Consistency in medium.* For any specific artwork, a trained artist typically employs only a single consistent medium, e.g. pencil, charcoal, and rarely combine multiple mediums. For example, in the sketch in image a), our experts locate not only a "weird halo effect" (detail 1) due to the use of both pencil and charcoal, but also a "crunchiness" (detail 2) to the lines that associate with neither pencil nor charcoal. AI models are associating lines from multiple mediums with the same style and fail to differentiate between them. Similarly, oil and acrylic paintings can display messy or smooth styles but never both. In image image d), the perfectly round moon looks like a digital art (detail 2), inconsistent with the overall messy painting style. The white is "too clean" (detail 1) when transitioning into the blue background.
- *Intentionality in details.* In art featuring human figures, human artists dedicate considerable effort to convey precise details of human features. In image b), the light caught in the eyes do not match and the hair ends flow in opposite directions from the rest of the hair (detail 2). Similarly, in image c), hairs behind the neck are floating and doing completely different things from the rest of her hair. Human artists also avoid unusual tangents with the bangs and eyebrow, something that this image completely overlooked.
- *Limitations of medium.* Experienced artists know that certain patterns and details are impossible to produce in real life due to physical limits of the medium. In image e), since watercolor is transparent and it bleeds after each brush stroke, the "white over dark in the quills" (detail 1) is impossible to physically produce. The image is also too smooth to be hand painted on paper (detail 2), since watercolor bleeds in random directions.
- *Domain knowledge.* There are specific rules when drawing specific subjects that are easy to validate. Wet paintings have an order of application, from light to dark, transparent to opaque. Thus all white spaces must be subtractive. Yet in image f), the white highlight is added after a dark area, which is wrong. Also the water flow shows the wrong diffusion pattern. A trained artist should not make these mistakes.

We consider these techniques in aggregate, and note that most of them require significant training and external knowledge to apply. In this sense, factors such as intentionality and domain knowledge seem like the most difficult for AI models to capture from training data. But as a whole, these domain-specific filters clearly operate very differently from statistical approaches used in automated detectors. Thus we believe these methods will continue to be effective even as AI models continue to evolve.

6.5 Summary of Key Findings

Our multi-tier user study finds that general (non-artist) users are unable to distinguish between human art and AI-generated images. Professional artists are more confident and make more accurate decisions, and experienced experts are the most effective. Diving deeper into concrete examples, we learn that experts leverage extensive knowledge of art mediums and techniques to identify what features are physically impossible, and inconsistencies and mistakes that professional artists would avoid.

Metric	Human detector			Machine detector		
	General	Professional	Expert	Hive	Optic	Illuminarty
ACC (%) ↑	59.23	75.32	83.00	98.03	90.67	72.65
FPR (%) ↓	40.81	23.53	20.78	0.00	24.47	67.40
FNR (%) ↓	40.75	25.37	14.63	3.17	1.15	4.69

Table 8: Performance of human and machine detectors on unperturbed imagery.

7 Human vs. Classifier Detectors

In this section, we present our findings by comparing the performance of human and machine detectors. Our analysis starts from the baseline results, where all detectors are tested using the same set of usual images, covering unperturbed human artworks and AI-generated images. Next, we investigate how human detectors respond to Glazed images, for which machine detectors have struggled to handle (as shown in §5.4). Finally, we explore both human and machine detectors' responses to unusual images, including both hybrid images (when human edits AI-generated images) and upscaled human art (where artists polish the digital image of their artworks using enhancement tools).

7.1 Decision Accuracy and Confidence

We start from comparing human and machine detectors on the baseline task of differentiating

Detection accuracy. Table 8 lists the detection performance of both human detectors and classifier-based detectors, on unperturbed images. Among the six detectors, we have Hive > Optic > Expert Artist > Professional Artist > Illuminarty > Non-Artist.

Decision confidence. We are also interested in understanding the distribution of decision confidence among human and machine detectors. Figure 8 shows, for AI-generated images and human artworks, the distribution of the "raw" decision represented by the 5-point Likert score used by our study. For AI-generated images (shown by the top figure), the dark blue bar represents the ratio of correct decisions that also carry high confidence, while for human artworks (the bottom figure), the dark orange bar captures the ratio of correct decisions made with high confidence.

Across the three groups of human detectors, general users are the least accuracy and also show the lowest confidence in their decisions, while expert artists are the most accurate and the most confident. This is as expected. Across the three machine detectors, Hive is highly confident (and accurate), followed by Optic. Additionally, Optic is more confident when facing AI-generated images than human artworks. Overall, the two machine detectors (Hive and Optic) show higher confidence (and higher accuracy) than expert artists, while the third machine detector (Illuminarty) performs worse than both expert and professional artists.

7.2 Are Human Artists Better at Judging Glazed Images?

As shown by Table 4 in §5.4, machine detectors, especially Hive and Optic, are much less effective at judging Glazed versions of AI-generated images. Thus a natural question is "would artists who know or use Glaze be more effective at judging Glazed images?" To answer this question, we conduct an additional user study with our expert group, who all use Glaze on their published artworks.

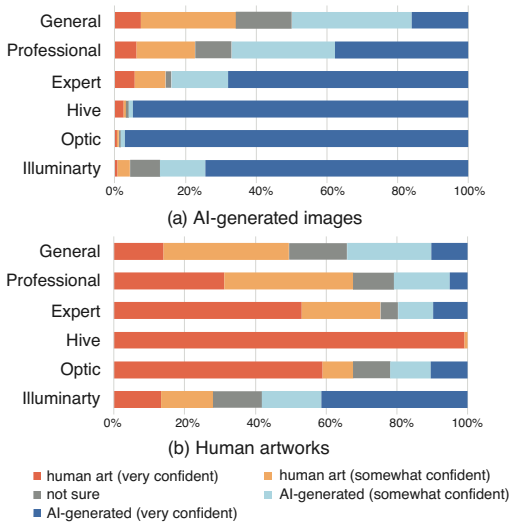


Figure 8: Distribution of detection decision represented by the 5-point Likert rating on (a) AI-generated images and (b) human artworks.

	ACC (%)↑	FPR (%) ↓	FNR (%) ↓
Expert Artist	83.44	23.53	8.97
Hive	87.76	4.04	20.62
Optic	61.20	38.30	39.33
Illuminarty	66.11	59.09	9.78

Table 9: Detection performance on Glazed version of human artworks and AI-generated images.

Here we randomly select 100 AI-generated images (20 images per AI generator) and 100 human artworks (15 images per style, except 10 images for cartoon), and use the WebGlaze tool and the medium intensity setting to Glaze all 200 images. For each expert artist, we randomly select 20 Glazed images and ask them to decide between human art and AI-generated.

Table 9 lists the overall performance across Glazed images. We see that human experts largely outperform both Optic and Illuminarty at judging Glazed images, whether it is human art or AI-generated. While Hive is still the best-performing detector in the overall accuracy (87.76%), our expert artists are not far off (83.44%). But more importantly, human experts achieve a much lower FNR (8.97%) compared to Hive (20.62%). This implies that **human artists do outperform machines at judging Glazed versions of AI-generated images**.

7.3 Judging Unusual Images

As discussed in §4.4, we also consider two unusual types of images: (i) hybrid images produced by human users editing AI-generated images, and (ii) upscaled human artworks where artists apply digital touchups to polish the image of their artworks. By evaluating them using both human and machine detectors, our goal is to identify, if any, notable differences in how they are evaluated by human and machine detectors.

In total, we collect 70 human artworks after applying upscaling and 40 hybrid images (see §4.4). Given the limited member size of our expert group, we do not have sufficient coverage on these

Unperturbed Human Artworks and AI Images			
Detector	ACC (%)↑	FPR (%) ↓	FNR (%) ↓
Hive	98.03	0.0	3.17
expert	83.00	20.78	14.63
Hive + expert	92.19	8.11	7.63

Table 10: Detection performance on unperturbed human artworks and AI-generated images. Three detectors: Hive, one expert (per image), Hive + expert (tiebreak=confidence).

Glazed Human Artworks and AI Images			
Detector	ACC (%)↑	FPR (%) ↓	FNR (%) ↓
Hive	87.12	6.06	19.70
expert	84.85	23.08	7.46
Hive + expert	92.54	6.06	8.82

Table 11: Detection performance on Glazed human artworks and AI-generated images. Three detectors: one expert (per image), Hive, Hive + expert (tiebreak = confidence).

images by expert artists and thus omit their results. Figure 10 in Appendix plots, for both types of images, the distribution of the "raw" detection decision represented by the 5-point Likert score, for both human and machine detectors.

Hybrid Images. For these images, the decision distribution is similar to that of AI-generated images shown in Figure 10(a), suggesting that both human and machine detectors frequently label these images as AI-generated.

Upscaled Human Art. The decision distribution is very similar to that of human artworks in Figure 10(b). This implies that upscaling does not have a significant impact of human artworks in terms of their decision outcomes from both human and machine detectors.

8 Combining Human and Automated Detectors

Our study shows that both human artists and automated detectors face challenges in distinguishing between human art and AI-generated images. Tools like Hive are highly effective at evaluating unperturbed images, but perform poorly when AI-generated images are intentionally perturbed (e.g., Glaze or image overlays) to evade detection. On the other hand, human experts can still identify perturbed AI-generated images. Thus we believe a mixed team of human artists and machine classifiers will be the most effective.

Teaming up Hive and Expert Artists. We evaluate a scenario that combines Hive scores with one expert artist. If Hive and the expert disagree, the score with higher confidence wins.

We evaluate the combined detector on both unperturbed artworks/images and their Glazed versions. Table 10 shows that, for unperturbed images, the combined detector exhibits a slightly lower accuracy compared to Hive. Next, Table 11 shows that the combined detector is highly effective in judging Glazed images, outperforming both Hive and expert. Notably, it outperforms Hive by lowering FNR from 19.70% down to 8.82%. Thus it is more effective at identifying AI-generated images that have applied Glaze in an attempt to evade detection. At the same time, the combined detector achieves a low FPR like Hive (6.06%), remarkably lower than that of expert (23.08%). Finally, the equilibrium between FPR and FNR values, on both Glazed and original images (Table 10), suggests "unbiased" detection accuracy for human artworks and AI-generated images.

9 Ethics

Our user study was reviewed and approved by our institutional review board (IRB). In our study, we prioritized consent and protection of all participants, especially human artists and their artworks.

Consent for Human Art. Our study necessitates the use of artwork by human artists. To obtain consent, we identified artists and reached out with request for permission. Many responded. We waited roughly 4 weeks, and reached out again to everyone else. Once downloaded, images are anonymized and stored on private, secure servers.

Since we asked for artwork from human artists with signatures/watermarks removed, we were extremely sensitive to potential unauthorized exposure. We took efforts to minimize the exposure of human artwork to external sources. They were available to participating crowdsourced workers only for a short time through the Prolific platform. In the regular artist user study, images were available to participants for a total of 14 hours, after which we shut down the study to minimize uncontrolled exposure.

Exposure to Web Services. We took careful steps to ensure that images of human art were not misused by external AI detection services. We reached out to both Optic and Illuminarty, and were assured that images are never used for training and deleted after process (with a max of 4 days for Optic). Hive's terms of service states they can train AI models using images uploaded via the free web service, but not images classified through paid APIs. Thus we obtained access to a paid Hive account, and ensured all images of human art were classified using this paid Hive account.

10 Discussion and Takeaways

As with any real-world study, there are limitations in our study that need to be considered.

- Category of styles: our dataset only included a few fixed art styles. More diverse styles might provide more comprehensive results.
- Cropping of images: cropping was applied to a small number of human artworks. We avoided samples with highly irregular aspect ratios and ensured meticulous cropping to minimize irregularities.
- Curating Likert Scale: we discarded "not sure" responses from the user study. This was done to maintain consistency with measuring the confidence level (40-60%) in automated detectors and to prevent guesses from affecting our metrics.

Our results also suggest takeaways for different audiences.

For artists. Proving human authenticity will be come increasingly important, and increasingly difficult. No single method wil be foolproof, and artists should consider incorporating work-in-progress (WIP) or timelapses into their process.

For researchers. Developing highly accurate detection tools requires continued investment in ethically obtaining diverse training sets. To enhance robustness, incorporating perturbed and adversarially altered images during training is crucial.

For policy makers. Implementing standards that mandate mixed detection teams in critical applications can enhance both detection accuracy and robustness.

As AI evolves, generative AI users seeking to avoid detection will adapt to exploit vulnerabilities. Some are using Adobe's Photoshop to add imperfections to avoid the smooth AI finishing look. Others

are using one model to generate foreground objects and another to create backgrounds [19]. These evolving techniques will continue to present challenges for future AI detection systems.

Acknowledgements

We thank our anonymous reviewers for their insightful feedback. Sincere thanks also go to the thousands of artists who participated in our user study. This work is supported in part by NSF grants CNS-2241303 and CNS-1949650. Opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of any funding agencies.

References

[1] Magnific AI. 2023. Magnific. https://magnific.ai.
[2] Sina Alemohammad et al. 2023. Self-Consuming Generative Models Go MAD. In *arXiv preprint:2307.01850*.
[3] ANDY BAIO. 2022. Invasive Diffusion: How one unwilling illustrator found herself turned into an AI model.
[4] Quentin Bammey. 2020. Synthbuster: Towards Detection of Diffusion Model Generated Images. *IEEE Open Journal of Signal Processing* 5 (2020), 1–9.
[5] Xiuli Bi et al. 2023. Detecting Generated Images by Real Images Only. *arXiv preprint:2311.00962* (2023).
[6] Jordan J. Bird and Ahmad Lotfi. 2023. CIFAKE: Image Classification and Explainable Identification of AI-Generated Synthetic Images. *arXiv preprint:2303.14126* (2023).
[7] Isabelle Bousquette. 2023. Companies Increasingly Fear Backlash Over Their AI Work. WSJ.
[8] G. W. Braudaway. 1997. Protecting publicly-available images with an invisible image watermark. In *Proc. of ICIP*.
[9] Sergi D. Bray et al. 2023. Testing human ability to detect 'deepfake' images of human faces. *Journal of Cybersecurity* (2023), 1–18.
[10] Martin Briesch, Dominik Sobania, and Franz Rothlauf. 2023. Large Language Models Suffer From Their Own Output: An Analysis of the Self-Consuming Training Loop. In *arXiv preprint:2311.16822*.
[11] Cara. 2023. https://cara.app/.
[12] Seung Youn Chyung et al. 2017. Evidence-Based Survey Design: The Use of a Midpoint on the Likert Scale. *Performance Improvement* (2017), 15–23.
[13] Civitai. 2022. What is Civitai? https://civitai.com/content/guides/what-is-civitai.
[14] Linda Codega. 2023. Dungeons & Dragons Updates Bigby to Replace AI-Enhanced Images. Gizmodo.
[15] Linda Codega. 2023. New Dungeons & Dragons Sourcebook Features AI Generated Art. Gizmodo.
[16] Riccardo Corvi et al. 2023. On The Detection of Synthetic Images Generated by Diffusion Models. In *Proc. of ICASSP*.
[17] CourtListener. 2024. Andersen v. Stability AI Ltd.(3:23-cv-00201). https://www.courtlistener.com/docket/66732129/andersen-v-stability-ai-ltd/.
[18] Davide Cozzolino et al. 2023. Raising the Bar of AI-generated Image Detection with CLIP. *arXiv preprint:2312.00195* (2023).
[19] Katie Deighton. 2024. How the Ad Industry Is Making AI Images Look Less Like AI. Wall Street Journal.
[20] Ambra Demontis et al. 2019. Why do adversarial attacks transfer? explaining transferability of evasion and poisoning attacks. In *Proc. of USENIX Security*.
[21] Prafulla Dhariwal and Alexander Nichol. 2021. Diffusion Models Beat GANs on Image Synthesis. In *Proc. of NeurIPS*.
[22] Maggie H. Dupre. 2023. Sports Illustrated Publisher Fires CEO After AI Scandal. Futurism.
[23] Salako Emmanuel. 2023. AI Tools for Combating Deepfakes. https://ijnet.org/en/story/ai-tools-combating-deepfakes.
[24] Joel Frank et al. 2024. A Representative Study on Human Detection of Artificially Generated Media Across Countries. In *Proc. of IEEE S&P*. San Francisco, CA.
[25] Ethan Gach. 2023. Amazon's First Official Fallout TV Show Artwork Is an AI-Looking Eyesore. Kotaku.com.
[26] Sara Ghazanfari et al. 2023. R-LPIPS: An adversarially robust perceptual similarity metric. *arXiv preprint:2307.15157* (2023).
[27] Paul Glynn. 2023. Sony World Photography Award 2023: Winner Refuses Award After Revealing AI Creation. BBC News.
[28] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. 2014. Explaining and harnessing adversarial examples. *arXiv preprint:1412.6572* (2014).
[29] Genki Hamano, Shoko Imaizumi, and Hitoshi Kiya. 2023. Effects of JPEG Compression on Vision Transformer Image Classification for Encryption-then-Compression Images. *Sensors* 23, 7 (2023).
[30] Hive. 2023. AI-Generated Content Classification. https://thehive.ai/apis/ai-generated-content-classification.
[31] Karen Ho. 2024. Database of 16,000 Artists Used to Train Midjourney AI, Including 6-Year-Old Child, Garners Criticism. https://www.artnews.com/art-news/news/midjourney-ai-artists-database-1234691955/.
[32] Ashish Hooda et al. 2024. D4: Detection of Adversarial Diffusion Deepfakes Using Disjoint Ensembles. In *Proc. of WACV*. IEEE.
[33] Illuminarty. 2023. Is an AI Behind Your Image? https://illuminarty.ai/.
[34] Heon Jae Jeong and Wui Chiang Lee. 2016. The level of collapse we are allowed: comparison of different response scales in safety attitudes questionnaire. *Biometrics Biostatistics International Journal* (2016), 128–134.
[35] Joseph Saveri Law Firm LLP. 2023. Class Action Filed Against Stability AI, Midjourney, and DeviantArt for DMCA Violations, Right of Publicity Violations, Unlawful Competition, Breach of TOS.
[36] Tero Karras et al. 2020. Analyzing and Improving the Image Quality of StyleGAN. *arXiv preprint:1912.04958* (2020).
[37] Alex Krizhevsky and Geoffrey Hinton. 2009. Learning multiple layers of features from tiny images. *University of Toronto* (2009).
[38] SAND Lab. 2023. Web Glaze. https://glaze.cs.uchicago.edu/webglaze.html.
[39] K. W. K. Lam, W. L. Lau, and Z. L. Li. 1999. Effects of JPEG compression on accuracy of image classification. In *Proc. of ACRS*.
[40] W-L Lau, Z-L Li, and KW-K Lam. 2003. Effects of JPEG compression on image classification. *Proc. of IJRS* (2003).
[41] Junnan Li et al. 2023. BLIP-2: bootstrapping language-image pre-training with frozen image encoders and large language models. In *Proc. of ICML*.
[42] Zeyu Lu et al. 2023. Seeing is not always believing: Benchmarking Human and Model Perception of AI-Generated Images. *arXiv preprint:2304.13023* (2023).
[43] Aleksander Madry et al. 2017. Towards deep learning models resistant to adversarial attacks. *arXiv preprint:1706.06083* (2017).
[44] Emanuel Maiberg. 2023. AI Images Detectors Are Being Used to Discredit the Real Horrors of War. 404Media.
[45] MaxonVFX. 2024. We extend our apologies to the community. https://twitter.com/MaxonVFX/status/1748826148858208286.
[46] Alex Nichol et al. 2022. Hierarchical text-conditional image generation with clip latents. *arXiv preprint:2204.06125* (2022).
[47] Travis Northup. 2023. Wizards of the Coast Repeats Anti-AI Art Stance After Player's Handbook Controversy. IGN.com.
[48] Jie Yee Ong. 2023. Scooby-Doo: Daphne Voice Actor Fell Victim To $1,000 AI Art Scam. The Chainsaw.
[49] Optic. 2023. AI or Not. https://www.aiornot.com.
[50] Kyle Orland. 2024. Magic: The Gathering Maker Admits it Used AI-generated Art Despite Standing Ban. Ars Technica.
[51] Susannah Page-Katz. 2023. Introducing Our AI Policy. Kickstarter.com.
[52] Dustin Podell et al. 2023. SDXL: Improving Latent Diffusion Models for High-Resolution Image Synthesis. *arXiv preprint:2307.01952* (2023).
[53] Alec Radford et al. 2021. Learning transferable visual models from natural language supervision. In *Proc. of ICML*.
[54] Aditya Ramesh et al. 2022. Hierarchical text-conditional image generation with clip latents. *arXiv preprint:2204.06125* (2022).
[55] Jonas Ricker, Simon Damm, Thorsten Holz, and Asja Fischer. 2023. Towards the Detection of Diffusion Model Deepfakes. *arXiv preprint:2210.14571* (2023).
[56] Robin Rombach et al. 2022. High-resolution image synthesis with latent diffusion models. In *Proc. of CVPR*.
[57] Kevin Roose. 2022. An A.I.-Generated Picture Won an Art Prize. Artists Aren't Happy. https://www.nytimes.com/2022/09/02/technology/ai-artificial-intelligence-artists.html.
[58] Mia Sato. 2023. How AI art killed an indie book cover contest. The Verge.
[59] Christoph Schuhmann et al. 2022. Laion-5b: An open large-scale dataset for training next generation image-text models. *arXiv preprint:2210.08402* (2022).
[60] Zeyang Sha, Zheng Li, Ning Yu, and Yang Zhang. 2023. DE-FAKE: Detection and Attribution of Fake Images Generated by Text-to-Image Generation Models. In *Proc. of CCS*.
[61] Shawn Shan, Jenna Cryan, Emily Wenger, Haitao Zheng, Rana Hanocka, and Ben Y. Zhao. 2023. Glaze: Protecting artists from style mimicry by text-to-image models. In *Proc. of USENIX Security*.
[62] "Eons Show". 2024. Eons Show apologies for violating own AI policy. Twitter. https://x.com/EonsShow/status/1751327424556544451.
[63] Ilia Shumailov et al. 2023. The Curse of Recursion: Training on Generated Data Makes Models Forget. *arXiv preprint:2305.17493* (2023).
[64] Zachary Small. 2023. As Fight Over A.I. Artwork Unfolds, Judge Rejects Copyright Claim. NY Times.
[65] Haixu Song, Shiyu Huang, Yinpeng Dong, and Wei-Wei Tu. 2023. Robustness and Generalizability of Deepfake Detection: A Study with Diffusion Models. *arXiv preprint:2309.02218* (2023).
[66] Kenneth R. Spring, John C. Russ, Matthew J. Parry-Hill, and Michael W. Davidson. 2016. JPEG Image Compression. National High Magnetic Field Laboratory.
[67] Stability AI. 2022. Stable Diffusion Public Release. https://stability.ai/blog/stable-diffusion-public-release.
[68] StabilityAI. 2022. Stable Diffusion v1-4 Model Card. https://huggingface.co/CompVis/stable-diffusion-v1-4.

[69] StabilityAI. 2022. Stable Diffusion v1-5 Model Card. https://huggingface.co/runwayml/stable-diffusion-v1-5.
[70] "Portal Staff". 2023. League of Legends AI-Generated LATAM Anniversary Video Gets Taken Down. ZLeague The Portal.
[71] Chandra Steele. 2023. How to Detect AI-Created Images. https://www.pcmag.com/how-to/how-to-detect-ai-created-images.
[72] Stuart A. Thompson and Tiffany Hsu. 2023. How Easy Is It to Fool A.I.-Detection Tools? NY Times.
[73] Sheng-Yu Wang et al. 2020. CNN-generated images are surprisingly easy to spot... for now. *arXiv preprint:1912.11035* (2020).
[74] Yabin Wang, Zhiwu Huang, and Xiaopeng Hong. 2023. Benchmarking Deepart Detection. *arXiv preprint:2302.14475* (2023).
[75] Zhendong Wang et al. 2023. DIRE for Diffusion-Generated Image Detection. In *Proceedings of ICCV*.
[76] Yuxin Wen, John Kirchenbauer, Jonas Geiping, and Tom Goldstein. 2023. Tree-Rings Watermarks: Invisible Fingerprints for Diffusion Images. In *Proc. of NeurIPS*.
[77] Cam Wilson. 2024. AI is producing 'fake' Indigenous art trained on real artists' work without permission. Crickey.com.au.
[78] YentaMagenta. 2023. Hive AI image "detection" is inaccurate and easily defeated. Reddit.
[79] Lijun Zhang et al. 2024. Robust Image Watermarking using Stable Diffusion. *arXiv preprint:2401.04247* (2024).
[80] Mingjian Zhu et al. 2023. GenImage: A Million-Scale Benchmark for Detecting AI-Generated Image. *arXiv preprint:2306.08571* (2023).

Appendix

Detection Results of DIRE Models. Table 12 shows the detection results using 6 DIRE checkpoints, using on our dataset of human artwork and AI-generated images. We test each model with the same set of 630 unperturbed images: 280 human artworks and 350 AI-generated images.

Additional Results on Hive. Table 13 shows the distribution of images across art style and generative model, which are "easy-to-detect" by Hive and unaffected by all five perturbations.

Additional Information on Data Collection. Table 14 lists the modifications made to the extracted BLIP captions, which are then used to prompt individual AI generators. Figure 9 shows examples of five types of perturbations considered in our study.

Results on Unusual Images. Figure 10 plots, for both hybrid images and upscaled human art, the distribution of the "raw" detection decision represented by the 5-point Likert score.

DIRE model		ACC(%)↑	FPR(%)↓	FNR(%)↓
Training dataset	Generation model			
ImageNet	ADM	54.44	96.43	4.86
LSUN	PNDM	54.76	100.00	1.43
LSUN	StyleGAN	55.40	99.29	0.86
LSUN	ADM	55.08	99.29	1.42
LSUN	iDDPM	54.45	97.86	3.71
CelebA-HQ	SD-v2	51.59	25.36	66.86

Table 12: Performance of six DIRE checkpoint models on our dataset.

	CIVITAI	DALL-E 3	Firefly	MJv6	SDXL	Total
Anime	4	4	9	2	6	25
Cartoon	5	6	8	1	6	26
Fantasy	10	4	5	0	6	25
Oil/Acrylic	6	6	1	0	1	14
Photography	1	3	0	0	2	6
Sketch	2	1	1	2	3	9
Watercolor	5	5	7	1	6	24
Total	33	29	31	6	30	129

Table 13: The number of AI-generated images for which Hive is >99% confident across all perturbations.

Modified Prompt for Each Style	
Style	Modified Prompt: added the phrase
Anime	"anime"
Cartoon	"a cartoon style image of"
Fantasy	"a fantasy style image of"
Oil/Acrylic	"an oil and acrylic painting of"
Photography	"a photography of"
Sketch	"a sketch drawing of"
Watercolor	"a watercolor painting of"

Table 14: Phrases added to BLIP captions extracted from human arts to correct art styles.

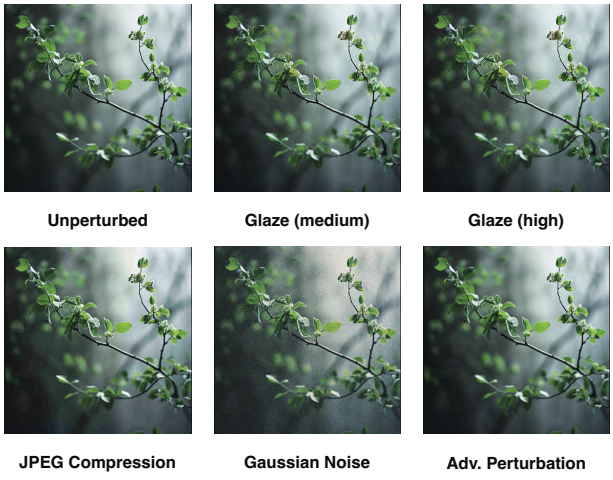


Figure 9: Samples of five different perturbations considered by our study.

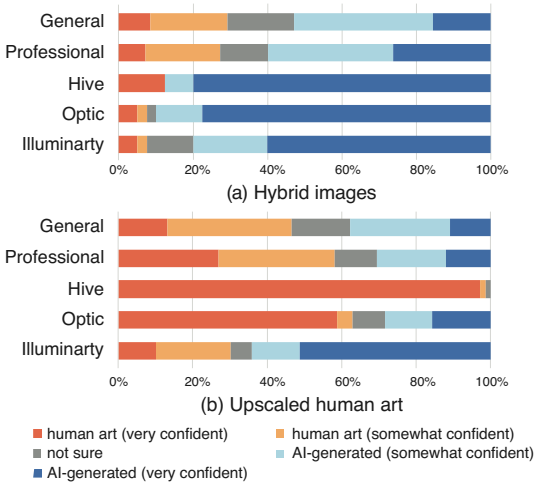


Figure 10: Distribution of detection decision represented by the 5-point Likert rating on (a) hybrid images and (b) upscaled human artworks.