Alison Sun

Hive's Best-in-Class AI-Generated Image Detection Outperforms Competitors: Leading the Way in Distinguishing Human Art from AI

Primary Keywords: Hive, AI-generated images, image detection, detection accuracy, human art

Secondary Keywords: AI art, generative AI, adversarial perturbations, automated detectors, human reviewers

Meta:

In a new study, Hive's AI-Generated Image Detector outshines human reviewers and competitors with a 98.03% accuracy rate. Discover key findings and the future of AI image detection.

Hive is thrilled to announce that in a recent study by the University of Chicago, our industry-leading AI-Generated Image Detector outperformed both human reviewers and commercial and research AI image detectors, achieving a nearly perfect success rate with no false positives. The paper, titled 'Organic or Diffused: Can We Distinguish Human Art from AI-generated Images?' explores the effectiveness of both AI image detectors and human reviewers in distinguishing AI-generated images from original artwork. Read on to learn some key findings and what they mean for the future of AI image detection.

At a Glance

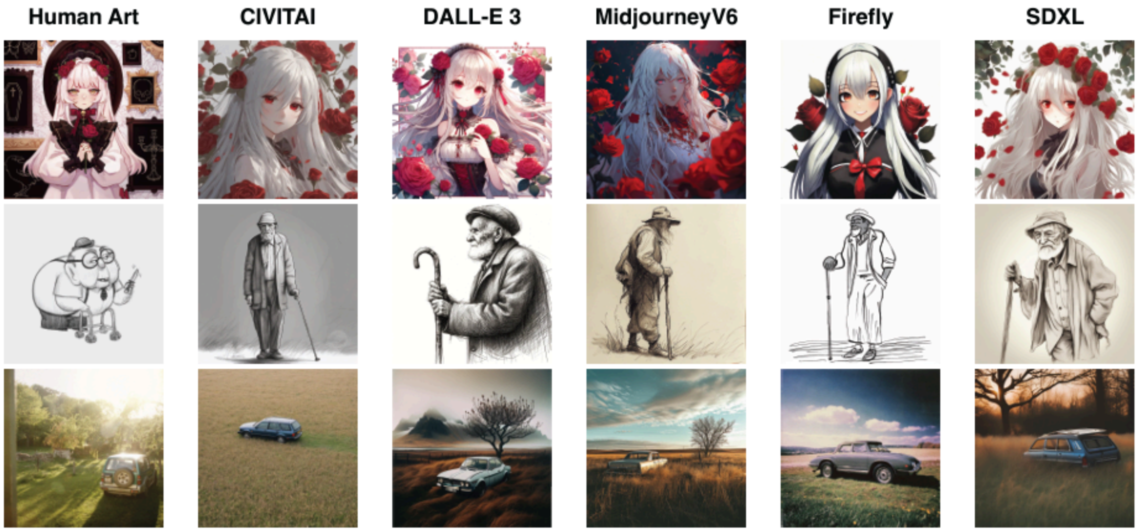
- With a 98.03% success rate, Hive performs leaps and bounds ahead of human reviewers and other AI image detectors. However, performance takes a noticeable hit when faced with adversarial perturbations and overlays.
- Non-artist users are generally unable to tell AI-generated imagery from human art. Artists do a bit better, while artists with specific experience in generative AI image detection are the best at spotting the difference.
- A mixed team of human artists working with machine classifiers will yield the best results at distinguishing original art from AI images.

The Importance of Detecting AI-Generated Images

Generative AI images have drastically shaken up the art world. With text prompts as short as a single word, AI models like Midjourney, Stable Diffusion and DALL-E 3 can generate stunning works of art that, to the untrained eye, appear professional. Unfortunately, that means AI images have often been passed off as human art to defraud unknowing customers and bypass copyright laws. There's also the worry of AI-generated imagery encroaching on human artistic expression. With these growing concerns, the ability to reliably detect AI images is more important than ever. But can our current systems really tell the difference between human artwork and AI-generated images?

Study Overview

Researchers curated a custom dataset of 280 original artworks across art styles ranging from anime to photography, along with 350 matching AI-generated images created using the models CIVITAI [13], DALL-E 3, MidjourneyV6, Adobe Firefly, and Stable Diffusion XL. (Several examples are shown below.) To diversify the data, researchers also included "hybrid" AI images altered by humans and human artworks upscaled with AI.

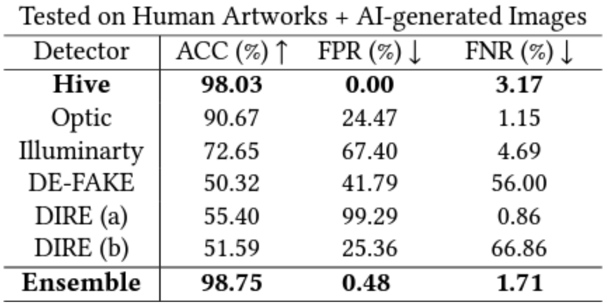


Alt text: Three rows of six images, where each row contains one original artwork created by a human artist, and five matching images sourced via generative AI models. The first row contains anime-style portraits of a girl with long white hair and red flowers in the background. The second row consists of black-and-white sketches of an elderly man with a cane. The third row contains photographs of an old blue car in a field of grass.

Images were evaluated with three commercial supervised learning classifiers (Hive, Optic, and Illuminarty) and two research-based detectors targeting diffusion models (DIRE and DE-FAKE). Researchers also ran a user study in which images were evaluated on a six-point scale by three groups of reviewers: non-artist general users, artists, and expert artists with specific skill in identifying AI-generated images. Each group looked at randomly selected batches of 20 images.

Automated Detector Results

As shown below, Hive performed the best overall, with 98.03% accuracy (ACC) and a 3.17% false negative rate (FNR). Notably, Hive was the only model with a zero false positive rate (FPR) — even expert human artists tended to slip up and mistakenly flag real human artworks as AI-generated. It was followed by Optic and Illuminarty, which performed a bit worse than Hive with high FPR. DE-FAKE and DIRE both performed poorly with low accuracy rates (< 55.5%) and high false positive and negative rates, which researchers attributed to the fact that their training data didn't include much artwork.



Tested on Human Artworks + AI-generated Images

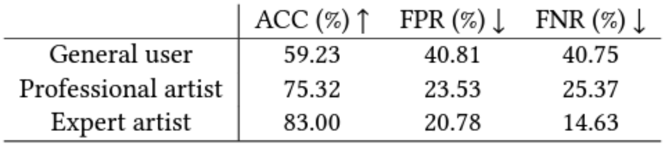
Detector	ACC (%) ↑	FPR (%) ↓	FNR (%) ↓
Hive	**98.03**	**0.00**	**3.17**
Optic	90.67	24.47	1.15
Illuminarty	72.65	67.40	4.69
DE-FAKE	50.32	41.79	56.00
DIRE (a)	55.40	99.29	0.86
DIRE (b)	51.59	25.36	66.86
Ensemble	**98.75**	**0.48**	**1.71**

Alt text: Table of detection accuracies, false positive rates, and false positive rates across all six AI image detectors' performance of human artworks and AI-generated images.

Human Detector Results

As shown in the table below, general users, with a 59.23% detection accuracy, weren't very good at telling human and AI-generated images apart. Professional artists did a bit better, with 75.32% detection accuracy overall. Accuracy discrepancies between artists and non-artists can be attributed to greater exposure to different art styles, and more knowledge and experience in art creation.

As to be expected, the expert artists performed the best with 83% accuracy. Despite their skill in analyzing nitty-gritty details and inconsistencies in AI-generated images, expert artists were still likely to make false positive errors, often mistaking art mistakes for evidence of AI.

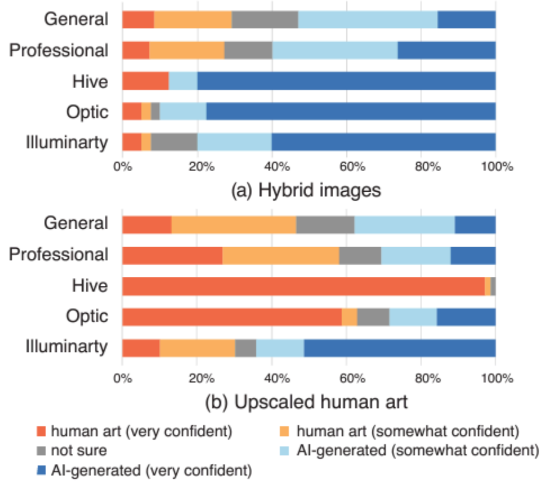


	ACC (%) ↑	FPR (%) ↓	FNR (%) ↓
General user	59.23	40.81	40.75
Professional artist	75.32	23.53	25.37
Expert artist	83.00	20.78	14.63

Alt text: Table of detection accuracies, false positive rates, and false negative rates across all three groups of human reviewers where detection accuracy rises and false positive and negative rates fall with increased reviewer experience.

Detection Accuracy on Unusual Images

Researchers also wanted to see how detectors would treat images that were a mix of human and AI art — whether these were hybrid AI-generated images edited by human users, or human artworks upscaled with AI tools. For the most part, the hybrid artworks were more likely to be flagged as AI art by both human and AI detectors, while upscaled images tended to be labeled as human art. From these results, we can see that extra edits, whether done by AI or humans, don't really affect the final decisions of either human or machine detectors.



Alt text: Bar chart of AI image detectors' judgments on hybrid and human art, ranked on a five-point Likert scale.

The Impacts of Adversarial Perturbations

Another goal of the study was to see how well detectors could handle AI-generated images altered to avoid detection using several types of perturbations: JPEG compression, Gaussian noise, adversarial perturbation via CLIP, and Glaze applied at medium and high intensities. For this comparison, we created a table that defines each perturbation type and reports how it affects the performance of each automated detector.

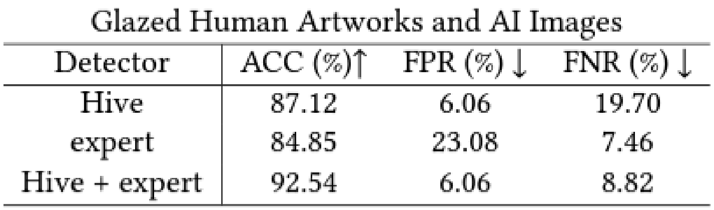
Perturbation Type	What it Does	The Result
JPEG compression	Reduces file sizes by changing some color values and grouping together blocks of like-colored pixels.	Minimal impact on performance; detection accuracy levels decrease but remain above 91% overall
Gaussian noise	Applies Gaussian signal noise to each pixel value.	Significant drops in detection accuracy for Optic and Illuminarty, minimal impact on Hive
CLIP-based adversarial perturbations	Creates and applies pixel-level perturbations designed to confuse the CLIP AI detection model.	The smallest drops in detection accuracy across all AI detectors
Glaze	Introduces imperceivable perturbations to each artwork via a popular tool for protecting art from plagiarism. Applied at medium and high intensities.	~6% drop in detection accuracy for Illuminarty, ~30% for Hive and Optic

Though Hive's model had the best detection rates for unperturbed images, adversarial perturbations applied with Glaze resulted in the largest drop in detection accuracy across all four perturbation types. Interestingly enough, adding Glaze to human artworks didn't affect classification success, which is probably due to differences in the availability of Glazed human artwork vs. Glazed AI images online — Glaze is super popular among artists, but it's not often used on AI-generated images.

A follow-up to the main study looked into a claim made on Reddit that overlaying a semi-transparent image of a white wall over an AI-generated image could trick Hive's detector. This trick was indeed proven to be valid — the researchers found that typically, overlaying the wall image at 60-80% intensity would change Hive's classification. The method of application mattered as well; using Adobe Photoshop, rather than the researchers' scripted blending algorithm, had a stronger effect.

Human Reviewers and AI Image Detectors

Clearly, both humans and automated detection tools face unique challenges in telling the difference between human art and AI-generated imagery. Researchers also looked into a scenario combining Hive accuracy scores with that of an expert human artist, which aimed to combine the strengths of both classification methods. This found that the best detection accuracy can be achieved when both Hive and an expert artist detector team up to review Glazed human artworks and AI images.



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

Alt text: Table of detection accuracies, false positive rates, and false negative rates between Hive, an expert artist, and the combined efforts of both Hive and the artist in assessing Glazed human artworks and AI images.

Final Takeaways

In light of this study, we're so excited to be able to continuously say our models are really and truly best-in-class. Nevertheless, the findings clearly outline areas for future improvement, specifically in detecting adversarial perturbations. Here at Hive, we're committed to continuously improving our models to stay ahead of the game and ensure that they're the best they can be. Research like this greatly informs the way we build and refine our products.

Finally, the best results come from the combined efforts of human reviewers and automated detectors, showing just how important it is for us to work together with AI to ensure that we have the tools to reliably and consistently tell apart human-created art and AI-generated imagery.

Read the full paper here, and learn more about Hive's best-in-class AI-generated image detection models here.

Ha, A. Y. J., Passananti, J., Bhaskar, R., Shan, S., Southen, R., Zheng, H., & Zhao, B. Y. (2024). Organic or Diffused: Can We Distinguish Human Art from AI-generated Images? arXiv preprint arXiv:2402.03214. Retrieved from https://arxiv.org/abs/2402.03214.