DeepMind

RESEARCH

Identifying AI-generated images with SynthID

29 AUGUST 2023

Sven Gowal, Pushmeet Kohli

Share

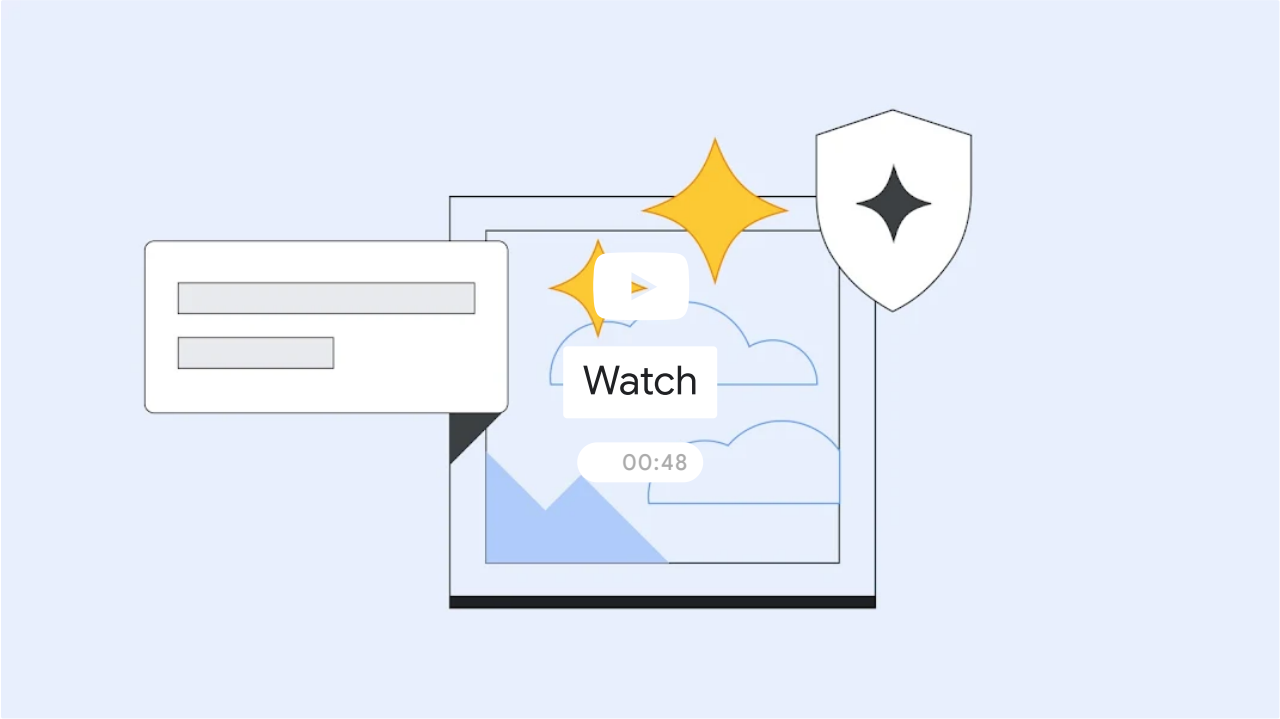
New tool helps watermark and identify synthetic images created by Imagen

AI-generated images are becoming more popular every day. But how can we better identify them, especially when they look so realistic?

Today, in partnership with Google Cloud, we're launching a beta version of SynthID, a tool for watermarking and identifying AI-generated images. This technology

DeepMind

SynthID is being released to a limited number of Vertex AI customers using Imagen, one of our latest text-to-image models that uses input text to create photorealistic images.



Generative AI technologies are rapidly evolving, and computer generated imagery, also known as 'synthetic imagery', is becoming harder to distinguish from those that have not been created by an AI system.

While generative AI can unlock huge creative potential, it also presents new risks, like enabling creators to spread false information — both intentionally or unintentionally. Being able to identify AI-generated content is critical to empowering people with knowledge of when they're interacting with generated media, and for helping prevent the spread of misinformation.

We're committed to connecting people with high-quality information, and upholding trust between creators and users across society. Part of this responsibility is giving users more advanced tools for identifying AI-generated images so their images — and even some edited versions — can be identified at a later date.

SynthID generates an imperceptible digital watermark for AI-generated images.

Google Cloud is the first cloud provider to offer a tool for creating AI-generated images responsibly and identifying them with confidence. This technology is grounded in our approach to developing and deploying responsible AI, and was developed by Google DeepMind and refined in partnership with Google Research.

SynthID isn't foolproof against extreme image manipulations, but it does provide a promising technical approach for empowering people and organisations to work with AI-generated content responsibly. This tool could also evolve alongside other AI models and modalities beyond imagery such as audio, video, and text.

New type of watermark for AI images

Watermarks are designs that can be layered on images to identify them. From physical imprints on paper to translucent text and symbols seen on digital photos today, they've evolved throughout history.

Traditional watermarks aren't sufficient for identifying AI-generated images because they're often applied like a stamp on an image and can easily be edited out. For example, discrete watermarks found in the corner of an image can be cropped out with basic editing techniques.

Finding the right balance between imperceptibility and robustness to image manipulations is difficult. Highly visible watermarks, often added as a layer with a name or logo across the top of an image, also present aesthetic challenges for creative or commercial purposes. Likewise, some previously developed

DeepMind

The watermark is detectable even after modifications like adding filters, changing colours and brightness.

We designed SynthID so it doesn't compromise image quality, and allows the watermark to remain detectable, even after modifications like adding filters, changing colours, and saving with various lossy compression schemes — most commonly used for JPEGs.

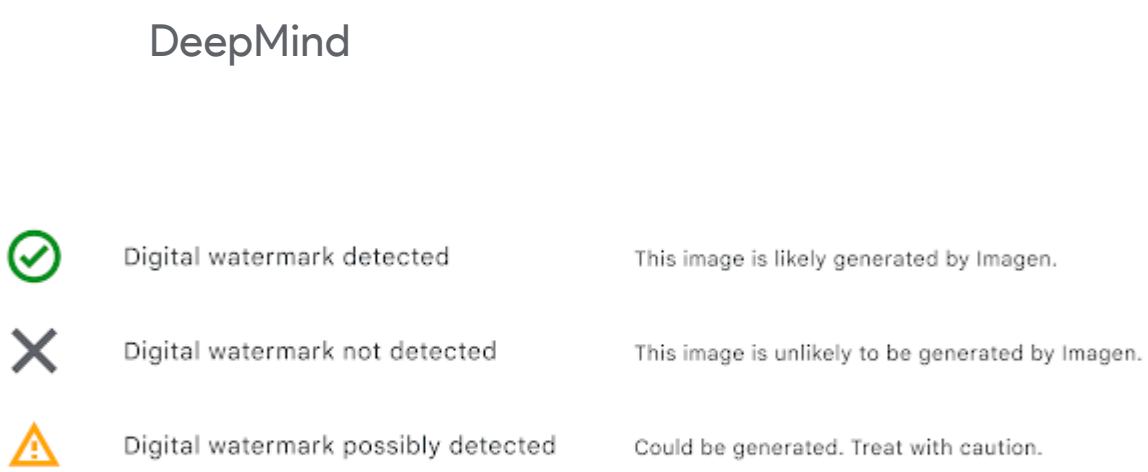
SynthID uses two deep learning models — for watermarking and identifying — that have been trained together on a diverse set of images. The combined model is optimised on a range of objectives, including correctly identifying watermarked content and improving imperceptibility by visually aligning the watermark to the original content.

Robust and scalable approach

SynthID allows Vertex AI customers to create AI-generated images responsibly and to identify them with confidence. While this technology isn’t perfect, our internal testing shows that it’s accurate against many common image manipulations.

SynthID's combined approach:

- **Watermarking**: SynthID can add an imperceptible watermark to synthetic images produced by Imagen.
- **Identification**: By scanning an image for its digital watermark, SynthID can assess the likelihood of an image being created by Imagen.



SynthID can help assess how likely it is that an image was created by Imagen.

This tool provides three confidence levels for interpreting the results of watermark identification. If a digital watermark is detected, part of the image is likely generated by Imagen.

SynthID contributes to the broad suite of approaches for identifying digital content. One of the most widely used methods of identifying content is through metadata, which provides information such as who created it and when. This information is stored with the image file. Digital signatures added to metadata can then show if an image has been changed.

When the metadata information is intact, users can easily identify an image. However, metadata can be manually removed or even lost when files are edited. Since SynthID's watermark is embedded in the pixels of an image, it's compatible with other image identification approaches that are based on metadata, and remains detectable even when metadata is lost.

What's next?

To build AI-generated content responsibly, we're committed to developing safe, secure, and trustworthy approaches at every step of the way — from image generation and identification to media literacy and information security.

These approaches need to be robust and adaptable as generative models advance and expand to other mediums. We hope our SynthID technology can work together with a broad range of solutions for creators and users across society, and we're

DeepMind

SynthID could be expanded for use across other AI models and we're excited about the potential of integrating it into more Google products and making it available to third parties in the near future — empowering people and organisations to responsibly work with AI-generated content.

Note: The model used for producing synthetic images in this blog may be different from the model used on Imagen and Vertex AI.

Learn more about SynthID

Read the Google Cloud announcement

Acknowledgements

This project was led by Sven Gowal and Pushmeet Kohli, with key research and engineering contributions from (listed alphabetically): Rudy Bunel, Jamie Hayes, Sylvestre-Alvise Rebuffi, Florian Stimberg, David Stutz, and Meghana Thotakuri.

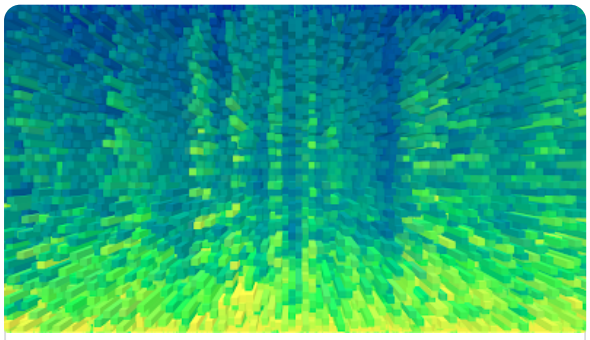
Thanks to Nidhi Vyas and Zahra Ahmed for driving product delivery; Chris Gamble for helping initiate the project; Ian Goodfellow, Chris Bregler and Oriol Vinyals for their advice. Other contributors include Paul Bernard, Miklos Horvath, Simon Rosen, Olivia Wiles, and Jessica Yung. Thanks also to many others who contributed across Google DeepMind and Google, including our partners at Google Research and Google Cloud.

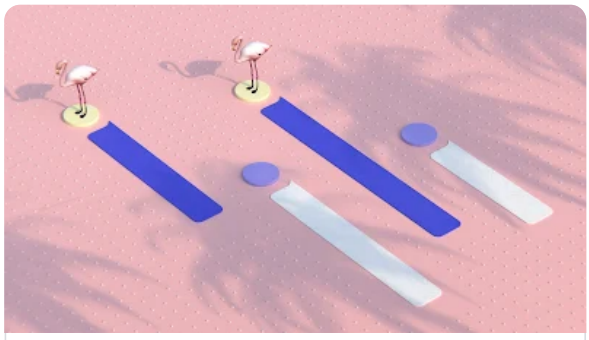
DeepMind

Watermarked image of a metallic butterfly with prismatic patterns on its wings

Related posts

View all posts





DeepMind

Robust and scalable tool for watermarking and identifying A...

tackling multiple tasks with a single visual language model

We introduce Flamingo, a single visual language model (VLM)...

28 APRIL 2022

COMPANY

Building a culture of pioneering responsibly

When I joined DeepMind as COO, I did so in large part...

24 MAY 2022

Follow us





Sign up for updates on our latest innovations

Sign up

I accept Google's Terms and Conditions and acknowledge that my information will be used in accordance with Google's Privacy Policy.

Google DeepMind

Google About Google Google products Privacy Terms