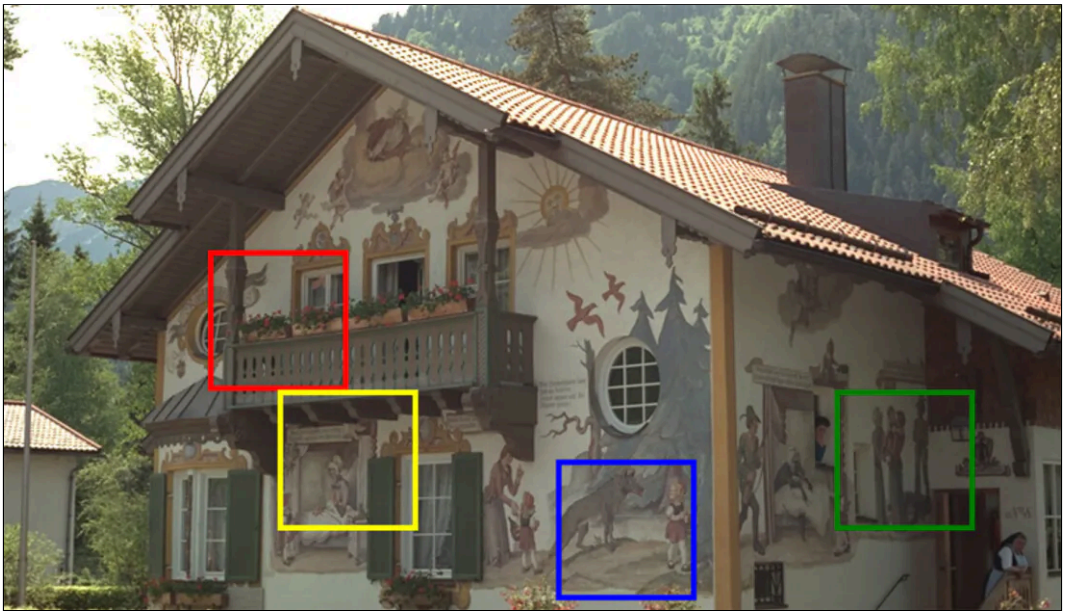
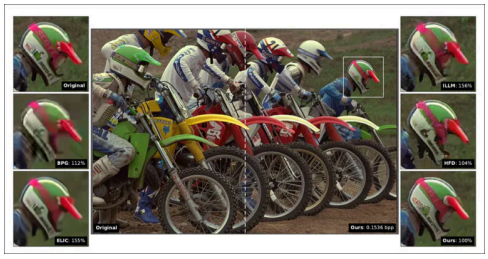
# Disney Research Offers Improved AI-Based Image Compression – But It May Hallucinate Details

Originally published at https://www.unite.ai/disney-research-offers-improved-ai-based-image-compression-but-it-may-hallucinate-details/



Disney's Research arm is offering a new method of compressing images, leveraging the open source Stable Diffusion V1.2 model to produce more realistic images at lower bitrates than competing methods.



*The Disney compression method compared to prior approaches. The authors claim improved recovery of detail, while offering a model that does not require hundreds of thousands of dollars of training, and which operates faster than the nearest equivalent competing method.* Source: https://studios.disneyresearch.com/app/uploads/2024/09/Lossy-Image-Compression-with-Foundation-Diffusion-Models-Paper.pdf

The new approach (defined as a 'codec' despite its increased complexity in comparison to traditional codecs such as JPEG and AV1) can operate over any Latent Diffusion Model (LDM). In quantitative tests, it outperforms former methods in terms of accuracy and detail, and requires significantly less training and compute cost.

The key insight of the new work is that *quantization error* (a central process in all image compression) is similar to *noise* (a central process in diffusion models).

Therefore a 'traditionally' quantized image can be treated as a noisy version of the original image, and used in an LDM's denoising process instead of random noise, in order to reconstruct the image at a target bitrate.


*Further comparisons of the new Disney method (highlighted in green), in contrast to rival approaches.*

The authors contend:

> *'[We] formulate the removal of quantization error as a denoising task, using diffusion to recover lost information in the transmitted image latent. Our approach allows us to perform less than 10% of the full diffusion generative process and requires no architectural changes to the diffusion model, enabling the use of foundation models as a strong prior without additional fine tuning of the backbone.*

> *'Our proposed codec outperforms previous methods in quantitative realism metrics, and we verify that our reconstructions are qualitatively preferred by end users, even when other methods use twice the bitrate.'*

However, in common with other projects that seek to exploit the compression capabilities of diffusion models, the output may hallucinate details. By contrast, lossy methods such as JPEG will produce clearly distorted or over-smoothed areas of detail, which can be recognized as compression limitations by the casual viewer.

Instead, Disney's codec may alter detail from context that was not there in the source image, due to the coarse nature of the Variational Autoencoder (VAE) used in typical models trained on hyperscale data.

> *'Similar to other generative approaches, our method can discard certain image features while synthesizing similar information at the receiver side. In specific cases, however, this might result in inaccurate reconstruction, such as bending straight lines or warping the boundary of small objects.*
>
> *'These are well-known issues of the foundation model we build upon, which can be attributed to the relatively low feature dimension of its VAE.'*

While this has some implications for artistic depictions and the verisimilitude of casual photographs, it could have a more critical impact in cases where small details constitute essential information, such as evidence for court cases, data for facial recognition, scans for Optical Character Recognition (OCR), and a wide variety of other possible use cases, in the eventuality of the popularization of a codec with this capability.

At this nascent stage of the progress of AI-enhanced image compression, all these possible scenarios are far in the future. However, image storage is a hyperscale global challenge, touching on issues around data storage, streaming, and electricity consumption, besides other concerns. Therefore AI-based compression could offer a tempting trade-off between accuracy and logistics. History shows that the best codecs do not always win the widest user-base, when issues such as licensing and market capture by proprietary formats are factors in adoption.

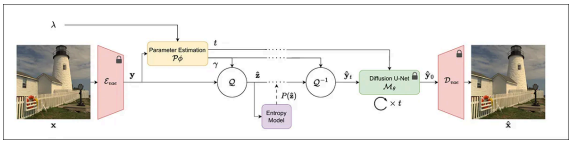
Disney has been experimenting with machine learning as a compression method for a long time. In 2020, one of the researchers on the new paper was involved in a VAE-based project for improved video compression.

The  new Disney paper was updated in early October. Today the company released an accompanying YouTube video. The project is titled *Lossy Image Compression with Foundation Diffusion Models*, and comes from four researchers at ETH Zürich (affiliated with Disney's AI-based projects) and Disney Research. The researchers also offer a supplementary paper.

## Method

The new method uses a VAE to encode an image into its compressed latent representation. At this stage the input image consists of derived features – low-level vector-based representations. The latent embedding is then quantized back into a bitstream, and back into pixel-space.

This quantized image is then used as a template for the noise that usually seeds a diffusion-based image, with a varying number of denoising steps (wherein there is often a trade-off between increased denoising steps and greater accuracy, vs. lower latency and higher efficiency).




*Schema for the new Disney compression method.*

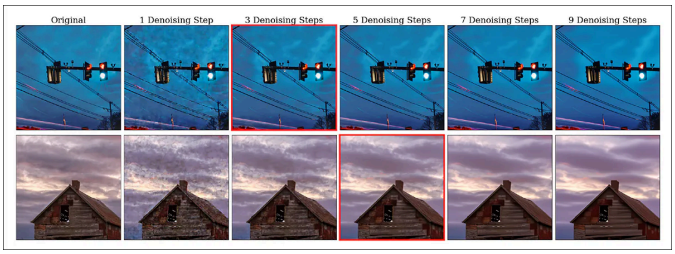
Both the quantization parameters and the total number of denoising steps can be controlled under the new system, through the training of a neural network that predicts the relevant variables related to these aspects of encoding. This process is called *adaptive quantization*, and the Disney system uses the Entroformer framework as the entropy model which powers the procedure.

The authors state:

> *'Intuitively, our method learns to discard information (through the quantization transformation) that can be synthesized during the diffusion process. Because errors introduced during quantization are similar to adding [noise] and diffusion models are functionally denoising models, they can be used to remove the quantization noise introduced during coding.'*

Stable Diffusion V2.1 is the diffusion backbone for the system, chosen because the entirety of the code and the base weights are publicly available. However, the authors emphasize that their schema is applicable to a wider number of models.

Pivotal to the economics of the process is *timestep prediction*, which evaluates the optimal number of denoising steps – a balancing act between efficiency and performance.




*Timestep predictions, with the optimal number of denoising steps indicated with red border. Please refer to source PDF for accurate resolution.*

The amount of noise in the latent embedding needs to be considered when making a prediction for the best number of denoising steps.

## Data and Tests

The model was trained on the Vimeo-90k dataset. The images were randomly cropped to 256x256px for each epoch (i.e., each complete ingestion of the refined dataset by the model training architecture).

The model was optimized for 300,000 steps at a learning rate of 1e-4. This is the most common among computer vision projects, and also the lowest and most fine-grained generally practicable value, as a compromise between broad generalization of the dataset's concepts and traits, and a capacity for the reproduction of fine detail.

The authors comment on some of the logistical considerations for an economic yet effective system*:

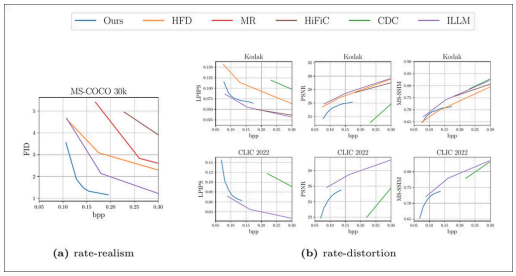
> *'During training, it is prohibitively expensive to backpropagate the gradient through multiple passes of the diffusion model as it runs during DDIM sampling. Therefore, we perform only one DDIM sampling iteration and directly use [this] as the fully denoised [data].'*

Datasets used for testing the system were Kodak; CLIC2022; and COCO 30k. The dataset was pre-processed according to the methodology outlined in the 2023 Google offering *Multi-Realism Image Compression with a Conditional Generator*.

Metrics used were Peak Signal-to-Noise Ratio (PSNR); Learned Perceptual Similarity Metrics (LPIPS); Multiscale Structural Similarity Index (MS-SSIM); and Fréchet Inception Distance (FID).

Rival prior frameworks tested were divided between older systems that used Generative Adversarial Networks (GANs), and more recent offerings based around diffusion models. The GAN systems tested were High-Fidelity Generative Image Compression (HiFiC); and ILLM (which offers some improvements on HiFiC).

The diffusion-based systems were *Lossy Image Compression with Conditional Diffusion Models* (CDC) and *High-Fidelity Image Compression with Score-based Generative Models* (HFD).



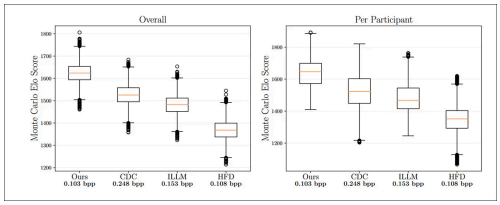

*Quantitative results against prior frameworks over various datasets.*

For the quantitative results (visualized above), the researchers state:

> *'Our method sets a new state-of-the-art in realism of reconstructed images, outperforming all baselines in FID-bitrate curves. In some distortion metrics (namely, LPIPS and MS-SSIM), we outperform all diffusion-based codecs while remaining competitive with the highest-performing generative codecs.*
>
> *'As expected, our method and other generative methods suffer when measured in PSNR as we favor perceptually pleasing reconstructions instead of exact replication of detail.'*

For the user study, a two-alternative-forced-choice (2AFC) method was used, in a tournament context where the favored images would go on to later rounds. The study used the Elo rating system originally developed for chess tournaments.

Therefore, participants would view and select the best of two presented 512x512px images across the various generative methods. An additional experiment was undertaken in which *all* image comparisons from the same user were evaluated, via a Monte Carlo simulation over 10,0000 iterations, with the median score presented in results.




*Estimated Elo ratings for the user study, featuring Elo tournaments for each comparison (left) and also for each participant, with higher values better.*

Here the authors comment:

> *'As can be seen in the Elo scores, our method significantly outperforms all the others, even compared to CDC, which uses on average double the bits of our method. This remains true regardless of Elo tournament strategy used.'*

In the original paper, as well as the supplementary PDF, the authors provide further visual comparisons, one of which is shown earlier in this article. However, due to the granularity of difference between the samples, we refer the reader to the source PDF, so that these results can be judged fairly.

The paper concludes by noting that its proposed method operates twice as fast as the rival CDC (3.49 vs 6.87 seconds, respectively). It also observes that ILLM can process an image within 0.27 seconds, but that this system requires burdensome training.

## Conclusion

The ETH/Disney researchers are clear, at the paper's conclusion, about the potential of their system to generate false detail. However, none of the samples offered in the material dwell on this issue.

In all fairness, this problem is not limited to the new Disney approach, but is an inevitable collateral effect of using diffusion models – an inventive and interpretive architecture – to compress imagery.

Interestingly, only five days ago two other researchers from ETH Zurich produced a paper titled *Conditional Hallucinations for Image Compression*, which examines the possibility of an 'optimal level of hallucination' in AI-based compression systems.

The authors there make a case for the desirability of hallucinations where the domain is generic (and, arguably, 'harmless') enough:

> *'For texture-like content, such as grass, freckles, and stone walls, generating pixels that realistically match a given texture is more important than reconstructing precise pixel values; generating any sample from the distribution of a texture is generally sufficient.'*

Thus this second paper makes a case for compression to be optimally 'creative' and representative, rather than recreating as accurately as possible the core traits and lineaments of the original non-compressed image.

One wonders what the photographic and creative community would make of this fairly radical redefinition of 'compression'.

**My conversion of the authors' inline citations to hyperlinks.*

*First published Wednesday, October 30, 2024*

Writer on machine learning, domain specialist in human image synthesis. Former head of research content at Metaphysic.ai, until its dissolution into DNEG's Brahma.ai.
Portfolio site: martinanderson.ai
Contact: martin@martinanderson.ai