Are Images Indistinguishable to Humans Also Indistinguishable to Classifiers?

Zebin You [1] Xinyu Zhang [2] Hanzhong Guo [1] Jingdong Wang [3] Chongxuan Li [1]

[1]Gaoling School of Artificial Intelligence, Renmin University of China, Beijing, China [2]The University of Adelaide [3]Baidu VIS

Australian Institute for Machine Learning

Funded by the Centre for Augmented Reasoning

Abstract

The ultimate goal of generative models is to perfectly capture the data distribution. For image generation, common metrics of visual quality (e.g., FID) and the perceived truthfulness of generated images seem to suggest that we are nearing this goal. However, through distribution classification tasks, we reveal that, from the perspective of neural network-based classifiers, even advanced diffusion models are still far from this goal. Specifically, classifiers are able to consistently and effortlessly distinguish real images from generated ones across various settings. Moreover, we uncover an intriguing discrepancy: classifiers can easily differentiate between diffusion models with comparable performance (e.g., U-ViT-H vs. DiT-XL), but struggle to distinguish between models within the same family but of different scales (e.g., EDM2-XS vs. EDM2-XXL). Our methodology carries several important implications. First, it naturally serves as a diagnostic tool for diffusion models by analyzing specific features of generated data. Second, it sheds light on the model autophagy disorder and offers insights into the use of generated data: augmenting real data with generated data is more effective than replacing it. Third, classifier guidance can significantly enhance the realism of generated images.

Game: Name-That-True

Figure 1. **Four-way distribution classification tasks:** *Which one is real in each row?* The samples are from real images or generated from state-of-the-art diffusion models. Notably, classifiers consistently and effortlessly distinguish between real and generated images in all settings.

Highlights

- We show that classifiers easily distinguish between diffusion-generated and real distributions, despite diffusion models achieving low FID scores and generating lifelike images (see Fig. 1).
- We reveal the intriguing contradictions between classifier performance and widely used evaluation metrics, such as FID and human judgments.
- We demonstrate that our approach complements traditional metrics like FID and can serve as a diagnostic tool to provide deeper insights into diffusion models.
- We provide a reasonable explanation for the model autophagy disorder phenomenon and show that augmenting real data with generated data is more effective than replacing it in supervised and semi-supervised learning.
- We demonstrate that classifier guidance can be used to enhance the realism of generated images.

Generated distributions are easily classified as generated

Distribution combinations	Model	Accuracy
U-H/2, *D*	ResNet-50	99.66
	ConvNeXt-T	97.59
	ViT-S	86.13
U-H/2, *D*, *I*-256	ResNet-50	99.87
	ConvNeXt-T	99.77
	ViT-S	95.90
U-H/2, *D*, *I*-256, *E*2-XXL	ResNet-50	99.91
	ConvNeXt-T	99.92
	ViT-S	98.13

Table 1. **Distribution classification accuracy for various combinations of diffusion models with similar FIDs and ImageNet**. Notably, classifier accuracy improves as more distributions are added. Here, combinations refer to multiple distributions, where the number of distributions corresponds to the number of classes in distribution classification.

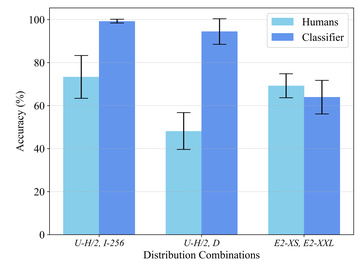


Figure 2. **User study.** Classifiers can easily distinguish between diffusion models with similar FIDs (U-H/2, D) but struggle with models from the same family that differ in parameters (E2-XS, E2-XXL). In contrast, humans show the opposite trend.

More training data improves distinguishing different generated images

Training samples	Model	Accuracy (%)
5k	ResNet-50	88.28
10k	ResNet-50	93.45
5k	ViT-S	76.33
10k	ViT-S	83.53
5k	ConvNeXt-T	80.53
10k	ConvNeXt-T	85.13

Table 2. **Four-way distribution classification on *text-to-image***. All classifiers yield high accuracy in distinguishing four distributions: COCO, Pixart-α, SDXL, and Playground-v2.5, using only 5k or 10k training samples per dataset. Notably, as the number of training samples increases, the accuracy of distribution classification consistently improves.

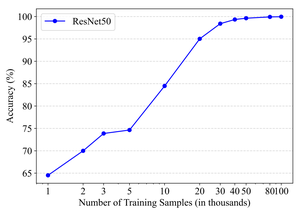


Figure 3. **Binary distribution classification on *label-to-image***. A positive correlation is observed between accuracy and the number of training samples. With only 50k samples per distribution, classifiers achieve over 99.5% accuracy.

Diagnosing problem within diffusion models

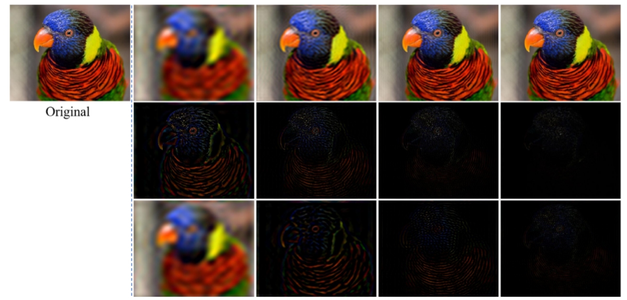
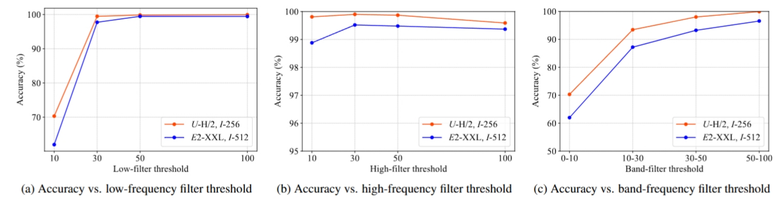


Figure 4. **Visualization of frequency domain processing.** *Top*: Low-pass filters, *Middle*: High-pass filters, each processed with increasing thresholds: 10, 30, 50, 100. *Bottom*: Band-pass filters, processed with band thresholds: 0-10, 10-30, 30-50, 50-100.



(a) Accuracy vs. low-frequency filter threshold (b) Accuracy vs. high-frequency filter threshold (c) Accuracy vs. band-frequency filter threshold

Figure 5. **Classifier achieves high accuracy across various combinations with limited frequency components,** except when only minimal low-frequency components are present (in subfig (a) low-pass filter threshold = 10, in subfig (c) band-pass filter threshold: 0-10).

Diffusion models **effectively learn low-frequency features**, making it more difficult for classifiers to distinguish real from generated images in this range. However, in higher frequency bands, diffusion models perform worse, making it easier for classifiers to identify generated images.

Evaluating the use of generated data

Method given # labels per class	Error rate ↓ 4 (0.08%)
FlexMatch*	4.97±0.06
FlexMatch†	5.85±0.02
FreeMatch*	4.90±0.04
FreeMatch†	5.94±0.10
FreeMatch‡	4.68±0.17

Table 3. **Augmenting real data with generated data is more effective than replacing it.** * indicates results from training with CIFAR-10 (real data); † and ‡ indicate results where CIFAR-10 is replaced or augmented with EDM_S, respectively.

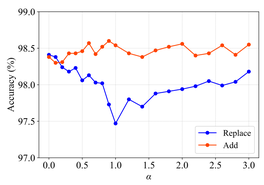


Figure 4. **Top-1 accuracy on the CIFAR-10 test set: Replacement vs. Augmentation.** α represents the ratio of EDM_S mixed with CIFAR-10.

Due to **distribution mismatch**, replacing real data with generated data degrades classification task performance, whereas **augmenting real data with generated data** generally improves performance.