ECE 580 Final Project
GenEvalDetect: Image Generation, Evaluation, and Detection via Diffusion Models

Yiming Mao, Enmu Liu

Abstract: In this project, we explore high-quality image generation using the Stable Diffusion model. While existing evaluation metrics such as Inception Score (IS) [1] and Fréchet Inception Distance (FID) [2] are widely used to assess generative models, they often rely on single-layer features, which may overlook multi-scale perceptual quality. To address this limitation, we propose a novel evaluation metric, Multi-Scale Fréchet Inception Distance (MSFID), which computes Fréchet distances across multiple intermediate layers of the Inception-V3 network. MSFID captures both low-level textures and high-level semantics, providing a more comprehensive assessment of generated image quality. In addition, we introduce a ViT-based binary classifier trained on real and diffusion-generated images from ImageNet. The model leverages the strong representation capabilities of pretrained ViTs and achieves high accuracy in distinguishing synthetic images, providing a reliable tool for diffusion image detection.

1. Introduction

The recent advancement in diffusion models has revolutionized the field of image generation, enabling the synthesis of high-resolution, photo-realistic images from textual descriptions. Among these, Stable Diffusion stands out due to its efficiency, open availability, and strong performance across diverse prompts. However, evaluating the perceptual quality of generated images remains a challenging task. Traditional metrics such as Inception Score (IS) and Fréchet Inception Distance (FID) offer limited insight, as they typically rely on single-layer feature representations and may not fully capture multi-scale perceptual fidelity.

In this project, we aim to address both the evaluation and authenticity detection aspects of diffusion-based image generation. First, we propose a novel metric called Multi-Scale Fréchet Inception Distance (MSFID), which extends FID by computing Fréchet distances across multiple layers of the Inception-V3 network. This allows MSFID to assess low-level texture fidelity and high-level semantic alignment simultaneously, yielding a more comprehensive evaluation of generative quality.

Second, we introduce a Vision Transformer (ViT)-based binary classifier for distinguishing between real and AI-generated images. Leveraging the strong global representation capabilities of ViTs pretrained on ImageNet, our model is fine-tuned on a curated dataset of real and synthetic images, achieving strong performance in detecting diffusion-generated content.

Together, MSFID and ViT-based detection form a robust framework for generation, evaluation, and authenticity verification of images produced by diffusion models.

2. Related Work

2.1. Personalized Image Generation with DreamBooth

DreamBooth [3] is a fine-tuning method for personalization in diffusion-based image synthesis. It allows a pretrained diffusion model, such as Stable Diffusion, to generate novel images of a specific subject given only a few (typically 3–5) reference images. By associating a unique identifier token with the subject and incorporating class prior preservation during training, DreamBooth enables the generation of diverse, identity-preserving images under varied prompts. It has become a widely adopted technique for subject-driven generation in both research and creative applications.

2.2. Evaluation Metrics for Generative Models

The Fréchet Inception Distance (FID) [2] and Inception Score (IS) [1] are two of the most popular quantitative metrics for evaluating the performance of generative models. FID computes the Fréchet distance between the

feature distributions of real and generated images using the Inception-V3 network. While FID captures both the quality and diversity of generated images, it relies on a single high-level feature layer, potentially missing low-level perceptual discrepancies. IS, on the other hand, measures both the clarity and class diversity of generated images but does not compare them to real data. These limitations have motivated the development of our Multi-Scale Fréchet Inception Distance (MSFID), which computes FID scores across multiple intermediate layers of Inception-V3 to better reflect both semantic and perceptual fidelity.

2.3. Detection of AI-Generated Images

As diffusion models become increasingly realistic, detecting AI-generated images has emerged as a critical research problem in digital media forensics and trustworthiness. Several recent works have proposed CNN- and transformer-based classifiers for real-vs-fake detection [4, 5]. Vision Transformer (ViT)-based models [6] have shown particular promise due to their ability to capture global structure and texture statistics. Fine-tuning pre-trained ViT backbones on curated datasets of real and synthetic images enables accurate detection, even when synthetic content mimics real-world statistics. In this project, we leverage a ViT-based binary classifier to identify images synthesized by diffusion models and demonstrate high detection accuracy under realistic test conditions.

3. Multi-Scale Fréchet Inception Distance

3.1. Definition

The Fréchet Inception Distance (FID) is a widely used metric for evaluating the quality of images generated by generative models. It measures the distance between the distributions of real and generated images by computing the Fréchet distance between two multivariate Gaussians fitted to the features extracted from the Inception-V3 network.

However, conventional FID is limited in that it only considers a single high-level feature layer (typically `Mixed_7c`). To address this, we propose the **Multi-Scale Fréchet Inception Distance (MSFID)**, which leverages multiple layers of the Inception-V3 network—such as `Mixed_5d`, `Mixed_6e`, and `Mixed_7c`—to capture both low-level textures and high-level semantics.

Given the feature activations from each layer ℓ, we compute the Fréchet distance as:

$$\mathrm{FID}_\ell = \left\| \mu_\ell^{(r)} - \mu_\ell^{(g)} \right\|^2 + \mathrm{Tr}\left(\Sigma_\ell^{(r)} + \Sigma_\ell^{(g)} - 2\sqrt{\Sigma_\ell^{(r)} \Sigma_\ell^{(g)}} \right)$$

The final MSFID score is the average across L selected layers:

$$\mathrm{MSFID} = \frac{1}{L} \sum_{\ell=1}^{L} \mathrm{FID}_\ell$$

where $\mu_\ell^{(r)}, \Sigma_\ell^{(r)}$ are the mean and covariance of real image features at layer ℓ, and $\mu_\ell^{(g)}, \Sigma_\ell^{(g)}$ are those of the generated images.

3.2. Interpretation and Rating Criteria

The MSFID score provides a multi-level evaluation of image fidelity and diversity. A lower MSFID indicates that the generated images are closer in distribution to real images, across both semantic and perceptual levels.

We propose the following rating criteria based on empirical observations on ImageNet-scale datasets:

MSFID Range	Rating	Quality Description
$[0, 20)$	5 / 5	Nearly indistinguishable from real images
$[20, 40)$	4 / 5	High quality, perceptually consistent
$[40, 70)$	3 / 5	Moderate quality, some artifacts visible
$[70, 100)$	2 / 5	Low quality, noticeable distribution gap
$100+$	1 / 5	Poor quality, significant artifacts or failures

4. ViT-Based Detection of AI-Generated Images

In this work, we develop a binary classification framework based on the Vision Transformer (ViT) architecture to distinguish AI-generated images from real-world photographs. Our approach leverages the strong representation

learning capabilities of ViT models, particularly the `vit-base-patch16-224-in21k` variant pretrained on ImageNet-21K, and fine-tunes it for the detection task on a curated dataset consisting of synthetic and authentic images.

4.1. Data Collection and Preprocessing

We organize the dataset in a class-balanced imagefolder format with two categories: `adm` for AI-generated images and `real` for natural images. The dataset is split into training (800 samples), validation (100 samples), and test (100 samples) sets. To ensure consistency with the ViT input expectations, each image is resized and center-cropped to 224×224 pixels, then normalized using the ViT-specific mean and standard deviation values.

4.2. Model Fine-tuning

We initialize the ViT model using the pretrained weights of `vit-base-patch16-224-in21k`. The original classifier head is replaced to support binary classification, with label mappings defined as `adm` → 0 and `real` → 1. The model is fine-tuned using the Hugging Face `Trainer` API with the following configuration: batch size of 16, AdamW optimizer, and a training schedule of 3 epochs. During training, we monitor the accuracy metric on the validation set and retain the best-performing model checkpoint.

4.3. Inference

For prediction, we apply the same preprocessing pipeline and feed the transformed image into the trained ViT model. The predicted class label is determined by applying `argmax` over the model logits. The prediction framework supports both single image and directory-level inference and outputs human-readable labels indicating whether an image is classified as AI-generated or real.

5. Experiments

5.1. Diffusion-Based Image Generation

To generate subject-specific images, we fine-tune the pre-trained Stable Diffusion v1.5 model using the DreamBooth method. DreamBooth enables the synthesis of personalized images based on a small number of reference images (typically 3–5) of a target concept, such as a pet or object. It achieves this by introducing a unique identifier token into the text encoder and optimizing the diffusion model to associate that token with the subject's visual appearance.

Our implementation follows the DreamBooth training pipeline, including class prior preservation to avoid overfitting and semantic drift. We fine-tune all parameters of the U-Net while freezing the VAE and text encoder. The training data includes both instance images of the subject and regularization images sampled from a general category (e.g., "a cat"). The final model is capable of generating high-quality, diverse images of the subject in novel scenes and poses based on prompt conditioning.

In this experiment, we fine-tuned the pre-trained Stable Diffusion v1.5 model using the DreamBooth method to achieve personalized image generation for a specific cat subject. DreamBooth enables subject-driven generation by learning a new concept (such as a pet) from a few instance images, while maintaining the model's generalization ability.

Instance Images: We selected several photos of the target cat subject as shown in Figure 1. These instance images were resized to 512×512 to meet the input requirement of Stable Diffusion. They serve as the core concept that the model learns to reproduce in novel scenes.

Class Prior Preservation: To prevent overfitting and semantic drift, we introduced additional regularization images belonging to the same class (cats) but unrelated to the target subject (see Figure 2). These regularization images help the model retain general knowledge about cats and avoid collapsing to the instance concept exclusively.

Prompt Engineering: During training and inference, each image is paired with a prompt containing a custom token (e.g., "aaa cat"), where "aaa" is a unique identifier for the subject. For example, we used prompts such as "a aaa cat sitting in Times Square, New York City, with buildings and billboards around." This allows the model to generate images with the personalized cat in different contexts.

Generated Results: After training, we generated a batch of images using variations of the prompt. As shown in Figure 3, the model successfully reconstructs the identity of the target cat while adapting it to novel settings like Times Square.

Fig. 1. Instance image of the target cat

Fig. 2. Class prior image for regularization

Fig. 3. Generated images using the prompt "a aaa cat sitting in Times Square, New York City, with buildings and billboards around."

5.2. Quantitative Comparison of ADM and DreamBooth Models using FID and MSFID

To evaluate the quality of image generation under different model settings, we conducted a quantitative comparison using Fréchet Inception Distance (FID) and Multi-Scale Fréchet Inception Distance (MSFID). Specifically, we compare the baseline ADM model (a denoising diffusion probabilistic model trained on ImageNet) with a DreamBooth-finetuned Stable Diffusion model that incorporates subject-specific generation using personalized prompts.

ADM on ImageNet. Table 1 reports the FID and MSFID scores of ADM-generated images evaluated against real ImageNet samples across the train, test, and validation splits. The dataset is obtained from https://github.com/ZhendongWang6/DIRE. The FID ranges from 136.86 on the training set to 270.38 on the validation set, showing notable degradation in quality as the model is evaluated on out-of-distribution data. A similar trend is observed for MSFID, suggesting that the alignment between the real and generated distributions becomes increasingly difficult in unseen domains.

Table 1. FID and MSFID results of ADM-generated images compared with ImageNet real images

Split	Model	FID	MSFID
Train	ADM (Stable Diffusion)	136.86	45.95
Test	ADM (Stable Diffusion)	259.49	87.38
Val	ADM (Stable Diffusion)	270.38	90.99

DreamBooth Fine-tuning with Subject Guidance. Table 2 presents the results of a DreamBooth-finetuned model evaluated against ImageNet real images. Unlike the generic ADM model, DreamBooth is trained to capture the visual identity of a specific subject (e.g., a user-defined cat) and generate images under customized prompt contexts. As a result, it achieves very low MSFID scores on early Inception layers (Mixed_5d: 3.90, Mixed_6e: 2.24), indicating high perceptual quality at the local level. However, the FID score is significantly higher (182.17),

and the MSFID score at the deepest semantic layer (Mixed_7c) is also large, reflecting that DreamBooth-generated images deviate substantially from the original ImageNet class distribution in high-level semantic space.

Table 2. FID and MSFID results of DreamBooth-generated images

Data	Model	FID	MSFID	Mixed_5d	Mixed_6e	Mixed_7c
DreamBooth (cat)	DreamBooth fine-tuned	182.17	62.77	3.90	2.24	182.17

This result is expected: DreamBooth is not designed to reproduce generic class-level content but rather to synthesize subject-specific outputs based on personalized textual prompts. While it performs exceptionally well in preserving low-level perceptual realism, its high-level semantic features diverge from the ImageNet class prior. This explains the high FID and deep-layer MSFID values. In this context, MSFID provides a more informative breakdown across network layers, revealing that DreamBooth maintains local realism while sacrificing global distributional alignment.

5.3. Evaluation of Binary Classification for Diffusion Detection

To evaluate the effectiveness of our ViT-based binary classifier for detecting AI-generated (fake) versus authentic (real) images, we performed inference on a held-out test set comprising 10 fake images generated via diffusion models and 10 authentic images sourced from the ImageNet validation set. These images were completely unseen during training or validation.

The prediction outcomes are summarized below:

- **Fake (adm)**: All 10 synthetic images were correctly classified as `adm`, indicating the model's high sensitivity to diffusion-specific visual patterns.
- **Real (real)**: 9 out of 10 real images were correctly identified as `real`, with a single misclassification as `adm`.

The corresponding confusion matrix is:

True / Pred	adm (Fake)	real (Real)
adm (Fake)	10	0
real (Real)	1	9

This yields an overall detection accuracy of **95%**, with a precision of 90% and recall of 100% for the `real` class. Notably, the model achieves perfect detection on AI-generated content, demonstrating strong generalization to synthetic distributions unseen during training.

Misclassification Analysis. The only misclassified real image exhibits low contrast and slight over-smoothing, resembling characteristics commonly present in diffusion-generated samples. This highlights a known limitation of image detectors: when diffusion-generated content achieves high perceptual realism, the boundary between real and fake becomes visually ambiguous even to human observers.

Robustness and Generalization. Despite being fine-tuned on a relatively small dataset (800 training images), the model generalizes remarkably well, likely due to the strong inductive bias and pretrained visual features of the ViT backbone. The robustness against unseen generation styles suggests that ViT-based detectors can effectively capture global and structural cues that distinguish synthetic distributions from natural ones.

Future work may include expanding the test set, introducing adversarially generated examples, and incorporating interpretability tools (e.g., Grad-CAM) to visualize the attention regions contributing to each decision.

6. Conclusion and Limitations

In this project, we propose a unified framework for diffusion-based image generation, evaluation, and detection. Our contributions include MSFID—a perceptually grounded metric for multi-scale evaluation—and a ViT-based binary classifier for robust detection of synthetic images. Empirical results demonstrate the effectiveness of both components, with the detector achieving 95% accuracy and MSFID providing interpretable quality scores.

However, our method is currently limited to binary detection and may be sensitive to generation domains outside of ImageNet. Future work may expand to multi-class detection and explore adversarial robustness under diverse generation styles.

References

1. T. Salimans, I. Goodfellow, W. Zaremba, V. Cheung, A. Radford, and X. Chen, "Improved techniques for training GANs," in *Adv. Neural Inf. Process. Syst.*, vol. 29 (2016).
2. M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter, "GANs trained by a two time-scale update rule converge to a local Nash equilibrium," in *Adv. Neural Inf. Process. Syst.*, vol. 30 (2017).
3. N. Ruiz, Y. Li, V. Jampani, Y. Pritch, M. Rubinstein, and K. Aberman, "DreamBooth: Fine tuning text-to-image diffusion models for subject-driven generation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pp. 22500–22510 (2023).
4. S. Y. Wang, O. Wang, R. Zhang, A. Owens, and A. A. Efros, "CNN-generated images are surprisingly easy to spot... for now," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pp. 8695–8704 (2020).
5. A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929* (2020).
6. Z. Zhang, X. Lu, G. Cao, Y. Yang, L. Jiao, and F. Liu, "ViT-YOLO: Transformer-based YOLO for object detection," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021, pp. 2799–2808.