AI-Generated Image Detection

Project Background introduction

The proliferation of AI-generated images poses an unprecedented challenge to information integrity. High-fidelity synthetic images now permeate digital ecosystems, with the most sophisticated examples evading human detection. This capability has enabled malicious actors to manipulate public discourse for strategic and financial gain. Consequently, robust and efficient methods for distinguishing AI-generated from authentic image have become critically necessary.

This study systematically evaluates existing detection approaches, identifies their limitations, and explores avenues for enhancement. We establish a comprehensive benchmark and propose incremental improvements to advance the state of detection capabilities.

Team Information

Member	Student ID	Primary Responsibility
DOU,Mingze	21211793	Contribute to Transformer Solutions Analysis
SHEN, Wenyi	21205744	Contribute to Diffusion Solutions Analysis
XU, Zeling	21214680	Contribute to Frequency Domain Solutions Analysis
ZHANG, Shengqi	21209697	Contribute to CNN Solutions Analysis

Methodologies

Self Attention

The Vision Transformer (ViT), renowned for its powerful global modeling capability, has gained widespread recognition in AI-generated image detection due to its high accuracy and robustness. Various ViT variants—such as those trained on different datasets, DeiT (designed for small-sample scenarios), Swin Transformer, and CvT—are fundamentally built upon the original ViT architecture with structural modifications. Given the high cost and complexity of training from scratch, and considering the availability of well-established pretrained models, this study adopts the ViT-B/16 model pretrained on ImageNet-21k as the backbone network. Input images are resized to 224×224 RGB format, and the model leverages self-attention mechanisms to capture long-range dependencies across image patches. A single fully connected layer serves as the classification head, mapping the CLS token output to a binary classification space. During fine-tuning, the backbone is frozen while only the classification head is trained. Should performance prove insufficient, additional feature extraction layers may be introduced atop the backbone for further refinement.

Detection Model for Images Generated by Diffusion

The rise of diffusion models has spurred the development of specialized detection methods. Taking DIRE as an example, its core idea is to leverage a pre-trained diffusion model to perform "inversion-reconstruction" on input images and compute the reconstruction error. Research reveals that images generated by diffusion models can be well reconstructed with minimal error, while real images exhibit significantly larger reconstruction errors. However, this model requires large datasets and has slow inference speed. In contrast, the lightweight NPR approach starts from the generator architecture, focusing on the strong local pixel correlations introduced by upsampling operations in generated images - correlations absent in real images. NPR constructs feature maps based on neighboring pixel relationships and trains a CNN for classification. These detectors designed for diffusion models demonstrate excellent cross-model generalization capability in practice and remain robust to common image perturbations. Both models represent viable deployment solutions, and through parameter tuning and CNN structure fine-tuning, they may achieve enhanced generalization across diverse datasets.

Frequency Domain Analysis

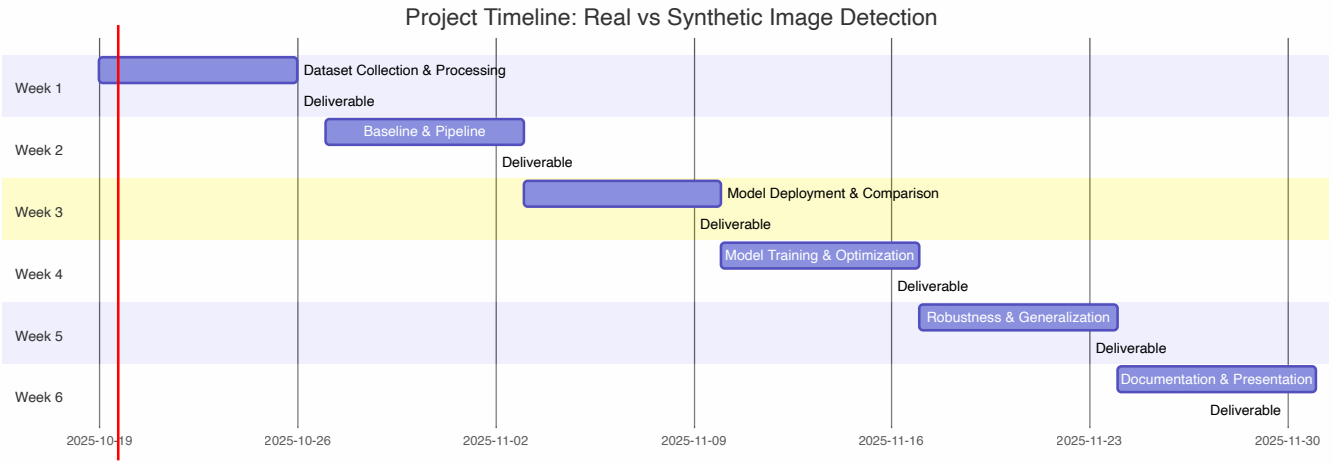
While spatial-domain methods such as CNNs and Vision Transformers have demonstrated success in detecting AI-generated images, they face fundamental challenges when confronting the latest generation of sophisticated generative models. Recent research has revealed that traditional approaches often struggle against cutting-edge diffusion transformers, with detection accuracy sometimes dropping to near-random levels. This limitation has prompted exploration of alternative detection paradigms that can capture more persistent artifacts in synthetic imagery.

Frequency domain analysis emerges as a promising approach because it targets the fundamental mathematical signatures left by the generation process itself. Regardless of how photorealistic an AI-generated image appears, the architectural constraints of generative models inevitably introduce characteristic patterns in the frequency spectrum. These patterns arise from upsampling operations inherent to both GAN and diffusion-based generators, creating spectral fingerprints that persist even as visual quality improves.

Convolutional Neural Networks (CNNs)

CNNs definitely revolutionized the computer vision field when they were raised since their hierarchical feature extraction naturally aligns with image detection tasks by progressively learning from pixel-level patterns to high-level semantics. Generated images, on the other hand, are prone to anomalies in details such as edges, lighting, color, and local texture; which can be used to distinguish generated images from real ones. Obviously, they have the benefits like Simplicity and Efficiency, however they have drawbacks such as weak generalization ability and lack of explainability. These make it is not the best choice for image detecting task today.

Project Timeline



Week 1: Dataset Collection & Processing (Oct 19 - Oct 26)

We will collect real image datasets and generate synthetic images using multiple state-of-the-art generative models. The collected data will be preprocessed and split into training, validation, and test sets with proper stratification. We will also define evaluation metrics including accuracy, precision, recall, F1-score, etc. The deliverable for this week is a structured dataset with complete metadata and defined evaluation criteria.

Week 2: Baseline & Data Pipeline (Oct 27 - Nov 2)

We will implement a baseline classifier and establish a robust data loading pipeline with standard augmentation techniques. Initial training experiments will be conducted to verify the pipeline functionality. The deliverable includes a working training pipeline and baseline performance results.

Week 3: Model Deployment & Comparison (Nov 3 - Nov 9)

This week focuses on deploying and comparing multiple detection approaches. We will implement CNN-based models, diffusion-based detectors, Vision Transformer variants, and frequency domain analysis methods. Preliminary comparative experiments will be conducted to identify the most promising architectures. The deliverable is a comprehensive comparison of different model architectures with initial performance benchmarks.

Week 4: Model Training & Optimization (Nov 10 - Nov 16)

We will perform hyperparameter tuning and train selected models on the full dataset. This includes optimizing learning rates, batch sizes, augmentation strategies, and architecture-specific parameters. Model checkpoints will be saved at regular intervals. The deliverable is a set of well-trained model checkpoints ready for comprehensive evaluation.

Week 5: Robustness & Generalization Testing (Nov 17 - Nov 23)

We will evaluate trained models on the held-out test set and assess generalization capability on images from unseen generators. Robustness tests will be conducted under various perturbations (compression, noise, blur, etc.). Based on evaluation results, we will perform targeted optimization and fine-tuning to improve model performance. Interpretability analysis will be performed to understand model decision-making. The deliverable is a comprehensive evaluation report with optimized models.

Week 6: Documentation & Presentation (Nov 24 - Nov 30)

The final week is dedicated to writing the technical report and preparing presentation materials. We will complete code documentation and repository cleanup to ensure reproducibility. The deliverables include the final report, presentation slides, and a well-documented code repository.