ISSN 2616-5767 Vol.9, Issue 3: 66-70, DOI: 10.25236/AJETS.2026.090309

Generative-Model-Based AI Image Forgery Forensics and Efficient Detection Framework

Christina Yanlin Chiu[1,*], Andrew Chiu[2], Xu Xiaorui[3], Chen Ziyi[4], Fun Ming Yang[5]

[1]*Wyoming Seminary, Kingston, USA*
[2]*Wyoming Seminary, Kingston, USA*
[3]*Basis International School, Shenzhen, China*
[4]*Jinan Thomas School, Jinan, China*
[5]*St Michaels University School, British Columbia, Canada*

Abstract: *The rapid advancement of generative artificial intelligence has brought significant convenience to image processing but has also led to a surge in AI-forged images. In e-commerce, malicious actors exploit these tools to fabricate "damaged goods" images for refund fraud. To address this, we propose an efficient image forensics framework based on generative models to identify and trace AI-forged images. By integrating reverse forensics of generative models, this framework leverages a "magic defeats magic" approach. We introduce feature trajectory prediction and multimodal feature fusion to enhance the detection of subtle forgery traces in low-quality images. Furthermore, an efficient batch detection system using a fast-screening and fine-detection cascade is developed to meet the real-time processing demands of large-scale e-commerce platforms. The framework provides not only binary classification but also explainable forensics via heatmaps and frequency anomaly visualizations. Our approach demonstrates strong robustness and high accuracy, offering a scalable technical path for cross-domain applications including social media content verification, judicial authentication, and copyright protection.*

Keywords: *AI image forensics, Generative models, Feature trajectory, Multimodal fusion, E-commerce fraud*

1. Introduction

In recent years, the rapid development of Generative Artificial Intelligence (Generative AI) technologies has brought immense convenience to creative design, image processing, and e-commerce marketing. However, with the widespread availability of these tools, AI-generated or tampered images have been increasingly abused in real life. A particularly concerning trend has emerged in the e-commerce refund process, where AI forgery has become a novel form of fraud. For instance, consumers utilize AI inpainting or generation tools to digitally alter normal product photos into images showing "damage," "stains," or "quality issues." These fabricated images are then used to fraudulently claim "refund-only" compensations from merchants. Such behaviors not only cause direct financial losses to merchants but also severely damage the trust mechanisms and transaction fairness of e-commerce platforms. Traditional image moderation methods often rely on human experience or simple image feature comparisons, such as metadata checking. These conventional techniques struggle to cope with the highly realistic forgeries produced by modern generative models, meaning there is an urgent need for new technological means to enhance detection capabilities.

Based on this premise, this project proposes an efficient framework for forgery detection and AI image forensics based on generative models. The core philosophy of this framework is "using magic to defeat magic," which entails utilizing the inherent features and trajectories of generative models for reverse forensics to identify potential forgery traces within images. Unlike traditional methods that focus solely on pixel-level anomalies, this research introduces feature trajectory prediction and comparison technology. By modeling the latent space distribution during the generation process, the system significantly enhances its sensitivity to subtle manipulations.

ISSN 2616-5767 Vol.9, Issue 3: 66-70, DOI: 10.25236/AJETS.2026.090309

2. Related Work

Internationally, AI image forensics has become a prominent research hotspot in computer vision and digital forensics. Academic institutions and tech enterprises in Europe and the US are actively exploring methods based on Generative Adversarial Network (GAN) trace analysis, diffusion model anti-forensics, and frequency domain feature detection[1]. Tech giants like Meta and Google have established deepfake detection platforms to curb the spread of false content on social media. Emerging methods attempt to improve detection accuracy by predicting latent feature trajectories, demonstrating high efficiency and scalability advantages.

Domestically, research is primarily concentrated on e-commerce and judicial application scenarios. E-commerce platforms are collaborating with universities to promote the application of forged image detection and trusted computing technologies. Judicial authentication agencies are also beginning to explore AI image forensics for evidence review. However, overall domestic research remains application-driven. There is a relative lack of academic exploration regarding the reverse forensics of generative models and large-scale real-time detection, necessitating a comprehensive integration of international cutting-edge methods with local application needs[2-3].

3. Core Technologies

3.1. Generative Model Reverse Forensics

Generative models (such as Diffusion models and GANs) inevitably leave unique implicit traces during the image generation and editing process. These traces include noise residual distributions, spectral fingerprint anomalies, and inconsistencies in latent space trajectories. By conducting reverse analysis on these features, tampered or forged images can be effectively identified. This "AI identifying AI" approach is not only more robust than traditional pixel-level detection but also maintains high recognition rates in highly realistic forgery scenarios. The framework extracts hyperparameters and synthesizes real and fake image fingerprints directly from the model parsing flow to determine authenticity, as shown in Fig. 1.

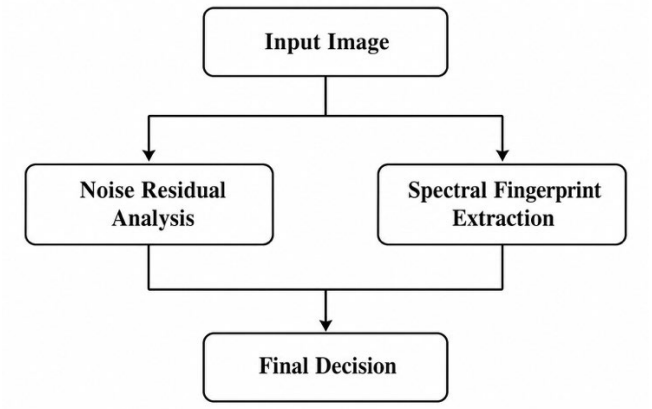

Figure 1: Architecture of the proposed generative model reverse forensics framework.

3.2. Feature Trajectory Prediction and Contrast

Real images and forged images often exhibit significant differences in their evolutionary trajectories within the latent space or diffusion process. By constructing the feature trajectory of an image and predicting its trend across different noise levels or latent dimensions, deviations in consistency and smoothness between tampered regions and normal regions can be captured. This method can discover subtle forgery traces that are imperceptible to the human eye, serving as a key technology for enhancing detection sensitivity. Typical errors include text spelling mistakes or missing generation details, which become evident through temporal trajectory tracking, as shown in Fig. 2.

ISSN 2616-5767 Vol.9, Issue 3: 66-70, DOI: 10.25236/AJETS.2026.090309

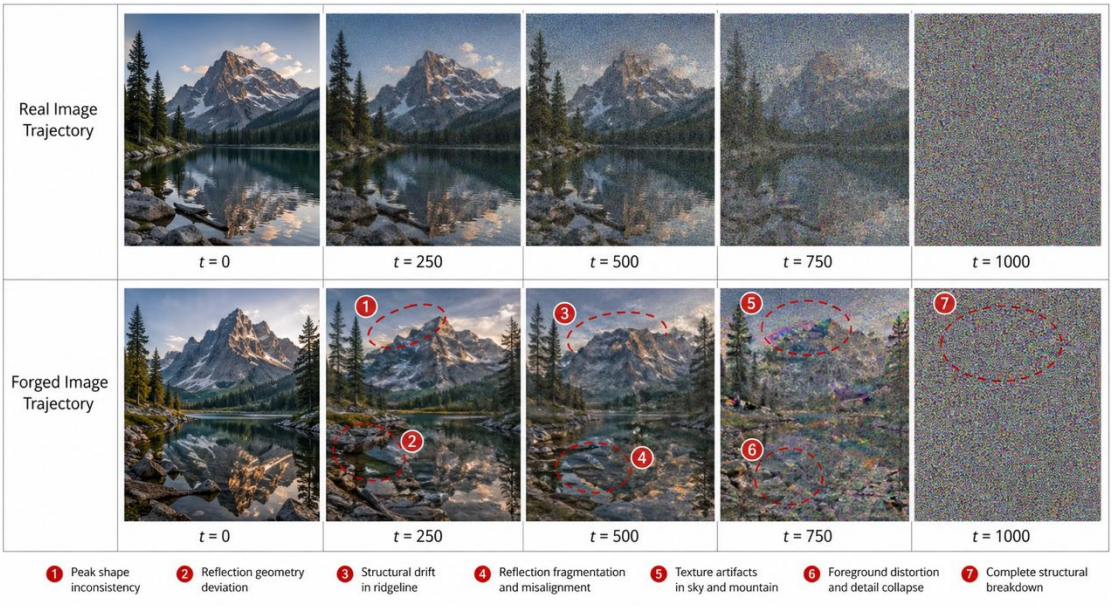

Figure 2: Illustration of feature trajectory deviations in forged regions compared to original image evolutionary paths across different noise levels.

3.3. Multimodal Feature Fusion

Detection methods relying on a single dimension are susceptible to interference from compression, scaling, or secondary editing. Therefore, integrating multiple features is essential to enhance robustness. Specifically, pixel-level anomaly detection (e.g., unnatural edges, local blurriness), frequency domain feature analysis (e.g., abnormal power spectrum distribution), and generative trace modeling (e.g., noise residuals or inversion errors) are combined. Through multimodal fusion, the detection system can identify forged images more comprehensively, improving stability across different domains and tasks.

3.4. Efficient Detection and Acceleration Mechanisms

In high-concurrency environments, such as massive e-commerce platforms, the detection system must be both accurate and highly efficient. To achieve this, lightweight technologies such as model distillation, pruning, and quantization are introduced. These are combined with a hierarchical screening strategy to realize a cascade architecture of "fast screening followed by fine detection." Furthermore, leveraging accelerated inference engines, batch processing, and multi-stream parallelism significantly reduces latency while maintaining detection accuracy, thereby supporting real-time and large-scale applications[4].

3.5. Explainable Forensics

Providing merely a "true/false" judgment is often insufficient as a basis for arbitration or legal evidence; thus, explainable forensics is crucial. By generating heatmaps, localizing suspicious regions, or visualizing spectral anomalies, the system clearly presents the likely forged areas. This approach bridges the gap between deep learning black-box predictions and human-interpretable forensic reporting, as shown in Fig.3.

ISSN 2616-5767 Vol.9, Issue 3: 66-70, DOI: 10.25236/AJETS.2026.090309

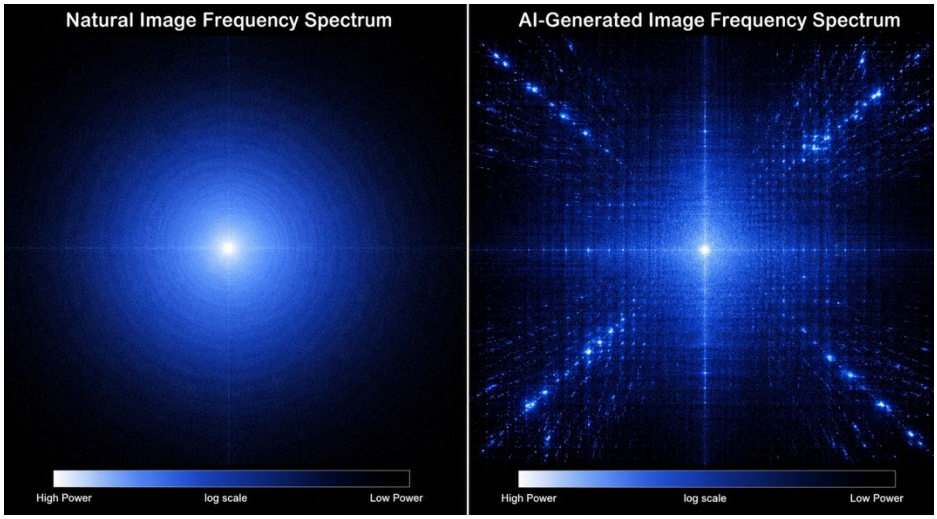

Figure 3: Comparison of Fourier Transform power spectrum between natural images and AI-generated images with high-frequency artifacts.

4. Implementation Details

4.1. Overall Architecture Design

The overall system adopts a two-stage cascaded architecture consisting of a fast screen layer and a fine detection layer. The frontend utilizes lightweight models and frequency domain heuristics for millisecond-level fast screening, rapidly diverting obviously normal or overtly suspicious samples. Samples falling within the intermediate threshold range proceed to the fine detection layer. The fine detection layer extracts features via three parallel branches (pixel anomalies, frequency domain features, forensic traces), followed by attention fusion. Simultaneously, a feature trajectory prediction module is integrated to measure the evolutionary consistency of the image. The final output provides a suspicious area heatmap, an evidence summary, and threshold-calibrated confidence scores to integrate with platform interception strategies.

4.2. Data and Annotation

The dataset comprises both real operational tickets (desensitized) and controllably generated forged samples. The forgeries include edited types (added stains, damages, molds) and generated types (diffusion/GAN local inpainting, global synthesis). Each image is provided with image-level authenticity labels and regional masks. Prior to training, common distortion augmentations (multi-level JPEG compression, scaling/sharpening/blurring, color shifts) are applied to construct robust data slices[5].

4.3. Tri-Branch Model Design

The model consists of three independent channels. The pixel channel analyzes image details and edges to determine unnatural blurs. The frequency channel converts the image to a spectral form to inspect high and low-frequency energy distributions. The forensics channel focuses on camera-specific fingerprints, such as sensor noise and color processing pipelines, which are difficult to maintain in forged images. The trajectory prediction module is incorporated to simulate the image's evolutionary path during generation, predicting the next step's features and comparing them with actual results to quantify anomalies.

4.4. Multi-task Training Process

During training, classification, localization, trajectory consistency, and forensic consistency losses are jointly optimized. This encourages the model to evaluate, locate, and explain the anomalies simultaneously. Online hard example mining enhances robustness against boundary samples and adversarial post-processing. Temperature scaling and isotonic regression are used for pre-threshold calibration, ensuring comparability of scores across different categories and resolutions.

ISSN 2616-5767 Vol.9, Issue 3: 66-70, DOI: 10.25236/AJETS.2026.090309

5. Evaluation and Experimental Results

The reliability of the results is analyzed and evaluated using standard metrics, along with robustness across JPEG, resolution, and edit type slices. The explainability of the model and its results is intuitively presented through heatmaps, spectral anomaly graphs, and inversion residual graphs, showing precisely where the image is suspicious and the underlying reasons, as is shown in Fig.4.

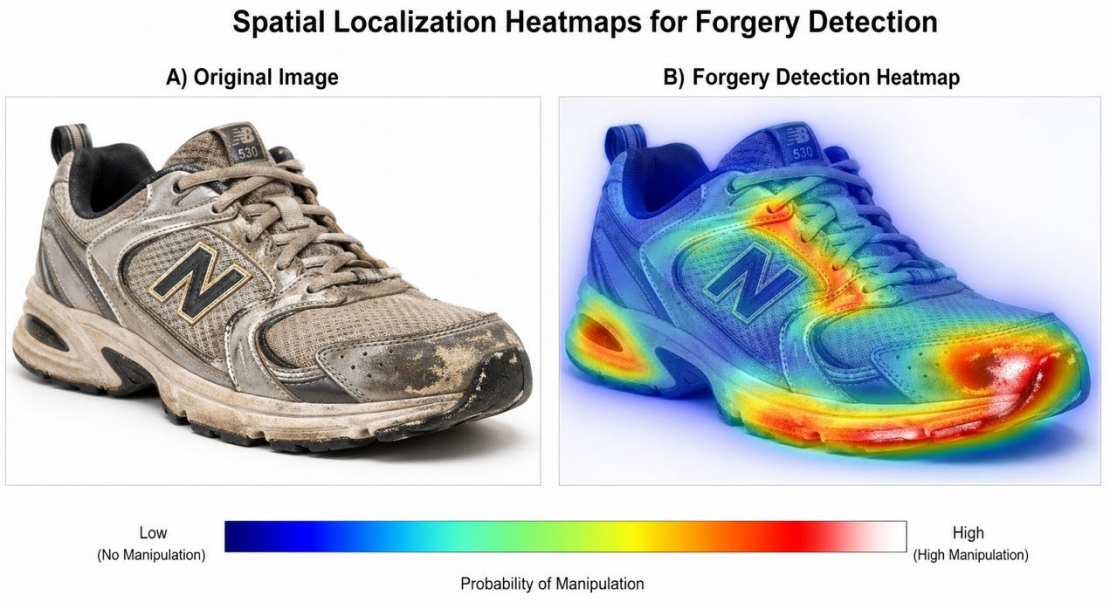

Figure 4: Spatial localization heatmaps demonstrating the system's capability to isolate and highlight AI-manipulated regions in e-commerce product imagery.

6. Conclusions

This project successfully constructs a highly efficient image forensics framework based on generative models, realizing the rapid identification and tracing of AI-forged images. By integrating feature trajectory prediction and multimodal fusion, the system significantly improves the detection capabilities for subtle forgery traces. Validated within the e-commerce refund fraud scenario, the system effectively intercepts malicious "P-shopped" refunds. Its scalable architecture and explainable outputs ensure its broad application potential across news media, judicial forensics, and copyright protection

References

[1] Wang H , Cheng R , Han C ,et al. Attribution as Retrieval: Model-Agnostic AI-Generated Image Attribution[J]. 2026.
[2] Li R , Wang X , Cui Y ,et al. A Semi-Supervised Diffusion-Based Framework for Weed Detection in Precision Agricultural Scenarios Using a Generative Attention Mechanism[J].AGRICULTURE-BASEL, 2025, 15(4):434.DOI:10.3390/agriculture15040434.
[3] Corvi, R., Cozzolino, D., Giudice, G., Poggi, G., Sansone, C. and Verdoliva, L. On the detection of synthetic images generated by diffusion models. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), (2023) 1-5.
[4] Wang, S.Y., Wang, O., Zhang, R., Owens, A. and Efros, A.A. CNN-generated images are surprisingly easy to spot... for now. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), (2020) 8695-8704.
[5] Rossler, A., Cozzolino, D., Verdoliva, L., Riess, C., Thies, J. and Nießner, M. FaceForensics++: Learning to detect manipulated facial images. Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), (2019) 1-11.