

Adaptive Multi-Generator Deep Learning for AI-Generated Image Detection

Aarti D. Galdhar¹, Urmila B. Shelke², Dr. H. B. Jadhav³

Student, Department of Computer Engineering¹

Guide, Department of Computer Engineering²

HOD, Department of Computer Engineering³

Adsul Technical Campus Chas, Ahilyanagar, India

Abstract: *Rapid progress in generative text-to-image architectures has significantly narrowed the perceptual gap between machine-generated visuals and authentic camera captures. Such rapid proliferation introduces critical vulnerabilities across digital forensics, automated content filtering, identity verification, and public media trust. A recurrent shortcoming across contemporary forensic classifiers is their narrow domain specificity: while they reliably identify synthetic images produced by architectures present within their training distributions, classification accuracy deteriorates when analysing samples from unfamiliar or recently introduced engines. Option 1 (Methodology-focused and precise): This study reviews key developments in multi-generator detection and presents an adaptive framework that pairs specialized CNN backbones with transfer learning, dynamic gating weights, and systematic degradation testing. Synthesizing peer-reviewed benchmark findings from 2023 through 2026 as an empirical foundation, this study focuses on architectural formulation rather than reporting an unverified experimental trial. Moving beyond conventional gated-expert designs, the proposed system incorporates broader generator clustering, systematic resistance to image degradation, calibrated prediction confidence, and dedicated out-of-distribution benchmarks. This strategy aims to attenuate single-generator overfitting, reinforce real-world resilience, and outline an efficient route toward modular deployment. Core data demands, operational limitations, and key research horizons are likewise addressed.*

Keywords: AI-generated images, image forensics, convolutional neural network, transfer learning, ensemble learning, gated network, deepfake detection, generative AI

I. INTRODUCTION

Text-to-image generative modelling has transitioned digital rendering from an intricate manual workflow into an accessible, prompt-driven utility. Current systems produce complex photorealistic portraits, commercial media, scenic landscapes, and physical documentation with minimal human intervention. Because modern synthesis models suppress perceptible boundary flaws and anatomical distortions, human inspection is no longer dependable. Establishing dependable computational techniques to differentiate real camera imagery from synthetic creations has consequently become a pressing focus in digital forensics.

The downstream risks of unchecked synthetic generation extend well beyond consumer platforms. Inconspicuous synthetic media can undermine judicial forensics, distort digital evidence, dilute journalistic credibility, compromise automated verification checkpoints, and complicate online provenance. Fabricated imagery can readily impersonate real individuals or substantiate fictional events. For these reasons, an effective forensic pipeline cannot rely solely on standard in-distribution validation; it must maintain predictable discriminative capability when confronted with unmodeled generators or real-world post-processing steps like resizing, compression, blurring, and spatial downscaling. A common pitfall in earlier deepfake detection literature involved training static binary classifiers over a homogeneous set of generative architectures. Under this setup, backbones frequently latch onto transient, engine-specific frequency



fingerprints rather than universal signatures of generative synthesis. Earlier work by Ojha et al. revealed that forensic models tuned on an isolated synthesis family generalize poorly to unfamiliar generative systems, prompting their use of pretrained multimodal representations to stabilize out-of-domain detection. The release of the GenImage benchmark that same year cemented the necessity of multi-generator and degradation-aware validation across balanced corpora.

Pursuing a distinct approach, a 2024 study by Saskoro, Yudistira, and Fatyanosa implemented parallel ResNet-50 sub-networks, each individually tuned on a specific generator and combined their intermediate features via a learned gating layer. Although this setup demonstrated remarkable fidelity on in-domain benchmarks averaging beyond 96% accuracy on known categories, its discriminative capability declined noticeably when confronted with entirely unmodeled generative architectures. This behaviour underlines both the utility of modular specialization and the ongoing vulnerability of learned gating models to unseen distributions.

Building upon the conceptual foundation of that study, this work does not introduce novel empirical measurements or replicate existing benchmark datasets. Instead, it synthesizes three years of empirical findings, isolates core out-of-distribution bottlenecks, and proposes an architectural framework tailored for practical deployment. The proposed system emphasizes structural generator diversity, robust handling of real-world visual degradations, calibrated uncertainty boundaries, and independent zero-shot evaluation protocols.

Problem Statement :

The fundamental challenge in current forensic detection is the operational gap separating static training corpora from the fluid, noisy conditions of production environments. A model may effectively recognize artifact patterns from familiar generators, yet those markers often shift dramatically across revised sampling schedules, distinct prompt configurations, or novel diffusion formulations. Therefore, high benchmark figures regularly mask severe out-of-distribution failure rates. We propose an adaptive routing framework that harmonizes distinct forensic feature maps and aims to reduce out-of-distribution failure across novel generator classes.

Objectives :

Review recent state-of-the-art methods for differentiating authentic photography from synthetic imagery across diverse generative pipelines. Analyse the specific roles of deep residual networks, transfer learning, dynamic gating, and ensemble strategies in multi-generator environments. Isolate performance bottlenecks stemming from unmodeled generative frameworks, lossy compression, spatial transformations, and limited training diversity. Detail an adaptive multi-expert architecture featuring dedicated zero-shot holdout protocols for unbiased generalization tracking. Outline an operational evaluation protocol that tracks accuracy, precision, recall, F1 Score, and out-of-distribution transfer.

II. BACKGROUND AND RELATED WORK

Synthetic image detection operates simultaneously as a binary classification problem and an applied media forensics discipline. In its standard formulation, an input sample must be categorized as either an authentic photograph or a machine-generated image. Convolutional Neural Networks (CNNs) process these inputs hierarchically: localized receptive fields in earlier layers isolate minute pixel discrepancies, high-frequency anomalies, and boundary irregularities, whereas deeper layers synthesize contextual and semantic relationships. This structural depth enables CNNs to surface minute statistical artifacts that evade casual visual inspection.

Transfer learning mitigates the severe sample complexity required to optimize deep residual networks from scratch. Architectures such as ResNet-50 leverage pre-trained weights learned from broad visual databases before fine-tuning on task-specific forensic datasets. This transfer process provides stable feature representations even under modest sample sizes. However, transfer learning alone cannot resolve cross-domain brittleness: When fine-tuning corpora are restricted to an isolated class of generative models, the learned weights inevitably align with the idiosyncrasies of those specific models rather than broader generative anomalies.



Ensemble methodologies address this brittleness by aggregating outputs across multiple complementary models. Static ensembles traditionally pool predictions via arithmetic averaging or majority voting, providing stability against individual model noise. Nevertheless, uniform weighting relies on the premise that all models provide comparable utility for any arbitrary input. In real-world environments, an image may carry artifacts that align cleanly with the domain of one specialized detector while lying well outside the learned distribution of another.

The module inspects the input visual features to compute sample-specific contribution coefficients on the fly. In a four-expert setup, the gating module outputs four non-negative scalar weights constrained to sum to unity, combining the individual predictions into a single synthetic likelihood score via convex combination. This gating concept served as the core mechanism in the 2024 baseline study, which allocated dedicated ResNet-50 experts to BigGAN, GLIDE, VQDM, and Stable Diffusion.

Recent Research From 2023–2026 :

Between 2023 and 2026, the synthetic media forensics landscape increasingly shifted focus from in-domain optimization toward out-of-distribution evaluation. In 2023, Ojha et al. established that standard classifiers trained on a single generative family break down when applied to foreign synthesis paradigms, finding that frozen vision-language representations offer superior zero-shot transfer capabilities.

Simultaneously, the GenImage benchmark established a million-scale reference corpus of paired authentic and artificial visuals covering eight generator systems, setting a new precedent for evaluating detectors under cross-generator domain shifts and varied visual distortions.

During 2024, researchers continued refining architectures to mitigate domain overfitting. Rather than relying purely on final pixel values, the Time Step Generating (TSG) technique samples intermediate states along the diffusion timeline, disentangling the generative synthesis mechanism from visual subject matter. Simultaneously, the gated ResNet-50 framework proved that modular specialized backbones could be adaptively combined, although testing confirmed that zero-shot detection on unmodeled generators remained an open vulnerability.

By 2025, the AI-GenBench initiative highlighted the importance of temporal validation splits. Rather than assuming stationary generator distributions, this benchmark evaluates detectors against generative models in chronological sequence, testing whether historical training pipelines remain viable as synthetic synthesis evolves.

Most recently, an extensive 2026 meta-study evaluated 23 pretrained detector variants across 12 distinct datasets comprising 2.6 million samples and 291 unique generative models. The benchmark revealed wide performance discrepancies across models, observing steep accuracy drops when evaluating against modern commercial synthesis engines. This body of work confirms that evaluating cross-domain robustness must take precedence over optimizing closed-set benchmark scores.

Research Gap :

The current literature exposes three interlinked bottlenecks: first, deep forensic backbones remain heavily tied to their specific training distributions; second, optimistic benchmark scores frequently mask performance degradation on unmodeled architectures; and third, real-world media streams are rarely pristine.

III. PROPOSED SYSTEM

The proposed architecture is designed as an adaptive, multi-generator forensic pipeline engineered to distinguish natural photography from synthetic imagery while actively mitigating reliance on specific generator signatures. The complete design is presented here as a structured methodological framework derived from recent empirical literature, avoiding unsubstantiated claims of original empirical trials.

System Architecture :

The framework comprises five modular processing stages: data ingestion and transformation, expert feature extraction, adaptive gating, degradation robustness testing, and threshold inference. The experimental corpus balances native



photographic captures with computer-generated imagery sourced uniformly across multiple generative systems. Before inference, inputs are normalized to a uniform resolution and exposed to controlled transformations, such as JPEG quantization, Gaussian smoothing, spatial resizing, and additive noise, simulating the visual degradation introduced by consumer communications channels.

Proposed System: Adaptive Multi-Generator Deep Learning for AI-Generated Image Detection

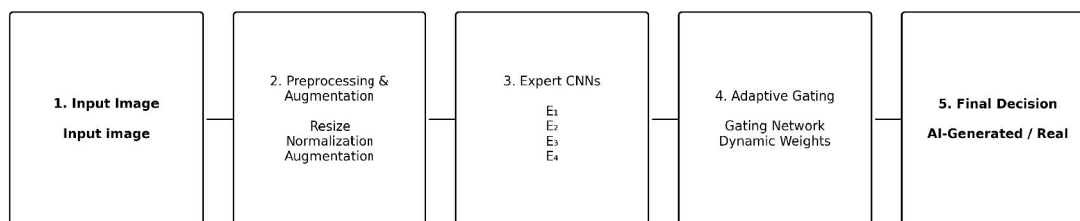


Fig. 1. Proposed workflow for adaptive multi-generator AI-image detection.

Expert Models :

Rather than expecting an isolated network to capture the nuanced signatures of every generative paradigm, the system distributes learning across several specialized CNN experts. Each back-end model trains on a specific generative sub-domain. ResNet-50 provides an effective baseline architecture due to its strong residual feature extraction and its direct alignment with previous gating studies. Initialized with ImageNet priors, every expert backbone starts from broad visual features before targeting the subtle, domain-level artifacts of synthetic synthesis.

The overall expert capacity remains modular rather than constrained to four static networks. Rather than remaining limited to four specific generators, the architecture can dynamically expand to support specialized experts across GAN, diffusion, autoregressive, and proprietary commercial classes. This decoupled architecture enables continuous expansion as new synthesis pipelines emerge.

Adaptive Gating :

When an input image passes through the network, each expert produces an independent confidence score indicating the probability that the sample is synthetic. The gating sub-network evaluates the input representation and assigns an adaptive routing weight to each expert. Applying a SoftMax normalization ensures all assigned weights remain non-negative and sum to one. Formally, letting x_j denote the classification prediction of expert j and z_j denote its associated normalized gating weight, the aggregated synthetic score y is expressed as:

$$y = \sum_{j=1}^n z_j x_j$$



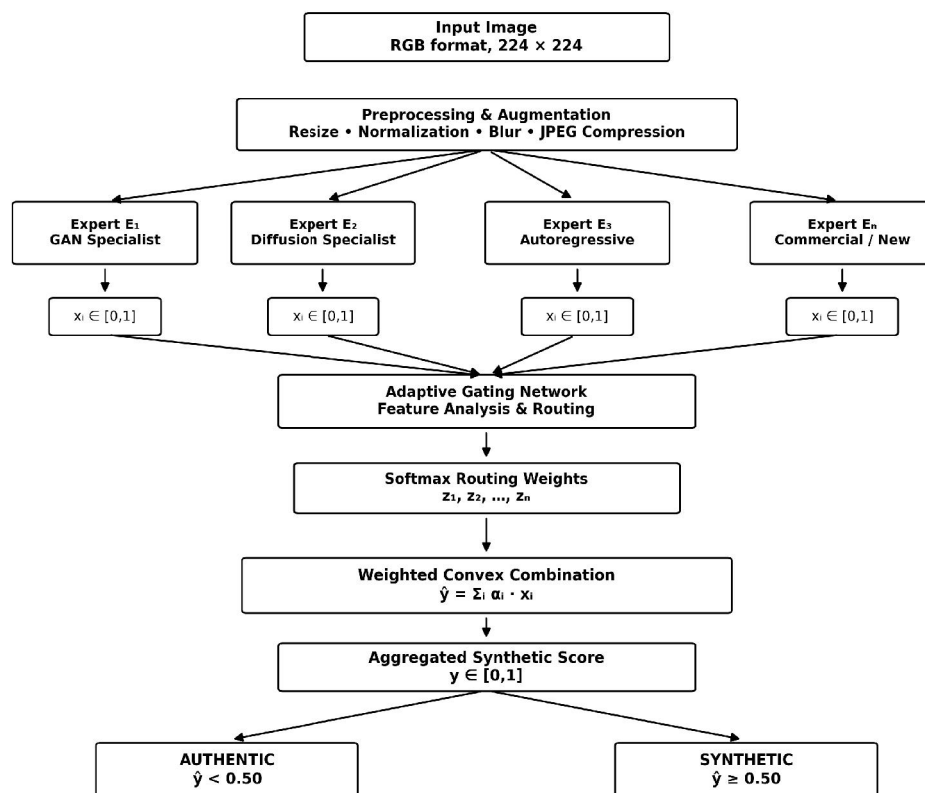


Fig 2. Proposed system architecture

A key distinction in this framework is that the gating layer is evaluated both on familiar generator profiles and on entirely unmodeled, holdout families. A routing network that overfits strictly to training distributions will exhibit sharp performance drops when handling novel inputs. Treating out-of-distribution performance as an independent, primary metric ensures a realistic assessment of cross-generator generalization.

Data Preparation :

An operationally viable dataset requires parity between authentic photography and synthetic media. Authentic samples can be curated from established photographic collections, while synthetic assets are compiled from public benchmarks such as GenImage. Training partitions must strictly isolate samples by generator rather than randomly dispersing all sources across splits, preventing generator-specific artifact leakage into the test partition.

We structure evaluation across three explicit data tracks: pristine inputs, degraded inputs, and unseen-generator inputs. Pristine inputs establish baseline discriminative capability; degraded inputs quantify resilience against lossy compression and transformations; and unseen-generator inputs assess zero-shot transfer onto unfamiliar generative architectures. This multi-tiered evaluation provides a more rigorous assessment than a single, aggregate accuracy metric.

IV. MODIFICATION OVER EARLIER APPROACH

The proposed architecture preserves the core principles of multi-generator dynamic gating while introducing explicit structural updates to counter the vulnerabilities documented in recent studies. These changes represent proposed



architectural refinements and should not be characterized as pre-validated empirical prior milestones experimental execution.

Wider Generator Diversity :

The initial gated-expert implementation grouped data into four static categories: BigGAN, GLIDE, VQDM, and Stable Diffusion. While valuable as early baselines, these targets no longer reflect the variety of modern synthesis engines. The revised architecture groups inputs by broader technological design and chronological release periods, enabling the framework to expand as new models emerge.

Explicit Unseen-Generator Testing :

A primary methodological modification is the separation of familiar and unfamiliar generative distributions during testing. By withholding at least one generative family entirely from the training pipeline, the framework's zero-shot transfer capabilities can be evaluated directly on that unmodeled set. This design directly targets the generalization failures flagged in early universal detection research and subsequent temporal benchmarks. Crucially, out-of-distribution performance is tracked as an independent operational metric rather than hidden within an aggregated average.

Robustness to Image Processing :

In production environments, digital media undergoes repeated downloading, lossy re-encoding, resolution adjustments, and screenshotting, all of which attenuate or eliminate fragile high-frequency forensic artifacts. To gauge practical durability, the updated framework integrates controlled JPEG quantization, spatial downscaling, blurring operations, and additive noise into its evaluation regime. A dependable forensic tool must maintain stable classification boundaries through these transformations, aligning with the evaluation criteria popularized by the GenImage benchmark.

Confidence-Based Decision :

Binary decisions without associated probability bounds fail to convey classification uncertainty. The proposed architecture pairs each discrete classification with an explicit confidence metric. Scores hovering near the 0.50 decision boundary are flagged as indeterminate and routed for manual audit or multimodal review, avoiding overconfident classifications on unfamiliar samples. This rejection mechanism is especially valuable for novel generator inputs where the underlying artifact space is unknown.

Lightweight Deployment :

Operating multiple deep neural networks concurrently introduces notable memory and compute requirements. To facilitate real-world use, a later stage of the proposed research includes profiling experts by inference cost and pruning or distilling low-impact sub-networks. A lightweight primary model can then serve as a preliminary screen, reserving the broader adaptive ensemble for ambiguous inputs.

Explainability :

The updated framework incorporates interpretability mechanisms to clarify model decisions, surfacing relative gating contributions alongside gradient-weighted spatial heatmaps. While feature attribution cannot serve as incontrovertible proof of authenticity, it offers crucial diagnostic insight into whether the architecture relies on legitimate forensic indicators or merely exploits spurious training artifacts.

Comparison Plan :

To isolate the direct impact of the proposed architectural modifications from data distribution artifacts, the framework should be benchmarked alongside three baseline approaches using identical data partitions: a monolithic ResNet-50 baseline, a uniform ensemble averaging model, and the original gated-expert architecture.

TABLE I. COMPARISON OF METHODOLOGICAL APPROACHES

Approach	Main Idea	Main Limitation
Single CNN	One model learns real/fake patterns	May overfit to known generators
Voting/Average	Combines several experts	All experts receive similar



Ensemble	outputs equally	influence
Gated Experts	Learns different weights for experts	Still needs diverse training data
Proposed Modified Model	Gating + unseen tests + degradation + confidence	Requires careful benchmark design

Evaluation Metrics :

Accuracy provides an initial assessment of model performance, but it is insufficient on its own. Precision quantifies the ratio of true synthetic items among all positive calls, recall measures the capture rate of generated content, and the F1-score provides a balanced harmonic mean. In deployment scenarios, minimizing false negatives is critical: misclassifying a synthetic asset as authentic photography can cause far greater operational harm than flagging a sample as indeterminate.

Metrics should be recorded and reported across pristine, degraded, and novel-generator conditions. Maintaining visibility into individual generator performance ensures that high accuracy on a familiar model does not obscure poor detection on unfamiliar generative frameworks.

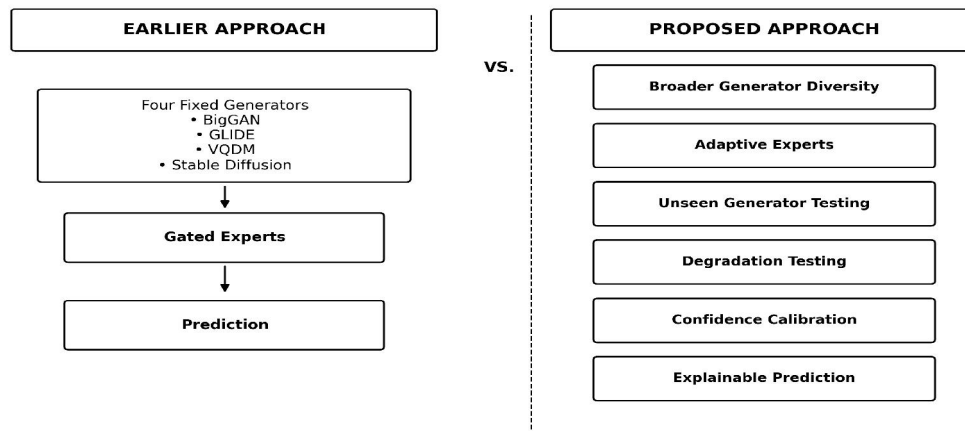


Fig 3. Earlier Approach Vs Proposed Approach

V. CONCLUSION

This paper reviewed current approaches for detecting AI-generated images across diverse generative engines and detailed an adaptive multi-expert deep learning framework. The primary operational challenge in modern forensics is not simply maximizing classification metrics on static datasets, but maintaining reliability across evolving generators, diverse compression profiles, and shifting data distributions.

Literature from 2023 through 2026 highlights the urgency of addressing these distribution shifts. Early studies demonstrated that classifiers struggle with newer generative paradigms; GenImage established standardized degradation and cross-model benchmarks; the 2024 gated-expert study showed the promise of combining specialized CNN backbones; AI-GenBench emphasized temporal validation; and a large 2026 meta-study documented significant performance drops on commercial synthesis engines.

Building on modular expert routing, our method couples adaptive multi-network aggregation with zero-shot generalization testing, degradation robustness, formal uncertainty calibration, diverse architectural families, and viable strategies for lightweight implementation. These additions are offered as a rigorous methodological blueprint rather



than an unverified empirical achievement. Subsequent research will focus on training the end-to-end framework, running the defined evaluation protocols on open benchmarks, and reporting empirical outcomes with complete transparency.

ACKNOWLEDGEMENT

I would like to sincerely thank our project guide and faculty members for their guidance, support, and valuable suggestions throughout this research work. Their encouragement and feedback helped us improve our work and complete the research successfully. We are also grateful to our department and college for providing us with the resources and facilities needed for this study. We would also like to thank our friends and family for their constant encouragement and support during the preparation of this research paper.

REFERENCES

- [1] S. Ren, Y. Zhou, X. Shen, K. Zewde, T. Duong, G. Huang, H. Tiangratanakul, T. Ng, E. Wei, and J. Xue, "How Well Are Open-Sourced AI-Generated Image Detection Models Out-of-the-Box: A Comprehensive Benchmark Study," arXiv:2602.07814, 2026.
- [2] L. Pellegrini, D. Cozzolino, S. Pandolfini, D. Maltoni, M. Ferrara, L. Verdoliva, M. Prati, and M. Ramilli, "AI-GenBench: A New Ongoing Benchmark for AI-Generated Image Detection," arXiv:2504.20865, 2025.
- [3] R. A. F. Saskoro, N. Yudistira, and T. N. Fatyanosa, "Detection of AI-Generated Images From Various Generators Using Gated Expert Convolutional Neural Network," IEEE Access, vol. 12, 2024, pp. 147772–147782.
- [4] Z. Zeng, H. Liu, D. Peng, L. Jing, and H. Watanabe, "Time Step Generating: A Universal Synthesized Deepfake Image Detector," arXiv:2411.11016, 2024.
- [5] M. Zhu, H. Chen, Q. Yan, X. Huang, G. Lin, W. Li.
- [6] Z. Tu, H. Hu, J. Hu, and Y. Wang, "GenImage: A Million-Scale Benchmark for Detecting AI-Generated Image," arXiv:2306.08571, 2023.
- [7] J. J. Bird and A. Lotfi, "CIFAKE: Image Classification and Explainable Identification of AI-Generated Synthetic Images," arXiv:2303.14126, 2023. U. Ojha, Y. Li, and Y. J. Lee, "Towards Universal Fake Image Detectors That Generalize Across Generative Models," Proc. IEEE/CVF CVPR, 2023, pp. 24480–24489.
- [8] A. Mohammed and R. Kora, "A Comprehensive Review on Ensemble Deep Learning: Opportunities and Challenges," J. King Saud Univ. Comput. Inf. Sci., vol. 35, no. 2, pp. 757–774, 2023.
- [9] S. Yang et al., "Image Data Augmentation for Deep Learning: A Survey," Artificial Intelligence Review, 2022.
- [10] O. Sagi and L. Rokach, "Ensemble Learning: A Survey," Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, vol. 8, no. 4, 2018.
- [11] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," Proc. IEEE CVPR, 2016, pp. 770–778

