IJIE (Indonesian Journal of Informatics Education)

Research Paper ISSN: 2549-0389

Artifact-Guided Logit Fusion of Fine-tuned CNN and ViT Branches for AI-Generated Image Detection: Implications for AI Literacy in Informatics Education

[*1]Mochamad Rizal Fauzan, [2]Yusuf Athallah Adriyansyah, [3]Feri Adriyanto

[1,2]Department of Electrical Engineering, College of Electrical Engineering and Computer Science, National Taipei University of Technology, Taipei 10608, Taiwan (R.O.C.)

[3]Departement of Electrical Engineering, Universitas Sebelas Maret, Surakarta 57126, Indonesia

Corresponding Email: t114998415@ntut.org.tw

Abstract:

Background: The rapid advancement of generative artificial intelligence has made it increasingly difficult to distinguish authentic images from synthetic visual content, raising concerns for digital and AI literacy and academic integrity in higher education. Informatics curricula increasingly need concrete, evaluated examples of how AI-generated image detectors are designed, tested, and limited.

Purpose: This study proposes Artifact-Guided EfficientViT-V3 (AG EfficientViT V3), a hybrid detection framework integrating fine-tuned EfficientNetB0 and ViT-Tiny branches with an artifact-oriented branch through learnable logit-level fusion, and discusses its potential relevance as an instructional case study and a supporting tool for AI/digital literacy and academic-integrity practice in informatics education.

Method: Instead of direct feature concatenation, the model preserves the decision behavior of separately fine-tuned branches while adding an artifact-sensitive correction signal at the decision level.

Results: On the CIFAKE benchmark, AG EfficientViT V3 achieves 98.865% accuracy, 98.938% precision, 98.790% recall, 98.864% F1-score, and 0.999116 ROC-AUC, the strongest clean-test performance among all evaluated baselines and variants. Ablation results show that naive CNN-Transformer hybridization alone does not outperform the strongest single-branch baseline, whereas fine-tuned branch initialization combined with logit-level fusion does. Robustness evaluation shows strong performance under clean and JPEG-compressed conditions, while severe blur, resize degradation, and additive noise remain challenging, illustrating that detector accuracy is condition-dependent rather than universal.

Conclusion/Implications: The proposed pipeline and evaluation protocol can be adapted as teaching material in deep learning and computer vision courses, while integration into digital-literacy or academic-integrity workflows is framed as a potential application requiring further validation rather than an established outcome.

Keywords: *AI-generated image detection, AI literacy, artifact-guided logit fusion, CIFAKE, informatics education.*

IJIE (Indonesian Journal of Informatics Education)
Vol 10 Issue 2 2026
DOI: https://doi.org/10.20961/ijie.v10i1.126345
Received: 26/05/26 Revised: 15/07/26 Accepted: 15/06/26 Online: 22/08/26

Introduction

The rapid development of generative artificial intelligence has significantly challenged the reliability and authenticity of digital visual content. Recent generative models can produce visually plausible synthetic images. These images are capable of visually plausible images. This issue has motivated growing research interest in AI-generated image detection, where the objective is to classify whether an input image originates from a real visual source or from an image synthesis model. The CIFAKE benchmark, for example, was introduced to support binary classification between real CIFAR 10 images and synthetic images generated using latent diffusion, reflecting the need for controlled datasets in this domain (Bird & Lotfi, 2023).

Beyond digital forensics and media verification, the growing accessibility of generative AI has direct implications for informatics education and higher education more broadly. University students increasingly encounter AI-generated visual content in coursework, multimedia assignments, design portfolios, and online learning materials, which raises concrete challenges for AI literacy, digital and visual media literacy, and academic-integrity procedures (Ng et al., 2021; Willie, 2025). Informatics curricula are also expected to prepare students to critically evaluate the reliability and limitations of AI systems rather than only to apply them, which requires accessible, well-documented case studies of how detection models are built, evaluated, and shown to fail (Najjar et al., 2025). Evidence from digital-literacy research further indicates that exposure to structured, technically grounded material on generative AI can support students' capacity to evaluate synthetic content critically (Wu & Zhang, 2025). AI-generated image detection therefore constitutes not only a technical forensics problem but also a concrete pedagogical opportunity: a transparent, reproducible detection pipeline can serve informatics education both as instructional material for deep learning and computer vision courses and as a potential supporting component for institutional academic-integrity and digital-literacy initiatives.

Despite the rapid progress of detection methods, AI-generated image detection remains a challenging task. Detection models must identify subtle evidence that may not be perceptually obvious, including local texture irregularities, residual artifacts, background inconsistencies, frequency related traces, and semantic anomalies. Recent studies on diffusion generated image detection have shown that detectors trained on earlier types of synthetic images may struggle when evaluated against newer generative mechanisms, indicating that the detection problem cannot be treated as fully solved by conventional binary classification alone (Wang et al., 2023). Similarly, generalization studies have reported that models trained on one type of generator may fail to identify images produced by unseen generative models, making robustness and transferability central concerns in this research area (Ojha et al., 2023).

Deep learning based detectors have commonly relied on convolutional neural networks because convolutional filters are effective for capturing local spatial patterns, texture cues, and artifact level evidence. However, artifact cues alone may not be sufficient when generated images exhibit high visual realism or when degradation such as compression, resizing, or blur alters the forensic traces. Recent artifact oriented methods, such as gradient based representation learning, have attempted to improve generalization by focusing on transferable synthetic patterns rather than only dataset specific visual cues (Tan et al., 2023). At the same time, large scale benchmarks such as GenImage emphasize the importance of evaluating detectors across diverse generators and degraded image conditions, suggesting that a reliable detector should be assessed not only on clean test accuracy but also on cross generator and degradation aware settings (Zhu et al., 2023).

Vision Transformer based approaches provide a complementary direction because they model images through patch level representations and can capture broader semantic relationships. Recent fake image detection research has shown that pretrained vision representations and Transformer based feature spaces can improve generalization across generative models, particularly when conventional classifiers overfit to generator specific artifacts (Yan et al., 2025). This suggests that local artifact sensitivity and global semantic reasoning should be considered jointly. However, simply combining convolutional and Transformer features does not guarantee superior performance, because hybrid models may suffer from feature imbalance, calibration mismatch, or ineffective fusion between branches.

Motivated by these issues, this work proposes Artifact-Guided EfficientViT-V3 (AG-EfficientViT-V3), abbreviated as AG EfficientViT V3, for AI-generated image detection. The proposed framework integrates a fine tuned EfficientNetB0 branch, a fine tuned ViT Tiny branch, and an artifact oriented branch through learnable logit level fusion. Unlike direct feature concatenation, the proposed fusion strategy operates at the decision level, allowing the model to preserve the strong decision behavior of separately fine tuned branches while incorporating artifact sensitive correction. This design is intended to address a specific limitation observed in the experimental development process: naive CNN and Transformer hybridization does not necessarily outperform a strong single branch baseline, whereas fine tuned branch initialization combined with logit level fusion provides a more stable and effective decision strategy.

The main contributions of this study are as follows. First, this work introduces an artifact guided hybrid detection framework that combines local texture evidence, global semantic representation, and artifact sensitive decision correction for AI-generated image detection. Second, it employs fine tuned branch initialization using separately trained EfficientNetB0 and ViT Tiny checkpoints to preserve strong branch specific decision boundaries. Third, it proposes learnable logit level fusion to combine CNN, Transformer, and artifact branch logits at the decision level rather than relying only on intermediate feature concatenation. Fourth, it presents a transparent ablation study comparing EfficientNetB0, ViT Tiny, EfficientViT Hybrid, AG EfficientViT V1, AG EfficientViT V2, and AG EfficientViT V3 on CIFAKE. Fifth, it reports degradation specific robustness analysis under JPEG compression, Gaussian blur, resize degradation, and additive noise, while avoiding unsupported claims of universal robustness.

The current evidence supports strong clean test and JPEG compressed performance on CIFAKE, but broader generalization still requires external dataset validation, multi seed evaluation, verified robustness metrics, and efficiency analysis. Sixth, this work discusses the potential educational relevance of the proposed framework and its evaluation protocol for informatics education, including their use as instructional material in deep learning and computer vision courses and their potential future integration into AI-literacy, digital-literacy, and academic-integrity practice; these educational applications are proposed as directions for future work and are not evaluated with actual students, instructors, or courses in the present study.

The remainder of this paper is organized as follows. Section 2 reviews related work on AI-generated image detection, CNN based detectors, Transformer based detectors, hybrid CNN and Transformer architectures, and artifact oriented learning. Section 3 presents the proposed AG EfficientViT V3 framework. Section 4 describes the dataset, compared models, implementation settings, and evaluation protocol. Section 5 reports clean test performance and ablation results. Section 6 discusses robustness, interpretability, and the implications of the proposed framework for informatics education and higher education practice. Section 7 concludes the study and outlines limitations and future work.

Related Work

AI-generated image detection aims to distinguish authentic images from synthetic images produced by generative models. As generated images become increasingly realistic, detection models must capture subtle evidence such as local texture irregularities, residual artifacts, edge inconsistencies, and semantic anomalies. Recent studies show that detector performance may decrease when models face images from different generators, domains, or post processing conditions, indicating that synthetic image detection remains challenging beyond clean in distribution testing (Cozzolino et al., 2025; Meng et al., 2024).

Convolutional neural networks are effective for extracting local spatial patterns, texture cues, edge structures, and high frequency information, while Vision Transformers provide stronger capacity for modeling broader semantic relationships through patch based representation. These complementary properties make CNN and Transformer based architectures relevant for AI-generated image detection. However, simply combining different feature representations does not necessarily guarantee better performance, because hybrid models may still suffer from calibration imbalance, ineffective fusion, or weak transferability across image conditions (Cozzolino et al., 2024).

Artifact oriented learning provides another important direction by focusing on residual traces or synthetic cues produced during image generation. Recent artifact purification based work shows that artifact related features can support cross domain detection when irrelevant visual information is reduced (Cozzolino et al., 2025). In AG EfficientViT V3, the artifact branch is not used as an independent detector, but as an auxiliary decision signal that supports fine tuned EfficientNetB0 and ViT Tiny branches through learnable logit level fusion.

Robustness is also an important concern because image degradation such as JPEG compression, blur, resizing, and additive noise can alter the visual traces used by detection models. Recent robustness benchmarks show that AI-generated image detectors may exhibit different sensitivities across augmentation types and data distributions (Pal et al., 2024). Therefore, clean test performance alone is insufficient to evaluate the reliability of an AI-generated image detector. The proposed method robustness under clean images, JPEG compression, Gaussian blur, resize degradation, and additive noise, while treating robustness as degradation specific rather than universal.

From an educational technology perspective, related work on AI-generated content detection has increasingly been framed as a component of academic-integrity and digital-literacy practice rather than solely as a forensic problem. Detection systems for AI-generated text have been proposed explicitly to support academic-integrity procedures in educational settings (Najjar et al., 2025), and broader socio-technical analyses emphasize that reliable verification of AI-generated images depends not only on algorithmic accuracy but also on institutional governance and user digital literacy (Willie, 2025). AI-literacy frameworks similarly call for concrete, technically grounded examples that help learners critically evaluate AI system behavior rather than treat AI outputs as inherently trustworthy (Ng et al., 2021). These perspectives motivate treating AG EfficientViT V3 not only as a forensic classifier but also as a candidate case study for informatics-education contexts, a direction developed further in the Discussion (Section 6).

Research Method

Research Design

This study was conducted as an experimental deep learning research for binary AI-generated image detection. The objective was to evaluate whether an artifact guided hybrid architecture can improve the classification of real and

AI-generated images by integrating local convolutional evidence, global Transformer representation, and artifact sensitive decision cues. The proposed model, AG EfficientViT V3, was designed as the final architecture and evaluated against EfficientNetB0, ViT Tiny, EfficientViT Hybrid, AG EfficientViT V1, and AG EfficientViT V2. This design follows the general structure of controlled experimental research described by Wohlin et al. (2012), in which a proposed treatment is scoped, planned, executed, and compared against defined baselines under a consistent evaluation protocol.

The research procedure consisted of four main stages. First, the CIFAKE dataset was prepared as the benchmark dataset for real versus AI-generated image classification. Second, baseline and variant models were evaluated to establish the comparative performance of CNN based, Transformer based, simple hybrid, and artifact guided architectures. Third, AG EfficientViT V3 was developed by integrating fine tuned EfficientNetB0, fine tuned ViT Tiny, and an artifact oriented branch through learnable logit level fusion. Fourth, the model was evaluated using clean test performance, ablation analysis, robustness testing under image degradation, qualitative examples, and Grad CAM visualization. This staged dataset preparation, comparative baseline evaluation, model development, and multi perspective evaluation procedure follows the scoping to result presentation structure recommended for experimental research in computing (Wohlin et al., 2012), and mirrors the dataset preparation, baseline and ablation comparison, and multi condition evaluation stages used in recent AI-generated image detection studies (Bhattacharjee et al., 2026; Zhu et al., 2023).

Dataset

The dataset used in this study was CIFAKE, which contains real and AI-generated images for binary classification. CIFAKE was constructed by pairing the 60,000 real photographic images of the CIFAR 10 benchmark with 60,000 synthetic counterparts generated using a latent diffusion model conditioned on the same ten CIFAR 10 classes, so that real and fake images share equivalent resolution, class distribution, and subject matter and differ primarily in their generative origin (Bird & Lotfi, 2023). This design isolates generation artifacts as the primary discriminative signal and makes CIFAKE well suited to binary real versus AI-generated classification, since class-level confounds such as differing categories or image quality between the two sets are minimized. The dataset consists of 120,000 images in total, with 60,000 real images and 60,000 AI-generated images. The training set contains 100,000 images, while the test set contains 20,000 images. Table 1 presents the dataset distribution used in this study.

Table 1. CIFAKE dataset split used in this study.

Split	Real Images	Fake Images	Total
Train	50,000	50,000	100,000
Test	10,000	10,000	20,000
Total	60,000	60,000	120,000

As shown in Table 1, the dataset is balanced between real and AI-generated images in both the training and test sets. This balanced distribution supports direct evaluation using standard classification metrics. In this study, class 0 represents an AI-generated image, while class 1 represents a real image. Each image was resized to 224 × 224 pixels before being processed by the model.

Compared Models

Several baseline and variant models were evaluated to examine the contribution of each architectural component. This follows the same single branch baseline versus progressively combined variant comparison strategy used in recent CNN and Transformer hybrid detectors for synthetic image and deepfake detection (Bhattacharjee et al., 2026), and is consistent with the classical principle that a combined decision rule should be evaluated against its constituent single classifiers to establish whether fusion provides genuine benefit (Kittler et al., 1998). Table 2 summarizes the evaluated models used in this study. EfficientNetB0 was used as the CNN baseline, while ViT Tiny was used as the Transformer baseline. EfficientViT Hybrid represented a simple CNN and Transformer hybrid model based on feature concatenation. AG EfficientViT V1 and AG EfficientViT V2 were included as earlier artifact guided variants, while AG EfficientViT V3 was used as the final proposed model.

Table 2. Evaluated baselines and AG EfficientViT variants.

Model	Description
EfficientNetB0	CNN baseline for local spatial and texture cues.
ViT Tiny	Transformer baseline for global semantic representation.
EfficientViT Hybrid	Simple CNN and Transformer hybrid using feature concatenation.
AG EfficientViT V1	Initial artifact guided hybrid variant using weighted fusion.

AG EfficientViT V2	Revised artifact aware variant using gated concatenation and interaction fusion.
AG EfficientViT V3	Final proposed model using fine tuned branches and logit level fusion.

The comparison in Table 2 was designed to determine whether performance improvement is caused by the CNN branch, the Transformer branch, the artifact branch, or the final logit level fusion strategy. This structure also allows AG EfficientViT V3 to be compared with both single branch and hybrid baselines.

Proposed AG EfficientViT V3 Architecture

The proposed AG EfficientViT V3 architecture consists of three parallel decision branches: a fine tuned EfficientNetB0 branch, a fine tuned ViT Tiny branch, and an artifact oriented branch. The EfficientNetB0 branch is used to capture local spatial patterns, texture cues, edge structures, and high frequency evidence. The ViT Tiny branch is used to capture global semantic representation through patch based visual modeling. The artifact oriented branch is used to extract residual and artifact sensitive cues that may be associated with synthetic image generation.

Figure 1 depicts the overall architecture of AG EfficientViT V3. The input image is processed by the three branches in parallel, and each branch produces a two class logit vector. These logits are then combined using learnable logit level fusion to produce the final prediction. This design differs from simple feature concatenation because the fusion is performed at the decision level rather than at the intermediate feature level.

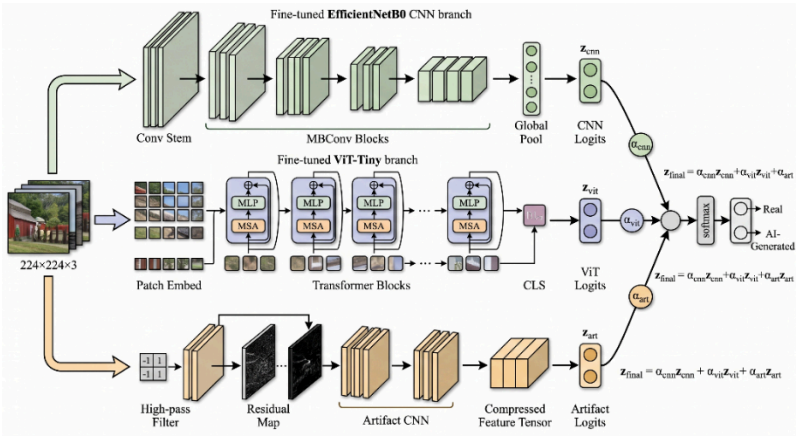

Figure 1. Overall architecture of AG EfficientViT V3.

The proposed framework integrates fine tuned EfficientNetB0, fine tuned ViT Tiny, and artifact oriented branches. Each branch produces class logits, which are combined through learnable logit level fusion to generate the final real versus AI generated prediction.

As shown in Figure 1, AG EfficientViT V3 is designed to preserve the decision behavior of fine-tuned CNN and Transformer branches while allowing the artifact oriented branch to provide an additional correction signal. This design is motivated by the observation that simple CNN and Transformer hybridization does not necessarily outperform a strong single branch baseline.

Problem Formulation

Let $x \in R^{H\times W\times 3}$ represent an input RGB image and $y \in 0,1$ represent the ground truth label. In this study, $y = 0$ denotes an AI-generated image, while $y = 1$ denotes a real image. The objective is to learn a classifier f_θ that maps the input image x into a predicted class label $\hat{y}$.

The model produces a two dimensional logit vector:

$$z = \left[z_0, z_1\right] \in R^2 \quad (1)$$

The posterior probability for class k is computed using the softmax function:

$$p(x) = \frac{\exp exp\left(z_k\right)}{\sum_{j=0}^{1} \exp exp\left(z_j\right)} \tag{2}$$

The predicted class is obtained by selecting the class with the highest posterior probability:

$$\hat{y} = \arg arg\, p\ (x) \tag{3}$$

This formulation is used consistently across all evaluated models, enabling direct comparison using accuracy, precision, recall, F1 score, and ROC AUC.

Fine Tuned EfficientNetB0 Branch

The EfficientNetB0 branch serves as the CNN based decision component of AG EfficientViT V3. The EfficientNetB0 backbone itself is an established compound-scaled convolutional architecture (Tan & Le, 2019) and is adopted here rather than proposed as a new contribution of this study; the contribution lies in how it is fine tuned, initialized, and combined with the other branches. This branch is initialized from a separately fine tuned EfficientNetB0 checkpoint and is used to extract local visual evidence, including spatial patterns, texture information, edge structures, and high frequency cues. These cues are relevant because AI-generated images may contain subtle local artifacts that are difficult to identify from semantic information alone.

The output of the EfficientNetB0 branch is represented as:

$$z_{cnn} = f_{cnn}\left(x; \theta^{*}_{cnn}\right) \tag{4}$$

where f_{cnn} denotes the EfficientNetB0 branch and θ^{*}_{cnn} represents the fine tuned EfficientNetB0 parameters. The output z_{cnn} is a two class logit vector that contributes to the final decision fusion.

Fine Tuned ViT Tiny Branch

The ViT Tiny branch serves as the Transformer based decision component of the proposed model. The underlying Vision Transformer architecture originates from Dosovitskiy et al. (2021), while the compact Tiny scale configuration used here follows the data efficient image transformer variants introduced by Touvron et al. (2021); both are adopted as established backbones rather than proposed as new contributions of this study. This branch is initialized from a separately fine tuned ViT Tiny checkpoint and is used to capture global contextual and semantic representation. Unlike the CNN branch, which emphasizes local patterns, the ViT Tiny branch models broader visual relationships across image regions through patch based representation.

The output of the ViT Tiny branch is represented as:

$$z_{vit} = f_{vit}\left(x; \theta^{*}_{vit}\right) \tag{5}$$

where f_{vit} denotes the ViT Tiny branch and θ^{*}_{vit} represents the fine tuned ViT Tiny parameters. In the final fusion design, this branch receives the largest initial fusion weight because ViT Tiny achieves the strongest performance among the single branch baselines in the experiment.

Artifact Oriented Branch

The artifact oriented branch is designed to provide additional evidence related to synthetic image traces. The use of high pass or noise residual filtering to expose manipulation and generation artifacts is an established forensic technique, following two dimensional variants of this idea such as the noise stream of Zhou et al. (2018) and the constrained convolutional layer of Bayar and Stamm (2018); this branch adapts that general principle to the AI-generated image detection setting rather than introducing a new filtering paradigm. This branch extracts residual and artifact sensitive cues using high pass filtering and lightweight convolutional feature extraction. The branch is not designed as a standalone classifier. Instead, it provides an auxiliary decision signal that supports the CNN and Transformer branches during final fusion.

The artifact feature representation is defined as:

$$h_{art} = g_{art}(x; \theta_{art}) \tag{6}$$

where g_{art} denotes the artifact feature extractor and θ_{art} represents its learnable parameters. The artifact feature vector is then projected into a two class logit vector:

$$z_{art} = W_{art} h_{art} + b_{art} \tag{7}$$

where W_{art} and b_{art} denote the classification weight and bias of the artifact branch. The resulting z_{art} is used as one of the three inputs in the final logit level fusion stage.

Artifact Guided Logit Level Fusion

The main methodological component of AG EfficientViT V3, and the primary novel contribution of this study, is artifact guided logit level fusion. Combining classifier outputs at the decision level rather than the feature level follows the general theoretical framework for classifier combination established by Kittler et al. (1998), and weighted ensembling of separately trained CNN and Transformer branches has recently been used for related deepfake detection tasks (Bhattacharjee et al., 2026). The contribution of AG EfficientViT V3 is the specific artifact guided instantiation of this general principle: a learnable, softmax normalized three way fusion of fine tuned CNN, fine tuned Transformer, and artifact branch logits, initialized asymmetrically toward the strongest single branch baseline, which to the authors' knowledge has not been used in this combination for AI-generated image detection. Instead of concatenating intermediate features, the proposed model combines the output logits from the CNN, Transformer, and artifact branches. The final logit vector is computed as:

$$z_{final} = \alpha_{cnn} z_{cnn} + \alpha_{vit} z_{vit} + \alpha_{art} z_{art} \tag{8}$$

The fusion coefficients are normalized using the softmax function:

$$[\alpha_{cnn}, \alpha_{vit}, \alpha_{art}] = softmax(w) \tag{9}$$

The initial fusion logits are defined as:

$$w = [1.0, 2.5, -2.0] \tag{10}$$

This initialization produces the following approximate fusion weights:

$$\alpha_{cnn} \approx 0.1808, \quad \alpha_{vit} \approx 0.8102, \quad \alpha_{art} \approx 0.0090 \tag{11}$$

The initialization is intentionally ViT dominant because ViT Tiny is the strongest single branch baseline in the reported experiment. At the same time, the CNN and artifact branches remain active in the learnable fusion mechanism, allowing their contributions to be adjusted during training. The final class probability is computed as:

$$p(x) = softmax(z_{final})_k \tag{12}$$

This fusion strategy preserves strong branch level decision boundaries while allowing artifact sensitive correction to influence the final prediction.

Training Objective

The proposed model is optimized using cross entropy loss. This is the standard objective for binary classification and is adopted without modification rather than proposed as a new contribution of this study. For N training samples, the loss function is defined as:

$$L_{CE} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{k=0}^{1} 1\left(y_i = k\right) \log log\, p\left(x_i\right) \quad (13)$$

where $1\left(y_i = k\right)$ is an indicator function that equals 1 when the ground truth label of sample i belongs to class k, and 0 otherwise. This objective encourages the model to assign higher probability to the correct class while optimizing the trainable components and the learnable fusion weights.

Implementation Details

The implementation was conducted using PyTorch. The input image size was set to 224 × 224 pixels. AdamW was used as the optimizer with a weight decay of 0.0001. This follows Loshchilov and Hutter (2019), who showed that decoupling weight decay from the gradient based update of Adam improves generalization compared with conventional L2 regularized Adam, making it a common default optimizer for fine tuning CNN and Transformer based vision models. The batch size was set to 128 for most models, while EfficientViT Hybrid used a batch size of 64 when required. The baseline models were trained for 20 epochs, whereas the final AG EfficientViT V3 run was trained for 10 epochs. Mixed precision training was enabled when CUDA was available. The random seed was set to 42. Table 3 summarizes the implementation and training settings used in this study.

Table 3. Implementation and training settings.

Item	Setting
Framework	PyTorch
Input size	224 × 224
Optimizer	AdamW
Weight decay	0.0001
Batch size	128 for most models; 64 for EfficientViT Hybrid if needed
Main training epochs	20 for baselines; 10 for AG EfficientViT V3 final run
Mixed precision	Enabled when CUDA is available
Random seed	42
Metrics	Accuracy, precision, recall, F1 score, ROC AUC

As presented in Table 3, the same evaluation metrics were used across all compared models to ensure a consistent performance comparison. The available implementation record includes the framework, input size, optimizer, weight decay, batch size, epoch configuration, mixed precision setting, random seed, and evaluation metrics.

Evaluation Protocol

The evaluation was structured to examine the proposed model from three perspectives: clean classification performance, architectural contribution, and robustness under image degradation. This three part structure follows established practice in the AI-generated image detection literature, where clean in distribution accuracy alone has been shown to be an insufficient indicator of detector quality: component wise ablation is needed to attribute performance gains to specific architectural choices (Kittler et al., 1998; Bhattacharjee et al., 2026), and degradation aware testing is needed because detectors can behave very differently under compression, blur, resizing, and noise than on clean images (Hendrycks & Dietterich, 2019; Pal et al., 2024; Zhu et al., 2023). Clean test performance was measured on the CIFAKE test set using accuracy, precision, recall, F1 score, and ROC AUC. These metrics were selected to provide a balanced view of classification correctness, class level reliability, and separability between real and AI-generated images.

To analyze the contribution of the proposed design, AG EfficientViT V3 was compared with EfficientNetB0, ViT Tiny, EfficientViT Hybrid, AG EfficientViT V1, and AG EfficientViT V2. This comparison was intended to show whether the final performance was mainly influenced by the CNN branch, the Transformer branch, the artifact branch, or the logit level fusion mechanism. By including both single branch baselines and earlier artifact guided variants, the evaluation provides a clearer view of how the final architecture improves over simpler design choices.

Robustness was evaluated by testing the models under several degraded image conditions. The evaluated conditions included clean images, JPEG compression at Q70, Q50, and Q30, Gaussian blur with radius 1 and 2, resize degradation to 112 × 112 followed by upsampling, and additive noise with intensities 0.03 and 0.05. This evaluation was used to determine whether the model maintains its detection ability when image quality changes, while avoiding the assumption that strong performance under one degradation type implies robustness under all conditions.

Qualitative examples and Grad CAM visualization (Selvaraju et al., 2017) were also included to support the interpretation of the quantitative results. The selected examples covered easy real images, easy AI-generated images, cases where AG EfficientViT V3 corrected at least one baseline error, failure cases of AG EfficientViT V3, and hard or ambiguous samples. This analysis provides additional insight into how the model behaves across different prediction scenarios and helps identify cases where the decision remains uncertain or incorrect.

Result and Discussion

Clean Test Performance

The clean test evaluation was conducted on the CIFAKE test set to measure the classification performance of AG EfficientViT V3 against the evaluated baseline and variant models. Table 4 presents the clean test results in terms of accuracy, precision, recall, F1 score, and ROC AUC. The evaluated models include EfficientNetB0, ViT Tiny, EfficientViT Hybrid, AG EfficientViT V1, AG EfficientViT V2, and the proposed AG EfficientViT V3.

Table 4. Clean test performance comparison on CIFAKE.

Rank	Model	Best Epoch	Accuracy (%)	Precision (%)	Recall (%)	F1 score (%)	AUC
1	AG EfficientViT V3	10	98.865	98.938	98.790	98.864	0.999116
2	ViT Tiny	20	98.750	98.809	98.690	98.749	0.998620
3	EfficientViT Hybrid	15	98.710	98.778	98.640	98.709	0.998759
4	AG EfficientViT V1	13	98.670	98.807	98.530	98.668	0.998713
5	AG EfficientViT V2	18	98.625	98.225	99.040	98.631	0.998733
6	EfficientNetB0	19	98.065	98.214	97.910	98.062	0.997674

As shown in Table 4, AG EfficientViT V3 achieves the highest clean test performance among all evaluated models, with 98.865% accuracy, 98.938% precision, 98.864% F1 score, and 0.999116 ROC AUC. ViT Tiny is the strongest single branch baseline, reaching 98.750% accuracy and 98.749% F1 score. EfficientViT Hybrid also performs strongly, but its result remains slightly below ViT Tiny and AG EfficientViT V3. This indicates that simply combining CNN and Transformer representations through hybridization does not automatically produce the best result.

Figure 2 visualizes the clean test comparison across the evaluated models. The performance gap between the models is relatively small because all methods already reach high accuracy on CIFAKE. Nevertheless, AG EfficientViT V3 consistently obtains the strongest overall result, especially in accuracy, precision, F1 score, and ROC AUC. This pattern suggests that the proposed logit level fusion strategy provides a measurable improvement over both single branch and earlier hybrid variants. Table 5 summarizes the direct performance difference between AG EfficientViT V3 and ViT Tiny, which is the strongest single branch baseline.

As presented in Table 5, the improvement of AG EfficientViT V3 over ViT Tiny is numerically small but consistent across the reported metrics. This result should be interpreted carefully because the baseline performance is already close to saturation on CIFAKE. Rather than indicating a large performance leap, the result suggests that fine tuned branch initialization and logit level fusion can provide additional gains in a high performance classification setting.

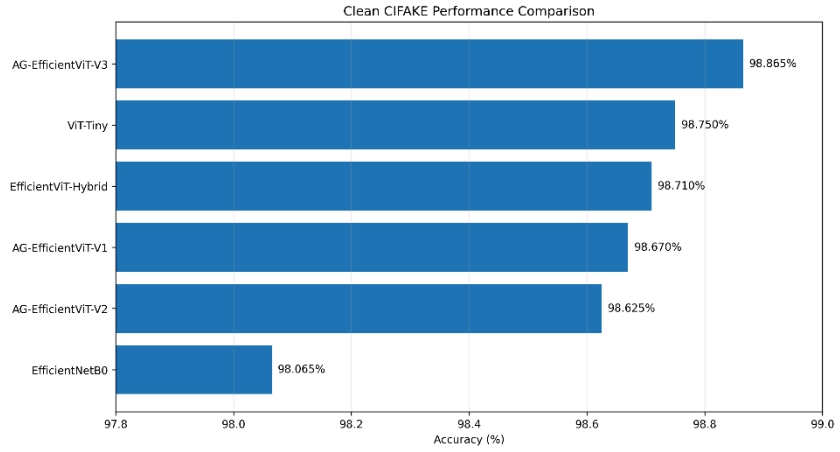

Figure 2. Clean CIFAKE performance comparison across evaluated models.

Table 5. Improvement of AG EfficientViT V3 over ViT Tiny.

Metric	ViT Tiny	AG EfficientViT V3	Difference
Accuracy (%)	98.750	98.865	+0.115
F1 score (%)	98.749	98.864	+0.115
AUC	0.998620	0.999116	+0.000496

Ablation Study

The ablation study was conducted to examine the contribution of each model component and fusion strategy. Table 6 compares the evaluated variants based on the presence of CNN, ViT, and artifact branches, as well as the fusion mechanism used by each model.

Table 6. Ablation study of model variants.

Variant	CNN Branch	ViT Branch	Artifact Branch	Fusion Strategy	Accuracy (%)	F1 score (%)	AUC
EfficientNetB0	✓	✗	✗	None	98.065	98.062	0.997674
ViT Tiny	✗	✓	✗	None	98.750	98.749	0.998620
EfficientViT Hybrid	✓	✓	✗	Feature concatenation	98.710	98.709	0.998759
AG EfficientViT V1	✓	✓	✓	Weighted artifact guided fusion	98.670	98.668	0.998713
AG EfficientViT V2	✓	✓	✓	Gated concatenation and interaction fusion	98.625	98.631	0.998733
AG EfficientViT V3	✓	✓	✓	Logit level fusion with fine tuned branch initialization	98.865	98.864	0.999116

Table 6 shows that ViT Tiny outperforms EfficientNetB0 as a single branch model, suggesting that global representation contributes strongly to this classification task. EfficientViT Hybrid performs better than EfficientNetB0 but does not outperform ViT Tiny, which indicates that adding a CNN branch through feature concatenation is not sufficient to surpass the strongest single branch baseline. AG EfficientViT V1 and AG EfficientViT V2 also do not exceed AG EfficientViT V3, even though they include artifact guided components. This result highlights that the presence of an artifact branch alone is not enough; the fusion strategy and branch initialization are central to the final performance.

Figure 3 illustrates the ablation comparison across the evaluated models. The trend supports the main methodological argument of this study: AG EfficientViT V3 performs best because it combines fine tuned branch initialization with logit level fusion. This design preserves the decision behavior of the separately trained EfficientNetB0 and ViT Tiny branches while allowing the artifact branch to contribute an additional correction signal at the decision level.

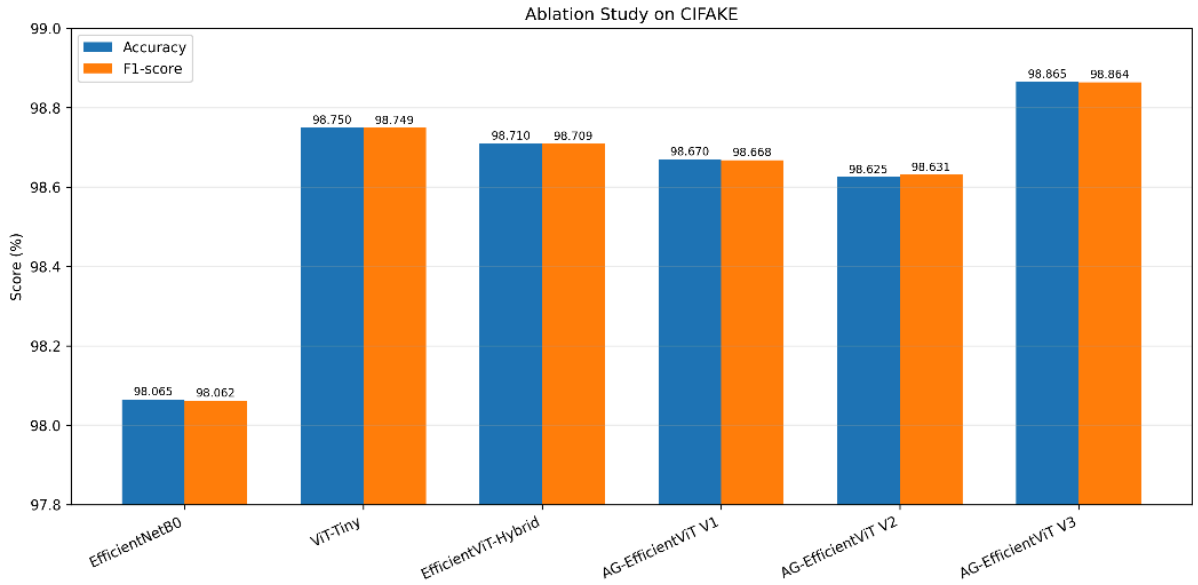

Figure 3. Ablation comparison across baselines and AG EfficientViT variants.

Robustness Analysis

Robustness evaluation was conducted to examine whether model performance remains stable under image degradation. The tested conditions include clean images, JPEG compression at Q70, Q50, and Q30, Gaussian blur with radius 1 and 2, resize degradation to 112 × 112 followed by upsampling, and additive noise with intensities 0.03 and 0.05. Table 7 presents the robustness results across ViT Tiny, EfficientViT Hybrid, and AG EfficientViT V3.

Table 7. Robustness results under image degradation.

Model	Condition	Accuracy	Precision	Recall	F1	AUC
ViT Tiny	clean	0.98750	0.98809	0.98690	0.98749	0.99862
EfficientViT Hybrid	clean	0.98705	0.98778	0.98630	0.98704	0.99876
AG EfficientViT V3	clean	0.98865	0.98938	0.98790	0.98864	0.99912
ViT Tiny	jpeg_q70	0.98535	0.98366	0.98710	0.98538	0.99839
EfficientViT Hybrid	jpeg_q70	0.98405	0.98371	0.98440	0.98406	0.99849
AG EfficientViT V3	jpeg_q70	0.98700	0.98506	0.98900	0.98703	0.99909
ViT Tiny	jpeg_q50	0.98480	0.98143	0.98830	0.98485	0.99807
EfficientViT Hybrid	jpeg_q50	0.98365	0.98216	0.98520	0.98368	0.99849
AG EfficientViT V3	jpeg_q50	0.98600	0.98214	0.99000	0.98606	0.99897
ViT Tiny	jpeg_q30	0.98240	0.97687	0.98820	0.98250	0.99800
EfficientViT Hybrid	jpeg_q30	0.98245	0.97857	0.98650	0.98252	0.99837
AG EfficientViT V3	jpeg_q30	0.98345	0.97701	0.99020	0.98356	0.99834
ViT Tiny	blur_r1	0.97810	0.96166	0.99590	0.97848	0.99712
EfficientViT Hybrid	blur_r1	0.97925	0.96452	0.99510	0.97957	0.99836
AG EfficientViT V3	blur_r1	0.97705	0.95875	0.99700	0.97750	0.99726
ViT Tiny	blur_r2	0.87565	0.80117	0.99930	0.88933	0.96620
EfficientViT Hybrid	blur_r2	0.88455	0.81287	0.99910	0.89642	0.98988
AG EfficientViT V3	blur_r2	0.84585	0.76447	0.99970	0.86640	0.96131
ViT Tiny	resize_112	0.96860	0.94333	0.99710	0.96947	0.99513
EfficientViT Hybrid	resize_112	0.96650	0.94018	0.99640	0.96747	0.99759
AG EfficientViT V3	resize_112	0.96235	0.93142	0.99820	0.96365	0.99531
ViT Tiny	noise_003	0.91930	0.86284	0.99710	0.92513	0.98857
EfficientViT Hybrid	noise_003	0.72510	0.72230	0.73140	0.72682	0.69263
AG EfficientViT V3	noise_003	0.76965	0.71776	0.88880	0.79417	0.00000*
ViT Tiny	noise_005	0.83710	0.75522	0.99750	0.85962	0.97011
EfficientViT Hybrid	noise_005	0.54075	0.56770	0.34170	0.42662	0.00000*
AG EfficientViT V3	noise_005	0.50240	0.50504	0.24070	0.32602	0.00000*

* The AUC values reported as 0.00000 require verification because they may indicate a metric computation issue, degenerate prediction behavior, invalid probability values, or an exception fallback during evaluation.

Table 7 shows that AG EfficientViT V3 achieves the highest accuracy under the clean, jpeg_q70, jpeg_q50, and jpeg_q30 conditions. This indicates that the proposed model remains strong when the input images are affected by JPEG compression. However, the same pattern does not hold under stronger blur, resize degradation, and additive noise. Under blur_r2, EfficientViT Hybrid achieves the highest accuracy and F1 score among the compared models. Under additive noise, ViT Tiny is more stable than the hybrid models, particularly at noise_003 and noise_005.

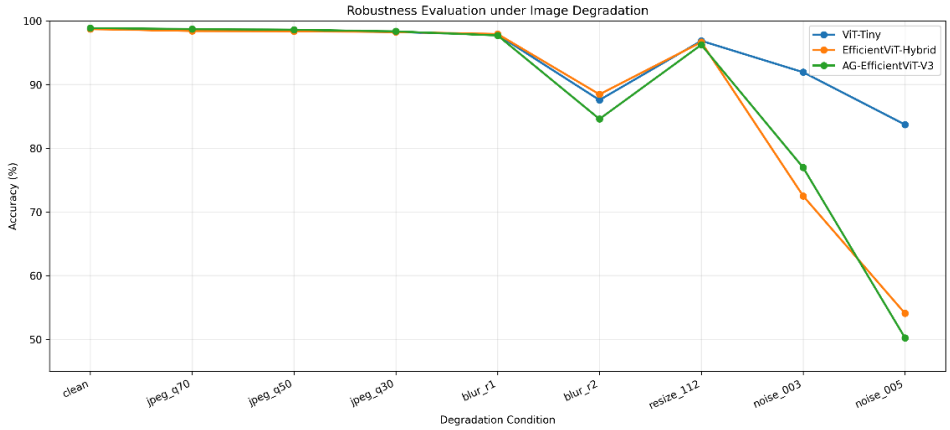

Figure 4. Robustness comparison of key models under image degradation.

Figure 4 visualizes the robustness trend across degradation conditions. The results show that AG EfficientViT V3 should not be interpreted as universally robust. Its strongest robustness behavior appears under clean and JPEG compressed conditions, while severe blur, resize degradation, and additive noise remain challenging. This finding is important because it shows that high clean test performance does not necessarily guarantee consistent robustness across all image corruptions.

Qualitative and Grad CAM Analysis

Qualitative analysis was included to provide visual evidence of model behavior beyond aggregate numerical metrics. Figure 5 presents representative CIFAKE examples, including easy real images, easy AI-generated images, cases where AG EfficientViT V3 improves over at least one baseline, failure cases of AG EfficientViT V3, and hard or ambiguous samples.

Figure 5. Representative qualitative examples on CIFAKE, including easy real and fake samples, V3 improvement cases, and remaining failure cases.

Table 8 summarizes the qualitative case categories used in Figure 5. These categories were selected to make the qualitative analysis more interpretable. Easy real and easy AI generated cases show where the evaluated models agree. V3 improvement cases show examples where the final fusion design provides a benefit over at least one baseline. Failure and hard cases are included to avoid presenting only successful predictions and to clarify the remaining limitations of the model.

Table 8. Qualitative case categories used in Figure 5.

Case ID	Case Category	Selection Criterion	Purpose
Q1	Easy real	Real image correctly classified by all evaluated models	Confirms agreement on clear real samples
Q2	Easy fake	AI-generated image correctly classified by all evaluated models	Confirms agreement on clear synthetic samples
Q3	V3 improvement	AG EfficientViT V3 correct while at least one baseline is incorrect	Shows benefit of final fusion design
Q4	V3 improvement	AG EfficientViT V3 correct while at least one baseline is incorrect	Highlights artifact guided correction
Q5	V3 failure	AG EfficientViT V3 incorrect	Shows remaining failure mode
Q6	Hard or ambiguous	Difficult sample from remaining predictions	Indicates need for further analysis

Grad CAM visualization was also used to inspect the regions that contributed to AG EfficientViT V3 predictions. Figure 6 presents the Grad CAM examples for the proposed model.

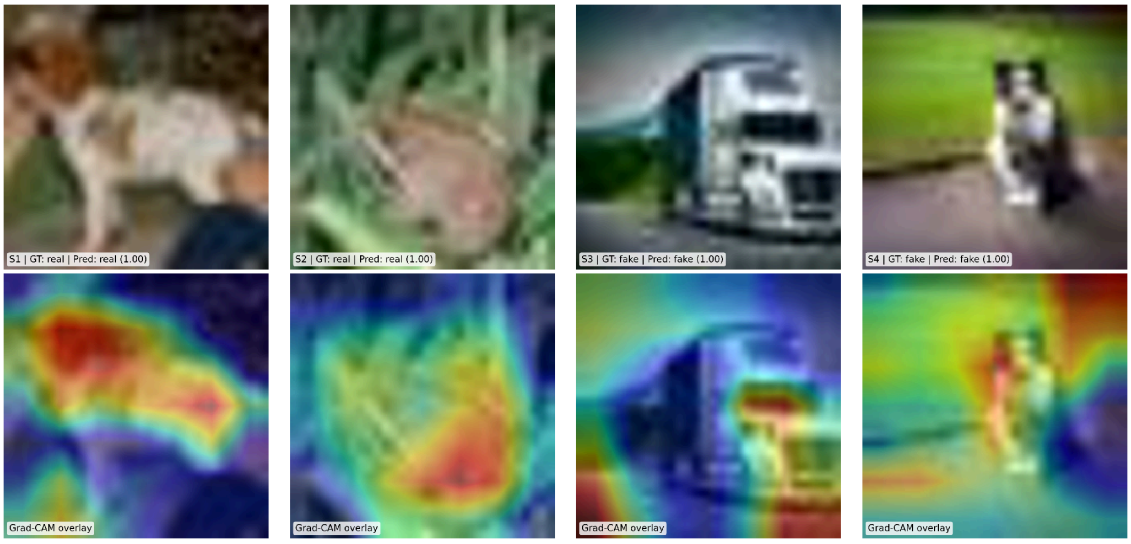

Figure 6. Grad CAM examples for AG EfficientViT V3.

Figure 6 provides an initial visual interpretation of the model decision behavior. The visualization can help examine whether the model focuses on object regions, local texture areas, background artifacts, or misleading regions in failure cases. Although Grad CAM does not fully explain the decision process, it provides useful supporting evidence for understanding how the model responds to representative samples.

Discussion

The results indicate that AG EfficientViT V3 achieves the strongest clean test performance among all evaluated models on CIFAKE. Its advantage comes from the combination of fine tuned EfficientNetB0 logits, fine tuned ViT Tiny logits, artifact branch logits, and learnable logit level fusion. The ablation results support this interpretation because neither simple CNN and Transformer feature concatenation nor earlier artifact guided variants outperform the final V3 design.

The comparison with ViT Tiny is especially important. ViT Tiny is the strongest single-branch baseline, and its high performance shows that global semantic representation is highly effective for this task. AG EfficientViT V3

improves over ViT Tiny by a small but consistent margin, suggesting that artifact guided decision fusion can add value when strong branch decisions are preserved rather than replaced. This also explains why the fusion initialization is ViT dominant in the proposed architecture.

At the same time, the robustness results show that the proposed model has clear limitations. AG EfficientViT V3 performs strongly under clean and JPEG compressed images, but its performance drops under severe blur, resize degradation, and additive noise. ViT Tiny remains more stable under additive noise, while EfficientViT Hybrid performs better under stronger blur. Therefore, the most defensible interpretation is that AG EfficientViT V3 provides strong performance on clean test data and degradation specific robustness, particularly under JPEG compression, rather than universal robustness across all corruptions.

Overall, the findings support the value of logit level fusion with fine tuned branch initialization for AI-generated image detection. The results also show that future improvement should focus on broader external validation, more stable robustness oriented training, multi seed evaluation, and efficiency analysis. Beyond these technical findings, the proposed framework, its ablation methodology, and its degradation-aware evaluation protocol also carry potential relevance for informatics education and higher education practice, discussed below as potential applications that would require dedicated pedagogical validation, since the present study did not involve students, instructors, courses, or measured learning outcomes.

Relevance to AI and digital literacy curricula. The evaluation protocol used in this study offers a concrete example of how AI system claims can be critically examined rather than accepted at face value, a core competency in AI-literacy frameworks (Ng et al., 2021). The ablation study (Table 6) shows that combining CNN and Transformer architectures does not automatically improve performance, illustrating the value of evidence-based evaluation over intuitive assumptions about model design. The robustness analysis (Table 7), including the explicit acknowledgment of degraded performance under blur, resize degradation, and additive noise, models the kind of cautious, condition-aware interpretation of AI performance claims that AI-literacy and digital-literacy curricula aim to instill in learners who must evaluate AI system reliability in their own coursework or future careers (Wu & Zhang, 2025). Both the ablation table and the robustness table are proposed here as reusable teaching material for illustrating these concepts in deep learning or AI-ethics courses.

Relevance to academic integrity and authenticity verification. A transparent, benchmarked AI-generated image detection pipeline is also potentially relevant to institutional efforts to verify the authenticity of visually submitted educational materials, such as design coursework, multimedia assignments, or AI-assisted portfolio submissions, in the same spirit as text-based AI detection systems that have been proposed to support academic-integrity procedures (Najjar et al., 2025). Socio-technical analyses of AI-generated image verification similarly emphasize that effective authenticity checks depend on combining technical detection with institutional governance and user literacy rather than on algorithmic accuracy alone (Willie, 2025). However, CIFAKE consists of curated, low-resolution benchmark images rather than real student-submitted content, and the present study did not evaluate AG EfficientViT V3 on institutional academic-integrity workflows; any such application would therefore require domain adaptation, evaluation on realistic educational submissions, and institutional piloting before deployment could be responsibly recommended.

Relevance to curriculum and classroom use. The overall study design, comprising problem formulation, fine-tuned branch initialization, logit-level fusion, a transparent ablation study, degradation-aware robustness testing, and Grad-CAM based interpretability analysis, provides a complete, reproducible worked example of responsible experimental practice in applied deep learning. This structure may be adapted as instructional material for undergraduate or graduate coursework, laboratory exercises, or capstone and thesis projects in informatics and computer science programs, illustrating practices such as baseline comparison, systematic ablation, robustness testing, and cautious interpretation of performance improvements that students are expected to apply in their own research.

Scope and limitations of the educational claims. These educational implications are proposed directions rather than demonstrated outcomes. This study did not involve students, instructors, classrooms, or curricula, and no learning outcomes were measured. Establishing the actual educational value of the proposed framework, whether as teaching material, an AI-literacy case study, or a component of academic-integrity workflows, would require dedicated future work, including classroom pilots, curriculum-integration studies, and evaluation of student learning outcomes and instructor feedback.

Conclusion

This study presented AG EfficientViT V3, an artifact-guided hybrid framework for AI-generated image detection that integrates fine tuned EfficientNetB0, fine tuned ViT Tiny, and an artifact oriented branch through learnable logit level fusion. Experiments on CIFAKE showed that the proposed model achieved the strongest clean test performance among the evaluated models, with 98.865% accuracy, 98.938% precision, 98.790% recall, 98.864% F1 score, and 0.999116 ROC AUC. The ablation results indicate that the improvement was not obtained merely by combining CNN and Transformer features, but by preserving strong fine tuned branch decisions and adding artifact sensitive correction at the logit level. Beyond its technical contribution, AG EfficientViT V3 and its evaluation protocol carry potential relevance for informatics education. The architecture, ablation methodology, and degradation-aware robustness protocol are presented in a form that can be adapted as instructional material for university courses in deep learning, computer vision, and AI ethics, and the framework may, in principle, support future AI-literacy, digital-literacy, and academic-integrity initiatives that require verification of visually submitted or AI-assisted educational content. These educational applications are proposed as directions for future work rather than established outcomes, since the present study did not involve students, instructors, curricula, or measured learning outcomes; realizing this potential will require dedicated classroom pilots, curriculum-integration studies, and evaluation of learning impact. Beyond education, the proposed approach may also have long-term relevance for other authenticity-sensitive domains, such as media verification and content moderation, though such applications likewise require further domain-specific validation and fall outside the scope of the present study.

References

Bayar, B., & Stamm, M. C. (2018). Constrained convolutional neural networks: A new approach towards general purpose image manipulation detection. *IEEE Transactions on Information Forensics and Security*, *13*(11), 2691–2706. https://doi.org/10.1109/TIFS.2018.2825953

Bhattacharjee, A., Islam, K., Anan, K., Intesher, A., Fuad, A. A., Saha, U., & Imtiaz, H. (2026). CAE-Net: Generalized deepfake image detection using convolution and attention mechanisms with spatial and frequency domain features. *Journal of Visual Communication and Image Representation*, *115*, 104679. https://doi.org/10.1016/j.jvcir.2025.104679

Bird, J. J., & Lotfi, A. (2023). CIFAKE: Image Classification and Explainable Identification of AI-Generated Synthetic Images. *IEEE Access*, *12*, 15642–15650. https://doi.org/10.1109/ACCESS.2024.3356122

Cozzolino, D., Poggi, G., Corvi, R., Nießner, M., & Verdoliva, L. (2024). Raising the Bar of AI-generated Image Detection with CLIP. *IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, 4356–4366. https://doi.org/10.1109/CVPRW63382.2024.00439

Cozzolino, D., Poggi, G., Nießner, M., & Verdoliva, L. (2025). Zero-Shot Detection of AI-Generated Images. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *15076 LNCS*, 54–72. https://doi.org/10.1007/978-3-031-72649-1_4

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., & Houlsby, N. (2021). An image is worth 16x16 words: Transformers for image recognition at scale. *Proceedings of the 9th International Conference on Learning Representations (ICLR 2021)*. https://doi.org/10.48550/arXiv.2010.11929

Hendrycks, D., & Dietterich, T. (2019). Benchmarking neural network robustness to common corruptions and perturbations. *Proceedings of the 7th International Conference on Learning Representations (ICLR 2019)*. https://doi.org/10.48550/arXiv.1903.12261

Kittler, J., Hatef, M., Duin, R. P. W., & Matas, J. (1998). On combining classifiers. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *20*(3), 226–239. https://doi.org/10.1109/34.667881

Loshchilov, I., & Hutter, F. (2019). Decoupled weight decay regularization. *Proceedings of the 7th International Conference on Learning Representations (ICLR 2019)*.https://doi.org/10.48550/arXiv.1711.05101

Meng, Z., Peng, B., Dong, J., Tan, T., & Cheng, H. (2024). Artifact feature purification for cross-domain detection of AI-generated images. *Computer Vision and Image Understanding*, *247*. https://doi.org/10.1016/j.cviu.2024.104078

Najjar, A. A., Ashqar, H. I., Darwish, O. A., & Hammad, E. (2025). Detecting AI-generated text in educational content: Leveraging machine learning and explainable AI for academic integrity. *arXiv preprint*.

https://doi.org/10.48550/arXiv.2501.03203

Ng, D. T. K., Leung, J. K. L., Chu, S. K. W., & Qiao, M. S. (2021). Conceptualizing AI literacy: An exploratory review. *Computers and Education: Artificial Intelligence*, *2*, 100041. https://doi.org/10.1016/j.caeai.2021.100041

Ojha, U., Li, Y., & Lee, Y. J. (2023). Towards Universal Fake Image Detectors that Generalize Across Generative Models. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, *2023-June*, 24480–24489. https://doi.org/10.1109/CVPR52729.2023.02345

Pal, A., Kruk, J., Phute, M., Bhattaram, M., Yang, D., Chau, D. H., & Hoffman, J. (2024). Semi-Truths: A Large-Scale Dataset of AI-Augmented Images for Evaluating Robustness of AI-Generated Image detectors. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, & C. Zhang (Eds.), *Advances in Neural Information Processing Systems* (Vol. 37, pp. 118025–118051). Curran Associates, Inc. https://doi.org/10.52202/079017-3748

Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., & Batra, D. (2017). Grad-CAM: Visual explanations from deep networks via gradient-based localization. *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 618–626. https://doi.org/10.1109/ICCV.2017.74

Tan, C., Zhao, Y., Wei, S., Gu, G., & Wei, Y. (2023). Learning on Gradients: Generalized Artifacts Representation for GAN-Generated Images Detection. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, *2023-June*, 12105–12114. https://doi.org/10.1109/CVPR52729.2023.01165

Tan, M., & Le, Q. V. (2019). EfficientNet: Rethinking model scaling for convolutional neural networks. *Proceedings of the 36th International Conference on Machine Learning (ICML)*, *97*, 6105–6114.https://doi.org/10.48550/arXiv.1905.11946

Touvron, H., Cord, M., Douze, M., Massa, F., Sablayrolles, A., & Jégou, H. (2021). Training data-efficient image transformers & distillation through attention. *Proceedings of the 38th International Conference on Machine Learning (ICML)*, *139*, 10347–10357.https://doi.org/10.48550/arXiv.2012.12877

Wang, Z., Bao, J., Zhou, W., Wang, W., Hu, H., Chen, H., & Li, H. (2023). DIRE for Diffusion-Generated Image Detection. *Proceedings of the IEEE International Conference on Computer Vision*, 22388–22398. https://doi.org/10.1109/ICCV51070.2023.02051

Willie, M. M. (2025). Verifying artificial intelligence-generated images: Socio-technical approaches to authenticity. *Computing and Artificial Intelligence*, *3*(4). https://doi.org/10.59400/cai3893

Wohlin, C., Runeson, P., Höst, M., Ohlsson, M. C., Regnell, B., & Wesslén, A. (2012). *Experimentation in software engineering*. Springer. https://doi.org/10.1007/978-3-642-29044-2

Wu, D., & Zhang, J. (2025). Generative artificial intelligence in secondary education: Applications and effects on students' innovation skills and digital literacy. *PLOS ONE*, *20*(5), e0323349. https://doi.org/10.1371/journal.pone.0323349

Yan, S., Li, O., Cai, J., Hao, Y., Jiang, X., Hu, Y., & Xie, W. (2025). a Sanity Check for Ai-Generated Image Detection. *13th International Conference on Learning Representations, ICLR 2025*, 64081–64099.https://doi.org/10.48550/arXiv.2406.19435

Zhou, P., Han, X., Morariu, V. I., & Davis, L. S. (2018). Learning rich features for image manipulation detection. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 1053–1061. https://doi.org/10.1109/CVPR.2018.00116

Zhu, M., Chen, H., Yan, Q., Huang, X., Lin, G., Li, W., Tu, Z., Hu, H., Hu, J., & Wang, Y. (2023). GenImage: A Million-Scale Benchmark for Detecting AI-Generated Image. *Advances in Neural Information Processing Systems*, *36*(NeurIPS).https://doi.org/10.48550/arXiv.2306.08571