Published by :
https://www.ijert.org/
An International Peer-Reviewed Journal

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 11 , November - 2026

Robust AI-Generated Image Detection Using Pretrained CNN and Vision Transformer Models

Atharv Anilrao Kulkarni
Dept. of Artificial Intelligence and Machine Learning
Vishwakarma Institute of Information Technology
Pune, India

Grishma Bobhate
Dept. of Artificial Intelligence and Machine Learning
Vishwakarma Institute of Information Technology
Pune, India

***Abstract*—The rapid development of Artificial Intelligence (AI) has made it possible to generate highly realistic images using modern generative models. These images are increasingly difficult to distinguish from authentic photographs, creating challenges in digital content verification and image forensics. This study evaluates and compares two pretrained architectures, a Convolutional Neural Network (ResNet50) and a Vision Transformer (ViT-B/16), for detecting AI-generated images. Both models were fine-tuned for binary classification on a balanced subset of the GenImage benchmark containing real images and images generated by the ADM diffusion model (3,500 training, 750 validation and 750 test images per class). They were evaluated on the original test images and under JPEG compression (quality 90, 70 and 50) and image resizing (112 px and 64 px, restored to 224 px). On the original test images, ResNet50 reached 99.00% accuracy and ViT-B/16 reached 100.00%. JPEG compression had little effect on either model: the worst-case accuracy drops were 0.20 points for ResNet50 and 0.07 points for ViT-B/16. In contrast, downsampling the images to 112 px or 64 px and restoring them to 224 px reduced both models to near-chance accuracy (49.67%–50.80%), because almost all AI-generated images were then classified as real. ViT-B/16 was marginally more stable under JPEG compression, while neither architecture was robust to resizing. Because all generated images come from a single generator, the results describe robustness to post-processing and do not establish generalization to unseen generators.**

***Index Terms*—Artificial Intelligence, AI-Generated Images, Deep Learning, Convolutional Neural Network, Vision Transformer, ResNet50, Image Forensics, Computer Vision.**

I. Introduction

Artificial Intelligence has experienced significant development in recent years, particularly in the field of image generation. Modern generative models can produce realistic images using text descriptions and other input data. Technologies such as Generative Adversarial Networks (GANs) and diffusion models have significantly improved the quality of artificially generated visual content.

AI-generated images are increasingly used in digital media, entertainment, advertising, education, and content creation. However, these technologies can also be misused to create misleading or manipulated visual information. As generated images become more realistic, distinguishing between authentic photographs and artificially generated images becomes increasingly difficult.

Wang et al. demonstrated that deep-learning classifiers can identify certain patterns in images generated by different convolutional generative models. Their research showed that a classifier trained on one generator could also detect images from several other generators, although the generalization of detection methods remains an important challenge [1].

To address this problem, researchers have investigated deep-learning-based image detection techniques. Convolutional Neural Networks (CNNs) are widely used in computer vision because they automatically learn visual features from image data. ResNet introduced residual connections to make the training of deeper neural networks more effective [5]. Another approach is the Vision Transformer (ViT), which processes an image as a sequence of smaller patches and uses self-attention mechanisms to learn relationships between different image regions. Dosovitskiy et al. demonstrated that Transformer architectures can perform effectively on image-classification tasks [4].

Although these architectures are useful for image classification, their performance may change when images undergo transformations. For example, images uploaded to social media platforms are frequently compressed or resized. Such operations may remove or modify visual information that a detection model uses to distinguish real images from AI-generated images. Grommelt et al. highlighted that JPEG compression and image-size differences can introduce unwanted biases into AI-generated image detection datasets. Their findings emphasize the importance of evaluating detection models under different image conditions [3].

Therefore, this research compares pretrained ResNet50 and Vision Transformer models for AI-generated image detection. The study examines their classification performance on original images and transformed images to understand their robustness under controlled experimental conditions.

II. Literature Review

A. CNN-Generated Image Detection

Wang et al. (2020) proposed a method for detecting images generated by convolutional neural networks. Their study investigated whether a universal detector could distinguish real images from generated images produced by different architectures. The researchers demonstrated that a standard

(This work is licensed under a Creative Commons Attribution 4.0 International License.)

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 11 , November - 2026

image classifier trained using images from one generator could generalize to several unseen generators. This work established an important foundation for deep-learning-based image detection [1].

B. Large-Scale AI-Generated Image Detection

Zhu et al. (2023) introduced GenImage, a large-scale benchmark dataset for detecting AI-generated images. The dataset contains more than one million real and AI-generated images produced using different generative models. The researchers also investigated degraded-image classification and cross-generator detection, making GenImage useful for evaluating detection performance under different image conditions [2].

C. Dataset Bias in AI-Generated Image Detection

Grommelt et al. (2024) presented a study titled *Fake or JPEG? Revealing Common Biases in Generated Image Detection Datasets*. The researchers investigated how JPEG compression and image-size differences can influence AI-generated image detection. Their findings demonstrated that detection models may learn dataset-specific characteristics instead of relying only on features associated with image generation, highlighting the necessity of applying consistent preprocessing and transformations during model evaluation [3].

D. Vision Transformer for Image Classification

Dosovitskiy et al. (2021) introduced the Vision Transformer architecture for image recognition. Instead of processing images using traditional convolution operations, the architecture divides images into smaller patches and applies Transformer-based self-attention. Their research demonstrated that a pure Transformer architecture can achieve strong image-classification performance when pretrained on large datasets and transferred to other recognition tasks [4].

E. ResNet Architecture for Image Classification

He et al. (2016) introduced Deep Residual Learning for Image Recognition. Their proposed ResNet architecture uses residual connections that allow information to pass across network layers, helping to train deeper neural networks. ResNet became an important architecture for computer-vision applications and provides the CNN-based baseline selected for comparison in this research [5].

F. Summary of Existing Research

A summary is presented in Table I.

TABLE I
SUMMARY OF RELATED RESEARCH

Reference	Research Focus	Relevance to this Study
Wang et al. [1]	Detection of CNN-generated images	Foundation for generated-image detection
Zhu et al. [2]	GenImage benchmark dataset	Source of the experimental data
Grommelt et al. [3]	JPEG and image-size bias	Robustness evaluation protocol
Dosovitskiy et al. [4]	Vision Transformer	ViT architecture
He et al. [5]	Deep Residual Learning	ResNet50 architecture

III. Problem Statement

With the development of generative artificial intelligence, distinguishing authentic photographs from AI-generated images has become increasingly challenging. Deep-learning models can learn patterns that help identify artificially generated images; however, their classification performance may decrease when images undergo transformations such as JPEG compression and resizing. These transformations can modify image details and affect the visual features learned by detection models. Previous research has also shown that image compression and image-size differences can introduce dataset-related biases into detection systems [3]. Therefore, it is critical to evaluate how pretrained CNN and Vision Transformer models perform under different image conditions and determine whether their detection performance remains consistent after transformations.

IV. Research Gap

1) Existing research demonstrates the broad application of deep learning for detecting AI-generated images [1], [2].
2) Image compression and resolution shifts influence detector accuracy and introduce evaluation bias [3].
3) CNN and Transformer architectures use fundamentally different inductive biases and feature-learning mechanisms [4], [5].
4) There is a lack of head-to-head empirical comparisons between selected pretrained architectures evaluated on identical dataset splits and strictly controlled image-transformation conditions.
5) The relative robustness of ResNet50 versus ViT-B/16 under degrading post-processing requires systematic comparative benchmarking.

V. Research Objectives

1) To investigate existing deep-learning architectures for AI-generated image detection.
2) To fine-tune pretrained ResNet50 and Vision Transformer (ViT-B/16) models for binary classification.
3) To evaluate and compare the baseline classification performance of both models on real and AI-generated images.
4) To investigate the degradation effects of JPEG compression on detection reliability.
5) To analyze the impact of image downsampling and resizing on classification performance.
6) To characterize the robustness and trade-offs of CNNs and Vision Transformers under degraded image conditions.

(This work is licensed under a Creative Commons Attribution 4.0 International License.)

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 11 , November - 2026

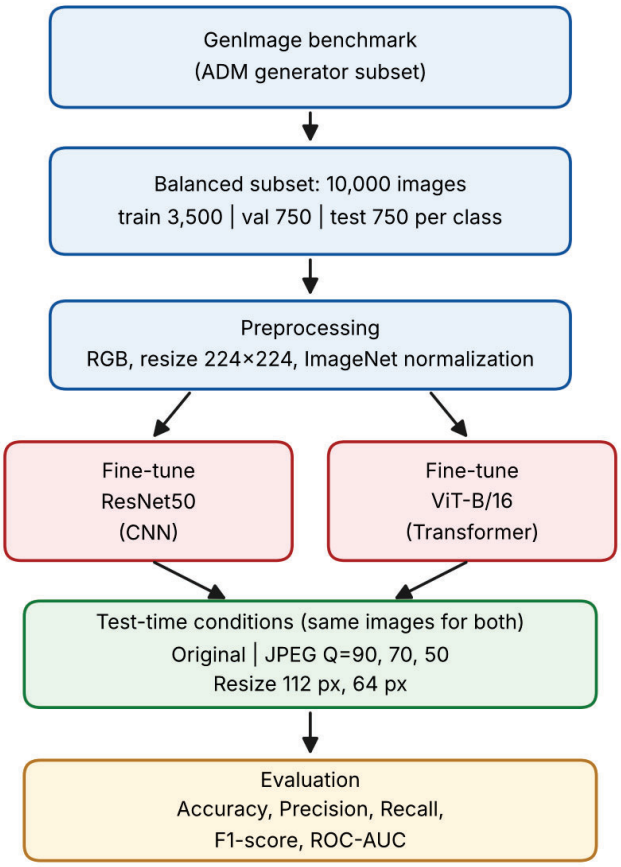

Fig. 1. Proposed methodology pipeline.

VI. Proposed Methodology

The study follows an experimental binary classification approach (Fig. 1). The system classifies input images into two categories: **Real**, an image obtained from an authentic photographic source, and **AI-generated**, an image synthesized by a generative model. AI-generated images are treated as the positive class in all metrics.

A. Dataset Collection

GenImage [2] serves as the data source. A class-balanced subset was curated to fit the available computational budget. It was streamed from the Hugging Face release of GenImage and uses only the ADM generator. Training and validation images were drawn from the ADM training split, and the test images were drawn from the official ADM test split, so the test set is disjoint from the training data by construction.

B. Image Preprocessing

1) Images were loaded as RGB and resized to 224×224 (bilinear interpolation).
2) Tensors were normalized using the ImageNet mean and standard deviation.

TABLE II
Selected Deep-Learning Architectures

Parameter	ResNet50	ViT-B/16
Architecture type	CNN	Vision Transformer
Main mechanism	Residual convolutions	Patches + self-attention
Pretrained weights	ImageNet	ImageNet
Task	Binary	Binary
Classes	Real / AI	Real / AI

TABLE III
Dataset Distribution

Split	Real	AI	Total	Source
Training	3,500	3,500	7,000	ADM train
Validation	750	750	1,500	ADM train (disjoint)
Testing	750	750	1,500	Official ADM test
Total	5,000	5,000	10,000	

3) Unreadable or corrupted files were skipped during collection.
4) Labels were assigned by folder: AI-generated $= 0$ and real $= 1$ in the implementation. AI-generated is the positive class for all reported metrics.
5) Images were kept in their original file formats. The effect of format and size differences between classes is discussed in Section VII-B.

C. Model Selection

Two architectures are evaluated (Table II): ResNet50, a 50-layer convolutional network with bottleneck residual blocks [5], and ViT-B/16, a self-attention architecture operating on non-overlapping 16×16 patches [4].

D. Dataset Splitting

The dataset distribution is given in Table III.

E. Model Training

Transfer learning was used. The original classification layer of each network was replaced with a two-output linear layer, and both models were fine-tuned end to end with softmax cross-entropy loss (the two-class form of binary cross-entropy) in PyTorch. Both models used the AdamW optimizer, batch size 32, 5 epochs, and automatic mixed precision. The learning rate was 1×10^{-4} for ResNet50 and 2×10^{-5} for ViT-B/16, since Transformer fine-tuning is typically unstable at higher rates. The checkpoint with the lowest validation loss was used for testing. For ViT-B/16, validation loss was lowest at epoch 4 and validation accuracy reached about 99.7% (Fig. 2).

F. Image Transformation

Transformations were applied to the test images only, before the standard resize and normalization, and both models were tested on identical transformed images (Table IV).

G. Model Testing and Comparison

Each model was evaluated separately under every condition using accuracy, precision, recall, F1-score and ROC-AUC.

(This work is licensed under a Creative Commons Attribution 4.0 International License.)

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 11 , November - 2026

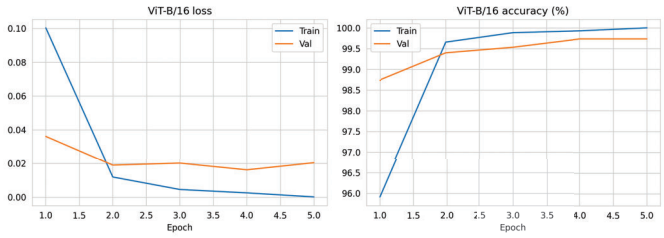

Fig. 2. Training and validation loss and accuracy of ViT-B/16.

TABLE IV
IMAGE-TRANSFORMATION CONDITIONS

Test Condition	Description
Original	Unmodified test images
JPEG $Q = 90$	High-quality lossy compression
JPEG $Q = 70$	Moderate compression
JPEG $Q = 50$	Severe lossy compression
Reduced Resolution 1	Downsampled to 112×112, then bilinearly upsampled to 224×224
Reduced Resolution 2	Downsampled to 64×64, then bilinearly upsampled to 224×224

TABLE V
DATASET CHARACTERISTICS

Property	Specification
Dataset name	GenImage (ADM subset)
Content	Real photographs and ADM-generated images
Task	Binary classification
Class 0 in code (positive)	AI-generated
Class 1 in code (negative)	Real photographs
Scope	Balanced subset, 10,000 images
Split	70% train, 15% val., 15% test
Source	Zhu et al. [2]

VII. DATASET DESCRIPTION

A. Dataset Characteristics

The experiment uses a curated subset of GenImage [2] (Table V).

B. Dataset Bias Analysis

Grommelt et al. [3] reported that GenImage contains systematic differences between its classes: real images are predominantly stored as JPEG while generated images are stored losslessly, and image sizes differ. A detector can exploit such cues instead of generation artifacts. Our subset keeps the original file formats, so the accuracy on original images should be read with this possibility in mind.

The JPEG experiments provide a partial check. Re-encoding every test image as JPEG removes any difference in container format between the classes, yet accuracy stayed at or above 98.8% for both models (Table VIII). Detection therefore does not depend on the file format alone. Differences in image size or in the compression history of real images cannot be ruled out by this test.

TABLE VI
EXPERIMENTAL CONFIGURATION

Component	Specification
Language	Python 3
Platform	Google Colab, NVIDIA Tesla T4 GPU
Framework	PyTorch, torchvision pretrained weights
Dataset	Curated GenImage (ADM) subset
Architectures	ResNet50, ViT-B/16
Input	RGB, 224×224, ImageNet normalization
Optimizer	AdamW (lr 10^{-4} ResNet50, 2×10^{-5} ViT-B/16)
Epochs / batch	5 / 32
Random seed	42
Distortions	JPEG ($Q = 90, 70, 50$), resizing (112 px, 64 px)
Metrics	Accuracy, precision, recall, F1, ROC-AUC

VIII. EXPERIMENTAL SETUP

The experimental framework is summarized in Table VI.

IX. EVALUATION METRICS

AI-generated images are the positive class and real photographs are the negative class.

Accuracy is the ratio of correct predictions:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{1}$$

Precision is the proportion of images predicted as AI-generated that are actually AI-generated:

$$\text{Precision} = \frac{TP}{TP + FP} \tag{2}$$

Recall is the proportion of actual AI-generated images that are detected:

$$\text{Recall} = \frac{TP}{TP + FN} \tag{3}$$

F1-score is the harmonic mean of precision and recall:

$$\text{F1} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{4}$$

ROC-AUC is the area under the receiver operating characteristic curve, computed from the predicted probability of the AI-generated class.

Here TP are AI images correctly identified as AI, TN are real images correctly identified as real, FP are real images incorrectly flagged as AI, and FN are AI images classified as real.

X. RESULTS AND DISCUSSION

All results come from a single training run per model (seed 42) on the 1,500-image test set.

(This work is licensed under a Creative Commons Attribution 4.0 International License.)

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 11 , November - 2026

TABLE VII
PERFORMANCE COMPARISON ON ORIGINAL IMAGES

Metric (%)	ResNet50	ViT-B/16
Accuracy	99.00	100.00
Precision	98.68	100.00
Recall	99.33	100.00
F1-score	99.00	100.00
ROC-AUC	99.78	100.00

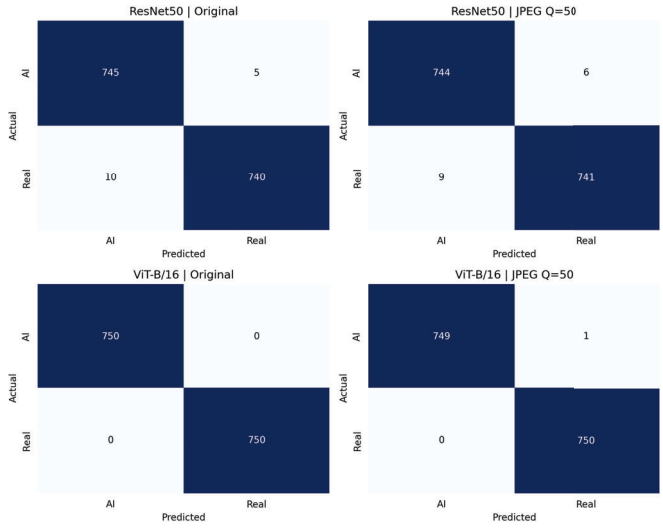

Fig. 3. Confusion matrices for both models on original and JPEG $Q = 50$ test images.

TABLE VIII
ACCURACY (%) UNDER JPEG COMPRESSION

Condition	ResNet50	ViT-B/16
Original	99.00	100.00
JPEG $Q = 90$	98.87	100.00
JPEG $Q = 70$	98.80	99.93
JPEG $Q = 50$	99.00	99.93

A. Performance on Original Images

Baseline performance is reported in Table VII.

Both models separate real from AI-generated ADM images almost perfectly. ResNet50 made 15 errors on the 1,500 test images: 5 AI-generated images were classified as real (false negatives) and 10 real images were flagged as AI (false positives), as shown in Fig. 3. ViT-B/16 made no errors. Since a zero-error result on 1,500 images is consistent with a true error rate of up to about 0.2%, and each model was trained only once, the 1-point gap should be read as a tendency in favor of ViT-B/16 rather than a firm difference. Such near-ceiling accuracy also reflects the fact that training and test images come from the same generator.

B. Effect of JPEG Compression

JPEG compression had almost no effect (Table VIII and Fig. 4). The largest accuracy drop was 0.20 points for

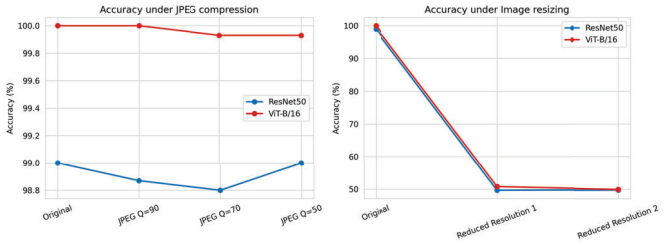

Fig. 4. Accuracy of ResNet50 and ViT-B/16 under JPEG compression and resizing.

TABLE IX
ACCURACY (%) UNDER IMAGE RESIZING

Condition	ResNet50	ViT-B/16
Original resolution	99.00	100.00
Reduced Res. 1 (112)	49.67	50.80
Reduced Res. 2 (64)	49.80	50.00

ResNet50 (3 images) and 0.07 points for ViT-B/16 (1 image). The ResNet50 curve is not monotonic: its accuracy at $Q = 50$ equals the original accuracy, which indicates that these small changes are within normal variation rather than a systematic degradation. ViT-B/16 kept a small lead at every quality level. A possible reason for this tolerance is that the cues both models rely on survive the mid-frequency content that moderate JPEG quantization preserves, but this was not tested directly.

C. Effect of Image Resizing

Resizing had a severe effect (Table IX). Both models fell to about 50%, which is chance level for a balanced test set. The class-level metrics show why: recall for AI-generated images dropped to 0.13% (ResNet50) and 1.87% (ViT-B/16) at 112 px, and to 0.00% and 0.27% at 64 px, so nearly every image was classified as real. Precision values in this regime (for example 87.5% for ViT-B/16 at 112 px) are based on only a handful of positive predictions and are not meaningful. The ROC-AUC values at 112 px (69.81% for ResNet50 and 58.95% for ViT-B/16) show that the models retain some ranking information at moderate downsampling, so part of the failure is a shift of the decision threshold, whereas at 64 px the AUC approaches chance (56.39% and 48.33%). The small differences between the two models (at most 1.1 points) are not meaningful at this level.

The likely cause is a distribution shift: downsampling followed by bilinear upsampling removes fine texture and introduces smoothing that neither model saw during training, since no resizing augmentation was used. The detectors appear to rely on fine-grained cues that this operation destroys. This explanation is consistent with the results but was not isolated experimentally.

(This work is licensed under a Creative Commons Attribution 4.0 International License.)

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 11 , November - 2026

TABLE X
OVERALL EXPERIMENTAL COMPARISON

Parameter	ResNet50	ViT-B/16
Original accuracy (%)	99.00	100.00
Worst-case JPEG drop (pts)	0.20	0.07
Worst-case resize drop (pts)	49.33	50.00
JPEG robustness	High	High
Resizing robustness	Low	Low

D. Overall Model Comparison

Robustness is High when the worst-case accuracy drop is under 1 percentage point, Moderate for 1 to 3 points, and Low above 3 points. Both architectures are highly robust to JPEG compression and neither is robust to resizing, so the architectural difference between a CNN and a Vision Transformer had little influence on robustness in this setting. The ViT-B/16 advantage is limited to slightly higher accuracy under JPEG compression and on original images. The expectation stated at the outset, that one architecture would hold up better than the other under degradation, is therefore only weakly supported for JPEG and not supported for resizing. With one run per model and 1,500 test images, differences of a few tenths of a percentage point should not be treated as conclusive.

XI. LIMITATIONS

1) Only two architectures (ResNet50 and ViT-B/16) are compared.
2) A 10,000-image subset of GenImage is used rather than the full corpus.
3) Only one generator (ADM) is used for both training and testing, so the results do not measure generalization to unseen generators.
4) Robustness tests are limited to JPEG compression and resizing.
5) Each model was trained once (single seed), so run-to-run variance is not estimated.
6) Real and generated images may differ in file format and size (Section VII-B), so part of the accuracy on original images may come from such cues rather than generation artifacts.
7) Neither model was trained with compression or resizing augmentation, so the collapse under resizing reflects an untreated distribution shift and not necessarily a limit of the architectures.
8) The test set contains 1,500 images, and the near-perfect results leave little room to separate the two models.
9) Detectors may overfit to high-frequency synthetic artifacts rather than semantically meaningful representations.

XII. FUTURE SCOPE

1) Benchmarking hybrid architectures (e.g., Swin Transformer and ConvNeXt).
2) Training with JPEG and resizing augmentation and measuring whether robustness to downsampling can be recovered.
3) Evaluating other transformations such as Gaussian blur, motion blur and cropping.
4) Zero-shot cross-generator evaluation on unseen generative models.
5) Explainable AI (Grad-CAM and Attention Rollout) to localize the artifacts the models use.
6) Repeating training over several seeds and reporting mean and standard deviation.
7) Extending detection to multimodal and deepfake video applications.

XIII. CONCLUSION

This study compared a pretrained ResNet50 and a ViT-B/16 for detecting AI-generated images using a balanced subset of the GenImage ADM data, evaluated on original images and under JPEG compression and resizing. ViT-B/16 reached 100.00% accuracy on original images against 99.00% for ResNet50, and it was marginally more stable under JPEG compression, where both models lost less than 0.25 percentage points even at quality 50. Resizing was the decisive weakness: downsampling to 112 px or 64 px reduced both models to about 50% accuracy because almost all AI-generated images were classified as real. The results are limited to one generator and one training run per model, so they indicate robustness to post-processing rather than general detection ability. In practice, a detector trained like these would tolerate recompression but fail on images that have been rescaled, for example by social media platforms. Training with resizing augmentation and testing on unseen generators are the most important next steps.

REFERENCES

[1] S.-Y. Wang, O. Wang, R. Zhang, A. Owens, and A. A. Efros, "CNN-Generated Images Are Surprisingly Easy to Spot...for Now," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2020, pp. 8695–8704.
[2] M. Zhu, H. Chen, Q. Yan, X. Huang, G. Lin, W. Li, Z. Tu, H. Hu, J. Hu, and Y. Wang, "GenImage: A Million-Scale Benchmark for Detecting AI-Generated Image," in *Adv. Neural Inf. Process. Syst. (NeurIPS)*, vol. 36, 2023.
[3] P. Grommelt, L. Weiss, F.-J. Pfreundt, and J. Keuper, "Fake or JPEG? Revealing Common Biases in Generated Image Detection Datasets," in *Eur. Conf. Comput. Vis. Workshops (ECCVW)*, 2024, pp. 80–95.
[4] A. Dosovitskiy et al., "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale," in *Int. Conf. Learn. Represent. (ICLR)*, 2021.
[5] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2016, pp. 770–778.

(This work is licensed under a Creative Commons Attribution 4.0 International License.)