iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 14 **Issue:** III **Month of publication:** March 2026

DOI: https://doi.org/10.22214/ijraset.2026.78396

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 14 Issue III Mar 2026- Available at www.ijraset.com

A Unified Multi-Task Vision Transformer for AI-Generated Image Detection and Generator Attribution

Dineshkumar Kumawat[1], Asst. Prof. Mamta Ramesh Devre[2]
Department of Information Technology and Mathematics, The S.I.A. College of Higher Education, Thane, India

Abstract: ***The rapid advancement of diffusion-based generative models such as Stable Diffusion, DALL·E 3, and Midjourney has significantly reduced the visual distinction between synthetic and real imagery. While most existing work focuses primarily on binary AI-image detection, relatively limited attention has been given to unified detection and generator attribution within a single framework. This paper presents a multi-task Vision Transformer (ViT) approach that jointly performs AI-generated image detection and source attribution using a shared transformer backbone with task-specific classification heads. This multi-task approach allows the model to understand general characteristics of AI-generated images while also recognizing patterns linked to particular generators. To evaluate reliability under practical conditions, a robustness-aware evaluation protocol is introduced that measures prediction variance and decision-flip rate under typical image transformations such as JPEG compression and resizing. In addition, a gradient-free occlusion-based explainability method is integrated to support interpretable and CPU-compatible inference.***

Experimental results demonstrate that the proposed framework achieves high detection reliability and strong attribution performance while maintaining stable behaviour under common image transformations. The overall system provides a unified and deployment-oriented approach for AI-image authenticity assessment in real-world environments.

Keywords: Vision Transformer; multi-task learning; source attribution; diffusion models; synthetic media forensics; robustness evaluation; confidence calibration

I. INTRODUCTION

In recent years, image generation models have progressed from visibly synthetic outputs to images that are, in many cases, difficult to distinguish from real photographs. Diffusion-based systems such as Stable Diffusion and DALL·E 3 produce high-resolution images with coherent structure, lighting, and texture patterns that closely match natural image statistics [1], [2]. Consequently, manual visual inspection is no longer sufficient for reliable authenticity verification. The detection of AI-generated images has therefore become a central problem in digital media forensics. Prior work has explored convolutional classifiers, frequency-domain analysis, and GAN fingerprinting methods [3]-[6]. These studies demonstrate that generative models often leave subtle statistical and distributional artifacts that can be exploited for detection. However, most existing approaches focus primarily on binary classification (real versus fake) without addressing the question of which generative model produced the image. In investigative or accountability contexts, source attribution can be as important as detection itself. Another practical concern is robustness. Images circulating online are frequently compressed, resized, or re-encoded by social platforms. Even well-performing detectors may exhibit unstable predictions under such transformations. Despite this, robustness under realistic perturbations remains insufficiently examined in many current detection frameworks.

Vision Transformers (ViTs) provide a compelling alternative to convolution-based detectors. Their global self-attention mechanism models long-range dependencies across image patches, which may enable improved capture of spatially distributed generative artifacts [7]. However, transformer-based detection combined with attribution and robustness analysis has not been systematically investigated. In this paper, we develop a multi-task ViT architecture that performs both AI-image detection and generator attribution within a shared representation space. We further incorporate a perturbation-based robustness evaluation protocol and a lightweight occlusion-based explanation method suitable for CPU-only deployment. Rather than optimizing solely for classification accuracy, the objective of this work is to design a stable, interpretable, and practically deployable authenticity assessment framework for real-world scenarios.

II. RELATED WORK

A. Generative image models and the need for forensic detection

Modern text-to-image generation has rapidly shifted from early GAN-based synthesis to diffusion-based pipelines that produce high-resolution outputs with consistent global structure and local texture realism. Latent diffusion models popularized efficient high-quality synthesis at scale [1], and recent proprietary systems further improved photorealism and prompt faithfulness [2]. As generative quality increases, the forensic problem is no longer limited to spotting obvious artifacts; detectors must instead identify subtle statistical inconsistencies that survive common post-processing.

B. AI-generated image detection: spatial, frequency, and fingerprint-based approaches

A large body of work in image forensics and manipulation detection motivates learning-based detectors for synthetic media. Classical forensic perspectives emphasize that manipulations often alter low-level statistics in ways that can be exploited algorithmically [3]. Building on this, deep learning detectors commonly train convolutional classifiers to discriminate real versus generated images by learning discriminative cues directly from pixel space. In parallel, frequency-domain approaches argue that generative pipelines can introduce characteristic spectral irregularities and periodic artifacts that are less obvious visually but detectable computationally [6]. Another influential direction is "fingerprinting," where generative models are shown to leave model-specific traces that can support attribution beyond binary detection. Studies on GAN fingerprints and artificial traces provide evidence that synthesis models imprint consistent signatures that can be learned and used for detection or source inference [4], [5]. While these families of approaches are effective under controlled settings, many works focus primarily on the binary detection task (real vs synthetic). In practical deployments, however, the source of generation (which model family produced the content) can be equally important for accountability, incident response, and provenance analysis especially when multiple generators coexist in the ecosystem.

C. Robustness under real-world transformations

A key limitation across many detection pipelines is sensitivity to distribution shift introduced by common image handling: JPEG compression, resizing, re-encoding, and platform-specific processing. These transformations can attenuate weak synthetic cues, distort frequency signatures, or change the statistics a detector relies on. Although robustness is often mentioned as a requirement in forensic settings, systematic evaluation under realistic perturbations is less consistently incorporated into detector design and reporting. This gap is particularly important for deployment, where reliability under routine post-processing can matter more than peak accuracy on clean benchmarks.

D. Transformer-based visual detectors and multi-task learning

Vision Transformers (ViTs) introduced a scalable alternative to convolutional backbones by modeling images as sequences of patches and learning global interactions through self-attention [7]. Compared with local receptive fields in CNNs, ViTs can integrate long-range dependencies more directly, which is potentially useful when generative artifacts are spatially dispersed or appear as weak correlations across regions.

Despite this potential, transformer-based detectors that jointly address detection and source attribution—and that report robustness behaviour—remain less explored relative to conventional CNN-centric baselines.

Multi-task learning provides a natural mechanism to unify binary detection and generator attribution by sharing a representation while keeping task-specific heads. In the context of synthetic image forensics, such a formulation can encourage features that generalize for detection while retaining generator-discriminative information for attribution, which is beneficial when the same image cues support both decisions.

E. Interpretability for authenticity assessment

Forensic tools increasingly benefit from interpretability to support user trust and debugging. Post-hoc explanation methods can highlight regions that influence a detector's decision, helping practitioners judge whether a prediction is driven by meaningful evidence or spurious correlations. Lightweight occlusion-style explanations are attractive for deployment because they do not require modifying the model and can be implemented reliably across backbones. However, explanations must be framed carefully: they provide diagnostic insight rather than proof of authenticity.

F. Research Gap and Motivation

Although significant advances have been made in detecting AI-generated images, a number of challenges still affect how effectively these methods can be used in real-world settings. Many existing methods based on convolutional models, frequency-domain analysis, and generative fingerprinting have shown that synthetic images often carry subtle statistical traces that can be used for classification [3]-[6]. While these approaches demonstrate promising results, they also highlight the growing need for authenticity verification systems that are reliable and applicable beyond controlled experimental settings. Most current detection frameworks continue to treat the problem mainly as a binary task, focusing on distinguishing real images from synthetic ones, while giving comparatively less attention to identifying the specific generative source. In real-world investigative and monitoring contexts, knowing which model produced an image can be just as important as determining whether it is synthetic. Another important concern is robustness under everyday image transformations. Images shared online are frequently compressed, resized, or re-encoded, and such modifications can influence detection performance. In addition, many existing systems provide limited insight into how decisions are made and are not always designed for deployment in environments with restricted computational resources. These limitations indicate the need for a unified, robust, and interpretable framework capable of supporting practical and dependable AI-image authenticity assessment.

III. METHODLOGY

This section describes the proposed AI Image Authenticity Analyzer, a multi-task Vision Transformer–based framework designed for simultaneous authenticity detection and source attribution of AI-generated images. The proposed model is trained and evaluated on the Defactify AI-Generated Image Veracity Dataset, containing both real and AI-generated images from multiple generative models including Stable Diffusion variants, DALL·E 3, and Midjourney. The dataset provides two labels for each image: (i) authenticity label for real versus AI-generated classification and (ii) generator source label for multi-class attribution. The dataset consists of 42,000 training images and 9,000 validation images and supports both detection and source identification tasks. The dataset is publicly available for research use [8].

A. Problem Formulation

Let an input image be denoted as:

$$\boldsymbol{x} \in \boldsymbol{R}^{\boldsymbol{H}*\boldsymbol{W}*\mathbf{3}}$$

where $\boldsymbol{H}$ and $\boldsymbol{W}$ represent image height and width, and $\mathbf{3}$ corresponds to RGB channels.

Each image is associated with two labels:

- Task A (Authenticity detection):

$$\boldsymbol{y_A} \in \{\mathbf{0}, \mathbf{1}\}$$

where $\mathbf{0}$ denotes real images and $\mathbf{1}$ denotes AI-generated images.

- Task B (Generator attribution):

$$\boldsymbol{y_B} \in \{\mathbf{0}, \mathbf{1}, \dots, \boldsymbol{K}\}$$

where $\boldsymbol{K}$ represents the number of generator classes (SD21, SDXL, SD3, DALLE3, Midjourney, and real).

The objective is to learn a function:

$$\boldsymbol{f}(\boldsymbol{x}) \rightarrow (\hat{\boldsymbol{y}}_{\boldsymbol{A}}, \hat{\boldsymbol{y}}_{\boldsymbol{B}})$$

that simultaneously predicts image authenticity and source model.

B. Data Preprocessing and Representation

All images are resized to a fixed spatial resolution:

$$\boldsymbol{x}' \in \mathbf{R}^{\mathbf{224}\times\mathbf{224}\times\mathbf{3}}$$

and normalized to a standardized distribution. Data augmentation is applied during training to improve generalization, including horizontal flipping and mild color perturbations.

The resized image is partitioned into fixed-size patches:

$$x' = \{p_1, p_2, \dots, p_n\}$$

where each patch

$$p_i \in R^{P \times P \times 3}$$

and $P = 16$ for ViT-Base. Each patch is flattened and linearly projected into a token embedding:

$$z_i = E(p_i)$$

where $E(\cdot)$ denotes the patch embedding function.

C. Multi-Task Vision Transformer Architecture

The backbone of the proposed framework is a Vision Transformer (ViT-Base), which models global image relationships through self-attention. Given tokenized image patches, the transformer encoder produces a shared feature representation:

$$h = f_\theta(x')$$

where:

- h is the learned global embedding,
- θ denotes model parameters,
- f_θ represents the transformer backbone.

This shared representation enables simultaneous optimization of authenticity detection and generator attribution.
Two task-specific classification heads operate on the shared embedding:

- Task A: Authenticity detection

$$\hat{y}_A = W_A h + b_A$$

- Task B: Generator attribution

$$\hat{y}_B = W_B h + b_B$$

Softmax probabilities are computed as:

$$p_i = \frac{e^{z_i}}{\sum_j e^{z_j}}$$

where p_i denotes predicted class probability.

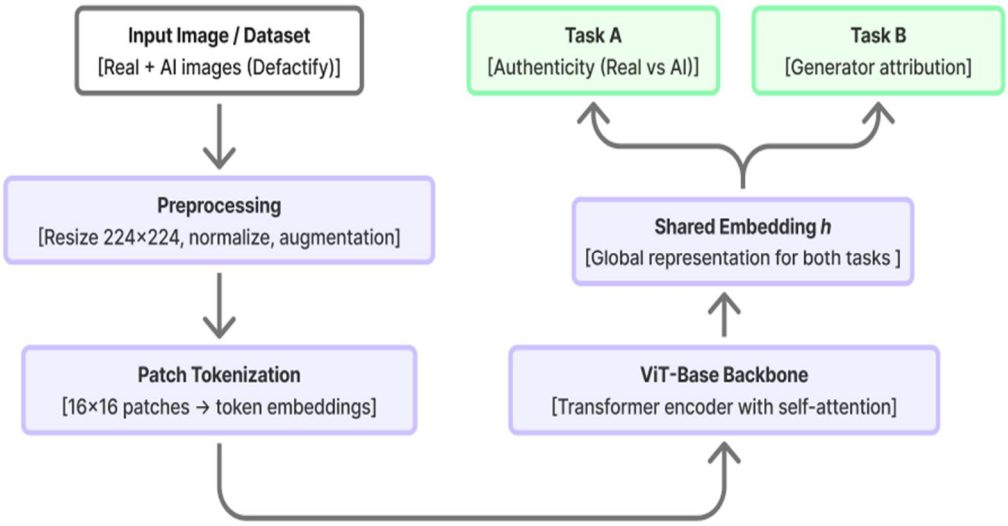


Fig. 1 Multi-Task Vision Transformer Architecture

D. Multi-Task Optimization

The model is trained using a combined cross-entropy loss.

For Task A:

$$\mathcal{L}_A = -\sum y_A \log(p_A)$$

For Task B:

$$\mathcal{L}_B = -\sum y_B \log(p_B)$$

The overall objective function is:

$$\mathcal{L} = \mathcal{L}_A + \lambda \mathcal{L}_B$$

where λ controls the contribution of generator attribution to total loss. In this work, $\lambda = 0.5$ is used to balance both tasks. This multi-task formulation encourages shared feature learning while preserving generator-specific discriminative cues.

E. Training Strategy

The model is trained using the AdamW optimizer with weight decay regularization. Automatic mixed precision is employed to improve computational efficiency while maintaining numerical stability. Training proceeds for multiple epochs using mini-batch gradient descent.

To monitor performance, the following metrics are computed:

- Task A accuracy
- Task B accuracy
- ROC–AUC for authenticity detection

ROC–AUC is computed using rank statistics:

$$AUC = \frac{R_p - \frac{n_p(n_p + 1)}{2}}{n_p n_n}$$

where:

- R_p = sum of ranks of positive samples
- n_p = number of AI images
- n_n = number of real images

This metric evaluates threshold-independent detection capability.

F. Robustness Evaluation

To simulate real-world conditions, robustness is evaluated using common image perturbations:

- JPEG compression
- Downscaling and resizing
- Re-encoding

Let predicted AI probabilities across perturbed versions be:

$$P = \{p_1, p_2, \ldots, p_n\}$$

Prediction variance is computed as:

$$\sigma^2 = Var(P)$$

Prediction flip rate is defined as:

$$FlipRate = \frac{\text{number of label changes}}{\text{total perturbations}}$$

A robustness score is derived:

$$Score = 100 - (\alpha\sigma^2 + \beta FlipRate)$$

where α and β control sensitivity to instability. This provides a stability-aware reliability measure.

G. Explainability via Occlusion Sensitivity

To enhance interpretability, an occlusion-based explanation method is integrated. Given an input image, small regions are masked sequentially and prediction change is measured:

$$\Delta = s(x) - s(x_{masked})$$

where:

- $s(x)$ is prediction confidence,
- x_{masked} is occluded image.

Regions causing larger confidence drops are considered more influential. The resulting heatmap is normalized:

$$H = \frac{H - H_{min}}{H_{max} - H_{min}}$$

and overlaid on the original image to visualize decision-relevant regions.

H. Deployment Design

The framework is designed for practical deployment using CPU-optimized inference and lightweight visualization. A Streamlit interface allows users to upload images and obtain:

- authenticity prediction
- generator attribution
- robustness analysis
- visual explanation heatmap

This design enables accessibility in resource-constrained environments while maintaining analytical transparency.

IV. EXPERIMENTAL SETUP

A. Dataset and Splits

All experiments use the Defactify AI-Generated Image Veracity Dataset, which provides two supervised labels per image: Label_A for authenticity (Real vs AI-generated) and Label_B for generator attribution (six-way classification including Real, SD21, SDXL, SD3, DALL·E 3, Midjourney). We use the dataset's predefined train and validation splits and load them directly from cached HuggingFace Arrow shards to ensure reproducible, lock-free data access during training.

B. Preprocessing and Augmentation

All images are converted to RGB and resized to 224×224 to match the ViT-Base input resolution.

- Training transforms: Resize(224), RandomHorizontalFlip(p=0.5), ColorJitter(brightness=0.10, contrast=0.10, saturation=0.10, hue=0.02), ToTensor, Normalize(mean=[0.5,0.5,0.5], std=[0.5,0.5,0.5]).
- Validation transforms: Resize(224), ToTensor, same normalization (no augmentation).

This setup enforces a consistent tokenization grid for the transformer while using mild photometric augmentation to improve generalization.

C. Model Configuration

We implement a multi-task Vision Transformer with a shared backbone and two task-specific heads:

- Backbone: vit_base_patch16_224 (ViT-Base, patch size 16), initialized with ImageNet-pretrained weights via timm.
- Shared representation: the backbone outputs a global embedding vector.
- Head A: 2-way softmax for authenticity (Real vs AI).
- Head B: 6-way softmax for generator attribution.
- Regularization: dropout 0.10 applied before both heads.

D. Training Protocol

Training is performed on CUDA only (the training script hard-fails if CUDA is unavailable), while inference is designed to run on CPU (deployment section below).

- Optimizer: AdamW
- Learning rate: 2×10^{-5}
- Weight decay: 0.01
- Epochs: 12
- Batch size: 4
- Loss: multi-task cross entropy

$$\mathcal{L} = \mathcal{L}_A + \lambda\mathcal{L}_B, \lambda = 0.5$$

- Mixed precision: enabled using torch.amp.autocast("cuda") and GradScaler
- Gradient clipping: global norm clip at 1.0
- Gradient accumulation: disabled (set to 1 in code)
- DataLoader: shuffle on train, pin_memory=True, num_workers=0 (reproducible and Windows-safe)

E. Checkpointing, Model Selection, and Logging

We maintain two checkpoints:

- LAST: saved every epoch (last_multitask_vit.pth) with model + optimizer + scaler + epoch and best-metric tracking.
- BEST: saved when validation improves, using Task A ROC–AUC as the default selection criterion (best_metric_mode="auc").

Training logs are stored in both:

- CSV history: checkpoints/history.csv
- TensorBoard: scalar logs under runs/multitask_vit
- Plots: loss curve, validation accuracy curves, and validation AUC curve exported as PNGs into checkpoints/.

F. Evaluation Metrics

We report:

- Task A Accuracy: fraction of correctly classified Real/AI samples.
- Task B Accuracy: fraction of correctly classified generator classes.
- Task A ROC–AUC: computed from P(AI) using a rank-based (Mann–Whitney) implementation, enabling threshold-free detection comparison.

G. Robustness Evaluation Protocol

To evaluate real-world stability, we test prediction consistency under common platform transformations using the same trained checkpoint:

Perturbation set (per image):

- JPEG recompression at qualities 95, 75, 50
- Resize down then up using bicubic interpolation at scales 75% and 50%
- Plus the original

For each variant, we collect $\mathbf{P(AI)}$ and compute:

- Variance: $\mathbf{Var}(\boldsymbol{P}(\boldsymbol{AI}))$
- Flip rate: fraction of variants whose binary decision differs from the original under a user-defined threshold $\mathbf{t}$ (default UI threshold set to 0.80).

We additionally expose a robustness score (0–100) that penalizes both instability sources (variance and flips), matching the deployment implementation.

H. Explainability and Deployment Configuration

For interpretability in CPU-only environments, we integrate occlusion sensitivity:

1) The input is resized to 224×224
2) A sliding occluder masks patches and measures confidence drop for one of:
- **P(AI)** (probability-based)
- Task A decision confidence (predicted Real/AI confidence)
- Task B generator confidence (predicted generator confidence)

3) Three compute modes trade off speed vs fidelity by changing occluder size/stride:
- Fast: 56 / 28
- Medium: 48 / 24
- Strong: 40 / 20

4) A time guard (default 12s) stops early to keep latency bounded on CPU while still returning a partial but informative heatmap.

Deployment is implemented via a Streamlit application that runs inference on CPU (CUDA disabled via environment settings), supports checkpoint selection, shows confidence outputs for both tasks, runs robustness checks, and produces occlusion heatmaps for user-facing interpretability.

V. RESULTS

A. Quantitative Performance Evaluation

The proposed multi-task Vision Transformer was evaluated on the Defactify validation split for both authenticity detection and generator attribution. The best model was selected using validation ROC–AUC for the authenticity detection task, as this metric provides a threshold-independent view of how well the model separates real and AI-generated images.

Table 1 Validation performance of the proposed multi-task Vision Transformer on the Defactify dataset

Model Checkpoint Selection	Epoch	Task A Accuracy (Real vs AI)	Task A ROC–AUC	Task B Accuracy (Generator Attribution)
Best ROC–AUC (Selected Model)	9	97.92%	0.9964	94.54%
Highest Task A Accuracy	12	98.01%	0.9954	94.33%
Highest Task B Accuracy	2	97.58%	0.9950	94.94%

The model achieved a maximum ROC–AUC of 0.9964, indicating that it can clearly distinguish between real and synthetic images. In practical terms, this means the detector maintains strong ranking ability even when the decision threshold changes. The validation accuracy for authenticity detection remained above 97.9% across training and reached its highest value at the final epoch, confirming that the predictions are both accurate and stable. For generator attribution, the model consistently achieved validation accuracy above 94%. This suggests that the shared transformer backbone is able to capture generator-specific visual patterns while still learning general cues that separate real and synthetic images. The results show that training the model jointly on both tasks does not reduce performance; instead, it allows the network to learn features useful for both detection and attribution simultaneously.

B. Training Convergence and Learning Behaviour

The training process was monitored using loss, ROC–AUC, and validation accuracy curves. The training loss decreased steadily from the first epoch to the last, indicating stable optimization and effective learning. No sudden spikes or irregular behaviour were observed, suggesting that the model converged smoothly under the multi-task objective.

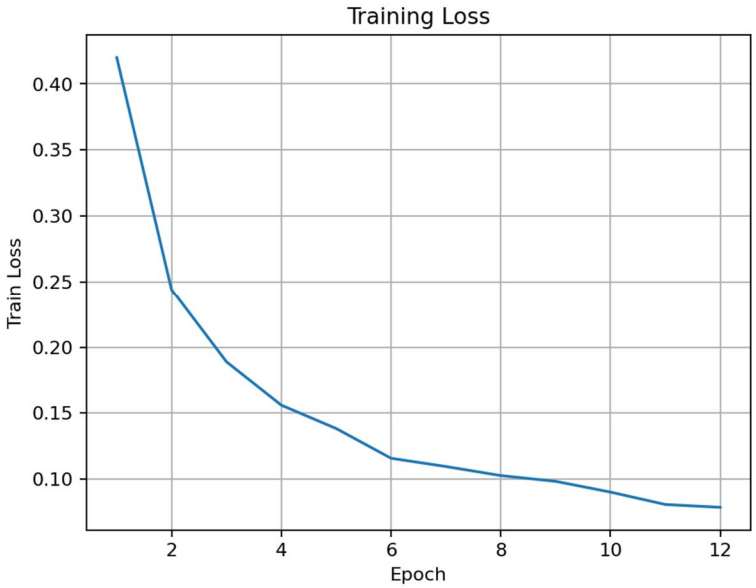


Fig. 2 Training loss across epochs showing stable convergence of the multi-task model.

The ROC–AUC for authenticity detection improved quickly during the early epochs and remained above 0.99 for the remainder of training. This shows that the model was able to learn meaningful authenticity-related features early on. Validation accuracy for both tasks followed a similar trend, remaining consistently high with only minor variation between epochs.

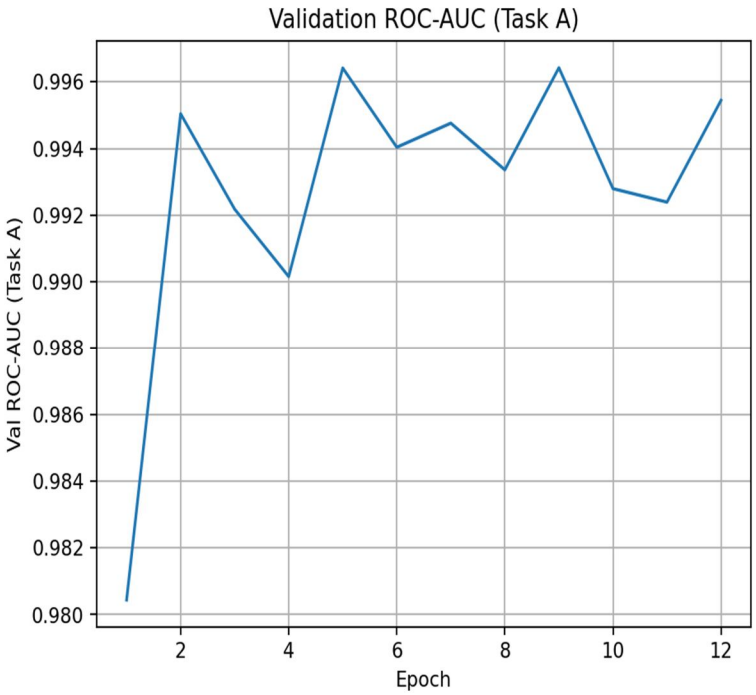


Fig. 3 Validation ROC–AUC for authenticity detection across training epochs. Peak performance occurs at epoch 9.

An interesting observation is that authenticity detection performance stabilized earlier than generator attribution. This is expected, as identifying whether an image is synthetic is generally easier than determining which model generated it. Attribution accuracy continued to improve gradually across later epochs, indicating that the model refined generator-specific representations over time. Overall, the training curves reflect stable learning and good generalization without noticeable overfitting.

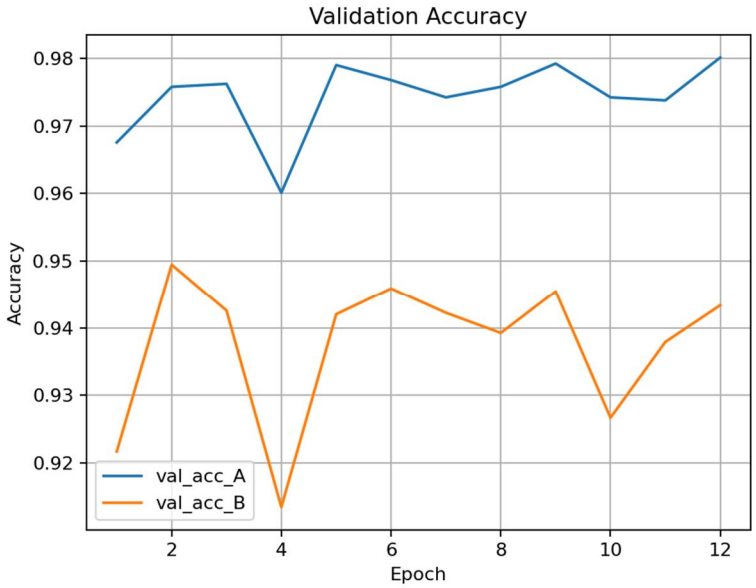


Fig. 4 Validation accuracy trends for authenticity detection (Task A) and generator attribution (Task B).

C. Efficiency and Training Stability

The training process progressed in a stable and predictable manner across all epochs. Loss values decreased gradually, and validation metrics remained consistent without sudden fluctuations. This indicates that combining authenticity detection and generator attribution within a single framework did not create optimization conflicts. Instead, both tasks benefited from learning through a shared representation built by the transformer backbone. Using the chosen configuration, the model processed approximately 20-21 images per second during training. This shows that the framework can be trained efficiently on a single GPU without requiring heavy computational infrastructure. Even with the multi-task design, the training pipeline remained straightforward to manage and reproduce, which is important for both research and practical deployment settings. Another encouraging observation was the consistency of prediction confidence as training progressed. Features learned for authenticity detection also contributed to improvements in generator attribution, while attribution learning helped refine authenticity predictions. This mutual reinforcement between tasks resulted in stable validation performance and reduced variability across epochs. Overall, the training behaviour suggests that the multi-task formulation supports efficient and dependable learning without introducing instability.

D. Result Interpretation

The overall performance demonstrates that the proposed framework can serve as a reliable tool for AI-image authenticity analysis. The very high ROC–AUC values indicate that the model is able to clearly distinguish between real and AI-generated images, even when evaluated under different decision thresholds. This makes the system more adaptable to real-world scenarios where strict threshold settings may vary depending on the application. In addition, the strong generator attribution accuracy shows that the model successfully retains generator-specific patterns within its learned representations. Rather than treating detection and attribution as separate tasks, the multi-task approach allows both to be learned together within a shared representation space. This enables the model to capture both general authenticity cues and subtle generator-specific characteristics without sacrificing performance. The training curves further support the reliability of the approach, showing steady convergence and balanced improvements across both tasks. Taken together, these results suggest that the proposed multi-task Vision Transformer provides a practical and dependable solution for detecting AI-generated images and identifying their likely source models in real-world environments.

VI. DISCUSSION

The results show that the proposed multi-task Vision Transformer works effectively as a unified framework for AI-image authenticity analysis. Instead of separating detection and generator attribution into different models, both tasks are handled within a shared representation.

This design helps the model learn general visual cues of synthetic images while also capturing generator-specific characteristics, leading to balanced and dependable performance across tasks. The high ROC–AUC values indicate that the model can clearly differentiate between real and AI-generated images across a range of decision thresholds. Such consistency is important in practical environments where images may be compressed, resized, or re-encoded before analysis. A system that maintains reliable separation under these variations is more suitable for real-world use than one tuned only for fixed evaluation conditions. Performance on generator attribution also remains strong, suggesting that the shared transformer backbone retains meaningful patterns associated with different generation pipelines. Even with the growing realism of modern diffusion models, subtle visual signatures remain detectable, allowing the model to identify likely source models alongside authenticity classification. Learning both tasks together encourages the network to preserve features that support both decisions within a single representation space.

The training process was stable and computationally manageable. Convergence occurred smoothly without noticeable instability, and the model was trained successfully on a single GPU with moderate settings. This indicates that the framework can be reproduced and adapted without requiring extensive computational infrastructure, which is beneficial for both academic and applied settings. From a practical standpoint, using a single model for both detection and attribution simplifies deployment. It reduces system complexity and avoids the need for multiple processing pipelines. Combined with robustness checks and interpretability components, the framework offers a transparent and accessible approach for authenticity verification. Overall, the study demonstrates that a multi-task transformer-based approach provides a reliable direction for modern AI-image forensics. By combining detection and attribution within one coherent system, the proposed framework supports accurate, stable, and practical authenticity assessment in real-world scenarios.

VII. LIMITATIONS AND FUTURE WORK

The proposed framework demonstrates strong performance; a few limitations should be considered. This model has been trained and evaluated on a specific dataset that includes selected diffusion-based generators. As latest generative models continue to evolve, their visual patterns and synthesis techniques may differ from those seen during training. This could influence how well the model generalizes to unseen generators. Expanding the scope of dataset to include a wider variety of generative models and real-world image sources would help improve adaptability and long-term reliability. The current robustness evaluation focuses mainly on everyday image transformations such as JPEG compression and resizing. However, images shared around various digital platforms often undergo additional transformations, including cropping, filtering, and repeated re-encoding. Exploring more broader range of real-world distortions in future experiments would provide a more comprehensive understanding of model reliability under practical conditions. Another aspect to consider is training efficiency; while inference can be performed on standard CPU systems, training the transformer backbone still requires GPU resources. Future research could explore lightweight model variants or knowledge distillation approaches to reduce training cost and make the system more accessible and right. Further improvements may also include integrating additional signals such as metadata or adopting continual learning strategies to keep pace with emerging generative technologies.

VIII. CONCLUSION

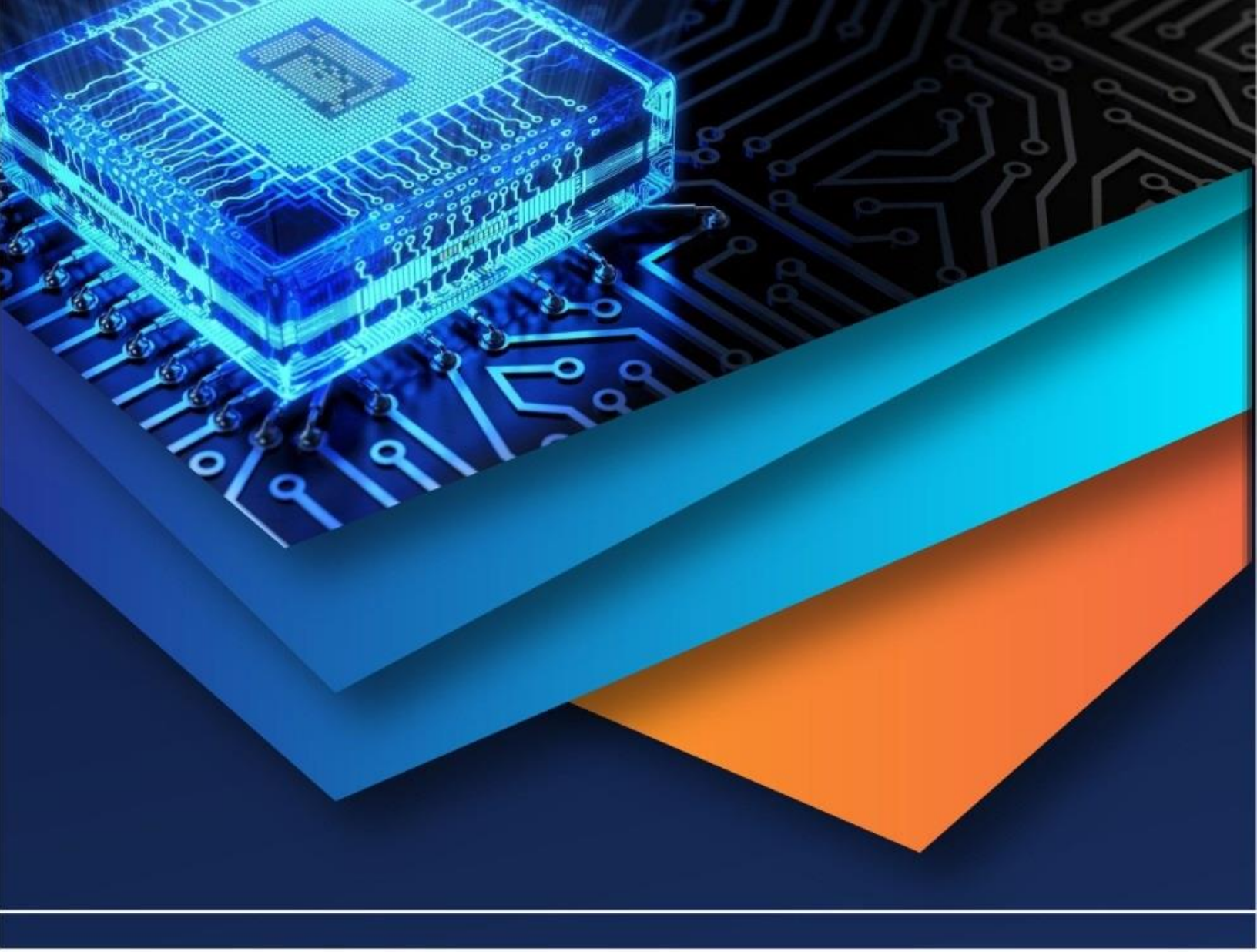
This paper presented a multi-task Vision Transformer framework for unified AI-generated image detection and generator attribution. By combining both tasks within a shared representation, the proposed approach demonstrates that authenticity detection and source identification can be performed together without sacrificing performance. Experimental results show high detection reliability and strong attribution accuracy, supported by stable training behaviour and robustness under typical image transformations. The study also highlights the importance of moving beyond simple binary detection toward more informative authenticity analysis that includes generator-level insights. The integration of robustness evaluation and interpretability further strengthens the practical usefulness of the framework in real-world scenarios.

The proposed system offers a balanced and deployable solution for modern synthetic media forensics. As generative models continue to improve in visual realism, unified approaches that combine detection, attribution, and reliability assessment will play an increasingly important role in maintaining trust in digital visual content.

REFERENCES

[1] R. Rombach, A. Blattmann, D. Lorenz, P. Esser and B. Ommer, "High-Resolution Image Synthesis with Latent Diffusion Models," 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA, 2022, pp. 10674-10685, doi: 10.1109/CVPR52688.2022.01042.
[2] OpenAI, “DALL·E 3 Technical Report,” 2023. Available: https://openai.com/research/dall-e-3

[3] H. Farid, "Image forgery detection," in IEEE Signal Processing Magazine, vol. 26, no. 2, pp. 16-25, March 2009, doi: 10.1109/MSP.2008.931079.
[4] L. Guarnera, O. Giudice and S. Battiato, "Fighting Deepfake by Exposing the Convolutional Traces on Images," in IEEE Access, vol. 8, pp. 165085-165098, 2020, doi: 10.1109/ACCESS.2020.3023037. arXiv:2008.04095
[5] F. Marra, D. Gragnaniello, L. Verdoliva and G. Poggi, "Do GANs Leave Artificial Fingerprints?," 2019 IEEE Conference on Multimedia Information Processing and Retrieval (MIPR), San Jose, CA, USA, 2019, pp. 506-511, doi: 10.1109/MIPR.2019.00103. arXiv:1812.11842
[6] R. Durall, M. Keuper, and J. Keuper, "Watch Your Up-Convolution: CNN Based Generative Deep Neural Networks Are Failing to Reproduce Spectral Distributions," Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020. arXiv:2003.01826
[7] A. Dosovitskiy et al., "An Image is Worth 16×16 Words: Transformers for Image Recognition at Scale," International Conference on Learning Representations (ICLR), 2021. arXiv:2010.11929
[8] R. Roy et al., "Defactify: AI-Generated Image Veracity Dataset," Hugging Face, 2024. [Online]. Available: https://huggingface.co/datasets/Rajarshi-Roy-research/Defactify_Image_Dataset



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY