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

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 05 , May - 2026

Content-Agnostic Artifact-Based Detection of AI-Generated Images and Videos with Incremental Learning

Lourdu Radjou A[1], P. Salini[2], Khalid Ahmad[3], Aditya Singh[4], K R Rangarajan[5]
Department of Computer Science and Engineering
Puducherry Technological University (PTU)

Abstract - Generative AI for creating images and videos is growing rapidly with images and videos indistinguishable from human eyes, which creates problems in how we interpret real and AI generated content in day to day life. This creates fake news, fake manipulated images and videos of people and things which is hard to detect with new models coming each day. Our idea revolves around a new way to detect both images and videos using a single architecture which is easy to maintain, fine tune and scale in real time. A lot of methods which are already published doesn't utilize this scope to create a unified model with good amount of accuracy. We combine the methods used in multiple research papers to make a foundation model which will seamlessly work for both images and videos making a baseline model which could be iterative trained and fine tuned to achieve a significant deep learning model which is smaller in size and less preprocessing with the help of frequency artifacts of the images, which makes our title content agnostic in nature by extracting the frequency artifacts of the images or frames of the videos. Besides the model itself, we also built a web interface so anyone can upload an image or video and get a prediction on if it is real or artificial intelligence-made plus a confidence score. This part could make technology more useful to everyday people.

Index Terms—AI-generated image detection, Vision Transformer, Complex Steerable Pyramid, Incremental Learning, Catastrophic Forgetting, Frequency Artifacts, Content-agnostic Features, Visual media detection

I. INTRODUCTION

AI's development has been rapid lately; enhancements in Generative Adversarial Networks (GANs) and transformer architectures have changed the way media can be produced. Newer technologies, such as Google's Nanobanana and OpenAI's Sora, have enabled users to create videos and images that look extremely real, making it difficult to identify which is real and which is fake.

The creation of new technologies has made it easy for people to develop unique content in many different fields, including art, business, and industry; however, these technologies pose major ethical and security concerns. For instance, both traditional and new forms of media now have the ability to look real, allowing fake news stories or other types of disinformation to go viral on social media and even allowing some forms of scams. If action is not taken, people will lose trust in the internet and other digital platforms.

Currently, it is getting harder for people to tell whether a picture or video is real or not. Research shows that people often face trouble differentiating between online media. As the intelligence technology is getting better, this problem is becoming worse. So we really need to come up with ways to correctly detect media made with AI. The methods we have now are not very good. They often work well with one type of media, but not with others. They also have trouble with types of media that they have not seen before. This is a weakness in our ability to defend against media made using AI.

Fig. 1. Spatial Artifacts Vs Frequency Artifacts

The EchtAI project is being established because of this. The system will be able to recognize when images & videos are produced using artificial intelligence. It detects multiple signs of being fake across any type of media produced using AI instead of just identifying certain indicators for a single kind of fake media, such as an original photograph or a computergenerated video. An additional tool has been created to provide users with the ability to upload images & videos, which can determine whether the media has been artificially created. This tool is set up for researchers, institutions, and everyday users as a resource to help identify fake content from real content while browsing the Internet.

II. RELATED WORK

The rapid advancement of generative models, particularly GANs and diffusion-based architectures, has significantly increased the realism of synthetic media, necessitating robust and generalizable detection mechanisms. Existing research in AI-generated media detection can broadly be categorized into (i) spatial artifact-based methods, (ii) temporal dynamicsbased approaches, (iii) hybrid architectures, and (iv) generalizable and incremental learning frameworks.

Early works in AI-generated image detection primarily focused on identifying spatial artifacts introduced during the

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 05 , May - 2026

generation process. These approaches leveraged inconsistencies such as upsampling traces, color discrepancies, and frequency-domain anomalies. However, as generative models evolved, these artifacts became increasingly subtle, reducing the effectiveness of purely spatial detectors. To address this limitation, Tang *et al.* proposed a content-agnostic adapterbased framework that utilizes a Vision Transformer backbone to extract global forensic features independent of image content [1]. Their work highlights the importance of learning generalized representations that remain robust across different generative models. Furthermore, their incremental learning paradigm addresses the challenge of catastrophic forgetting, enabling the detector to adapt to newly emerging generative techniques without losing prior knowledge.

In the domain of video deepfake detection, temporal inconsistencies have been widely explored as a more reliable signal compared to static spatial artifacts. Prashnani *et al.* introduced a phase-based motion analysis approach that leverages temporal phase variations in frequency bands to capture subtle inconsistencies in facial dynamics [2]. Unlike traditional methods that rely on explicit motion estimation or landmark tracking, their approach directly models temporal phase changes, resulting in improved robustness against compression artifacts and adversarial perturbations. As highlighted in their results, phase-based representations achieve superior cross-dataset generalization, demonstrating the effectiveness of frequency-domain analysis for detecting synthetic media.

Hybrid approaches have also gained attention due to their ability to combine the strengths of multiple architectures. Odeh *et al.* proposed a hybrid CNN–Vision Transformer model that integrates local spatial feature extraction with global contextual understanding [3]. Their findings show that such hybrid models significantly outperform human perception in detecting deepfake images, achieving over 91% accuracy. This line of work emphasizes the complementary nature of convolutional and transformer-based architectures, suggesting that multi-scale and multi-representation learning is essential for robust detection systems.

Another important research direction focuses on generalization and robustness across domains. Many existing detectors perform well on known datasets but fail under domain shifts, compression, or adversarial conditions. Jadhav and Bartere addressed these limitations by proposing a semantically disentangled and temporally-aware framework that separates contentrelated features from manipulation-specific features [4]. Their architecture incorporates a memory-guided temporal transformer to capture long-range dependencies in videos and introduces privacy-preserving federated learning mechanisms for real-world deployment. Notably, their approach demonstrates significant improvements in adversarial robustness and crossdomain performance, highlighting the importance of modular and scalable architectures.

Despite these advancements, several challenges remain. First, many methods are modality-specific, focusing either on images or videos but not both. Second, models often rely on dataset-specific artifacts, limiting their ability to generalize to unseen generative techniques. Third, the increasing complexity of detection frameworks makes them difficult to maintain and deploy in real-time applications.

Motivated by these limitations, recent research trends emphasize content-agnostic feature extraction, frequency-domain analysis, and unified architectures. Frequency-based methods, in particular, have shown strong potential due to their ability to capture intrinsic generation patterns that are less dependent on visual content. Additionally, incremental and continual learning strategies are becoming increasingly important for adapting to rapidly evolving generative models.

In contrast to prior work, our approach aims to unify image and video detection within a single scalable architecture. By leveraging frequency artifacts extracted from both images and video frames, our method remains content-agnostic while maintaining efficiency and adaptability. Furthermore, our framework is designed to be lightweight, easily finetunable, and suitable for real-time applications, addressing key gaps in existing literature.

Additional Recent Advances: Beyond the aforementioned foundational works, recent studies have further expanded the landscape of AI-generated media detection by addressing emerging challenges such as generalization, data scarcity, explainability, and efficiency. In particular, the increasing dominance of diffusion-based generative models has motivated the development of specialized detection techniques. Xu *et al.* proposed a hybrid attention-based and Vision Transformer framework to detect images generated by text-to-image diffusion models, demonstrating strong robustness by capturing longrange dependencies and global representations [10]. Similarly, Bammey introduced Synthbuster, which specifically targets diffusion-generated images by exploiting intrinsic generation artifacts, highlighting the growing need for detectors tailored to new generation paradigms [7]. Complementing these efforts, Tran *et al.* proposed DiffCoR, which leverages reconstruction discrepancies from stable diffusion processes combined with frequency-domain analysis and representation learning to achieve strong cross-dataset generalization [16]. These works collectively emphasize the importance of adapting detection strategies to evolving generative mechanisms.

Another significant direction focuses on improving generalization under limited data conditions. Xu *et al.* introduced FAMSeC, a few-shot learning-based detection method that utilizes a forgery-aware module and contrastive learning to achieve strong performance even with minimal training samples [6]. This is particularly important in real-world scenarios where access to large-scale datasets of generated content is restricted. In a related vein, Li *et al.* proposed an image transformation-based framework (SAFE) that improves generalization by preserving artifact features and introducing invariant augmentations, thereby mitigating overfitting and domain bias [15]. These approaches highlight a shift toward designing detectors that are not only accurate but also adaptable to unseen generative models and data distributions.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 05 , May - 2026

Several works have also explored artifact-focused and frequency-domain representations to improve cross-domain detection performance. Meng *et al.* proposed an artifact purification network that explicitly separates and enhances manipulation-specific features while suppressing contentrelated information, leading to improved cross-generator and cross-scene generalization [12]. Their findings reinforce the importance of disentangling content and forgery features, aligning with the broader trend toward content-agnostic detection. Additionally, lightweight detection models have been explored to facilitate real-world deployment. Ghosh and Naskar proposed a minimalist approach using chrominance-based statistical features, achieving competitive performance with significantly reduced computational complexity [11]. Similarly, Li *et al.* developed lightweight artifact-based detection methods that balance efficiency and accuracy for scalable applications [14].

Explainability and interpretability have also emerged as critical aspects of modern detection systems. Dwivedi *et al.* introduced an ensemble explainable AI framework that combines saliency maps, CAM, and Grad-CAM to provide insights into model decision-making, thereby increasing trust and transparency in biometric forgery detection systems [9]. As noted in their study, understanding model behavior is essential for deploying detection systems in high-stakes environments such as security and authentication [9]. Furthermore, Khan and Khan explored the intersection of generative AI and biometric systems, emphasizing the security implications of deepfake technologies and the necessity for robust and interpretable detection mechanisms in identity verification applications [5].

Finally, recent works have explored alternative learning paradigms and architectural innovations. Pintelas and Livieris proposed a diffusion-based CNN framework that leverages the diffusion process itself to reveal hidden manipulation patterns, offering a novel perspective on feature extraction [13]. Lamichhane investigated Vision Transformer-based architectures for improved detection of AI-generated images, demonstrating the effectiveness of transformer models in capturing global dependencies [8]. These studies collectively indicate a trend toward integrating novel learning paradigms, including diffusion processes and transformer-based architectures, into detection pipelines.

Overall, the recent literature highlights a transition from task-specific and modality-specific detectors toward more generalizable, efficient, and interpretable frameworks. Despite these advancements, most existing methods still focus on either images or videos independently and often rely on complex or resource-intensive architectures. This reinforces the need for unified, lightweight, and content-agnostic models capable of handling both modalities effectively, which motivates the approach proposed in this work.

III. PROPOSED METHODOLOGY

A. System Overview

The EchtAI framework implements a unified, end-to-end pipeline to seamlessly process both static images and dynamic video content. As shown in Figure 2, the core intuition behind the system is that frequency-domain phase representations capture structural information that remains invariant to visual content such as color and texture. This enables the extraction of intrinsic forensic signals that are difficult for generative models to replicate.

Building upon this insight, the complete pipeline (illustrated in Figure 3) integrates frequency-domain feature extraction with transformer-based representation learning. The system first converts input media into phase-based representations using a Complex Steerable Pyramid (CSP), and then processes these features through a frozen Vision Transformer (ViT) augmented with lightweight adapter modules. This design enables efficient, scalable, and real-time detection of AI-generated media while maintaining strong generalization across unseen generative models.

B. Input Preprocessing

Images are resized to 224 × 224 and normalized using ImageNet statistics. Videos are decomposed into frames using OpenCV, where every *N*-th frame is sampled to reduce computational overhead. Each frame is treated identically to a static image and passed through the same processing pipeline.

C. CSP Phase Extraction

As illustrated in Figure 2, the Complex Steerable Pyramid decomposes an input image into multiple frequency bands across different scales and orientations. Specifically, each image is decomposed into 4 scales and 2 orientations, producing multiple bandpass components.

Only mid-frequency (bandpass) components are retained, as they contain the most informative forensic cues. Lowfrequency components capture global illumination, while highfrequency components are often dominated by noise and compression artifacts.

From these bandpass components, the phase information is extracted. Unlike magnitude, phase encodes the geometric structure of the image. As shown in the phase reconstruction in Figure 2, structural details such as edges and object boundaries are preserved even when magnitude information is discarded.

AI-generated images tend to introduce subtle but consistent inconsistencies in these phase patterns due to upsampling, convolution artifacts, and synthesis priors. In contrast, real images exhibit natural phase distributions governed by physical imaging processes. This makes phase a robust, contentagnostic signal for forgery detection.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 05 , May - 2026

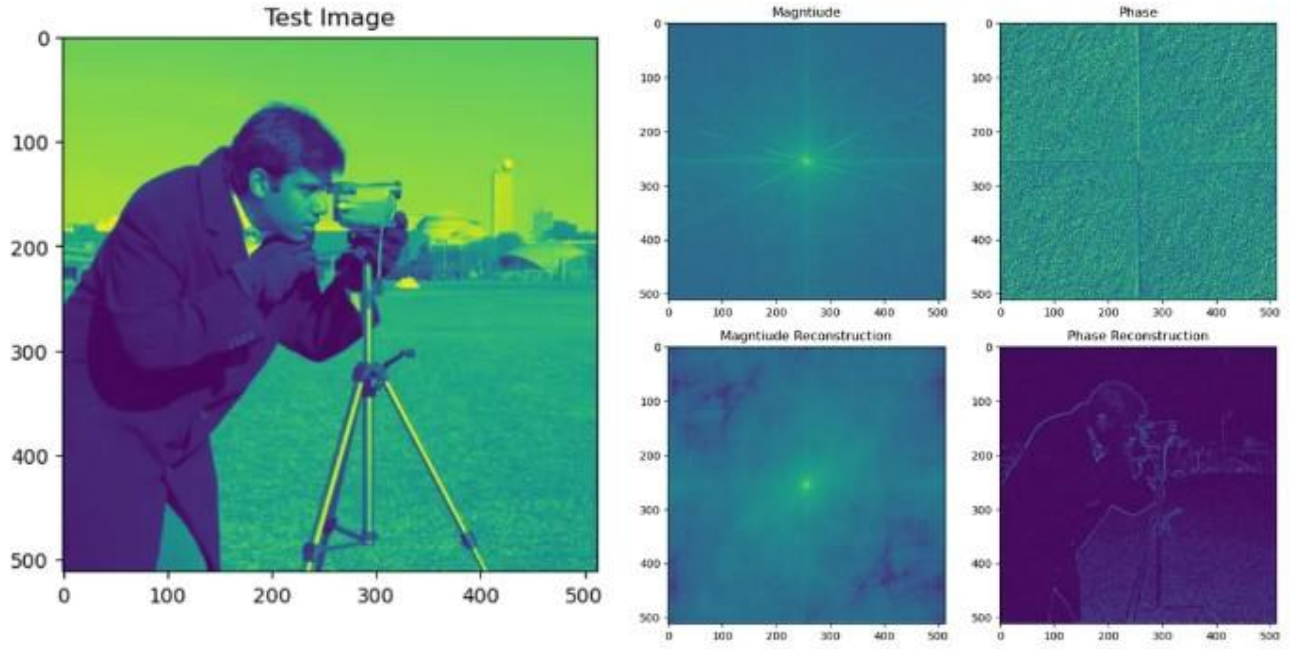

Fig. 2. Frequency-domain decomposition of an input image using Complex Steerable Pyramid (CSP). The magnitude and phase representations highlight structural information, where phase reconstruction preserves geometric details critical for detecting synthetic artifacts.

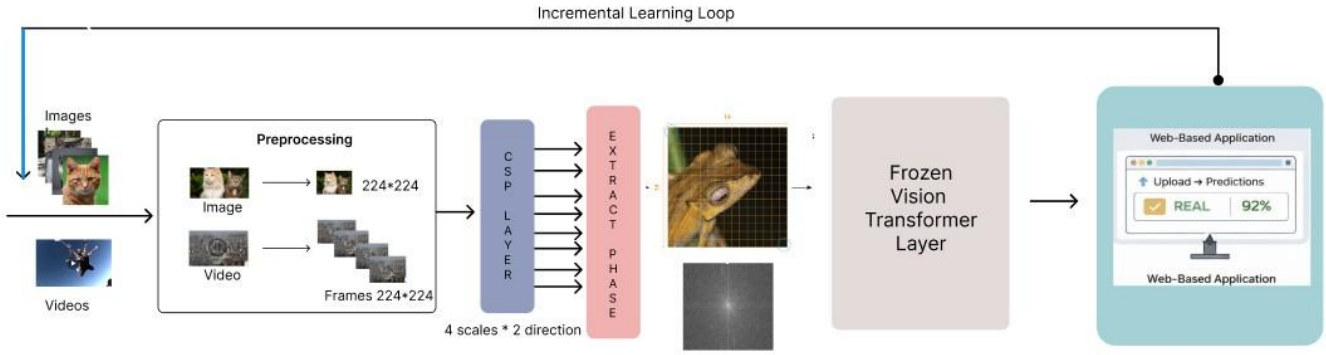

Fig. 3. EchtAI pipeline: input → CSP phase extraction → frozen Vision Transformer with adapters → MLP head → web-based prediction output.

D. Vision Transformer Backbone

EchtAI employs a ViT-Base model with 12 encoder layers and 768-dimensional embeddings. The backbone is initialized from pretrained weights and remains frozen during training to preserve general visual representations.

Each 224 × 224 phase map is divided into 196 nonoverlapping patches of size 16×16. A learnable class token is prepended to the sequence, enabling the model to aggregate global forensic evidence through multi-head self-attention.

E. Adapter Modules

To enable efficient learning, lightweight adapter modules are inserted into the frozen ViT backbone. Early layers (1–10) use convolutional adapters to capture localized artifact patterns in a bottleneck structure. Later layers (11–12) employ dualstream shuffle adapters, where one stream processes tokens in original order and another processes randomly shuffled tokens.

A cosine similarity constraint is applied between the two streams to enforce content invariance, ensuring that the model focuses on artifact patterns rather than semantic structure. Only adapter parameters are trained, resulting in approximately 3% of total parameters being updated.

F. Incremental Learning Setup

The model is trained in sequential sessions to handle evolving generative models. The initial session focuses on GANgenerated data, followed by subsequent sessions incorporating diffusion-based models and other emerging generators.

To mitigate catastrophic forgetting, knowledge distillation is applied by aligning CLS token outputs between successive model versions. Both point-wise and structure-wise distillation strategies are used. Additionally, category-aware domain alignment is applied to maintain separation between real and synthetic distributions across sessions.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 05 , May - 2026

G. Loss Function

The training objective consists of three components:

- Cross-Entropy Loss for binary classification
- Cosine Shuffle Loss to enforce content invariance
- Knowledge Distillation Loss to preserve prior knowledge

H. Classification Head

The CLS token is passed through a two-layer MLP to produce a scalar output, which is converted to a probability using a sigmoid function. Outputs below 0.5 are classified as real, while outputs above 0.5 are classified as AI-generated, along with a confidence score.

IV. EXPERIMENTAL SETUP

A. Dataset

A total of approximately 5,000 images were used across four categories. Real images were sourced from COCO, ImageNet, LSUN, and CelebA-HQ. Synthetic images were generated using GAN-based models (ProGAN, StyleGAN, BigGAN, SAGAN), diffusion models (Stable Diffusion, Midjourney, ADM, DALL-E), and inpainting techniques.

The dataset was split into 70% training, 15% validation, and 15% testing. All images were resized to 224 × 224 and augmented using JPEG compression, Gaussian blur, and random cropping to simulate real-world conditions.

B. Implementation Details

The model was implemented in PyTorch using the timm library for ViT-Base initialization. Training was performed using the AdamW optimizer with a learning rate of 10^{-4} and batch size of 32.

The full model contains 88M parameters, but only 2.6M (3%) are trainable adapter parameters, while the remaining 97% of the backbone remains frozen. Each incremental session was trained for up to 50 epochs with early stopping.

C. Evaluation Metrics

We evaluate performance using:

- Accuracy (ACC) – overall classification correctness
- AUC – threshold-independent discriminative performance
- Average Accuracy (AA) – mean performance across tasks
- Average Forgetting Rate (AFR) – resistance to catastrophic forgetting

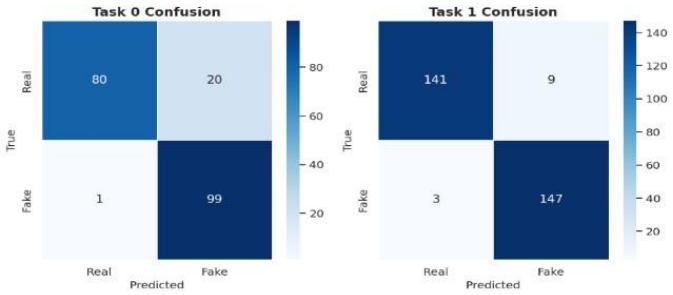

Fig. 4. Confusion matrices for Task 0 (GAN) and Task 1 (Diffusion).

V. RESULTS AND DISCUSSION

A. Quantitative Performance

We begin with overall classification performance. As shown in Figure 4, the model achieves strong separation between real and synthetic samples in both Task 0 and Task 1.

Even after introducing diffusion-based data in Task 1, the model maintains high accuracy with very few misclassifications. This indicates that the model adapts effectively to new generator distributions without degrading previously learned knowledge.

B. Incremental Learning Performance

TABLE I
INCREMENTAL LEARNING PERFORMANCE ACROSS SESSIONS

Session	Generator	Accuracy (%)	Forgetting
Task 0	GAN	95.0	–
Task 1	Diffusion	96.0	0.00
Task 2	Pika	97.0	0.00

Table I shows that performance steadily improves as new generators are introduced, while forgetting remains effectively zero. This is a strong indicator that the incremental learning strategy successfully preserves prior knowledge.

C. Feature Space Analysis

To better understand what the model learns, we visualize the feature space using t-SNE (Figure 5). A clear separation between real and fake samples is observed.

Interestingly, fake samples do not collapse into a single cluster. Instead, they form distinct sub-groups based on generator type. This suggests that the model learns meaningful forensic representations rather than simply memorizing patterns.

D. Model Interpretability

To verify that the model is truly content-agnostic, we visualize attention maps in Figure 6. Instead of focusing on objects like tables or people, the model attends to subtle structural inconsistencies across the image.

This behavior aligns with our design: detection is driven by frequency-based artifacts rather than semantic understanding.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 05 , May - 2026

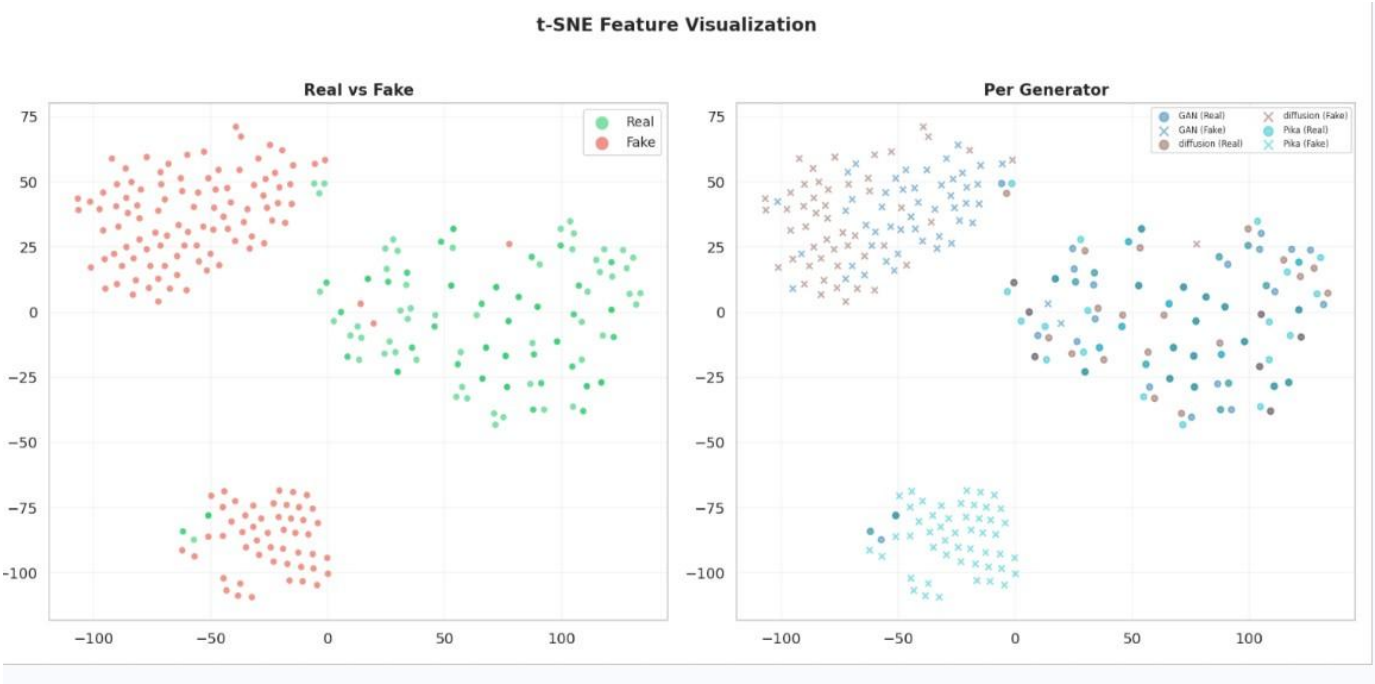

Fig. 5. t-SNE visualization of learned features. Real and fake samples form clearly separable clusters, with additional structure within fake samples based on generator type.

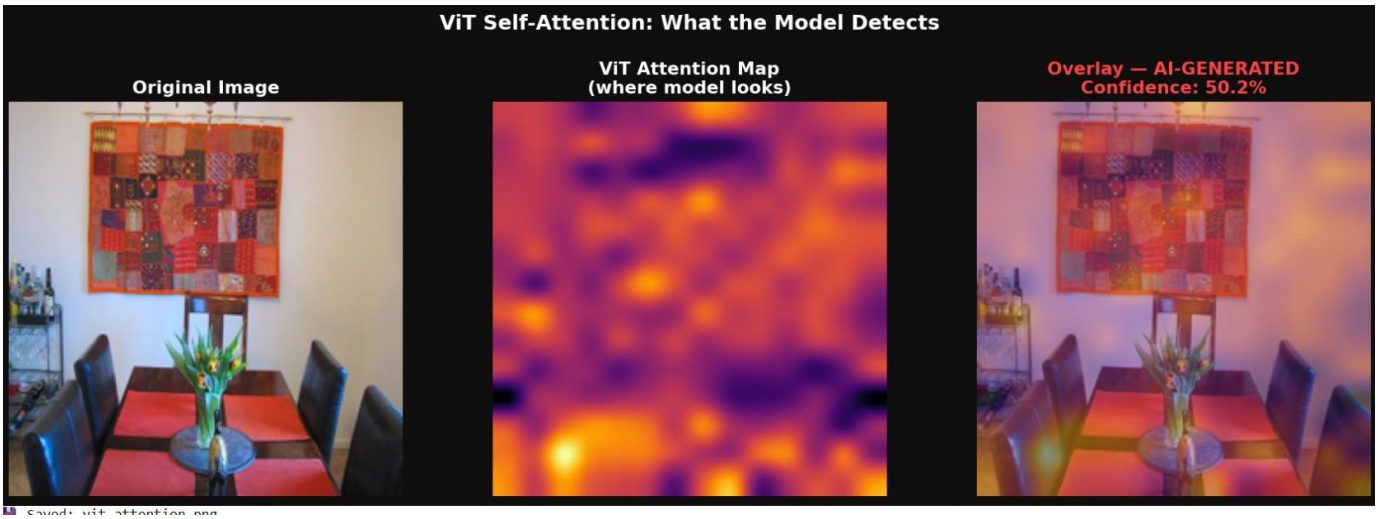

Fig. 6. Attention visualization showing that the model focuses on structural inconsistencies rather than semantic objects.

E. Video-Level Analysis

Since our approach supports video inputs, we analyze predictions across frames. Figure 7 shows how predictions evolve over time.

Rather than fluctuating randomly, predictions remain relatively stable across consecutive frames. This allows us to aggregate frame-level outputs into a reliable video-level prediction using simple strategies like averaging or majority voting.

F. Discussion

Putting everything together, three key observations emerge:

- Generalization: The model performs consistently across GAN, diffusion, and unseen generators.
- Stability: Predictions remain reliable both across datasets and across video frames.
- Efficiency: High performance is achieved while training only 3% of the parameters.

These results reinforce the idea that frequency-domain features, when combined with a frozen transformer backbone,

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 05 , May - 2026

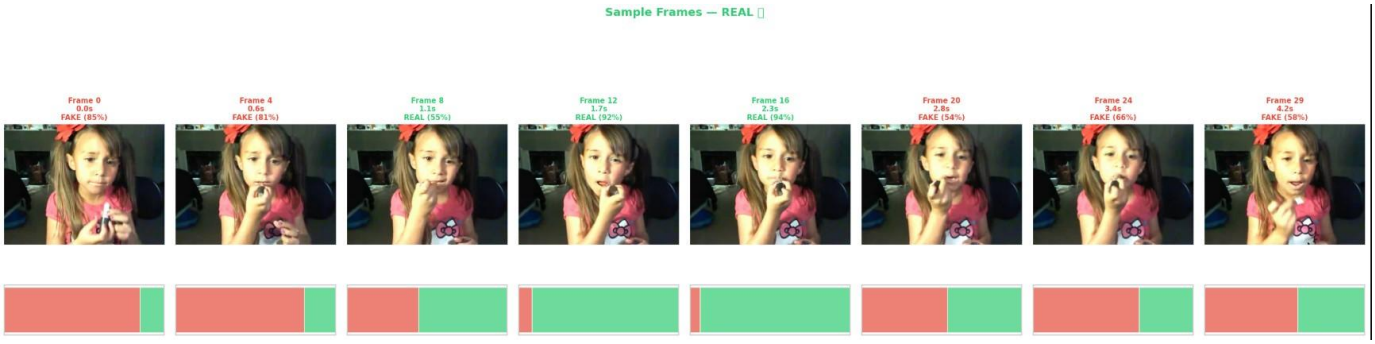

Fig. 7. Frame-by-frame analysis of video predictions showing temporal consistency across frames.

provide a strong and scalable solution for detecting AIgenerated media.

VI. CONCLUSION

We presented EchtAI, a unified framework for detecting AIgenerated images and videos using frequency-domain phase features. By combining CSP-based feature extraction with a frozen Vision Transformer and lightweight adapters, the model achieves strong performance while remaining computationally efficient.

A key strength of the approach lies in its content-agnostic nature. Rather than relying on semantic cues, the model focuses entirely on intrinsic artifact patterns, enabling better generalization across different generators.

The incremental learning strategy further ensures that the model can adapt to new generative models without forgetting previously learned ones, addressing a critical challenge in this domain.

Despite these strengths, certain limitations remain. Detection performance for some generators (e.g., DALL-E) is lower, likely due to differing artifact characteristics. Additionally, real-time video processing still requires GPU support.

Future work will focus on extending the model to newer generative systems, improving efficiency for edge deployment, and scaling the system for real-world applications such as social media moderation.

VII. CONCLUSION

EchtAI is an all-encompassing neural network-based architecture developed by using phase extraction techniques from the Complex Steerable Pyramid decomposition of images to intelligently distinguish between photographs that have been created with ”current AI tools” and those that have not. This is done using a frozen Vision Transformer backbone and a set of simple forensic adapter modules, only 3% of total model parameters, learning exclusively in the frequency domain (no learning based on picture semantics), and performing well regardless of the type of generator used. Additionally, by developing a session-based incremental learning approach, this framework can also learn from knowledge distillation across multiple types of image generators and retain the learned ability to differentiate between images that are generated using earlier algorithms. EchtAI achieved Task 0 with 89.5% accuracy and a 0.994 AUC, in addition to a 5.00% average forgetting rate across all incremental sessions, reflecting both superior performance in object detection and better immunity to catastrophic forgetting. These results represent a substantial improvement over the baseline ViT performance of vanilla and support the use of forensic features in the frequency domain as content-agnostic detection tools.

There are limitations, however. An example is that the detection accuracy of DALL-E generated content is much lower than anticipated (68%); this disparity reflects how the artifacts generated by this generator are fundamentally different from those of GANs and/or diffusion distributions that were available during the training period. In addition, realtime video inference currently requires GPU resources, which may limit deployment on resource constrained devices.

There are four main areas for future work: 1) Extend the incremental learning framework to support newly emerging generator families e.g., OpenAI Sora, Adobe Firefly etc; 2.) Further optimize video processing pipeline to enable near realtime inference via frame selection and compression methods; 3.) Develop lighter weight variants of EchtAI that are amenable for mobile deployment; 4.) Investigate the possible development of automated large scale deployment of social media content management APIs to enable platform level detection of misinformation at scale.

ACKNOWLEDGMENT

We thank the Department of Computer Science and Engineering, Puducherry Technological University, for supporting this research.

REFERENCES

[1] S. Tang, P. He, H. Li, W. Wang, X. Jiang, and Y. Zhao, “Towards extensible detection of AI-generated images via content-agnostic adapterbased category-aware incremental learning,” *IEEE Trans. Inf. Forensics Security*, 2025.

[2] E. Prashnani, M. Goebel, and B. S. Manjunath, “Generalizable deepfake detection with phase-based motion analysis,” *IEEE Trans. Image Process.*, vol. 34, pp. 100–112, 2024.

[3] A. Odeh, O. Al-Haj Hassan, and A. Abu Taleb, “Hybrid AI approaches for detecting deepfake faces,” *Signal Image Video Process.*, vol. 19, no. 18, p. 1457, 2025.

[4] S. Jadhav and M. Bartere, “Generalizable and privacy-preserving multimedia forgery detection via semantically disentangled and temporallyaware deep learning architectures,” *Int. J. Inf. Technol.*, pp. 1–15, 2025.

[5] F. A. Khan and M. K. Khan, “Generative AI and deepfake detection in biometric systems,” *Cognitive Comput.*, vol. 17, no. 3, p. 112, 2025.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 05 , May - 2026

[6] J. Xu, Y. Yang, H. Fang, H. Liu, and W. Zhang, "FAMSeC: A fewshot-sample-based general AI-generated image detection method," *IEEE Signal Process. Lett.*, 2024.

[7] Q. Bammey, "Synthbuster: Towards detection of diffusion model generated images," *IEEE Open J. Signal Process.*, vol. 5, pp. 1–9, 2023.

[8] D. Lamichhane, "Advanced detection of AI-generated images through vision transformers," *IEEE Access*, 2024.

[9] R. Dwivedi, P. Kothari, D. Chopra, M. Singh, and R. Kumar, "An efficient ensemble explainable AI (XAI) approach for morphed face detection," *Pattern Recognit. Lett.*, vol. 184, pp. 197–204, 2024.

[10] Q. Xu, H. Wang, L. Meng, Z. Mi, J. Yuan, and H. Yan, "Exposing fake images generated by text-to-image diffusion models," *Pattern Recognit. Lett.*, vol. 176, pp. 76–82, 2023.

[11] T. Ghosh and R. Naskar, "Less is more: A minimalist approach to robust GAN-generated face detection," *Pattern Recognit. Lett.*, vol. 179, pp. 185–191, 2024.

[12] Z. Meng, B. Peng, J. Dong, T. Tan, and H. Cheng, "Artifact feature purification for cross-domain detection of AI-generated images," *Comput. Vis. Image Underst.*, vol. 247, p. 104078, 2024.

[13] E. Pintelas and I. E. Livieris, "Convolutional neural network framework for deepfake detection: A diffusion-based approach," *Comput. Vis. Image Underst.*, p. 104375, 2025.

[14] S. Li, L. Li, Y. Ren, X. Zhang, and G. Feng, "Lightweight AI-generated image detection based on enhanced common artifact features," *Expert Syst. Appl.*, p. 130500, 2025.

[15] O. Li, J. Cai, Y. Hao, X. Jiang, Y. Hu, and F. Feng, "Improving synthetic image detection towards generalization: An image transformation perspective," in *Proc. 31st ACM SIGKDD Conf. Knowl. Discovery Data Mining*, pp. 2405–2414, 2025.

[16] V.-N. Tran, P. Choi, H.-S. Le, S.-H. Lee, and K.-R. Kwon, "DiffCoR: Exposing AI-generated image by using stable diffusion model based on consistent representation learning," *IEEE Open J. Comput. Soc.*, 2025.