DEEPFAKE FORENSICS USING VISION TRANSFORMER MODELS VIDEODr Brindha.S[1] , Ms. Abirami P[2], Mr. Kanishq C[3] , Mr. Nitheesh BR[4], Mr. Nitin UA[5], Mr. Revanth K[6], Mr. Sundar V[7]

[1]*Head of the Department, Computer Networking, PSG Polytechnic College, Coimbatore*

[2]*Lecturer, Computer Networking, PSG Polytechnic College, Coimbatore*

[34567]*Students, Computer Networking, PSG Polytechnic College, Coimbatore*

----------***----------

Abstract

Deepfake generation has grown at an unprecedented rate due to advancements in GANs, autoencoders, and diffusion models. These synthetic media pose significant threats to identity, privacy, and digital trust. This paper presents a comprehensive framework for Deepfake Forensics using Vision Transformer (ViT) models, integrating meta-learning–based generalization and 3D decomposition techniques. The hybrid approach combines EfficientNet-ViT backbone architectures with structured 3D face modeling, lighting estimation, and identity texture analysis to capture subtle inconsistencies in manipulated content. The resulting system significantly improves robustness against unseen manipulation methods, cross-domain shifts, and compression artifacts. This work synthesizes insights from MEViT and 3D decomposition-based forensic research to highlight an advanced and generalizable deepfake detection methodology.

Key Words: Deepfake Detection, Vision Transformer, Meta-Learning, 3D Decomposition, Face Forensics, Domain Generalization

1. INTRODUCTION

Deepfakes refer to AI-generated media where the facial identity, expression, or entire appearance of a person is digitally manipulated. While such technologies can be used for harmless entertainment, they pose growing concerns in misinformation, fraud, and cybercrime. Traditional detection methods rely on convolutional neural networks (CNNs), which suffer from poor generalization across datasets and fail when encountering novel manipulation techniques.

Recent research has explored Vision Transformers (ViTs) due to their global attention mechanisms, ability to capture long-range inconsistencies, and superior representation learning capabilities. Additionally, 3D decomposition–based forensic models evaluate physical face attributes—such as geometry, lighting, and identity texture—to reveal subtle inconsistencies hidden by generative models.

This paper integrates concepts from these two domains to develop a unified, generalizable deepfake forensic framework.

2. RELATED WORK

2.1 Deepfake Creation Techniques

Deepfakes are primarily generated using advanced deep learning techniques that focus on either identity swapping or expression manipulation. Identity-swapping methods replace the face of a target individual with another person's face while preserving head pose, facial expressions, and motion dynamics. Popular tools and frameworks such as DeepFake, FaceSwap, MegaFS, and FSGANv2 achieve this by learning facial representations and transferring identity-specific features between source and target faces.

Expression-manipulation techniques, such as Face2Face, modify facial expressions and lip movements without changing the underlying identity. These methods rely on facial landmark detection and 3D face modeling to realistically animate expressions. In addition, GAN-based models like StarGANv2 and SemanticStyleGAN introduce multi-domain image translation, enabling the manipulation of facial attributes such as age, gender, lighting, pose, and artistic style. The increasing realism produced by these models makes visual detection extremely challenging and motivates the need for advanced forensic analysis techniques.

2.2 Deepfake Detection Approaches

Early deepfake detection approaches relied on handcrafted features that captured visible inconsistencies in manipulated media. These included abnormal eye blinking patterns, unnatural head movements, inconsistent facial boundaries, and irregular lip synchronization. While effective against early-generation deepfakes, these methods failed as generation models became more sophisticated.

CNN-based detection approaches marked a significant improvement by automatically learning discriminative features from large datasets. These models focused on spatial artifacts, texture inconsistencies, and color mismatches. However, CNN-based detectors often suffer from domain-specific overfitting, meaning they perform well only on manipulation techniques seen during training. To address these limitations, recent research has explored frequency-domain analysis, self-supervised learning, and attention-based architectures, which aim to capture subtle and global forgery patterns that generalize better across datasets.

2.3 3D Decomposition and Physical Forensics

3D decomposition-based forensic techniques analyze deepfakes by examining the physical properties of the human face rather than relying solely on pixel-level artifacts. These methods decompose a facial image into its underlying components, including 3D facial geometry, lighting conditions, and texture information. Since most deepfake

generation models focus on visual realism, they often fail to maintain physically consistent lighting and geometry.

Fine-grained Morphing Networks (FM-Net) enhance shape reconstruction accuracy, allowing forensic systems to detect inconsistencies in facial contours, shadows, and reflectance. By analyzing lighting maps and texture residuals, 3D decomposition methods expose unnatural shading patterns and identity mismatches that are difficult for generative models to conceal. This physically grounded approach significantly improves robustness against high-quality deepfakes.

2.4 Meta-Learning and Vision Transformers

Vision Transformers (ViTs) have gained attention in deepfake detection due to their ability to model long-range dependencies using global self-attention mechanisms. Unlike CNNs, which focus on local features, ViTs divide images into patches and analyze relationships across the entire facial region. This enables the detection of subtle spatial anomalies distributed across the face.

Meta-learning techniques further enhance ViT-based models by improving generalization across unseen manipulation methods. Approaches such as MEViT simulate domain shifts during training by separating data into meta-training and meta-testing domains. Loss functions like Pair Discrimination Loss increase separability between real and fake samples, while Domain Adjustment Loss aligns feature distributions across domains. This combination allows the model to remain robust even when encountering novel deepfake generation techniques.

3. SYSTEM OVERVIEW

The proposed system integrates:
• Hybrid EfficientNet–Vision Transformer backbone
• Meta-learning for domain generalization
• 3D decomposition for geometry, lighting, and texture analysis
• Composition Search to identify optimal feature combinations

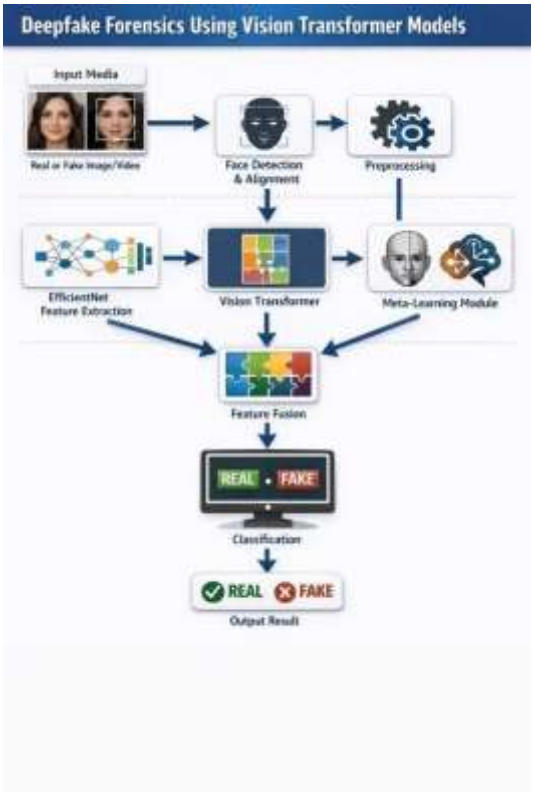

Fig. 1. Vision transformer model

This unified pipeline captures both learned and physically grounded deepfake inconsistencies.

4. WORKING PRINCIPLE

The working principle of the proposed Deepfake Forensics Using Vision Transformer Models is based on analyzing both visual and physical inconsistencies present in manipulated facial media. The system follows a structured workflow that integrates deep learning, transformer-based attention mechanisms, 3D facial analysis, and meta-learning for robust deepfake detection.

The process begins with the input image or video, which may contain either real or deepfake facial content. The system first performs face detection and alignment to accurately locate the facial region and standardize its orientation. This step ensures consistency across different inputs and improves detection accuracy.

Next, preprocessing is applied to the aligned face. This includes resizing the image to a fixed resolution, normalizing pixel values, across the entire facial region. Unlike convolutional networks, ViT effectively identifies subtle spatial inconsistencies that are distributed across different parts of the face.

In parallel, the system performs 3D face decomposition, where the facial image is analyzed based on physical attributes such as 3D shape, lighting estimation, and texture and identity information. This step exposes unnatural shading, geometric distortions, and and reducing noise. These operations enhance the quality of the input data and make it suitable for feature extraction.

The preprocessed face is then passed through an EfficientNet feature extraction module, which captures low-level and mid-level facial features such as edges, textures, and fine details. The extracted features are divided into fixed-size patches using patch embedding, preparing the data for transformer-based analysis.

These patches are processed by the Vision Transformer (ViT), which uses global self-attention to learn long-range dependencies

Fig. 2. Overview of Deepfake Generation and Detection Methods

texture mismatches that often occur in deepfake media.

A meta-learning module is incorporated to improve generalization across different deepfake generation techniques. By simulating domain shifts during training, the model learns domain-invariant features,

enabling it to detect previously unseen manipulation methods.

The features obtained from the Vision Transformer and the 3D decomposition

modules are then combined in the feature fusion stage, forming a comprehensive representation of both learned and physically grounded inconsistencies.

Finally, the fused feature vector is fed into the classification layer, which determines whether the input media is real or fake. The classification result is then displayed as the final output.

Thus, the proposed system effectively combines global attention, physical face modeling, and meta-learning to achieve accurate and robust deepfake detection.

5. IMPLEMENTATION

5.1 Vision Transformer Feature Learning

Image patches extracted from EfficientNet are fed into the ViT encoder, which applies multi-head self-attention to learn global contextual relationships. The CLS token summarizes the entire image and serves as the basis for binary classification (real vs fake).

5.2 Meta-Learning for Generalization

Meta-learning divides data into meta-train and meta-validation domains to simulate real-world domain shifts. Pair Discrimination Loss increases separability between real and fake samples, while Domain Adjustment Loss aligns feature distributions across domains.

5.3 3D Decomposition

Faces are decomposed into:
• 3D geometric shape
• Spherical harmonic lighting
• Common texture
• Identity texture

These components expose forgery clues that pixel-level detectors often miss, including unnatural shading and inconsistent microtextures.

5.4 Composition Search

Composition Search (CS) employs differentiable architecture search to determine the best combination of 3D features and transformer embeddings for classification, reducing the need for manual design.

5.5 Flow diagram explanation

The flow diagram represents the working process of the proposed Deepfake Forensics Using Vision Transformer Models. The system follows a sequential pipeline to analyze and classify media content as real or fake.

The process begins with the input image or video, which may contain either genuine or manipulated facial content. The input is first passed through a face detection and alignment stage, where the facial region is identified and aligned to a standard orientation for consistent analysis.

Next, preprocessing is performed to enhance the quality of the detected face. This includes resizing, normalization, and noise reduction, which help improve feature extraction accuracy.

The preprocessed face is then fed into the EfficientNet feature extraction module, which captures low-level spatial features such as texture, edges, and facial details. These features are divided into fixed-size patches through patch embedding, making them suitable for transformer-based processing.

The Vision Transformer (ViT) module analyzes these patches using global self-attention mechanisms. Unlike convolutional models, ViT captures long-range dependencies across the entire face, enabling the detection of subtle spatial inconsistencies introduced during deepfake generation.

Fig. 3. Deepfake Detection Architecture

Simultaneously, the system performs 3D face decomposition, where the face is analyzed based on physical properties such as 3D shape, lighting estimation, and texture and identity information. This step exposes unnatural shading, geometric distortions, and texture inconsistencies that are often present in deepfakes.

A meta-learning module is integrated to improve generalization across different manipulation techniques. By simulating domain shifts during training, this module enables the model to remain robust when detecting unseen or novel deepfake methods.

The features extracted from the Vision Transformer and 3D decomposition modules are combined in the feature fusion stage, creating a comprehensive representation of both visual and physical inconsistencies.

Finally, the fused features are passed to the classification layer, which determines whether the input media is real or fake. The output result is displayed, completing the deepfake detection process.

5. RESULTS

The proposed DeepFake Detection System was tested using both real and manipulated video samples. The system analyzes input media by extracting frames and applying a trained deep learning model to classify each frame as real or fake. The final result is generated based on the confidence scores and frame-level predictions.

Fake Video Detection Result

Fig. 4. Fake image 1

Fig. 5. Fake image 2

When a manipulated (deepfake) video was given as input, the system correctly detected the forgery.
The analysis result showed a 73.2% confidence score for Fake and 26.8% for Real.

The system extracted 12 frames from the video and analyzed each frame individually. All the sampled frames were classified as fake, indicating clear manipulation in the video content. The frame breakdown visually displayed the prediction for each frame, confirming the presence of synthetic facial modifications.

This demonstrates that the model can effectively identify facial inconsistencies and artifacts commonly present in deepfake videos.

Real Video Detection Result

Fig. 6. Real image 1

Fig. 7. Real image 2

When a genuine video sample was provided, the system successfully classified it as Real with 84.5% confidence.

The system again sampled 12 frames from the video and analyzed them individually. All the frames were predicted as real, confirming that the video did not contain any manipulated facial features.

The consistent classification across frames indicates the robustness of the model in distinguishing authentic videos from manipulated ones.

System Interface Output

Fig. 8. System interface output

The developed interface allows users to upload images or videos through a drag-and-drop feature. After processing, the system displays:

- Overall prediction result (Real or Fake)
- Confidence score for both classes
- Frame-by-frame analysis
- Visual indicators highlighting classification results

This user-friendly interface enables quick verification of media authenticity.

Performance Observation

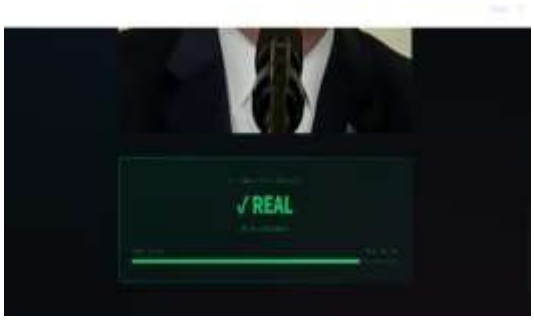

Fig. 9. Performance interface observation

The experimental results show that the proposed system is capable of accurately detecting deepfake content. By analyzing multiple frames and aggregating predictions, the model improves reliability and reduces the risk of misclassification.

Overall, the system demonstrates effective performance in identifying manipulated videos and distinguishing them from real media.

6. CONCLUSION

The integration of Vision Transformers, meta-learning, and 3D decomposition techniques offers a powerful solution for modern deepfake forensics. ViTs provide global spatial modeling, while physical 3D decomposition reveals lighting and geometric inconsistencies. Meta-learning enables strong domain generalization, ensuring stable detection even against unseen manipulations. This unified framework represents a major step toward reliable and future-proof deepfake detection systems.

7. ACKNOWLEDGMENT

We express our gratitude to all researchers whose work contributed to this combined forensic framework, and acknowledge the support of institutions and peers who provided insight and encouragement throughout this study.

8. REFERENCES

[1] Bennett, C. H., & Brassard, G. (1984). *Quantum cryptography: Public key distribution and coin tossing*. Proceedings of IEEE International Conference on Computers, Systems and Signal Processing, Bangalore, India.

[2] Heo, Y., Kim, Y., Kim, H., & Kim, J. (2023). *MEViT: Generalization of Deepfake Detection with Meta-Learning EfficientNet-Vision Transformer*. IEEE Access.

[3] Li, L., Bao, J., Yang, H., Chen, D., & Wen, F. (2020). *Face forgery detection by 3D decomposition and composition search*. IEEE Transactions on Pattern Analysis and Machine Intelligence.

[4] Li, T., & Zhang, H. (2020). *Deep learning applications in video security*. Journal of Information Security and Applications.

[5] Yuan, X., Li, Z., & Lyu, S. (2019). *3D-based face analysis for deepfake detection*. IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops.

[6] Afchar, D., Nozick, V., Yamagishi, J., & Echizen, I. (2018). *MesoNet: A compact facial video forgery detection network*. IEEE International Workshop on Information Forensics and Security.

[7] Rossler, A., Cozzolino, D., Verdoliva, L., Riess, C., Thies, J., & Nießner, M. (2019). *FaceForensics++: Learning to detect manipulated facial images*. IEEE International Conference on Computer Vision (ICCV).

[8] Li, Y., Chang, M., & Lyu, S. (2018). *In Ictu Oculi: Exposing AI generated fake face videos by detecting eye blinking*. IEEE International Workshop on Information Forensics and Security.

[9] Nguyen, H. H., Yamagishi, J., & Echizen, I. (2019). *Capsule-Forensics: Using capsule networks to detect forged images and videos*. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP).

[10] Chollet, F. (2017). *Xception: Deep learning with depthwise separable convolutions*. IEEE Conference on Computer Vision and Pattern Recognition (CVPR).

[11] Tan, M., & Le, Q. (2019). *EfficientNet: Rethinking model scaling for convolutional neural networks*. International Conference on Machine Learning (ICML).

[12] Dosovitskiy, A., et al. (2021). *An image is worth 16×16 words: Transformers for image recognition at scale*. International Conference on Learning Representations (ICLR).

[13] Verdoliva, L. (2020). *Media forensics and deepfake detection*. IEEE Journal of Selected Topics in Signal Processing.

[14] Mirsky, Y., & Lee, W. (2021). *The creation and detection of deepfakes: A survey*. ACM Computing Surveys.

[15] Zhou, P., Han, X., Morariu, V. I., & Davis, L. S. (2017). *Two-stream* ***neural networks for tampered face detection*****. IEEE Conference on Computer Vision and Pattern Recognition Workshops.**