IJARSCT

International Journal of Advanced Research in Science, Communication and Technology

International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal

ISSN: 2581-9429

Volume 5, Issue 2, November 2025

Impact Factor: 7.67

AI-Powered Detection of Deepfakes Using EfficientNet and Vision Transformer

Mr. Rushikesh Wagh' Mr. Sarthak Thorat, Mr. Rohan Pohakar,
Prof. S. Y. Mandlik, Dr. A. A. Khatri
Student, Computer Department
Jaihind College of Engineering Kuran, Pune, India
rushiwagh3829@gmail.com, Sarthakthorat5859@gmail.com, rohanpohakar77@gmail.com

Abstract: *Deepfakes are artificially tampered videos generated by GANs that often present realistic yet fake scenes[1]. Given the rapid advancement of deep generative models, both the sophistication and ease of access to manipulation technologies have increased significantly, thus increasing the difficulty of detection[1]. Most of the existing deepfake detection methods are built on top of CNNs, with very good performance in individual target datasets but poor generalization to unseen manipulation techniques due to overfitting[1]. In this paper, we develop a novel deepfake detection model, termed MEViT, that leverages a meta-learning framework to enhance generalization across unseen forgery types on top of EfficientNet Vision Transformer[1]. More concretely, MEViT introduces Pair-Discrimination Loss (PDL) that pushes away the feature representations of fake samples from the real ones at the embedding level[1]. On top of PDL, it introduces Domain Adjustment Loss (DAL) to reduce domain shifts across different manipulation methods by pushing all feature representations to a common embedding center[1]. Our extensive experiments on two widely benchmarked datasets, FaceForensics++ and CelebDF-v2, show that MEViT consistently outperforms the state- of-the-art approaches, especially in challenging cross-domain testing scenarios where generalization to unseen manipulations is most critical[1].*

Keywords*:* Deepfake detection, Meta-learning, Domain generalization, EfficientNet, Vision Transformer, GANs, Pair- Discrimination Loss, Domain Adjustment Loss

I. INTRODUCTION

Deepfakes, built upon GAN and VAE architectures, have posed serious threats related to misinformation, identity manipulation, and erosion of digital media. CNN-based detection methods achieve high accuracy for specific datasets but may fail when confronted with unseen manipulation techniques due to overfitting. In fact, this domain shift problem is fundamental, as different forgery methods will create varied data distributions that the existing models cannot generalize to without retraining.

Hybrid architectures that combine CNNs with Vision Transformers recently demonstrated their potential in capturing both local and global features with AUC scores of over 0.95 on benchmarks. These methods still are not well-equipped to deal with novel manipulation techniques and cannot adapt to new distributions without target domain data. Meta-learning allows models to simulate domain shifts during training, which can provide a pathway for generalization to unseen forgery types..

This paper proposes MEViT, a deepfake detection framework designed to generalize to unseen forgery types without accessing manipulated samples during training. In particular, MEViT integrates meta-learning with a hybrid EfficientNet Vision Transformer backbone. Two novel loss functions are further introduced to MEViT: Pair-Discrimination Loss (PDL) that aims to clearly separate real and fake feature representations in the embedding level and Domain Adjustment Loss (DAL) that reduces domain gaps by aligning the representations toward a common center. Extensive experiments on FaceForensics++ and CelebDF-v2 demonstrate superior performance in the cross-domain scenario where generalization to unseen manipulations is most critical..

DOI: 10.48175/IJARSCT-29737

II. PROBLEM STATEMENT

Since existing deepfake detection methods achieve high accuracy on their training datasets but fail when applied to unseen manipulation techniques, current architectures exhibit overfitting. CNN-based approaches tend to exploit domain-specific artifacts that do not generalize to new forgery methods, and as a result, these methods suffer severe performance degradation in cross-domain scenarios. The fundamental challenge of developing a deepfake detection system that generalizes to unseen forgery techniques without requiring target domain data during training remains an open problem..

III. OBJECTIVES

- To develop a deepfake detection framework based on meta-learning that achieves superior cross-domain generalization without access to the target domain data during training.
- To integrate EfficientNet and Vision Transformer architectures with meta-learning mechanisms that capture both local spatial features and global dependencies for robust forgery detection.
- To introduce new loss functions, namely Pair-Discrimination Loss (PDL) and Domain Adjustment Loss (DAL), for improving discriminative feature representations and reducing domain shifts among various manipulation methods.
- To validate the proposed MEViT model on several deepfake benchmarks including FaceForensics++, CelebDF-v2, and cross-domain evaluation settings, achieving superior performance compared to the current state-of-the-art approaches in deepfake detection.

IV. LITERATURE REVIEW

Deepfake technology has rapidly evolved with developments in generative models such as GANs and VAEs, enabling highly realistic manipulated videos that pose serious threats including misinformation and identity fraud. Early detection methods were mostly based on CNNs, focusing on a binary classification of real or fake content. However, most have been found to be suffering from overfitting due to dependence on dataset-specific artifacts and often failed against new or unseen manipulation methods.

Recent works have thus explored hybrid architectures, combining CNNs with Vision Transformers to effectively model both local spatial patterns and global contextual information, significantly boosting detection performance. Even with these advances, domain shift proves to be a serious challenge: deepfake generation methods can be so different that traditional models tend to fail on data distributions not seen before.

Meta-learning has recently been identified as a promising method to solve domain generalization by allowing models to simulate domain shifts at training time and adapt accordingly. The MEViT framework is no exception, which merges this concept with the hybrid EfficientNet-Vision Transformer backbone and introduces new loss functions that boost the discriminative ability in between real and fake features, further aligning domain representations. Extensive evaluations manifest that MEViT significantly outperforms the current state-of-the-art in cross-domain generalization, hence pushing toward reliable deepfake detection.

V. METHODOLOGY

Model Backbone: Combines EfficientNet and Vision Transformer to extract both local spatial features and global dependencies from input images.

Meta-Learning Framework: This employs a meta-learning setup that splits source domains into meta-train and meta-validation subsets in order to simulate domain shifts during training.

Pair-Discrimination Loss (PDL): It promotes the distinction between the embedding of fake and real samples by maximizing Euclidean distance at the feature level.

DAL: aligns feature representations of source domains to a common embedding center, reducing gaps across domains.

Training stages:

- Meta-Train: It optimizes model parameters using combined losses on meta-train domains.
- Meta-Test: Runs model on meta-validation domains; uses selected loss components.

- Meta-Optimization: Sums gradients from both meta-train and meta-validation, then updates model parameters.
- Transfer to Target Domain: This model generalizes to unseen manipulation methods without access to target domain data during training.
- Datasets Used: Validated on the datasets of FaceForensics++ and CelebDF-v2 for demonstrating strong cross- domain generalization.

VI. RESULT

- MEViT achieves state-of-the-art performance on cross-domain deepfake detection benchmarks.
- MEViT outperforms the baseline models consistently in both accuracy and AUC scores on the FaceForensics++ and CelebDF-v2 datasets.
- The model exhibits strong generalization capabilities in dealing with unseen forgery methods without target domain training data.
- MEViT learns discriminative and domain-invariant features with meta-learning and two loss functions: Pair-Discrimination Loss and Domain Adjustment Loss.
- These results confirm that MEViT effectively mitigates the domain shift issues which were common in previous detection approaches.
- MEViT's robustness across manipulation types and compression levels proves practical in real-world deployment.
- Ablation studies in the paper confirm the contribution of each component, especially the meta-learning framework and novel loss terms.
- The model balances computational efficiency with detection effectiveness using an EfficientNet B4 backbone combined with Vision Transformers.
- In particular, MEViT pushes the frontier of reliable deepfake detection to address cross-domain challenges by devising a novel meta-learning architecture and loss design.

The results suggest strong potential impact for scalable and generalized digital media verification systems.

VII. SYSTEM ARCHITECTURE

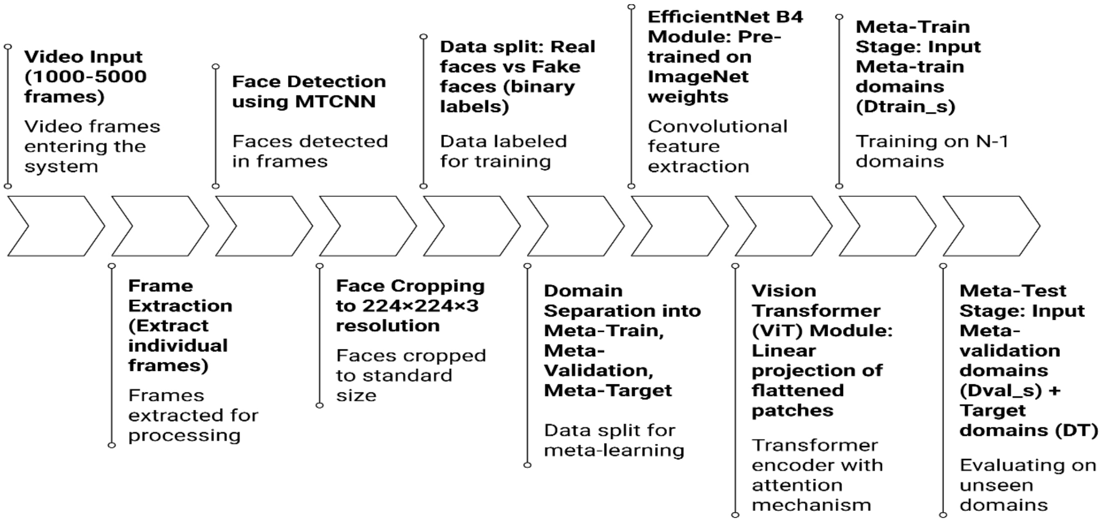
Video Input (1000-5000 frames)
Video frames entering the system
Frame Extraction (Extract individual frames)
Frames extracted for processing
Face Detection using MTCNN
Faces detected in frames
Face Cropping to 224×224×3 resolution
Faces cropped to standard size
Data split: Real faces vs Fake faces (binary labels)
Data labeled for training
Domain Separation into Meta-Train, Meta-Validation, Meta-Target
Data split for meta-learning
EfficientNet B4 Module: Pre-trained on ImageNet weights
Convolutional feature extraction
Vision Transformer (ViT) Module: Linear projection of flattened patches
Transformer encoder with attention mechanism
Meta-Train Stage: Input Meta-train domains (Dtrain_s)
Training on N-1 domains
Meta-Test Stage: Input Meta-validation domains (Dval_s) + Target domains (DT)
Evaluating on unseen domains

Fig.MEViT System Architecture for Deepfake detection.

1 Input & Preprocessing:

- Video Input & Frame Extraction: Process raw video into manageable frames for analysis.
- Data Split: Splitting into training, validation, and testing domains simulates real-world distribution shifts and limits overfitting.

- Data Split: Splitting into training, validation, and testing domains simulates real-world distribution shifts and limits overfitting.

2. Feature Extraction Backbone:
- EfficientNet: A CNN architecture optimized for balancing accuracy and computational efficiency; extracts local and detailed spatial features from images.
- Vision Transformer: This model uses self-attention to model global dependencies across image patches, adding holistic context awareness to local features learned by CNNs.

3. Meta-Learning Framework:
- Meta-Train and Meta-Test Stages: Simulate domain shifts by alternatively training and validating on different subsets of domains to improve cross-domain generalization.
- Meta-Optimization: jointly considers gradients from meta-train and meta-validation steps to update model parameters for robustness across domains.

4. Loss Functions:
- Pair-Discrimination Loss (PDL): maximizes the distance between embeddings of real and fake samples in order to increase discriminative power in feature representation.
- Domain Adjustment Loss: Reduces the disparity between each domain-specific feature center and the overall centroid to dynamically reduce the effects of the domain gap.
- BCE - Binary Cross-Entropy Loss: A standard classification loss for binary real/fake labeling.

5. Classification and Output:
- CLS Token Embedding: The special embedding for a classification token, used by ViT, summarizes feature information across the entire image for final binary classification.
- Sigmoid Activation: This processes the logit outputs to probability scores for real or fake detection.

VIII. BENEFITS TO SOCIETY

Due to robust detection, the MEViT architecture enhances the reliability and safety of digital media. Generalizing to unseen forgery techniques helps in preventing misinformation, identity theft, and erosion of trust in online content, while maintaining safety in communication and content verification across social media, news, and biometric systems. MEViT contributes to the protection of democratic processes, personal privacy, and intellectual property by reducing manipulated media. Because it does not have to be retrained with every new generation technology for fake media, it is practical for real-world deployment, ensuring ongoing protection against evolving digital threats.

IX. CONCLUSION

The MEViT framework presents a novel and effective approach to deepfake detection by combining meta-learning with a hybrid EfficientNet Vision Transformer backbone. It successfully addresses the critical challenge of generalizing detection across unseen forgery types without requiring access to manipulated samples during training. By integrating Pair-Discrimination Loss and Domain Adjustment Loss, the model learns discriminative and domain-invariant feature representations, achieving state-of-the-art performance on multiple benchmark datasets. This significantly advances the robustness and practical applicability of deepfake detection systems, helping to combat the growing threat of digital media manipulation.

X. FUTURE SCOPE

1. Extend MEViT for multimodal deepfake detection by incorporating audio and video cues to improve verification.
2. Develop continuous learning capabilities to adapt in real time to emerging forgery techniques without full retraining.

DOI: 10.48175/IJARSCT-29737

3. Optimize the model for deployment on edge devices to enable real-time detection on mobile and social media platforms.
4. Implementing explainability and interpretability methods to highlight regions triggering fake predictions will be important for user trust.
5. Improve adversarial robustness to prevent the detection model from being evaded or misled.
6. Federated learning frameworks that allow privacy-preserving training without data sharing across distributed data sources.
7. Explore lightweight model variants which decrease computational complexity but preserve detection accuracy.

REFERENCES

[1] V. N. Tran, H. S. Le, P. Choi, S. H. Lee, and K. R. Kwon, MEViT: Generalization of deepfake detection with meta-learning EfficientNet Vision Transformer, IEEE Open J. Comput. Soc. 6, 1 (2025).

[2] D. Coccomini, N. Messina, C. Gennaro, and F. Falchi, Combining EfficientNet and Vision Transformers for video deepfake detection, in Proceedings of the International Conference on Pattern Recognition (2022).

[3] A. H. Soudy, O. Sayed, H. Tag-Elser, R. Ragab, S. Mohsen, T. Mostafa, A. A. Abohany, and S. O. Slim, Deepfake detection using convolutional vision transformers and convolutional neural networks, Sci. Rep. 14, 18556 (2024).

[4] Y. J. Heo, Y. J. Choi, B. G. Kim, and Y. W. Lee, Deepfake detection scheme based on Vision Transformer and distillation, in Proceedings of IEEE International Conference on Image Processing (2021).

[5] A. Ashraf Bekheet, A. S. Ghoneim, and G. Khoriba, Unmasking the digital deception: A comprehensive survey of large vision models for deepfake detection, Informatics Bull.7, 2 (2025).

[6] W. Wodjao and S. Atnafu, Deepfake video detection using convolutional vision transformer, arXiv preprint arXiv:2102.11126 (2021).

[7] D. Cozzolino, A. Rössler, J. Thies, M. Nießner, and L. Verdoliva, Detecting face synthesis using convolutional neural networks and image quality assessment, in 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (IEEE, 2018), pp. 2537–2541.

[8] Y. Li, X. Yang, P. Sun, H. Qi, and S. Lyu, Celeb-DF: A large-scale challenging dataset for deepfake forensics, in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2020), pp. 3207– 3216.

[9] R. Sabir, E. Cheng, A. Jaiswal, W. AbdAlmageed, I. Masi, and P. Natarajan, Recurrent convolutional strategies for face manipulation detection in videos, Interfaces (GUI) 3, 80 (2019).

[10] L. Li, J. Bao, H. Yang, D. Chen, and B. Guo, Advancing high-fidelity identity-preserving face swapping, in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2022), pp. 5074– 5084.