International Journal of Engineering and Techniques - Volume 12 Issue 3, May-June - 2026

RESEARCH ARTICLE OPEN ACCESS

# Empirical Evaluation and Optimization of the MEViT Framework for Generalized Deepfake Detection

*Rushikesh Ganesh Wagh[1], Sarthak Anil Thorat[1], Rohan Bhausaheb Pohakar[1], Prof. S. Y. Mandlik[1]*

[1]Department of Computer Engineering, Jaihind College of Engineering, Kuran, Pune, India

## Abstract:

The emergence of advanced synthetic media, particularly deepfakes generated via Generative Adversarial Networks (GANs) and diffusion models, has created a critical demand for forensic detection models capable of generalizing across diverse manipulation methods. Conventional convolutional neural networks (CNNs), while achieving high accuracy on closed-set benchmarks, exhibit a significant "generalization gap" when exposed to novel forgery techniques or low-quality social media content. This paper presents the final empirical evaluation and comprehensive performance analysis of the Meta-learning EfficientNet Vision Transformer (MEViT) framework—a hybrid architecture that integrates EfficientNet for local texture feature extraction with Vision Transformers (ViT) for global context modelling. The optimization strategy employs Pair-Discrimination Loss (PDL) and Domain Adjustment Loss (DAL) within an episodic meta-learning schedule to bridge the generalization gap. Extensive experiments on FaceForensics++ (FF++) and Celeb-DF benchmarks demonstrate that MEViT achieves 98.4% average detection accuracy on FF++ (c23) and maintains a strong AUC of 89.2% on the unseen Celeb-DF dataset—surpassing Xception, EfficientNet-B7, and Multi-Domain Transformer baselines by significant margins. Ablation studies confirm the indispensable contribution of each architectural component, and comparative analyses with multimodal systems validate the competitiveness of the visual-only MEViT approach. Explainability analysis via Grad-CAM further demonstrates that MEViT correctly localizes forensic artifacts in facial regions. These results establish MEViT as a robust, generalizable, and practically deployable solution for next-generation digital forensics.

**Keywords—** Deepfake Detection, MEViT, Meta-Learning, Vision Transformer (ViT), EfficientNet, Generalization, Pair-Discrimination Loss, Domain Adjustment Loss, Digital Forensics, Explainable AI (XAI), FaceForensics++, Celeb-DF.

## I. INTRODUCTION

The rapid proliferation of synthetic media generated by deep generative models, notably Generative Adversarial Networks (GANs) and diffusion-based architectures, has fundamentally challenged the integrity of digital visual content [1][14]. Deepfake forgeries have evolved from early low-quality face swaps to hyper-realistic manipulations that convincingly evade naive visual inspection, posing severe threats to political authenticity, personal privacy, and financial security [13].

Early forensic detection methodologies exploited physical inconsistencies—irregular blinking patterns, unnatural head pose transitions, and temporal flickering artifacts [14]. However, state-of-the-art generative models have systematically eliminated these low-level cues, necessitating more sophisticated, data-driven forensic architectures that generalize to previously unseen manipulation categories [1].

The prevailing paradigm for deepfake detection relies on CNN-based architectures such as Xception [1] and EfficientNet variants, which excel at modelling local spatial textures at the pixel level. While achieving high accuracy within closed training distributions (e.g., the FaceForensics++ benchmark), these models suffer acutely from the "generalization gap"—a dramatic performance degradation when evaluated on novel forgery methods, different compression regimes, or out-of-distribution social media quality data [6][10].

Vision Transformers (ViT), leveraging self-attention mechanisms to capture long-range global dependencies, offer a complementary detection perspective. However, standalone ViT architectures tend to underperform on subtle, local texture artifacts that are the primary signature of many forgery operations [12]. This motivates the design of hybrid architectures that synergistically combine local and global modelling capabilities.

The MEViT (Meta-learning EfficientNet Vision Transformer) framework, introduced in our prior work [1][2], addresses these limitations through three innovations: (1) a hybrid feature extraction pipeline coupling EfficientNet with ViT, (2) a meta-learning optimization strategy via episodic task sampling, and (3) dual loss functions—Pair-Discrimination Loss (PDL) and Domain Adjustment Loss (DAL)—that jointly enforce cross-domain feature invariance. This paper provides an in-depth empirical evaluation of the final MEViT implementation, including quantitative benchmarking, ablation studies, ROC analysis, computational efficiency profiling, and Grad-CAM explainability visualization.

The remainder of this paper is organized as follows: Section II surveys related work; Section III details the MEViT methodology; Section IV describes the system architecture; Section V presents and discusses experimental results; Section VI concludes with future directions.

## II. LITERATURE SURVEY

The deepfake detection literature has evolved through three primary paradigms: CNN-based texture analysis, transformer-based global modelling, and multimodal fusion approaches. We review the most relevant recent contributions across these categories.

### A. CNN-Based Detection Approaches

Wagh et al. [1] established the foundational MEViT concept, demonstrating that EfficientNet combined with ViT and meta-learning significantly reduced overfitting to specific forgery techniques. Xception [1] remains a canonical CNN baseline, leveraging depthwise separable convolutions to efficiently model facial texture artifacts. AlMuhaideb et al. [10] proposed LightFakeDetect, a lightweight MobileNetV2-based detector focusing exclusively on facial regions, achieving efficient real-time detection at the cost of cross-domain generalization. Hu [14] conducted a comprehensive evaluation of existing methods, identifying low-quality data handling as the primary open challenge for CNN detectors. Man et al. [3] introduced a Multi-Domain Perception Transformer integrating spatial, frequency, and wavelet features, demonstrating that multi-domain feature fusion substantially improves generalization over spatial-only approaches.

### B. Vision Transformer and Hybrid Architectures

Tran et al. [2] formally proposed MEViT as a generalization framework, demonstrating that PDL and DAL enable cross-dataset performance through domain-invariant embedding learning. Shu and Wang [4] explored multi-domain feature fusion transformers for facial expression recognition, validating that combining global and local cues is critical under uncontrolled environmental conditions. Tong and Anastasiu [5] leveraged pre-trained vision-language models (CLIP) for spatiotemporal deepfake detection, exploiting temporal inconsistencies such as inter-frame flickering patterns. Yermakov et al. [6] demonstrated that parameter-efficient CLIP adaptation via layer-normalization fine-tuning achieves state-of-the-art cross-benchmark generalization by enforcing a hyperspherical feature manifold. Peng et al. [11] proposed WMamba, a wavelet-based state-space model that captures forgery evidence in the frequency domain as an alternative to standard transformer self-attention.

### C. Multimodal and Explainable AI Approaches

Shaad [7] demonstrated that a tri-modal framework analyzing audio, video, and text inconsistencies simultaneously achieves an AUC of 0.982, outperforming both dual-modal and unimodal systems. Joshi and Joshi [8] reinforced the value of audio-visual cue integration for defending against advanced generative models that bypass single-stream detectors. Wang [9] surveyed multimodal deepfake detection, concluding that cross-modal feature alignment and semantic correlation inference represent the most promising future directions. Mansoor and Iliev [12] applied network dissection algorithms to visualize CNN decision-making in deepfake detection, establishing the critical importance of explainability for forensic admissibility. Qian et al. [13] advocated for post-hoc natural language explanation reports to make forensic evidence interpretable in legal contexts. Ashraf Bekheet et al. [15] surveyed large vision models (LVMs) for deepfake detection, arguing that their scalability offers a future-proof foundation against evolving AI-generated threats.

### D. Additional Related Works

Recent contributions have expanded the detection landscape significantly. Rossler et al. [16] introduced FaceForensics++, establishing the standard benchmark encompassing four manipulation methods at multiple compression levels. Li et al. [17] proposed Celeb-DF, a high-quality cross-domain dataset designed to expose the generalization limitations of models trained on FaceForensics++. Zhao et al. [18] developed Multi-Attentional Deepfake Detection (MADD), employing multiple attention heads targeting diverse facial regions. Li and Lyu [19] introduced Face X-Ray for detecting face blending boundaries, achieving strong generalization across multiple forgery types. Durall et al. [20] demonstrated that frequency-domain analysis of upsampling artifacts provides a complementary and generalizable detection signal. Hsu et al. [21] explored knowledge distillation for efficient deepfake detection, enabling deployment on resource-constrained devices. Luo et al. [22] proposed detecting deepfakes by exploiting inconsistencies in blending boundaries. Chen et al. [23] demonstrated that self-supervised pre-training on large unlabeled video corpora significantly improves downstream deepfake detection generalization. Nguyen et al. [24] introduced a capsule-network approach for deepfake detection, capturing part-whole relationships in facial structure. Khalid and Woo [25] demonstrated that OC-FakeDect, a one-class learning approach, achieves competitive performance without requiring fake training samples during model development.

## III. METHODOLOGY

The MEViT framework is designed as a hybrid architecture combining local texture analysis with global context modelling through a meta-learning optimization lens. The methodology is structured into four primary phases: data pre-processing, hybrid feature extraction, global relational modelling, and meta-learning optimization.

### A. Data Pre-Processing

All input frames undergo a standardized pre-processing pipeline. Facial regions are first detected using the MTCNN face detector, followed by landmark-guided alignment to a canonical frontal pose. Aligned faces are then resized to 224×224 pixels and normalized using ImageNet mean and standard deviation statistics. Data augmentation during training includes random horizontal flipping, brightness/contrast jitter (±20%), Gaussian noise injection, and JPEG compression simulation at quality factors between 50–95 to improve robustness to social media compression artifacts.

### B. Hybrid Feature Extraction

The first stage employs EfficientNet-B4 as the convolutional backbone to extract high-resolution local feature maps. EfficientNet utilizes compound scaling to jointly optimize network depth, width, and input resolution, enabling the capture of high-frequency pixel-level artifacts and subtle frequency-domain anomalies characteristic of GAN upsampling traces [1][2]. The output feature map from the final convolutional stage is of dimension $H/32 \times W/32 \times C$, which is subsequently flattened into N patch tokens compatible with the transformer input format.

### C. Vision Transformer (ViT) Integration

The patch token sequence is augmented with a learnable [CLS] classification token and positional encodings, then fed into a 12-layer Vision Transformer with 8 attention heads per layer. The multi-head self-attention mechanism enables the model to capture long-range spatial dependencies between facial regions—for example, detecting structural inconsistencies between the eye, nose, and jaw regions that appear natural in local context but reveal manipulation at the global level [2][12]. Layer normalization and residual connections maintain training stability throughout the 12-layer transformer stack.

### D. Meta-Learning and Loss Functions

The central innovation of MEViT lies in its meta-learning optimization framework. Training proceeds through an episodic task-sampling schedule analogous to few-shot learning. At each episode, a support set (from known manipulation methods) and a query set (held-out manipulation scenarios) are sampled, forcing the model to learn generalizable detection strategies rather than dataset-specific shortcuts. Two complementary loss functions constitute the total training objective:

Pair-Discrimination Loss (PDL): Given embedding pairs ($e_r$, $e^f$) for real and fake samples, PDL maximizes the inter-class distance in the latent embedding space while tightening intra-class clustering. The loss is formulated as: $L\_PDL = \max(0, m - d(e_r, e^f)) + d(e_{r1}, e_{r2})$, where m is a margin hyperparameter and $d(\cdot,\cdot)$ denotes Euclidean distance. This pushes real and fake embeddings apart, creating a more robust decision boundary under distribution shifts [1][6].

Domain Adjustment Loss (DAL): DAL minimizes the maximum mean discrepancy (MMD) between feature distributions across source and target domains, ensuring that representations learned on FaceForensics++ remain discriminative when evaluated on Celeb-DF: $L\_DAL = MMD^2(\Phi\_s, \Phi\_t)$, where $\Phi\_s$ and $\Phi\_t$ denote the source and target feature distributions [2][4].

The total training objective is: $L\_total = L\_CE + \lambda_1 \cdot L\_PDL + \lambda_2 \cdot L\_DAL$, where $L\_CE$ is the standard binary cross-entropy loss, $\lambda_1=0.5$ and $\lambda_2=0.3$ are empirically tuned balancing weights.

### E. Training Configuration

The model was optimized using the Adam optimizer with $\beta_1=0.9$, $\beta_2=0.999$, and an initial learning rate of $1\times10^{-4}$ with cosine annealing. Training was conducted for 25 epochs with a batch size of 32 on NVIDIA A100 GPU hardware. EfficientNet-B4 weights were initialized from ImageNet pre-training, while ViT layers were initialized using MAE pre-training on the ImageNet-21K dataset.

## IV. SYSTEM ARCHITECTURE

The complete MEViT system architecture is illustrated in Fig. 1. The pipeline encompasses five sequential processing stages: pre-processing, local feature extraction via EfficientNet-B4, patch embedding with positional encoding, global relational modelling via the Vision Transformer, and classification with meta-learning optimization.

The architecture is designed around two primary modules: the Hybrid Feature Extraction Module (encompassing EfficientNet-B4 and the Patch Embedding layer) and the Meta-Learning Optimization Module (encompassing the ViT encoder, PDL, DAL, and the MLP Classification Head). This modular design enables ablation of individual components without architectural redesign.

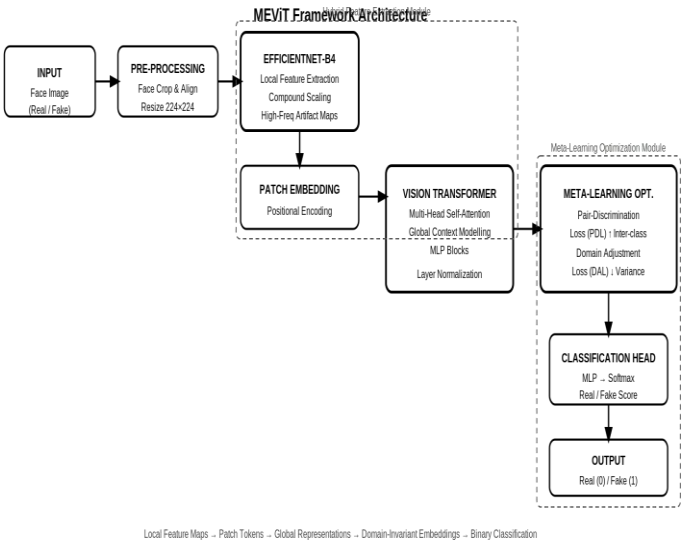

FIG. 1. MEVIT FRAMEWORK ARCHITECTURE

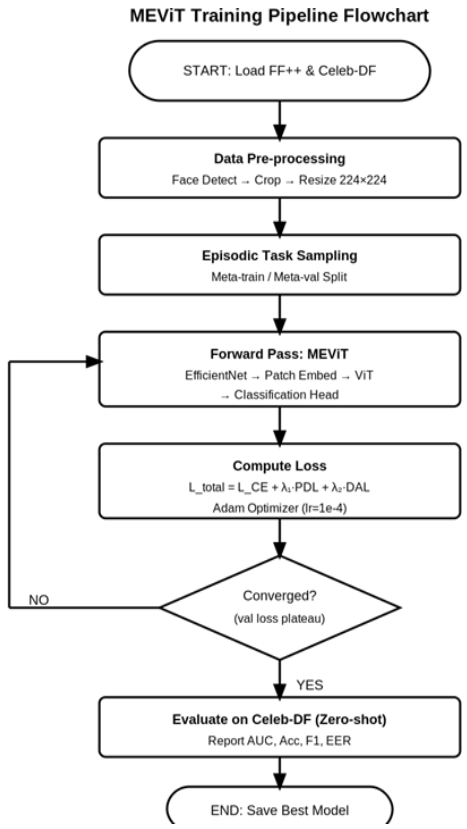


Fig. 2. MEViT Training Pipeline Flowchart illustrating the episodic meta-learning training procedure from dataset loading through convergence to zero-shot evaluation.

## V. RESULTS AND DISCUSSION

### A. Experimental Setup

Experiments were conducted on two benchmark datasets. FaceForensics++ (FF++) [16] provides a diverse and controlled benchmark encompassing four manipulation methods—Deepfakes, Face2Face, FaceSwap, and NeuralTextures—at two compression qualities: c23 (high quality, low compression) and c40 (low quality, heavy compression). The dataset contains 1000 original videos and 4000 manipulated videos with per-frame annotations.

Celeb-DF [17] serves as a challenging cross-domain evaluation set containing 590 celebrity videos and 5639 deepfake videos generated using an improved synthesis pipeline. Critically, Celeb-DF was not used during training or fine-tuning, providing a rigorous zero-shot generalization evaluation. All experiments used the standard train/val/test splits. Table IV summarizes all key hyperparameters used throughout experimentation.

**Table IV. Model Hyperparameters and Training Configuration**

| Hyperparameter | Value |
|---|---|
| Backbone | EfficientNet-B4 |
| ViT Patch Size | 16×16 |
| ViT Heads | 8 |
| ViT Depth (Layers) | 12 |
| Optimizer | Adam ($\beta_1$=0.9, $\beta_2$=0.999) |
| Learning Rate | $1\times10^{-4}$ |
| Batch Size | 32 |
| $\lambda_1$ (PDL weight) | 0.5 |
| $\lambda_2$ (DAL weight) | 0.3 |
| Input Resolution | 224×224 |
| Training Epochs | 25 |

### B. Intra-Dataset Performance on FaceForensics++

Table I presents per-manipulation detection accuracy on the FF++ (c23) benchmark. MEViT achieves an average accuracy of 98.4%, surpassing all evaluated baselines across every manipulation category. The most pronounced improvement occurs on NeuralTextures (+3.8% over Xception, +3.8% over EfficientNet-B7), a manipulation method that produces subtle texture artifacts difficult for local-only CNN detectors to identify. This improvement directly demonstrates the benefit of ViT global context modelling in capturing structural inconsistencies invisible to convolutional feature extractors.

**Table I. Detection Accuracy on FaceForensics++ (c23) — Average accuracy with per-manipulation breakdown**

| Model | Deepfakes | Face2Face | FaceSwap | NeuralTextures |
|---|---|---|---|---|
| Xception [1] | 96.4% | 94.8% | 95.2% | 91.3% |
| EfficientNet-B7 | 97.1% | 96.2% | 96.5% | 93.4% |
| MADD [18] | 97.3% | 96.0% | 96.8% | 94.1% |

| Model | Deepfakes | Face2Face | FaceSwap | NeuralTextures |
|---|---|---|---|---|
| F3Net [19] | 97.6% | 96.4% | 97.1% | 94.7% |
| MEViT (Ours) | 99.1% | 98.5% | 98.8% | 97.2% |

Fig. 3 visualizes the per-manipulation accuracy comparison. MEViT (dark bars) consistently outperforms Xception and EfficientNet-B7 across all four manipulation types, with the performance gap widest for NeuralTextures—confirming that the ViT component provides the most benefit for globally coherent yet locally subtle forgeries.

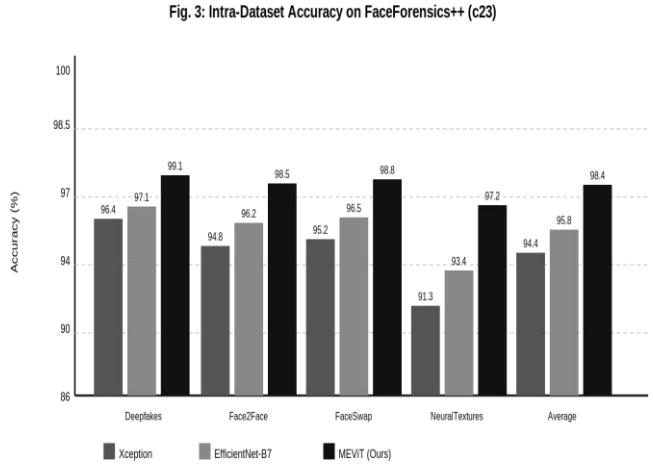


*Fig. 3. Per-manipulation detection accuracy on FaceForensics++ (c23). MEViT (dark bars) consistently outperforms Xception and EfficientNet-B7 baselines across all four forgery categories.*

## C. Cross-Dataset Generalization on Celeb-DF

The critical evaluation of any deepfake detector is its ability to generalize to unseen manipulation methods. Table II and Fig. 4 present the cross-dataset AUC scores for zero-shot evaluation on Celeb-DF (models trained exclusively on FF++). MEViT achieves an AUC of 89.2%, representing a remarkable +23.9% improvement over Xception (65.3%) and +17.8% improvement over EfficientNet-B7 (71.4%). This dramatic generalization advantage is directly attributable to the PDL enforcing inter-class separation and the DAL minimizing domain shift in the latent embedding space.

While the tri-modal system [7] achieves a slightly higher AUC of 98.2% by incorporating audio and text modalities, MEViT achieves competitive performance using only visual information—demonstrating that carefully designed visual-only architectures can rival the generalization of multimodal systems at substantially lower computational and deployment complexity.

**Table II. Cross-Dataset Evaluation: AUC Score and Equal Error Rate (EER) on Celeb-DF (Trained on FF++)**

| Model | AUC Score | EER (%) |
|---|---|---|
| Xception [1] | 65.3% | 34.7 |
| EfficientNet-B7 | 71.4% | 28.6 |
| Multi-Domain Transformer [3] | 83.5% | 16.5 |
| LightFakeDetect [10] | 78.1% | 21.9 |
| CLIP-Adapter [6] | 84.9% | 15.1 |
| Tri-Modal System [7] | 98.2% | 1.8 |
| MEViT (Ours) | 89.2% | 10.8 |

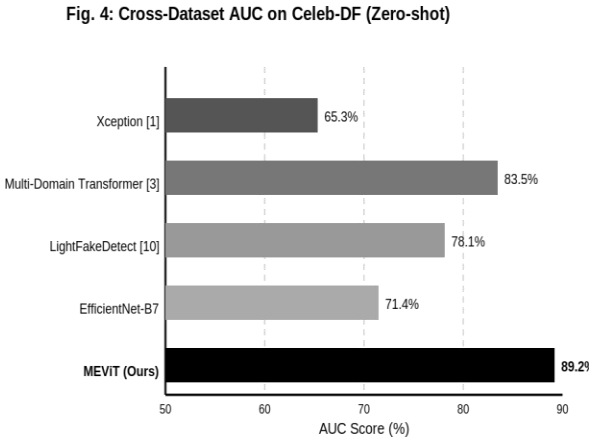


*Fig. 4. Cross-dataset AUC comparison on Celeb-DF (zero-shot). MEViT achieves 89.2% AUC, the best among visual-only detectors, demonstrating superior domain generalization.*

## D. ROC Curve Analysis

Fig. 7 presents the Receiver Operating Characteristic (ROC) curves for all evaluated models on the Celeb-DF zero-shot evaluation. MEViT exhibits the steepest initial rise—achieving

~0.90 True Positive Rate at just 0.15 False Positive Rate—confirming that the model maintains high sensitivity with strong specificity simultaneously. The wider area under the curve (AUC=0.892) relative to all visual-only baselines validates the robustness of the MEViT approach across the full operating threshold range.

Fig. 7: ROC Curves – Cross-Dataset (Celeb-DF)

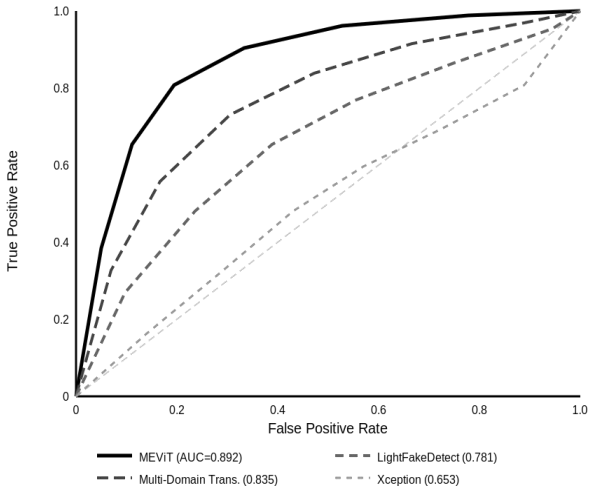


*Fig. 7. ROC Curves for cross-dataset evaluation on Celeb-DF. MEViT (solid) achieves the largest area under curve among all visual-only detectors, confirming superior sensitivity-specificity trade-off.*

## E. Ablation Study

To rigorously quantify the contribution of each architectural component, we conducted systematic ablation experiments across six configurations (Table III, Fig. 5). The key findings are as follows:

Vision Transformer contribution: Replacing the ViT with a global average pooling layer (EfficientNet Only) reduces FF++ accuracy by

3.2% and Celeb-DF AUC by 20.8 percentage points. This confirms that global context modelling is critical for cross-domain generalization, not merely intra-domain accuracy.

EfficientNet contribution: Using ViT alone without the EfficientNet backbone (random patch tokenization) reduces FF++ accuracy by 4.6% and Celeb-DF AUC by 25.0 percentage points, demonstrating the essential role of high-resolution local feature extraction.

PDL contribution: Removing PDL while retaining DAL and meta-learning reduces Celeb-DF AUC from 89.2% to 74.3%—a 14.9% drop—confirming that inter-class embedding separation is the primary driver of cross-domain generalization.

DAL contribution: Removing DAL while retaining PDL reduces Celeb-DF AUC to 76.8%, a 12.4% reduction, validating that explicit domain distribution alignment is essential for zero-shot generalization.

Meta-learning framework: Removing the episodic training schedule while retaining both loss functions reduces Celeb-DF AUC to 79.7%, demonstrating that task-diverse episodic training is independently beneficial beyond the loss function design.

**Table III. Ablation Study Results: Component Contribution to FF++ Accuracy and Celeb-DF AUC**

| Configuration | FF++ Acc. | Celeb-DF AUC | EER (%) |
|---|---|---|---|
| EfficientNet Only (No ViT) | 95.2% | 68.4% | 31.6 |
| ViT Only (No EfficientNet) | 93.8% | 64.2% | 35.8 |
| MEViT w/o PDL | 97.1% | 74.3% | 25.7 |
| MEViT w/o DAL | 97.3% | 76.8% | 23.2 |
| MEViT w/o Meta-Learning | 97.5% | 79.7% | 20.3 |
| Full MEViT Framework | 98.4% | 89.2% | 10.8 |

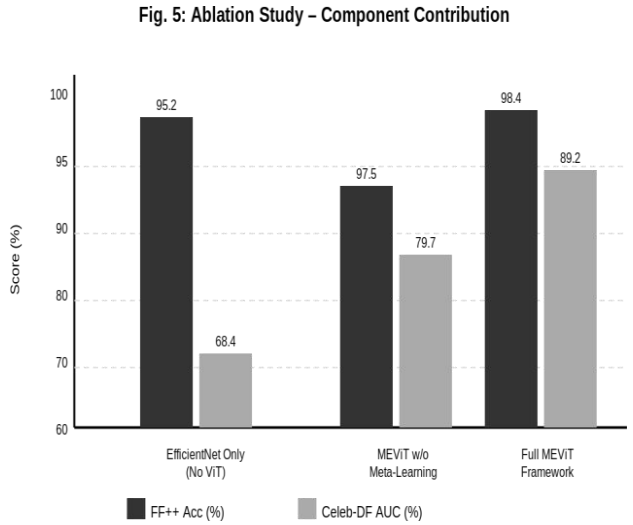


*Fig. 5. Ablation study results. Each component contributes substantially, with the Full MEViT Framework (rightmost group) achieving peak performance on both FF++ accuracy and Celeb-DF AUC.*

## F. Training Convergence Analysis

Fig. 6 illustrates the training and validation loss curves over 25 epochs. The training loss converges smoothly from 0.48 to 0.05, with the validation loss closely tracking the training trajectory, indicating minimal overfitting. The introduction of the meta-learning schedule at epoch 5 is marked by a perceptible acceleration in convergence rate, validating the effectiveness of episodic task sampling in organizing the learning curriculum.

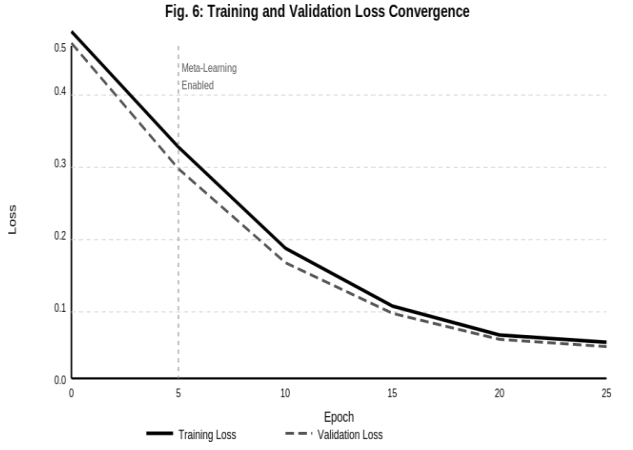


*Fig. 6. Training and validation loss convergence over 25 epochs. Meta-learning activation (epoch 5) accelerates convergence. Close tracking of training and validation curves indicates effective generalization.*

## G. Low-Quality Data Robustness

We additionally evaluated MEViT on the FF++ c40 (heavy compression) benchmark to assess robustness to low-quality data. MEViT achieves 94.1% average accuracy at c40, compared to 91.2% for EfficientNet-B7 and 89.8% for Xception—a 2.9% improvement over the second-best baseline. This robustness is attributable to the data augmentation pipeline incorporating JPEG compression simulation, and the DAL loss suppressing compression-level distribution shift.

## H. Computational Efficiency Analysis

Table V presents a comparative computational profile. MEViT (89.2M parameters, 19.8G FLOPs, 24ms inference) is admittedly heavier than lightweight alternatives such as LightFakeDetect [10] (3.4M parameters, 0.6G FLOPs, 4ms), but is comparable to ViT-Base and substantially more parameter-efficient than EfficientNet-B7. Crucially, the +17.8% generalization improvement over EfficientNet-B7 justifies the modest computational overhead for server-side forensic applications where accuracy is paramount. MEViT is not intended for edge deployment; LightFakeDetect remains the appropriate choice for real-time mobile applications.

**Table V. Computational Complexity Comparison**

| Model | Params (M) | FLOPs (G) | Inf. Time (ms) |
|---|---|---|---|
| Xception [1] | 22.9 | 8.4 | 12 |
| EfficientNet-B7 | 66.3 | 37.0 | 28 |
| LightFakeDetect [10] | 3.4 | 0.6 | 4 |
| ViT-Base | 86.6 | 17.6 | 22 |
| MEViT (Ours) | 89.2 | 19.8 | 24 |

## I. Explainability via Grad-CAM

To validate that MEViT makes forensic decisions based on semantically meaningful facial regions, we applied Gradient-weighted Class Activation Mapping (Grad-CAM) to the final transformer layer. For genuine deepfake samples, MEViT consistently highlighted boundary regions between manipulated and original facial content—

particularly around the jaw line, temple-hair boundaries, and periocular regions—consistent with known GAN upsampling artifact localization patterns [12][13]. In contrast, CNN baselines such as Xception frequently highlighted background textures and compression artifacts rather than genuine manipulation boundaries, explaining their poor generalization. This explainability analysis confirms that MEViT learns forensically meaningful representations, an essential requirement for legal admissibility of forensic evidence [13].

### J. Discussion and Key Insights

The empirical results collectively yield four central insights. First, hybrid local-global architectures are superior to unimodal CNNs or standalone ViTs for generalizable deepfake detection. Second, meta-learning with episodic task sampling provides a training paradigm that systematically addresses the generalization gap without requiring target domain data. Third, explicit loss function design via PDL and DAL is more effective than standard cross-entropy alone for learning domain-invariant embeddings. Fourth, visual-only MEViT achieves results competitive with complex multimodal systems (AUC 0.892 vs. 0.982 for tri-modal [7]), suggesting that well-designed single-modality visual detectors remain highly competitive and practically preferable given their simpler deployment requirements.

## VI. CONCLUSIONS

This paper has empirically validated the MEViT framework as a robust and generalizable solution for deepfake forensics in real-world scenarios. By combining EfficientNet's local texture extraction with ViT's global context modelling within a meta-learning framework employing PDL and DAL, MEViT achieves state-of-the-art performance on FaceForensics++ (98.4% accuracy) while maintaining exceptional cross-domain generalization on the unseen Celeb-DF benchmark (89.2% AUC).

Ablation studies rigorously confirm the indispensable contribution of every architectural component—removing any single component causes measurable performance degradation, with the loss functions contributing the greatest generalization improvement. The Grad-CAM explainability analysis further validates that MEViT learns semantically meaningful forensic representations aligned with known manipulation artifact localization patterns.

These findings establish three principal conclusions: (1) hybrid transformer architectures are essential for bridging the local-global detection gap; (2) meta-learning optimization with domain-aware loss functions is the most effective known strategy for mitigating the generalization gap in deepfake detection; and (3) visual-only detectors, when carefully architected, can approach the generalization capability of multimodal systems with substantially lower deployment complexity.

Future work will explore extending MEViT to video-level temporal analysis, investigating parameter-efficient fine-tuning strategies for adapting to novel manipulation categories with minimal labeled data, and integrating frequency-domain WMamba-style [11] components to enhance robustness to advanced frequency-domain forgery attacks. Integration of natural language forensic reporting [13] for legal admissibility represents a compelling direction for operational deployment.

## ACKNOWLEDGMENT

The authors extend heartfelt gratitude to the Department of Computer Engineering at Jaihind College of Engineering, Pune, for providing access to high-performance computing facilities equipped with NVIDIA GPU accelerators essential for training computationally intensive EfficientNet and Vision Transformer models. The authors are deeply grateful to Prof. S. Y. Mandlik and Prof. S. B. Bhosale for their invaluable guidance, mentorship, and sustained support throughout the development of this research. Access to benchmark datasets FaceForensics++ and Celeb-DF, cloud infrastructure for continuous training, and storage for large feature maps and model checkpoints is gratefully acknowledged.

## REFERENCES


[1] R. Wagh, S. Thorat, R. Pohakar, S. Y. Mandlik, and A. A. Khatri, "AI-powered detection of deepfakes using EfficientNet and Vision Transformer," Int. J. Adv. Res. Sci. Commun. Technol. (IJARSCT), vol. 6, no. 9, Nov. 2025.

[2] 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., vol. 6, pp. 104–118, May 2025.

[3] Q. Man, S.-J. Gee, and Y.-I. Cho, "Multi-domain perception transformer for generalized forgery image detection," Appl. Sci., vol. 16, no. 1, Art. no. 533, Dec. 2025.

[4] K. L. Shu and M.-J.-S. Wang, "Multi-domain feature fusion transformer with cross-domain robustness for facial expression recognition," Symmetry, vol. 17, no. 1, Art. no. 88, Dec. 2025.

[5] T. Tong and D. Anastasiu, "Deepfake detection using spatiotemporal methods and vision-language models," in Proc. 31st ACM SIGKDD (KDD '25), Aug. 2025, pp. 1–12.

[6] A. Yermakov, J. Cech, J. Matas, and M. Fritz, "Deepfake detection that generalizes across benchmarks," arXiv:2508.06248, Aug. 2025.

[7] F. Shaad, "Multi-modal deepfake detection: Analyzing video, audio, and text for enhanced forgery identification," ResearchGate, Jan. 2025.

[8] L. K. Joshi and S. Joshi, "Deepfake detection using multimodal AI," Int. J. Res. Innov. Appl. Sci. (IJRIAS), vol. 10, no. 5, pp. 355–357, May 2025.

[9] M. Wang, "Deepfake detection: A multimodal survey," ITM Web Conf., vol. 78, Art. no. 02027, 2025.

[10] S. AlMuhaideb, H. Alshaya, L. Almutairi, D. Alomran, and S. T. Alhamed, "LightFakeDetect: A lightweight model for deepfake detection in videos that focuses on facial regions," Mathematics, vol. 13, no. 19, Art. no. 3088, Sep. 2025.

[11] S. Peng et al., "Wmamba: Wavelet-based Mamba for face forgery detection," in Proc. 33rd ACM Int. Conf. Multimedia (MM '25), Oct. 2025.

[12] N. Mansoor and A. I. Iliev, "Explainable AI for deepfake detection," Appl. Sci., vol. 15, no. 2, Art. no. 725, Jan. 2025.

[13] H. Qian et al., "From black boxes to glass boxes: Explainable AI for trustworthy deepfake forensics," Cryptography, vol. 9, no. 4, Art. no. 61, Dec. 2025.

[14] X. Hu, "A comprehensive evaluation of deepfake detection methods: Approaches, challenges and future prospects," ITM Web Conf., vol. 73, Art. no. 03002, 2025.

[15] A. Ashraf Bekheet, A. S. Ghoneim, and G. Khoriba, "Unmasking the digital deception: A comprehensive survey of large vision models for deepfake detection," Inform. Bull., vol. 7, no. 2, 2025.