RESEARCH ARTICLE

Enhanced Deepfake Detection Through Multi-Attention Mechanisms: A Comprehensive Framework for Synthetic Media Identification

Farhan Ali[1,*] **and Zainab Ghazanfar**[2]

[1] Department of Computer Science, Graz University of Technology, Graz 8010, Austria
[2] Department of AI and Software, Gachon University, Seongnam-si 13120, Republic of Korea

Abstract

The proliferation of deepfake technology poses significant threats to digital media authenticity, necessitating robust detection systems to combat manipulated content. This paper presents a novel attention-based framework for deepfake detection that systematically integrates multiple complementary attention mechanisms to enhance discriminative feature learning. Our approach combines spatial attention, multi-head self-attention, and channel attention modules with a VGG-16 backbone to capture comprehensive representations across different feature spaces. The spatial attention mechanism focuses on discriminative facial regions, while multi-head self-attention captures long-range spatial dependencies and global contextual relationships. Channel attention further refines feature representations by emphasizing the most informative channels for detection. Extensive experiments on FaceForensics++ and Celeb-DF datasets demonstrate the effectiveness of our progressive attention integration strategy. The proposed framework achieves competitive performance with 92.67% accuracy and 99.30% Area Under the Curve (AUC) on FF++, while maintaining solid generalization capabilities with 82.35% accuracy and 82.7% AUC on the challenging Celeb-DF dataset. Comprehensive ablation studies validate the contribution of each attention component and justify key design choices, including the optimal 3×3 kernel size for spatial attention. Comparison with state-of-the-art methods demonstrates that our approach achieves competitive detection performance while maintaining architectural simplicity and computational efficiency. The modular design of our framework provides interpretability and flexibility for deployment across various computational environments, making it suitable for practical artificial media detection applications.

Keywords: deepfake detection, multi-head self-attention, synthetic media detection, facial manipulation detection.

Academic Editor:
Xue-Bo Jin

Submitted: 28 May 2025
Accepted: 03 July 2025
Published: 24 November 2025

Vol. 2, **No.** 4, 2025.
10.62762/TIS.2025.756872

***Corresponding author:**
✉ Farhan Ali
farhan.ali@student.tugraz.at

Citation
Ali, F., & Ghazanfar, Z. (2025). Enhanced Deepfake Detection Through Multi-Attention Mechanisms: A Comprehensive Framework for Synthetic Media Identification. *ICCK Transactions on Intelligent Systematics*, 2(4), 248–258.

1 Introduction

The rapid advancement of artificial intelligence and generative technologies has revolutionized digital media creation, making sophisticated content generation more accessible than ever before. While these technological developments have democratized creative expression and media production, they have simultaneously enabled the emergence of deepfake technology, raising critical concerns about the authenticity and trustworthiness of digital content. Deepfakes pose significant threats across multiple domains, affecting the privacy and security of individuals ranging from public figures, celebrities, and politicians to ordinary citizens. The proliferation of synthetic media facilitates the spread of misinformation, enables sophisticated fraud schemes, compromises personal security, damages reputations, and fundamentally undermines public trust in digital media [1]. Malicious actors increasingly exploit these technologies for misinformation campaigns, which fall under the broader umbrella of cybercrime [2]. Consequently, digital privacy preservation has become a critical concern in contemporary society [3, 4]. To address these challenges, robust deepfake detection (DD) systems are urgently needed across various platforms and applications. There is an imperative need to develop reliable methods for identifying synthetic content to safeguard security and maintain confidence in digital media consumption [5]. The emergence of deepfake technology has fundamentally transformed how facial features in images and videos can be manipulated and synthesized [6, 7], leading to a significant erosion of trust and safety in digital media ecosystems. As society becomes increasingly dependent on visual media for information dissemination, entertainment, and communication [8], the impact of deepfake technology on public perception and decision-making processes has become profound. Therefore, developing sophisticated neural networks capable of accurately identifying synthetic content is crucial for addressing the escalating security and trust challenges in our digital-first world.

Current deepfake detection methodologies predominantly focus on specific types of synthetic manipulations. For instance, Afchar et al. [9] proposed a specialized technique for detecting lip-syncing deepfakes by exploiting inconsistencies between phonemes (spoken sounds) and visemes (mouth shapes). Das et al. [10] introduced a data augmentation-based approach called *Face-Cutout*, which removes irrelevant facial landmark information and selectively extracts relevant features from video frames. Agarwal et al. [11] utilized the OpenFace2 toolbox to extract facial landmark features, combining them with action units to train a binary SVM classifier for deepfake detection. The MesoNet architecture [9] represents an end-to-end system designed to identify synthetic Face2Face videos and other deepfake variants. Furthermore, Bonettini et al. [12] employed ensemble learning with four different CNN architectures, incorporating data augmentation techniques such as downsampling, hue saturation adjustment, and JPEG compression during training and validation to enhance model robustness. Lee et al. [13] proposed the TAR (Transfer Learning-based Autoencoder with Residuals) approach, which leverages transfer learning and autoencoder architectures to detect various types of deepfakes. The effectiveness of autoencoders across multiple domains has been well-documented [14]. Despite these advances, evaluation on real-world datasets, such as training on FF++ and testing on approximately 200 real-world videos featuring 50 different celebrities, reveals the ongoing need for more sophisticated detection methods. The continuous evolution of generative technologies necessitates intelligent advancements in detection systems, incorporating large-scale training data, progressive integration of attention mechanisms, and effective fine-tuning strategies to enhance security and maintain trust in digital content consumption. To address these challenges and advance the SOTA in deepfake detection, this paper presents a novel attention-based framework that integrates multiple complementary attention mechanisms for enhanced synthetic content identification. The main contributions of this work are as follows:

- **Progressive Multi-Attention Framework:** We propose a novel deepfake detection architecture that sequentially integrates spatial attention, multi-head self-attention, and channel attention mechanisms with a VGG-16 backbone, enabling progressive feature refinement across multiple representation spaces for enhanced discrimination capability.
- **Dual-Pooling Spatial Attention Design:** We introduce an optimized spatial attention module that combines global max and average pooling with sequential 3×3 and 1×1 convolutions, specifically designed to capture subtle manipulation artifacts in facial regions that

traditional methods often overlook.

- **Systematic Design Validation:** We provide comprehensive ablation studies demonstrating the individual contribution of each attention component and validate key architectural choices, including optimal kernel size selection and attention module ordering, through rigorous experimental analysis.
- **Balanced Performance-Efficiency Trade-off:** Our framework achieves competitive detection performance (92.67% accuracy on FF++, 82.35% on Celeb-DF) while maintaining computational efficiency and architectural interpretability, making it suitable for practical deployment scenarios where resource constraints are critical.

2 Related Literature

The proliferation of synthetic media has sparked extensive research into deepfake detection methodologies, encompassing both conventional machine learning (ML) and advanced deep learning (DL) approaches [15]. Researchers have developed diverse strategies to address the multifaceted challenges inherent in synthetic content identification [16]. Early approaches focused on behavioral analysis and the handcrafted extraction of features. Agarwal et al. [11] developed a framework that analyzes individual-specific facial expressions and head movements from extended video sequences. Their method extracts 190-dimensional features from 10-second video segments processed frame by frame, subsequently employing Support Vector Machines (SVMs) to determine consistency with learned behavioral patterns. While innovative, this approach requires individual-specific model training and relies heavily on manually engineered features, which may limit its scalability and generalizability. The FakeCatcher system [17] introduced a novel approach by analyzing physiological signals extracted from three distinct facial regions in authentic videos. This methodology demonstrated particular promise for real-time applications where user interaction is paramount. Complementing this direction, Guera and Delp [18] proposed a hybrid architecture combining Convolutional Neural Networks (CNNs) for spatial feature extraction with Long Short-Term Memory (LSTM) networks to capture temporal dependencies. However, their evaluation methodology, utilizing randomly collected videos from diverse web sources rather than standardized datasets, raises concerns about the robustness and generalizability of their findings.

Advanced detection techniques have emerged, focusing on pixel-level analysis and architectural innovations. Li et al. [19] introduced Face X-ray, a sophisticated forgery detection method that analyzes grayscale representations to determine whether images result from blending multiple source images. The MesoNet architecture [9] represents a significant advancement in end-to-end deepfake detection, employing stacked convolutional layers designed explicitly for the identification of synthetic content. Building upon transformer architectures, Heo et al. [20] developed a Vision Transformer (ViT)-based detection system incorporating knowledge distillation techniques with pre-trained EfficientNetB7 backbones. Their approach processes patch-based representations extracted through EfficientNet and feeds them to transformer encoders trained on the DFDC dataset. Data augmentation and feature selection strategies have also received considerable attention. Das et al. [10] addressed limitations in existing detection methods through their Face-Cutout technique, which selectively removes irrelevant facial landmark information while preserving discriminative features from video frames. Similarly, Wang et al. [21] demonstrated the effectiveness of bilateral feature fusion with hexagonal attention under uncertain environments. The DefakeHop methodology [22] employs successive subspace learning for automated feature extraction from various regions within manipulated images. This approach utilizes channel-wise Saab transforms followed by feature distillation through spatial dimensionality reduction and soft classification mechanisms. Transfer learning approaches have shown promising results in cross-domain detection scenarios. Lee et al. [13] developed a transfer learning-based autoencoder incorporating residual connections to identify multiple deepfake variants, achieving 89.49% zero-shot accuracy on evaluation datasets. Wodajo and Melese [23] combined CNN-based feature extraction with ViT architectures for binary classification, where CNN-derived features serve as inputs to subsequent prediction networks. Nguyen et al. [24] explored capsule networks for detecting various forms of image and video manipulations.

Ensemble methods and multimodal approaches have emerged as effective strategies for improving detection robustness. Bonettini et al. [12] implemented ensemble learning utilizing four distinct CNN architectures enhanced with attention mechanisms.

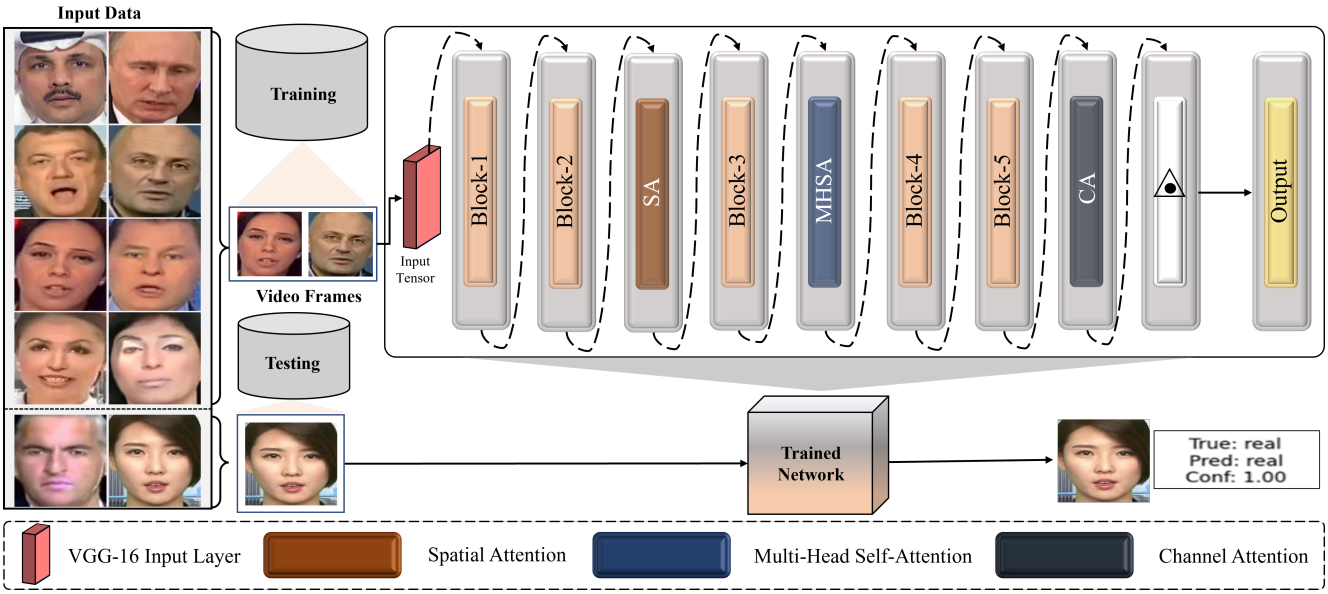

Figure 1. Overall architecture of the proposed deepfake detection framework. The system takes facial images as input and processes them through a VGG-16 backbone network, followed by three sequential attention mechanisms: Spatial Attention (SA), Multi-Head Self-Attention (MHSA), and Channel Attention (CA), which emphasize discriminative feature channels. The final classification output distinguishes between authentic and manipulated facial content.

Recent work has shown the importance of progressive feature refinement [26, 27], which shares conceptual similarities with the multi-scale attention mechanisms required for effective deepfake detection. Prajapati et al. [25] introduced MRI-GAN, a GAN-based detection framework analyzing perceptual differences between authentic and synthetic content. Their evaluation on DFDC demonstrated 91% accuracy for frame-based models and 74% for perceptual difference analysis. Multi-scale feature processing has proven effective across various computer vision applications. Researchers are now also focusing on multimodal fusion techniques. Mittal et al. [28] employed Siamese networks to simultaneously learn audio-visual feature representations, evaluating their approach on both DFDC and Deepfake-TMIT datasets. Spatiotemporal analysis has also gained traction in deepfake detection research. Montserrat et al. [29] leveraged spatiotemporal properties within video sequences for detection, demonstrating effectiveness on face-swap datasets like DFDC, albeit with significant computational overhead limiting practical deployment. De Lima et al. [30] focused on convolutional architectures designed to identify spatial and temporal artifacts specific to GAN-generated content, exploiting the subtle inconsistencies inherent in synthetically generated media. Despite these significant advances, substantial opportunities remain for developing optimized feature-driven visual intelligence systems capable of effectively capturing the increasingly sophisticated characteristics of modern deepfake technologies.

3 Proposed Methodology

This section outlines our comprehensive deepfake detection framework, which combines convolutional feature extraction with multiple attention mechanisms to accurately identify synthetic facial content. Our architecture follows a multi-stage pipeline, comprising hierarchical feature extraction, spatial attention, multi-head self-attention, and channel-wise refinement, to effectively capture both subtle and prominent artifacts introduced during deepfake generation. The complete framework is illustrated in Figure 1.

3.1 Feature Extraction with VGG-16

Our framework begins with feature extraction using the VGG-16 network, a renowned model for image classification. VGG-16 serves as a robust backbone for learning hierarchical representations of facial structures. Given an input image $I \in \mathbb{R}^{H\times W\times 3}$, it is processed through five convolutional blocks of VGG-16, each containing multiple 3×3 convolution layers followed by ReLU activations and max-pooling. These blocks extract features from low-level textures to high-level semantics:

$$\mathbf{F}_{conv} = \text{VGG-16}(I; \theta_{vgg})$$

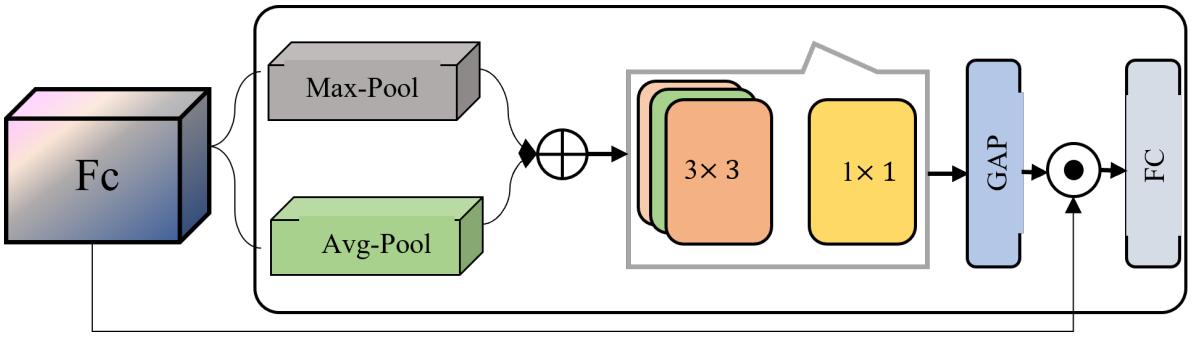

Figure 2. Detailed structure of the Spatial Attention module. The input feature maps undergo parallel global max pooling and global average pooling operations to capture the most salient feature responses across spatial dimensions.

where $\mathbf{F}_{conv} \in \mathbb{R}^{H' \times W' \times D}$ denotes the output feature maps, and θ_{vgg} represents the pre-trained parameters. The use of ImageNet-trained weights provides a strong initialization, aiding in the detection of pixel-level inconsistencies typical of deepfakes. VGG-16's architectural uniformity and small receptive fields help retain fine spatial details, making it ideal for detecting localized manipulations such as skin inconsistencies, boundary artifacts, and abnormal textural transitions.

3.2 Spatial Attention Module

We introduce a spatial attention mechanism to localize and emphasize discriminative regions within the facial image, as illustrated in Figure 2. This module enhances important regions while suppressing irrelevant or background areas. Given $\mathbf{F}_{conv}$, we compute the spatial attention map as:

$$\mathbf{A}_{spatial} = \sigma\left(\text{Conv}_{1\times1}(\text{ReLU}(\text{Conv}_{3\times3}(\mathbf{F}_{conv})))\right)$$

where $\text{Conv}_{3\times3}$ and $\text{Conv}_{1\times1}$ denote convolutional layers of the corresponding kernel sizes, and σ is the sigmoid activation function. The attended features are:

$$\mathbf{F}_{spatial} = \mathbf{A}_{spatial} \odot \mathbf{F}_{conv}$$

3.3 Multi-head self-attention (MHSA)

We incorporate a Multi-head self-attention (MHSA) module to model global relationships and long-range dependencies within facial regions. We reshape $\mathbf{F}_{spatial}$ into a sequence $\mathbf{X} \in \mathbb{R}^{N \times D}$, where $N = H' \times W'$. For each attention head h, query ($\mathbf{Q}^{(h)}$), key ($\mathbf{K}^{(h)}$), and value ($\mathbf{V}^{(h)}$) matrices are computed as:

$$\mathbf{Q}^{(h)} = \mathbf{X}\mathbf{W}_Q^{(h)}, \quad \mathbf{K}^{(h)} = \mathbf{X}\mathbf{W}_K^{(h)}, \quad \mathbf{V}^{(h)} = \mathbf{X}\mathbf{W}_V^{(h)}$$

with $\mathbf{W}_Q^{(h)}, \mathbf{W}_K^{(h)}, \mathbf{W}_V^{(h)} \in \mathbb{R}^{D \times d_k}$ and $d_k = D/H$ for H heads. The attention output per head is:

$$\text{Attention}^{(h)} = \text{softmax}\left(\frac{\mathbf{Q}^{(h)}\mathbf{K}^{(h)T}}{\sqrt{d_k}}\right)\mathbf{V}^{(h)}$$

All head outputs are concatenated and linearly projected:

$$\mathbf{F}_{mhsa} = \text{Concat}(\text{Attention}^{(1)}, \ldots, \text{Attention}^{(H)})\mathbf{W}_O$$

where $\mathbf{W}_O \in \mathbb{R}^{D \times D}$ is the output projection matrix. This module enables the model to understand semantic relationships across the face, which are critical for identifying synthetic inconsistencies.

3.4 Channel Attention Module

We employ a channel attention mechanism to refine features along the channel dimension. This module emphasizes the most relevant channels that contribute to deepfake detection, as shown in Figure 3. The MHSA output $\mathbf{F}_{mhsa}$ is reshaped back to $\mathbb{R}^{H' \times W' \times D}$. Global average pooling and max pooling are applied:

$$\mathbf{f}_{avg} = \text{GAP}(\mathbf{F}_{mhsa}), \quad \mathbf{f}_{max} = \text{GMP}(\mathbf{F}_{mhsa})$$

These are fed into a shared MLP with a reduction ratio r:

$$\mathbf{A}_{channel} = \sigma\left(\text{MLP}(\mathbf{f}_{avg}) + \text{MLP}(\mathbf{f}_{max})\right)$$

where the MLP structure is $\mathbb{R}^D \to \mathbb{R}^{D/r} \to \mathbb{R}^D$. The refined feature map is:

$$\mathbf{F}_{final} = \mathbf{A}_{channel} \odot \mathbf{F}_{mhsa}$$

This module allows the network to selectively amplify feature channels that encode key artifacts such as blending errors, texture mismatches, or abnormal lighting.

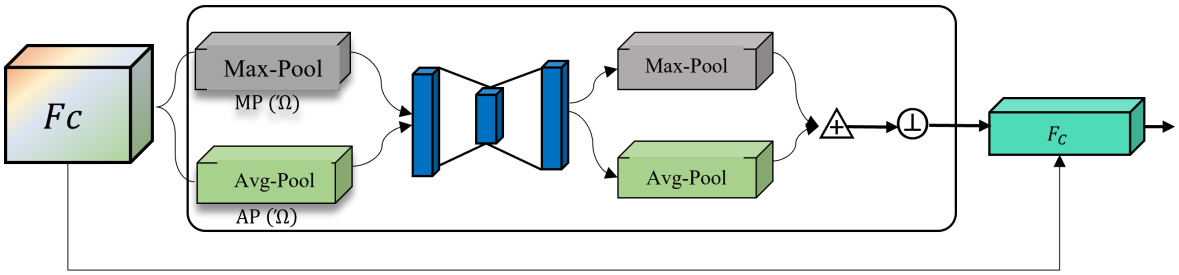

Figure 3. Channel Attention Module with Global Pooling and Shared MLP. The architecture employs both global max pooling and global average pooling followed by a shared multilayer perceptron to generate adaptive channel weights for feature refinement.

3.5 Classification and Training Strategy

The final feature representation $\mathbf{F}_{final}$ is pooled globally and passed through a fully connected layer for classification:

$$p = \sigma(\mathbf{W}_{fc} \cdot \text{GAP}(\mathbf{F}_{final}) + b)$$

where $\mathbf{W}_{fc}$ and b are the parameters of the classification layer. Training is conducted using the binary cross-entropy loss:

$$\mathcal{L} = -\frac{1}{N}\sum_{i=1}^{N}[y_i \log(p_i) + (1 - y_i)\log(1 - p_i)]$$

where y_i is the ground truth label for the i-th image and p_i is the corresponding prediction. Our approach effectively combines convolutional hierarchies with spatial and channel attention, augmented by MHSA for global context modeling. The synergy of these components allows the framework to detect a wide range of deepfake techniques, including face swapping and facial reenactment, by leveraging both local and global discriminative features.

4 Experiments

4.1 Experimental Setup

We evaluate our proposed method on two challenging datasets: **FaceForensics++ (FF++)** [31] and **Celeb-DF** [32]. FF++ is the most widely used forgery dataset, comprising 720 videos for training and 280 videos for validation or testing. Celeb-DF is generated via face swapping for 59 pairs of subjects and includes 590 real videos and 5,639 high-quality fake videos. All detected face images are cropped and resized to a fixed resolution of 224×224 for consistency.

4.2 Evaluation Metrics

We employ several common metrics to evaluate our method, including the Area Under the Receiver Operating Characteristic Curve (AUC), Equal Error Rate (EER), Accuracy (ACC), Precision, Recall, and F1-Score. The AUC value ranges from 0 to 1, with values closer to 1 indicating better performance. The EER represents the point on the ROC curve where the false acceptance rate equals the false rejection rate.

4.3 Implementation Details

We implemented our proposed network using the TensorFlow deep learning framework on an NVIDIA GeForce RTX 4090 GPU. The Adam optimizer was used with a momentum of 0.9, and the initial learning rate was set to 1×10^{-4}. The network was trained for 50 epochs with a batch size of 8.

4.4 Ablation Study

To comprehensively evaluate the effectiveness of our proposed framework and validate the design choices, we conduct extensive ablation studies examining both the individual contribution of each attention module and the impact of key hyperparameters on detection performance.

4.4.1 Module-wise Contribution Analysis

Table 1 presents a systematic analysis of the performance gains achieved by progressively integrating each attention mechanism into our detection framework. We evaluate four different configurations across both FF++ and CELEB-DF datasets to demonstrate the consistent effectiveness of our approach. Starting with the VGG-16 backbone as our baseline, we observe strong foundational performance with 85.72% accuracy on FF++ and 75.85% on CELEB-DF. The performance difference between datasets reflects the inherently challenging nature of CELEB-DF, which contains higher-quality deepfakes that are more difficult to detect. The incorporation of Spatial Attention (SA) yields substantial improvements, increasing accuracy by 2.73% on FF++ (88.45%) and 2.40% on CELEB-DF (78.25%). This demonstrates that focusing on

Table 1. Performance analysis on FF++ and CELEB-DF with integrated modules.

Dataset	Model	ACC	Precision	Recall	F1 Score	AUC
FF++	Backbone	85.72	85.18	86.28	85.73	93.15
	Backbone + SA	88.45	87.92	89.05	88.48	94.28
	Backbone + SA + MHSA	90.83	90.25	91.42	90.83	95.97
	Backbone + SA + MHSA + CA	**92.67**	**92.15**	**93.21**	**92.68**	**99.30**
CELEB-DF	Backbone	75.85	76.20	75.50	75.85	78.45
	Backbone + SA	78.25	78.65	77.85	78.25	79.82
	Backbone + SA + MHSA	80.42	80.85	80.00	80.42	81.15
	Backbone + SA + MHSA + CA	**82.35**	**82.78**	**81.92**	**82.35**	**82.70**

Table 2. Impact of spatial attention kernel size on detection performance.

Kernel Size	ACC (%)	Precision (%)	Recall (%)	F1 Score (%)	AUC (%)
1×1	86.82	86.35	87.31	86.83	93.45
3×3	**88.45**	**87.92**	**89.05**	**88.48**	**94.28**
5×5	88.12	87.68	88.58	88.13	94.02
7×7	87.89	87.41	88.39	87.90	93.84

discriminative spatial regions significantly enhances the model's ability to identify manipulation artifacts. The AUC improvements of 1.13% and 1.37% respectively indicate enhanced discriminative capability across different decision thresholds.

Adding Multi-Head Self-Attention (MHSA) to the SA-enhanced model provides additional performance gains of 2.38% on FF++ (90.83%) and 2.17% on CELEB-DF (80.42%). The MHSA mechanism enables the model to capture long-range spatial dependencies and global contextual relationships, which are crucial for identifying subtle inconsistencies across different facial regions that may indicate synthetic manipulation. The final integration of Channel Attention (CA) achieves the highest performance across all metrics, reaching 92.67% accuracy on FF++ and 82.35% on CELEB-DF. The complete framework demonstrates remarkable AUC scores of 99.30% and 82.70% respectively, indicating excellent discriminative ability between authentic and manipulated content. The consistent improvement pattern across both datasets validates the generalizability and robustness of our attention-based approach.

4.4.2 Spatial Attention Kernel Size Analysis

Table 2 presents a detailed analysis of different kernel sizes for the spatial attention mechanism to justify our architectural design choice. We evaluate various kernel configurations ranging from 1×1 to 7×7 convolutions to determine the optimal balance between computational efficiency and spatial modeling capability. The results reveal that 3×3 convolution kernels achieve optimal performance across all evaluation metrics, with 88.45% accuracy and 94.28% AUC. The 1×1 convolution, while computationally efficient, shows limited performance (86.82% accuracy) due to insufficient spatial context modeling. This confirms that purely point-wise operations are inadequate for capturing the spatial relationships necessary for effective attention weight generation.

Larger kernel sizes (5×5 and 7×7) show marginal performance degradation compared to 3×3, achieving 88.12% and 87.89% accuracy, respectively. This decline can be attributed to over-smoothing effects and increased parameter overhead without proportional benefits in spatial reasoning capability. The 3×3 kernel strikes an optimal balance by providing sufficient receptive field coverage for local spatial dependencies while maintaining computational efficiency. These findings validate our design choice of 3×3 convolution in the spatial attention module, demonstrating that this configuration maximizes performance while ensuring computational tractability. The results align with established practices in attention mechanism design, where moderate kernel sizes typically provide the best trade-off between modeling capacity and efficiency.

4.5 Comparison with SOTA methods

To demonstrate the effectiveness and competitiveness of our proposed attention-based framework, we conduct comprehensive comparisons with state-of-the-art deepfake detection methods across multiple evaluation scenarios. Table 3 presents a detailed cross-database evaluation where models are trained on FF++ and tested on both FF++

Table 3. Cross-database evaluation from FF++ to Celeb-DF in terms of AUC and EER.

Method	FF++		Celeb-DF	
	AUC (%)	EER (%)	AUC (%)	EER (%)
GFF [33]	98.4	3.8	75.7	32.2
Local-relation [34]	99.2	3.0	78.3	29.1
Face X-ray [35]	87.5	4.2	75.7	26.5
DCL [36]	99.3	3.2	82.3	26.4
EN-b4 [37]	99.2	3.3	68.6	35.3
MAT(EN-b4) [38]	99.3	3.3	76.6	32.5
UIA-ViT [39]	99.3	3.2	82.5	27.4
F3-Net [40]	98.1	3.5	71.2	33.7
MLDG [41]	98.6	3.4	74.5	30.8
Xception [42]	99.1	3.7	65.3	38.7
LTW [43]	99.2	3.3	77.2	29.3
DFNet [44]	99.8	2.9	87.8	22.8
The Proposed Model (Ours)	99.3	2.9	82.7	25.2

(intra-dataset) and Celeb-DF (cross-dataset) to assess generalization capabilities.

4.5.1 Intra-Dataset Performance Analysis

On the FF++ dataset, our proposed method achieves competitive performance with 99.3% AUC and 2.9% EER, positioning it among the top-performing approaches. While DFNet [44] achieves the highest AUC of 99.8%, our method demonstrates comparable performance to other state-of-the-art methods including DCL [36], MAT(EN-b4) [38], and UIA-ViT [39], all achieving 99.3% AUC. This demonstrates that our attention-based architecture effectively captures the discriminative patterns necessary for accurate deepfake detection within the training domain. The strong intra-dataset performance validates the effectiveness of our multi-attention design, where the combination of spatial attention, multi-head self-attention, and channel attention mechanisms successfully identifies manipulation artifacts across the four different deepfake generation techniques present in FF++. Our method's EER of 2.9% is particularly noteworthy, matching the best-performing methods and indicating excellent balance between false positive and false negative rates.

4.5.2 Cross-Dataset Generalization Performance

The cross-dataset evaluation on Celeb-DF reveals more nuanced performance characteristics that highlight the challenges of deepfake detection generalization. Our proposed method achieves 82.7% AUC and 25.2% EER on Celeb-DF, demonstrating solid generalization capabilities while revealing areas for improvement. Comparing with existing methods, our approach outperforms several established techniques including GFF [33] (75.7% AUC), Local-relation [34] (78.3% AUC), and LTW [43] (77.2% AUC). However, DFNet [44] achieves superior cross-dataset performance with 87.8% AUC, indicating potential for further architectural improvements in our attention mechanisms. Notably, our method shows competitive performance compared to UIA-ViT [39] (82.5% AUC) and DCL [36] (82.3% AUC), demonstrating that attention-based approaches can achieve comparable generalization to more complex architectural designs. The relatively small performance gap (0.2-0.4% AUC) with these methods suggests that our simpler attention framework provides an effective balance between performance and computational efficiency.

5 Conclusion

This work presents a novel attention-based framework for deepfake detection that strategically integrates spatial attention, multi-head self-attention, and channel attention mechanisms with a VGG-16 backbone. Our modular approach effectively identifies synthetic content across diverse generation techniques by capturing discriminative features at multiple representational levels. Systematic ablation studies confirm that each attention component makes a meaningful contribution to detection performance, with spatial attention focusing on critical facial regions, multi-head self-attention modeling long-range dependencies, and channel attention emphasizing informative feature channels. The

proposed method achieves competitive performance with 92.67% accuracy on FF++ while demonstrating solid cross-dataset generalization with 82.35% accuracy on the challenging Celeb-DF dataset. Our evaluation reveals both the potential and limitations of current deepfake detection approaches. While strong intra-dataset performance indicates effective feature learning, cross-dataset results highlight the ongoing challenge of generalization across different manipulation techniques and data distributions. The modular architecture offers advantages in interpretability and computational efficiency compared to more complex state-of-the-art methods. As synthetic media technology rapidly advances, our interpretable and extensible approach provides a robust foundation for future research directions. The demonstrated effectiveness across multiple datasets and the architectural flexibility for incorporating additional attention mechanisms position this work as a valuable contribution to maintaining digital content authenticity and trust.

Data Availability Statement

Data will be made available on request.

Funding

This work was supported without any funding.

Conflicts of Interest

The authors declare no conflicts of interest.

Ethical Approval and Consent to Participate

Not applicable.

References

[1] Wu, C. K., Cheng, C.-T., Uwate, Y., Chen, G., Mumtaz, S., & Tsang, K. F. (2022). State-of-the-art and research opportunities for next-generation consumer electronics. *IEEE Transactions on Consumer Electronics, 69*(4), 937–948. [CrossRef]

[2] Yamauchi, M., Ohsita, Y., Murata, M., Ueda, K., & Kato, Y. (2020). Anomaly detection in smart home operation from user behaviors and home conditions. *IEEE Transactions on Consumer Electronics, 66*(2), 183–192. [CrossRef]

[3] Parashar, A., Rida, I., Parashar, A., & Aski, V. (2022). Protecting the privacy of face by De-Identification Pipeline Based on Deep Learning. In *2022 16th International Conference on Signal-Image Technology & Internet-Based Systems (SITIS)* (pp. 409–416). [CrossRef]

[4] Kaur, J., Sharma, K., & Singh, M. P. (2024). Exploring the Depth: Ethical Considerations, Privacy Concerns, and Security Measures in the Era of Deepfakes. In *Navigating the World of Deepfake Technology* (pp. 141-165). IGI Global. [CrossRef]

[5] Zhang, G., Gao, M., Li, Q., Zhai, W., Zou, G., & Jeon, G. (2023). Disrupting deepfakes via union-saliency adversarial attack. *IEEE Transactions on Consumer Electronics, 70*(1), 2018–2026. [CrossRef]

[6] Whittaker, L., Mulcahy, R., Letheren, K., Kietzmann, J., & Russell-Bennett, R. (2023). Mapping the deepfake landscape for innovation: A multidisciplinary systematic review and future research agenda. *Technovation, 125*, 102784. [CrossRef]

[7] He, W., Fei, L., Zhao, S., Zhang, W., Teng, S., & Rida, I. (2023, December). Rrfae-net: Robust rgb-d facial age estimation network. In *Chinese Conference on Biometric Recognition* (pp. 119-128). Singapore: Springer Nature Singapore. [CrossRef]

[8] Ding, F., Zhu, G., Alazab, M., Li, X., & Yu, K. (2020). Deep-learning-empowered digital forensics for edge consumer electronics in 5G HetNets. *IEEE Consumer Electronics Magazine, 11*(2), 42–50. [CrossRef]

[9] Afchar, D., Nozick, V., Yamagishi, J., & Echizen, I. (2018, December). Mesonet: a compact facial video forgery detection network. In *2018 IEEE international workshop on information forensics and security (WIFS)* (pp. 1-7). IEEE. [CrossRef]

[10] Das, A., Das, S., & Dantcheva, A. (2021, December). Demystifying attention mechanisms for deepfake detection. In *2021 16th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2021)* (pp. 1-7). IEEE. [CrossRef]

[11] Agarwal, S., & Farid, H. (2021, June). Detecting Deep-Fake Videos from Aural and Oral Dynamics. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)* (pp. 981-989). IEEE. [CrossRef]

[12] Bonettini, N., Cannas, E. D., Mandelli, S., Bondi, L., Bestagini, P., & Tubaro, S. (2021, January). Video face manipulation detection through ensemble of cnns. In *2020 25th international conference on pattern recognition (ICPR)* (pp. 5012-5019). IEEE. [CrossRef]

[13] Lee, S., Tariq, S., Kim, J., & Woo, S. S. (2021, June). Tar: Generalized forensic framework to detect deepfakes using weakly supervised learning. In *IFIP International conference on ICT systems security and privacy protection* (pp. 351-366). Cham: Springer International Publishing. [CrossRef]

[14] Wu, J., Qiu, Z., Zeng, Z., Xiao, R., Rida, I., & Zhang, S. (2024). Graph Autoencoder Anomaly Detection for E-commerce Application by Contextual Integrating Contrast With Reconstruction and Complementarity. *IEEE Transactions on Consumer Electronics, 70*(1), 1623–1630. [CrossRef]

[15] Heidari, A., Jafari Navimipour, N., Dag, H., &

Unal, M. (2024). Deepfake detection using deep learning methods: A systematic and comprehensive review. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 14*(2), e1520. [CrossRef]

[16] Yoon, J. H., Panizo-LLedot, A., Camacho, D., & Choi, C. (2024). Triple-modality interaction for deepfake detection on zero-shot identity. *Information Fusion, 109*, 102424. [CrossRef]

[17] Ciftci, U. A., Demir, I., & Yin, L. (2020). Fakecatcher: Detection of synthetic portrait videos using biological signals. *IEEE Transactions on Pattern Analysis and Machine Intelligence*. [CrossRef]

[18] Güera, D., & Delp, E. J. (2018, November). Deepfake video detection using recurrent neural networks. In *2018 15th IEEE international conference on advanced video and signal based surveillance* (*AVSS*) (pp. 1-6). IEEE. [CrossRef]

[19] Li, Y., & Lyu, S. (2018). Exposing deepfake videos by detecting face warping artifacts. *arXiv preprint arXiv:1811.00656*.

[20] Heo, Y. J., Choi, Y. J., Lee, Y. W., & Kim, B. G. (2021). Deepfake detection scheme based on vision transformer and distillation. *arXiv preprint arXiv:2104.01353*.

[21] Wang, J., Lei, J., Li, S., & Zhang, J. (2025). STA-3D: Combining Spatiotemporal Attention and 3D Convolutional Networks for Robust Deepfake Detection. *Symmetry, 17*(7), 1037. [CrossRef]

[22] Chen, H. S., Rouhsedaghat, M., Ghani, H., Hu, S., You, S., & Kuo, C. C. J. (2021, July). DefakeHop: A Light-Weight High-Performance Deepfake Detector. In *2021 IEEE International Conference on Multimedia and Expo* (*ICME*) (pp. 1-6). IEEE. [CrossRef]

[23] Wodajo, D., & Atnafu, S. (2021). Deepfake video detection using convolutional vision transformer. *arXiv preprint arXiv:2102.11126*.

[24] Nguyen, T. T., Nguyen, Q. V. H., Nguyen, D. T., Nguyen, D. T., Huynh-The, T., Nahavandi, S., Nguyen, T. T., Pham, Q.-V., & Nguyen, C. M. (2022). Deep learning for deepfakes creation and detection: A survey. *Computer Vision and Image Understanding, 223*, 103525. [CrossRef]

[25] Prajapati, P., & Pollett, C. (2022). MRI-GAN: A Generalized Approach to Detect DeepFakes using Perceptual Image Assessment. *arXiv preprint arXiv:2203.00108*.

[26] Song, W., Guo, S., Gao, M., Li, Q., Zhu, X., & Rida, I. (2025). Deepfake detection via Feature Refinement and Enhancement Network. *Image and Vision Computing*, 105663. [CrossRef]

[27] Guan, W., Wang, W., Dong, J., & Peng, B. (2024). Improving generalization of deepfake detectors by imposing gradient regularization. *IEEE Transactions on Information Forensics and Security, 19*, 5345-5356. [CrossRef]

[28] Mittal, T., Bhattacharya, U., Chandra, R., Bera, A., & Manocha, D. (2020). Emotions don't lie: An audio-visual deepfake detection method using affective cues. In *Proceedings of the 28th ACM International Conference on Multimedia* (pp. 2823–2832). [CrossRef]

[29] Montserrat, D. M., Hao, H., Yarlagadda, S. K., Baireddy, S., Shao, R., Horváth, J., ... & Delp, E. J. (2020, June). Deepfakes Detection with Automatic Face Weighting. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops* (*CVPRW*) (pp. 2851-2859). IEEE. [CrossRef]

[30] De Lima, O., Franklin, S., Basu, S., Karwoski, B., & George, A. (2020). Deepfake detection using spatiotemporal convolutional networks. *arXiv preprint arXiv:2006.14749*.

[31] Rössler, A., Cozzolino, D., Verdoliva, L., Riess, C., Thies, J., & Niessner, M. (2019, October). FaceForensics++: Learning to Detect Manipulated Facial Images. In *2019 IEEE/CVF International Conference on Computer Vision* (*ICCV*) (pp. 1-11). IEEE. [CrossRef]

[32] Li, Y., Yang, X., Sun, P., Qi, H., & Lyu, S. (2020, June). Celeb-DF: A Large-Scale Challenging Dataset for DeepFake Forensics. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*) (pp. 3204-3213). IEEE. [CrossRef]

[33] Haliassos, A., Vougioukas, K., Petridis, S., & Pantic, M. (2021). Lips don't lie: A generalisable and robust approach to face forgery detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 5039–5049). [CrossRef]

[34] Chen, S., Yao, T., Chen, Y., Ding, S., Li, J., & Ji, R. (2021, May). Local relation learning for face forgery detection. In *Proceedings of the AAAI conference on artificial intelligence* (Vol. 35, No. 2, pp. 1081-1088). [CrossRef]

[35] Li, L., Bao, J., Zhang, T., Yang, H., Chen, D., Wen, F., & Guo, B. (2020, June). Face X-Ray for More General Face Forgery Detection. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*) (pp. 5000-5009). IEEE. [CrossRef]

[36] Sun, K., Yao, T., Chen, S., Ding, S., Li, J., & Ji, R. (2022, June). Dual contrastive learning for general face forgery detection. In *Proceedings of the AAAI conference on artificial intelligence* (Vol. 36, No. 2, pp. 2316-2324). [CrossRef]

[37] Tan, M., & Le, Q. (2019, May). Efficientnet: Rethinking model scaling for convolutional neural networks. In *International conference on machine learning* (pp. 6105-6114). PMLR.

[38] Zhao, H., Wei, T., Zhou, W., Zhang, W., Chen, D., & Yu, N. (2021, June). Multi-attentional Deepfake Detection. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*) (pp. 2185-2194). IEEE. [CrossRef]

[39] Zhuang, W., Chu, Q., Tan, Z., Liu, Q., Yuan, H., Miao, C., ... & Yu, N. (2022, October). UIA-ViT: Unsupervised inconsistency-aware method based on vision transformer for face forgery detection. In *European conference on computer vision* (pp. 391-407). Cham: Springer Nature Switzerland. [CrossRef]

[40] Wei, J., Wang, S., & Huang, Q. (2020, April). F^{3}Net: fusion, feedback and focus for salient object detection. In *Proceedings of the AAAI conference on artificial intelligence* (Vol. 34, No. 07, pp. 12321-12328). [CrossRef]

[41] Li, D., Yang, Y., Song, Y. Z., & Hospedales, T. (2018, April). Learning to generalize: Meta-learning for domain generalization. In *Proceedings of the AAAI conference on artificial intelligence* (Vol. 32, No. 1). [CrossRef]

[42] Chollet, F. (2017, July). Xception: Deep Learning with Depthwise Separable Convolutions. In *2017 IEEE Conference on Computer Vision and Pattern Recognition* (*CVPR*) (pp. 1800-1807). IEEE. [CrossRef]

[43] Sun, K., Liu, H., Ye, Q., Gao, Y., Liu, J., Shao, L., & Ji, R. (2021, May). Domain general face forgery detection by learning to weight. In *Proceedings of the AAAI conference on artificial intelligence* (Vol. 35, No. 3, pp. 2638-2646). [CrossRef]

[44] Usman, M. T., Khan, H., Singh, S. K., Lee, M. Y., & Koo, J. (2024). Efficient deepfake detection via layer-frozen assisted dual attention network for consumer imaging devices. *IEEE Transactions on Consumer Electronics*. [CrossRef]

Farhan Ali is a graduate student currently pursuing a Master's in Computer Science with a specialization in Data Science at Technische Universität Graz, Austria. He has a strong academic background and hands-on experience in data science, machine learning, and computer vision. He worked as an Associate Software Engineer at OpusAI, where he was involved in building user-centric web applications. In addition, he completed data science internships with Oasis Infobyte and Info AidTech, respectively, gaining valuable experience.

Zainab Ghazanfar received her M.S. degree in Computer Science from the University of Lahore. She served as a Lecturer in the Department of Computer Science at the University of Management and Technology (UMT), Lahore. She has also held a lecturer position in the Department of Computer Science at Lahore Garrison University. Currently, she is pursuing a Ph.D. in the Department of Software and Artificial Intelligence at Gachon University, South Korea. Her research interests include deep learning, medical image processing, image recognition, and the application of artificial intelligence in medical images.