IETE Technical Review

ISSN: 0256-4602 (Print) 0974-5971 (Online) Journal homepage: www.tandfonline.com/journals/titr20

GLAD-ViT: Global–Local Attention Duality in Vision Transformers for Deepfake Detection

Pawan Pandey, Arun Solanki & Sanjay Kumar Sharma

To cite this article: Pawan Pandey, Arun Solanki & Sanjay Kumar Sharma (09 Feb 2026): GLAD-ViT: Global–Local Attention Duality in Vision Transformers for Deepfake Detection, IETE Technical Review, DOI: 10.1080/02564602.2026.2621327

To link to this article: https://doi.org/10.1080/02564602.2026.2621327

Published online: 09 Feb 2026.

Submit your article to this journal

View related articles

View Crossmark data

Full Terms & Conditions of access and use can be found at
https://www.tandfonline.com/action/journalInformation?journalCode=titr20

IETE TECHNICAL REVIEW
https://doi.org/10.1080/02564602.2026.2621327

Check for updates

GLAD-ViT: Global–Local Attention Duality in Vision Transformers for Deepfake Detection

Pawan Pandey, Arun Solanki and Sanjay Kumar Sharma

Department of Computer Science and Engineering, SoICT, Gautam Buddha University, Gautam Budh Nagar, Greater Noida, UP201312, India

ABSTRACT

The rapid growth of deepfake creation techniques is an important challenge to digital assets and forensic analysis. Recent detection approaches show strong outcomes on visual recognition tasks. However, they exhibit notable limitations in accurately identifying deepfakes. To address these limitations, a novel GLAD-ViT (Global–Local Attention Duality in Vision Transformers) architecture is proposed for enhanced face forgery identification. It incorporates a dual attention mechanism that combines Multi-Head Attention (MHA) with a subsequent Convolutional Block Attention Module (CBAM) within the ViT encoder structure. The global context is modeled by MHA, then CBAM selectively refines local representations through spatial and channel attention mechanisms. The proposed architecture is assessed on several benchmarked deepfake datasets and achieves accuracies of 90.11%, 95.80%, 84.27%, 97.27%, and 99.84% on the DFDC, Celeb-DF (v2), FF++ (F2F), FF++ (FSh), and DF-TIMIT datasets, respectively. Ablation studies confirm the effectiveness of attention-based enhancements across various ViT configurations. Furthermore, the proposed model exhibits robustness against a variety of standard image degradations and significant performance improvement over state-of-the-art methods. This global–local attention duality introduced by GLAD-ViT contributes to advancing deepfake forensics toward more trustworthy media authentication.

KEYWORDS

CBAM; Channel attention; Deepfake detection; Dual attention; Spatial attention; Vision transformer

1. INTRODUCTION

Deepfake technology utilizes advanced Artificial Intelligence (AI) technologies to manipulate highly realistic synthetic media. It allows the generation of synthetic content that can alter voices, faces, or whole video clips [1]. These support artistic or entertainment work, such as movies, video games, and other entertainment projects. But they also produce significant harms to digital safety, personal freedom, and the spread of fake news. Audio and video deepfakes can be extremely convincing, and it becomes difficult to distinguish manipulation from genuine content. This results in severe threats such as scams, damage to reputation, and compromise of personal and public security [2]. Deepfakes can be used by hackers to deceive people into sharing their passwords, funds, or confidential information. This makes it easier for attackers to carry out fraud and other cybercrimes. Deepfakes are also capable of producing fake videos or audio of individuals without their consent. Such misuse may lead to emotional harm, damage reputations, and exposure of confidential information. Deepfakes can also employed to circulate rumors or misinformation that seems real. This misinformation has the potential to create misunderstandings, harm communities, and affect important decisions in society. So, Deepfake detection is highly important for media authentication, addressing digital fraud, and maintaining public trust to multimedia content.

Autoencoders [3] and Generative Adversarial Networks (GANs) [1] are two primary deep-learning architectures used for generating deepfake multimedia objects. Deepfake creation relies on several types of facial manipulation techniques. One of the most commonly used is Face Swapping. An individual's face is substituted with another while maintaining facial cues and movements. Another Face Reenactment technology exchanges the facial expressions and movements of one individual with another [4]. Some artificial editing can be performed on a person's smile or emotions, called Face Morphing [5]. In Identity Manipulation, the entire face is synthesized by deep learning techniques that do not belong to real people. Lip movements can also be changed to match a given audio track while keeping the rest of the face unchanged. Alterations in some attributes, like age, gender, hair, and skin tone, can also be done for age progression and gender transformation. Some aging or de-aging effects [6] produce wrinkles and inconsistencies in skin texture.

© 2026 IETE

Therefore, there is a necessity for an efficient and trustworthy deepfake detection system to identify such facial manipulations.

1.1 Motivation

Traditional CNN-based methods [7,8] are efficient in capturing local structural and textural patterns but inadequate in detecting global facial inconsistencies. It leads to incorrect predictions when manipulations are subtle and dispersed throughout the face. Conversely, Standard ViTs [9] are very efficient at capturing global dependencies among facial image patches. However, they often fail to detect fine-grained features such as inconsistency of boundaries, traces of blending, or mismatches in lighting. These features are essential in detecting synthetic media. Recent hybrid models like GenConViT [10] combine a convolution-based network with the Swin Transformer. It has an accuracy of 90.94% on the Celeb-DF (v2) dataset. Another, temporal Bi-LSTM + MHSA approach [11] also shows comparatively lower accuracy on the DFDC and Celeb-DF (v2) datasets. Some attention-based ViT models [12] attempt to overcome these limitations by incorporating additional attention mechanisms, such as Squeeze-and-Excitation (SE) [13] or Convolutional Block Attention Module (CBAM) [14]. These models typically apply the additional attention mechanism either before patch embeddings or outside the traditional transformer encoder layered structure. It filters raw embeddings at an early stage and weakens the model's ability to capture global facial inconsistencies. These deficiencies promote the need for a unified architecture that can easily accommodate both global dependencies and localized refinement within the transformer encoder structure. The proposed GLAD-ViT architecture is motivated by the "global-first, local-refinement" hypothesis. It first captures global dependencies and then refines the local artifacts to target more proficiently on deepfake-relevant artifacts.

1.2 Novelty and Key Contributions

Unlike existing convolution-heavy hybrids like GenConViT [10] or attention-based ViT models that apply attention modules externally only on 25 facial landmarks [12], the proposed GLAD-ViT introduces a dual-attention mechanism within each transformer encoder layer to process the entire facial region. In the proposed method, CBAM is integrated immediately after the Multi-Head Attention (MHA) block. MHA captures long-range dependencies from diverse facial patches, while CBAM highlights the most informative localized channel and spatial features. This allows for attention refinement while keeping global dependencies intact. Experimental conclusions show that the proposed architecture presents robust outcomes on standard datasets and achieves higher accuracy than existing deepfake detection systems. The primary contributions of this work are as follows:

- GLAD-ViT, a novel dual attention Vision Transformer architecture integrating CBAM after MHA within each encoder block, is proposed and implemented.
- A comprehensive performance evaluation is conducted on standard benchmark datasets along with compressed, noisy, and cross-domain datasets to validate robustness and generalization.
- An ablation study is performed to compare the proposed architecture with different CBAM placements and other attention mechanisms.
- The paper also presents an assessment of the proposed methodology against some baseline and state-of-the-art deepfake detection strategies.

Following the introduction in Section 1, the remaining paper is arranged as follows. Section 2 demonstrates related work, and Section 3 illustrates the proposed GLAD-ViT architecture to determine whether facial features are real or fake. Section 4 presents experimental work. In Section 5, the experimental findings are discussed. Finally, Section 6 outlines the conclusion & future scope of this study.

2. RELATED WORK

Here, various deepfake detection methods are discussed by categorizing them based on their underlying methodologies and key contributions.

2.1 Early Period

Early methods of deepfake detection used hand-crafted facial features to detect visual inconsistencies. Li *et al.* [15] designed a model that utilized eye flickering patterns in visual media. Matern *et al.* [16] applied an MLP classifier for deepfake detection with AUC values up to 0.866 through visual artifact inspection, *e.g.* eyes and teeth. Yang *et al.* [17] used an SVM classifier for the categorization of computed 3D head postures from facial photos. Jung *et al.* [18] obtained 87.5% accuracy by monitoring the alterations in the eye flickering patterns of visual data. Xu *et al.* [19] focused on texture properties by calculating the gradient domain, standard deviation, GLCM, and wavelet transform using an SVM classifier. These methods provided a foundation for deepfake detection. However, they had some scalability and robustness

issues, which led to the utilization of deep learning-based approaches.

2.2 Transitional Period

Researchers in deepfake detection have also explored several deep learning techniques. Recurrent Neural Networks (RNN) and Long Short-Term Memory (LSTM) are used for sequential data analysis. Convolutional Neural Networks (CNNs) are applied for texture pattern recognition. In addition, some advanced deep learning approaches, such as Graph Neural Networks (GNNs), Texture and Artifact Detectors (TAD), and Federated Deep Learning models, are employed to enhance detection performance.

Lin *et al.* [20] employed RNN & LSTM to identify fake videos by detecting inconsistencies between the successive frames with an improved accuracy of around 90%. Singh *et al.* [21] utilized spatio-temporal properties by identifying region-based inconsistencies across frames. This study was performed on the DeepFake Detection Challenge dataset with an accuracy score of 97.6%. Fei *et al.* [22] presented a motion discrepancy-based method. This method amplifies face motions to reveal distortions using CNN & LSTM on the FF++ dataset. Although these approaches enhanced the performance by including temporal features but they still faced challenges in capturing detailed spatial features. This led to the use of CNN-based architectures for more robust deepfake detection.

Kingra *et al.* [7] examined the Local Binary Pattern (LBP) of digitally manipulated images using a CNN-based network. LBPNet gets the accuracy of 80.04% on DFDC and 92.38% on the Celeb-DF (v2) datasets. Khalifa *et al.* [8] implemented multiple Gabor filters, generating linear, elliptical, and circular filters with Convolutional Neural Networks for deepfake recognition. Khalid *et al.* [23] proposed a pyramid-structured graph neural network for interpretable and generalizable deepfake detection. Heidari *et al.* [24] presented a blockchain-enabled federated learning architecture (BFLDL) combining SegCaps, CNNs, and capsule networks. Gao *et al.*[25] discussed Texture and Artifact Detector (TAD), applying texture-artifact synergy and ensemble learning. This method improves deepfake detection and generalization across diverse datasets and forgery types. These studies show significant progress in deepfake detection. However, many existing models still face limitations in capturing global dependencies and have less adaptability to diverse manipulations. This motivates the exploration of attention mechanisms and Vision Transformers (ViTs) hereafter.

2.3 Contemporary Period

The attention mechanism was suggested by Bahdanau *et al.* in 2015 [26]. It was later adapted to the transformer model by Vaswani *et al.* [27]. Dosovitskiy *et al.* [9] introduced the concept of ViT for image processing rather than Traditional CNNs in 2021. Heo *et al.* [28] proposed an enhanced ViT by combining vector-concatenated CNN features with patch & positional embeddings. Kumar *et al.* [29] implemented a Transformer Technique with Self-Attention Mechanism (T2SAM). Dagar *et al.* [30] combined ResNet-18 with a global texture extraction module based on Gram matrices and local binary patterns. This approach was incorporated within a cross-attention vision transformer. Ramadhani *et al.* [12] utilized Depthwise Separable Convolution and CBAM outside the transformer encoder layer on 25 facial landmarks of the input video. Deressa *et al.* [10] introduced a deepfake detection model, GenConViT, combining ConvNeXt, Swin Transformer, Autoencoder, and Variational Autoencoder. Xiong *et al.* [11] extracted 68 facial landmarks of video frames and implemented a Bi-LSTM with the multi-head self-attention mechanism (MHSA).

In summary, deepfake detection has evolved from CNN-based methods to transformer models with hybrid structures and attention mechanisms. A detailed comparison of existing deepfake detection approaches is discussed in Table 1. Based on the key findings and limitations of existing approaches, the GLAD-ViT architecture is proposed to increase the effectiveness and stability of deepfake detection systems.

3. PROPOSED ARCHITECTURE

This section outlines the proposed GLAD-ViT architecture for deepfake identification, as shown in Figure 1. The architecture is divided into four modules, *i.e.* Preprocessing Module, Patch & Position Embedding, GLAD-ViT Encoder, and MLP Head for final classification. A brief overview of each module is provided below:

3.1 Preprocessing Module

The preprocessing module is divided into four submodules as presented in Figure 1. First, Frames from each input video are captured using OpenCV based on uniform interval sampling. Then, the MTCNN (Multi-Task Cascaded Convolutional Networks) algorithm [31] is utilized to crop facial regions, due to its high precision in localizing facial features and identifying facial landmarks. It enables the model to emphasize only the essential attributes. The face region is resized to $\mathbb{R}^{72 \times 72 \times 3}$, ensuring uniform data input dimensions for

Table 1: Comparison of existing deepfake detection approaches

Study	Method/Approach	Datasets used	Key findings	Limitations
Li *et al.* [15]	CNN, LRCN LSTM-RNN	CEW	Detects fake face videos by analyzing eye blinking.	Need to explore other patterns beyond eye blinking.
Matern *et al.* [16]	MLP, Log. Regression	Deepfake, Face2Face	Detects deepfakes by analyzing eyes & teeth with AUC values up to 0.866.	Limited to certain artifacts, eyes & teeth.
Yang *et al.* [17]	SVM classifier	UADFV, DARPA-GAN	Analyze inconsistencies in 3D head poses from 68 facial landmarks	Difficulty in finding facial landmarks in blurry faces.
Jung *et al.* [18]	GAN	Deepfake	Analyze the human eye blinking pattern with an accuracy of 87.5%.	Need to explore more facial features.
Xu *et al.* [19]	SVM	DF-TIMIT, Celeb-DF, FF++, DFDC	Features are computed using the gradient domain, standard deviation, GLCM, and wavelet transform.	More informative texture features can be derived to improve the performance.
Lin *et al.* [20]	CNN, LSTM	Trump-Cage	Use CNN for LBP Histograms, then LSTM to check inconsistencies between successive frames.	A dense CNN architecture with LBP images can be implemented to increase the accuracy.
Singh *et al.* [21]	CNN, LSTM	DFDC	Analyze the spatio-temporal features of videos by taking sequences of frames with an accuracy of 97.60%.	The model can also be evaluated on other datasets to produce generalizability.
Fei *et al.* [22]	CNN, LSTM	FF++ (DF, F2F, NT, FS)	Use a deep CNN and LSTM on video-extracted frames.	Need for generalization on different datasets.
Kingra *et al.* [7]	CNN	DFDC, Celeb-DF, Deeper-Forensics, FF++ (DF, F2F, NT, FS)	Used LBP for feature extraction and CNN for classification.	Frame-level texture inconsistency can be expanded to temporal sequences.
Khalifa *et al.* [8]	CNN	DFDC, Celeb-DF(v2), Wilddeepfake, FF++ (DF, F2F, NT, FS)	Use Gabor Filters, DSLRFN to reduce the model size by 64.9%.	Computational costs like inference time & training time can also be analyzed.
Khalid *et al.* [23]	Graph Convolutional Neural Network	DFDC, Celeb-DF(v2), WLRD, FF++ (DF, F2F, NT, FS, FSh)	Implements pyramid ResNet structure with GraphNet and FFN.	Need to assess robustness on synthetic images.
Heidari *et al.* [24]	SegCaps, CNN, Capsule Network	DFDC, Celeb-DF, DF-TIMIT, FF++	Introduce SegCaps with CNN for feature extraction and a capsule network to improve generalization.	Need real-time deepfake detection for both audio-visual data.
Gao *et al.*[25]	TAD	FF++ (DF, F2F, NT, FS, FSh)	Proposed a single-model and ensemble TAD to improve cross-dataset generalization.	Need to improve generalization across different video formats.
Heo *et al.* [28]	CNN, ViT	DFDC, Celeb-DF(v2)	Combine the vector-concatenated CNN feature with the Vision Transformer to capture global dependency.	Spatial & temporal features can be combined for robust deepfake detection.
Dagar *et al.* [30]	ResNet-18, ViT	DFDC, Celeb-DF(v2), FF++ (DF, F2F, NT, FS)	Features are extracted with ResNet-18 and then used cross-attention mechanism.	Limited robustness on noisy datasets.
Ramadhani *et al.* [12]	ViT, DSC, CBAM	DFDC, Celeb-DF(v2), DF-TIMIT, FF++	Used Vision Transformer with Depthwise Separable Convolution and CBAM.	Need to implement CBAM within the transformer encoder layer to improve detection accuracy.
Deressa *et al.* [10]	Gen. Con. ViT	DFDC, Celeb-DF(v2), FF++	Utilizes Autoencoders to learn from latent data.	Need to incorporate additional attention for better generalization.
Xiong *et al.* [11]	BiLSTM, MHSA	DFDC, Celeb-DF(v2), UADFV, FF++	Implemented BiLSTM with MHSA to enhance deepfake detection.	Accuracy on the DFDC, Celeb-DF datasets is comparatively low.

the architecture. To enhance the architecture's ability on unseen data and reduce overfitting, some augmentation techniques (horizontal flip, random rotation, random zoom) are also implemented.

3.2 Patch and Position Embedding

After Preprocessing, the suggested architecture requires splitting input images into a set of unoverlapping patches

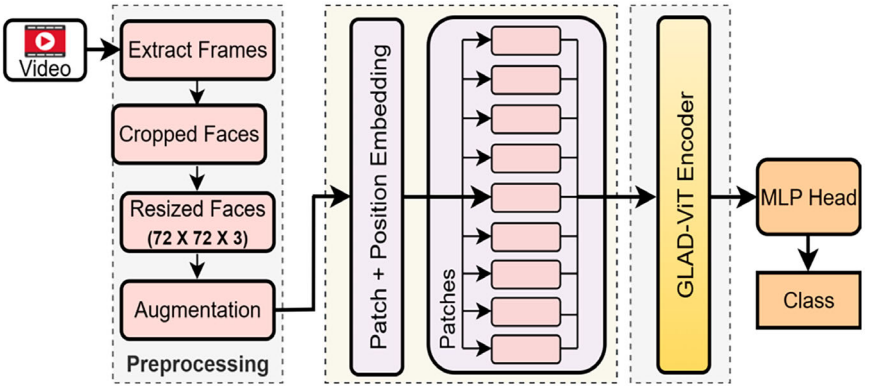


Figure 1: Proposed architecture for deepfake classification

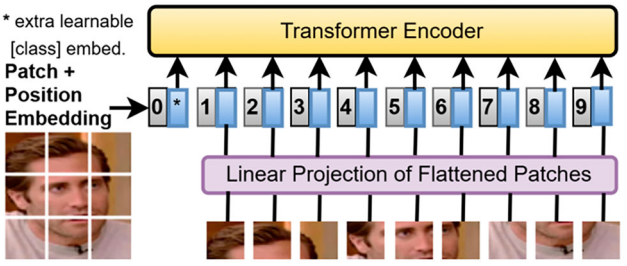


Figure 2: General form of patch and position embedding in conventional ViT

$\{p_1, p_2, \ldots, p_n\}$. A total of 144 patches, each of size $\mathbb{R}^{6\times6\times3}$, are generated from an augmented face. This provides a sufficient number of patches to extract fine spatial features and keeps the sequence length manageable for efficient training. These small patches are flattened into vectors and subsequently linearly projected into a high-dimensional space for generating patch encodings. To identify the respective arrangement of patches p_i, positional embeddings $E_{pos}(i)$ are combined with encoded patches as shown in Figure 2. Then feed the sequence X into the GLAD-ViT Encoder.

$$X = [x_1, x_2, \ldots, x_{144}],\ x_i = E(p_i) + E_{pos}(i) \quad (1)$$

In this implementation, the entire sequence of encoded patch embeddings is input to the GLAD-ViT Encoder. This allows the architecture to collectively process information from all patches instead of solely depending on an extra learnable class embedding.

3.3 GLAD-ViT Encoder

The GLAD-ViT Encoder serves as the backbone of the proposed architecture. Its key design principle involves integrating the CBAM directly after the MHA block, as illustrated in Figure 3. MHA computes self-attention responses through linear projections to capture global dependencies among facial patches. However, it often

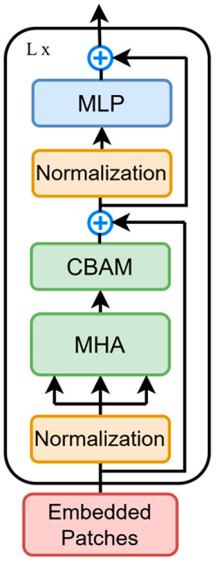


Figure 3: GLAD-ViT Encoder

does not detect minor fine-grained patterns like facial boundaries, light variations, and texture distortions. CBAM adjusts the weights of these local features to focus on important regions and ignore unnecessary information. This hierarchical integration improves feature representation, gradient flow, and captures fine-grained deepfake artifacts in the model. Placing CBAM before MHA degrades the learning of global relationships because it disrupts the statistical uniformity of patch embeddings and biases the self-attention mechanism.

The embedded patch sequence X is normalized to X′ using Layer Normalization and passed to the MHA layer for capturing global context among all patches throughout the entire image.

3.3.1 Multi-Head Attention (MHA)

MHA in ViT enables simultaneous attention to different patches of the input face by establishing multiple

attention heads. It calculates the attention scores by computing query (Q), keys (K), and values (V) vectors of dimension d_k from the normalized patch sequence X′. The query Q is dot-multiplied with all keys K^T to measure the similarity between queries & keys, and scaled by $\sqrt{d_k}$ [27]. Then, pass through a softmax to generate attention weights for the values and multiply by V to aggregate values based on attention weights.

$$Q = X'W^Q,\ K = X'W^K,\ V = X'W^V \tag{2}$$

$$Attention\ (Q,\ K,\ V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{3}$$

This is performed in parallel across multiple attention heads, and each head i has its own projection matrices W_i^Q, W_i^K, W_i^V. Then output from all attention heads (h) is concatenated for the final MHA output.

$$head_i = Attention\ (Q_i,\ K_i,\ V_i) \tag{4}$$

$$\text{MHA}\ (X') = \text{Concat}(\text{head}_1, \ldots, \text{head}_h)W^O \tag{5}$$

Where the projections are defined by parameter matrices:

$$W_i^Q \in \mathbb{R}^{d_{model} \times d_k},\ W_i^K \in \mathbb{R}^{d_{model} \times d_k},$$
$$W_i^V \in \mathbb{R}^{d_{model} \times d_v},\ W^O \in \mathbb{R}^{hd_v \times d_{model}}$$

The final multi-head attention output MHA (X′) is represented as F and passed to the CBAM module.

3.3.2 Convolutional Block Attention Module (CBAM)

CBAM applies attention in two sequential steps: identifying "what to focus" (Channel Attention) and "where to focus" (Spatial Attention) [14]. The input feature map F is multiplied element-wise by channel attention M_C, producing the channel-refined feature map F'_C. Next, F'_C is again multiplied element-wise by spatial attention M_S to find the final refined output feature map F'_S. The complete architecture of CBAM is shown in Figure 4, and the overall CBAM process is outlined as follows:

$$F'_C = M_C \otimes F \tag{6}$$

$$F'_S = M_S \otimes F'_C \tag{7}$$

A. Channel Attention

Throughout this attention module, the features acquire distinct weights through which essential channels become more prominent. In the channel attention module, Global Average Pooling (GAP) and Global Max Pooling (GMP) are utilized across the source features $F \in \mathbb{R}^{C \times H \times W}$ as shown in Figure 5(a). Both pooled outputs are passed through a common MLP to minimize the parameters and assist with regularization. Then, add both

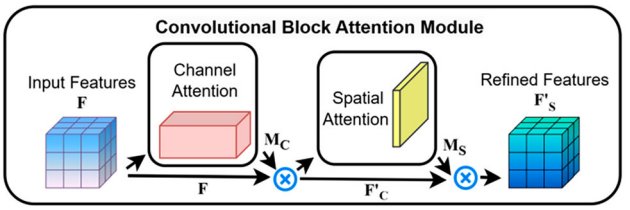


Figure 4: CBAM Architecture [14]

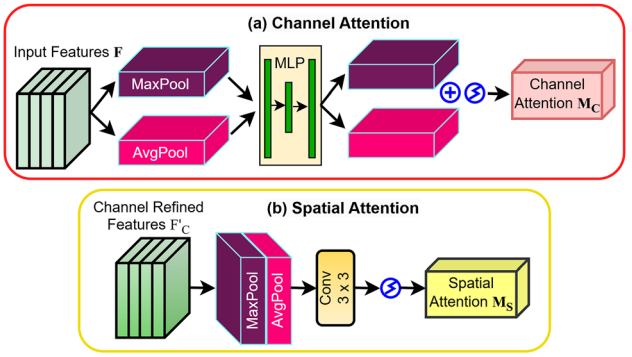


Figure 5: Channel and Spatial Attention [13]

output element-wise and use sigmoid activation to calculate the channel attention M_C. This procedure focuses on highlighting the most important channels or feature dimensions within the projection or feature map.

$$M_C = \sigma\ (\text{MLP}(\text{GAP}(F)) + \text{MLP}(\text{GMP}(F))) \tag{8}$$

B. Spatial Attention

A spatial attention map is generated through Max Pooling and Average Pooling computation executed along the channel axis by preserving key spatial information. The combined set of feature maps F'_C enters into a convolutional layer of kernel size 3 × 3, subsequently a sigmoid activation to yield a spatial attention M_S as in Figure 5(b). This operation highlights the most informative regions within every token embedding or patch.

$$M_S = \sigma\ (\text{Conv}_{3\times3}([\text{MaxPool}(F'_C); \text{AvgPool}(F'_C)])) \tag{9}$$

The CBAM output $F'_S$is added to the embedded patch sequence X for combining attention-refined information into the transformer's input. This combined representation X_1 is again normalized and proceeds to the MLP Layer. The MLP output is residually added with X_1, preserving both the original and attention-enhanced information. The final encoder output is obtained by repeating this process across L encoder layers. Here, 8 encoder layers are used to balance between model performance and computational cost with moderate-sized datasets of $\mathbb{R}^{6\times6\times3}$ patch size derived from $\mathbb{R}^{72\times72\times3}$ facial images.

3.4 MLP Head

Finally, the encoder output from the GLAD-ViT Encoder Block proceeds to an MLP head. It includes a final dense layer with sigmoid activation to produce the class prediction (Real or Fake). The entire process is described in Pseudocode 1 of the implemented GLAD-ViT architecture.

Pseudocode 1: GLAD-ViT

Input: Videos from the deepfake dataset
Output: output_class (*Real or Fake)*

```
# Step 1: Preprocessing
frames = extract_frames(videos, num_frames)
cropped_faces = MTCNN(frames)
resized_faces = resize_images(cropped_faces, size = (72,72,3))
augmented_faces = data_augmentation (resized_faces)

# Step 2: Patch and Position Embedding
patches = create_patches(augmented_faces, patch_size = (6,6,3))
encoded_patches = linear_projection(patches, output_dim = d)
position_ embedding = create_positional_embedding (
                      num_patches = 144, proj_dim = d)
input_patches = encoded_patches + position_ embedding

#Step 3: GLAD-ViT Encoder Block
X = input_patches
for i in range(L = 8):
    norm_patches = Layer_Norm (X)
    attention_output = MHA(norm_patches, num_heads = 4,
                       proj_dim = d)
    cbam_output = CBAM(attention_output, channels = d)
    X1 = cbam_output + X
    X2 = Layer_Norm(X1)
    X = MLP(x2) + X1
encoder_output = X

#Step 4: Classification
Y = Layer_Norm (encoder_output)
features = MLP (Y)
output_class = sigmoid(features)
```

4. EXPERIMENTAL SETUP

The novel architecture discussed in Section 3 is implemented on five datasets of different sizes. The corresponding implementation details are as follows:

4.1 Datasets Description

In this study, the implementation is performed on five benchmarked datasets, such as DF-TIMIT [32], DFDC [33], Celeb-DF (v2) [34], FF++ (F2F) [35], and FF++ (FSh) [35] of different sizes. As these datasets consist of videos, the frames are extracted from each one for further processing. Table 2 outlines the process of extracting facial images from the video datasets. The DF-TIMIT dataset is built on the VID-TIMIT dataset [36], containing 320 real and deepfaked videos. Here, 15 frames or faces are extracted from each video for classification purposes. Thereafter, the used dataset has 4,800 real & 4,800 fake face images. The DFDC dataset is very huge and organized into 50 parts of fake videos. Each part contains approximately 2,000 fake videos. Due to the large dataset's size and the associated computational cost, a randomly selected subset of 600 real and fake videos is used for computation. To maintain diversity, sampling was performed across different dataset partitions. To balance the dataset, 25 frames from each video are computed. Hence, the utilized dataset contains 15,000 real & 15,000 fake face images. Celeb-DF (v2) has 590 real and 5639 deepfake videos. This study uses 590 real and fake videos, with 30 frames cropped from an individual video. The employed dataset contains 17,700 authentic & 17,700 synthetic face images. In the Face2Face (F2F) and FaceShifter (FSh) datasets of FaceForensics++, 30 frames are extracted from each of the 1,000 videos. The utilized dataset has 30,000 real & 30,000 fake face images. Train, Val & Test ratio is 70%, 10% & 20% for DF-TIMIT, while 60%, 10% & 30% for rest of the datasets used.

4.2 Implementation Specifics

All experiments are conducted in a virtualized environment on an NVIDIA DGX A100 machine with 20 GB GPU configured using Multi-Instance GPU (MIG) architecture. The suggested architecture is implemented using the TensorFlow machine learning library. The Adam optimizer is applied during training at a learning rate of 0.0001 for FF++ (FSh) and 0.001 for other datasets with the sigmoid activation function. Dropout of 0.1 in Transformer layers & 0.5 before the classification head is applied to mitigate overfitting and improve generalization. The network is trained for 50 epochs using binary cross-entropy as the loss function. Table 3 provides additional technical details, including model hyperparameters and training configurations of the implemented architecture.

The computational performance comparison of the proposed architecture with some baseline methods is summarized in Table 4. All these methods are implemented using the same architectural parameters as discussed in Table 3. The training time and average inference time are calculated on the largest dataset, FF++ (FSh). The GLAD-ViT model contains 21,666,361 trainable parameters. It takes approximately 55 min for training and an average inference time of 0.51 milliseconds per image in a virtualized MIG environment on FF++ (FSh) dataset.

5. RESULTS & DISCUSSION

As depicted in Figure 7, the proposed architecture effectively learns from the facial features of various datasets and continually improves its performance with every

Table 2: Datasets description

	Total videos		Balancing required	Used videos		Frame count		Total faces	
Dataset	Real	Fake		Real	Fake	Real	Fake	Real	Fake
FF++ (F2F)	1,000	1,000	No	1,000	1,000	30	30	30,000	30,000
FF++ (FSh)	1,000	1,000	No	1,000	1,000	30	30	30,000	30,000
Celeb-DF (v2)	590	5,639	Yes	590	590	30	30	17,700	17,700
DFDC	19,154	100k	Yes	600	600	25	25	15,000	15,000
DF-TIMIT	320	320	No	320	320	15	15	4,800	4,800

Table 3: GLAD-ViT architecture parameters

Parameter	Value
Image Size	72
Patch Size	6
Projection Dimension	64
Transformer Encoder Layers	8
Attention Heads	4
Batch Size	64
Epochs	50

Table 4: Computational performance

Method	Trainable parameters	Training time	Average inference time
MesoInception-4	13,255	62 min	1.56 ms
EfficientNet-B0	4,008,829	61 min	1.34 ms
GLAD-ViT	21,666,361	55 min	0.51 ms

Table 5: Performance of the proposed architecture on each dataset

Dataset	Accuracy	Precision	Recall	F1-Score	AUC
DFDC	90.11	90.73	89.25	89.98	0.97
Celeb-DF (v2)	95.80	96.52	95.03	95.77	0.99
FF++ (F2F)	84.27	88.57	78.69	83.34	0.94
FF++ (FSh)	97.27	99.21	95.30	97.22	1.00
DF-TIMIT	99.84	99.69	100	99.84	1.00

Table 6: Ablation study method's description

Method	Description	Attention placement
Method 1	ViT (without CBAM)	No Attention Mechanism
Method 2	CBAM before Patch Embedding	Outside Encoder Module
Method 3	ViT with Channel Attention	Inside Encoder Module
Method 4	ViT with Squeeze-and-Excitation	Inside Encoder Module
Method 5	ViT with CBAM (GLAD-ViT)	Inside Encoder Module

epoch. Figure 7(a) showcases the variations in accuracy, and Figure 7(b) presents the variations in loss during each epoch on the validation portion of different datasets. During the training procedure, the architecture increasingly refines its understanding of identifying features between manipulated and real faces. The learning curves indicate stable learning and reliable performance of the GLAD-ViT architecture.

Performance of the GLAD-ViT architecture is assessed using metrics such as accuracy, F1-score, precision & recall by confusion matrix, and AUC value from ROC curves. The evaluation performance across different datasets is presented in Table 5, and a visual comparison of these metrics is shown in Figure 6. Specifically, achieves an accuracy of 90.11% on the DFDC, 95.80% on the Celeb-DF (v2), 84.27% on the FF++ (F2F), 97.27% on the FF++ (FSh), and 99.84% on the DF-TIMIT datasets. The architecture also attains an AUC of 0.97 on the DFDC, 0.99 on the Celeb-DF (v2), 0.94 on the FF++ (F2F), 1.00 on the FF++ (FSh), and 1.00 on the DF-TIMIT datasets, as illustrated in Figure 8. The results demonstrate consistent performance across multiple benchmarked datasets.

An ablation study is also performed to compare the results of the suggested architecture with different CBAM placements and other attention mechanisms, as illustrated in Table 6. Method 1 uses traditional ViT without CBAM. Method 2 integrates CBAM before patch embedding outside the encoder module. Method 3 utilizes ViT with only channel attention inside the encoder module, while Method 4 introduces a different Squeeze-and-Excitation (SE) mechanism at the same position. Finally, proposed Method 5 (GLAD-ViT) incorporates CBAM within the encoder module. The validation is performed on FF++ (F2F) and Celeb-DF (v2) datasets. These datasets include heterogeneous real-world deepfakes and more common manipulations. The empirical results are shown in Table 7. This emphasizes the effectiveness of using CBAM after MHA in the ViT encoder module for effective deepfake detection. However, placing CBAM before MHA significantly reduces performance. It reduces the accuracy from 84.27% to 50.01% on the FF++ (F2F) dataset, as observed in our study. The attention heatmap of MHA followed by CBAM is also shown in Figure 10. It visualizes the most influential patches that affect the model's predictions. MHA often spreads attention widely across the face, sometimes fixating on unimportant background areas. On the other hand, MHA with CBAM focuses more closely on important spots like eyes, mouth, and facial edges, where changes are most likely to happen.

To assess robustness, the performance of the proposed architecture is also assessed on compressed, noisy &

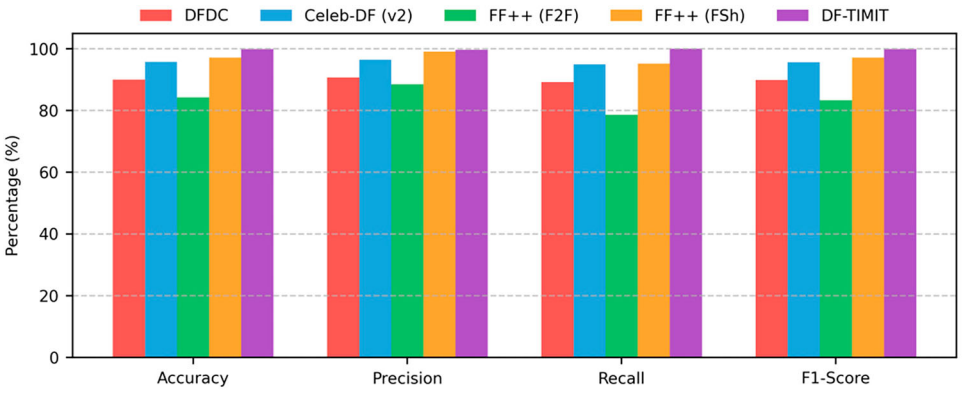


Figure 6: Performance metrics comparison for each dataset

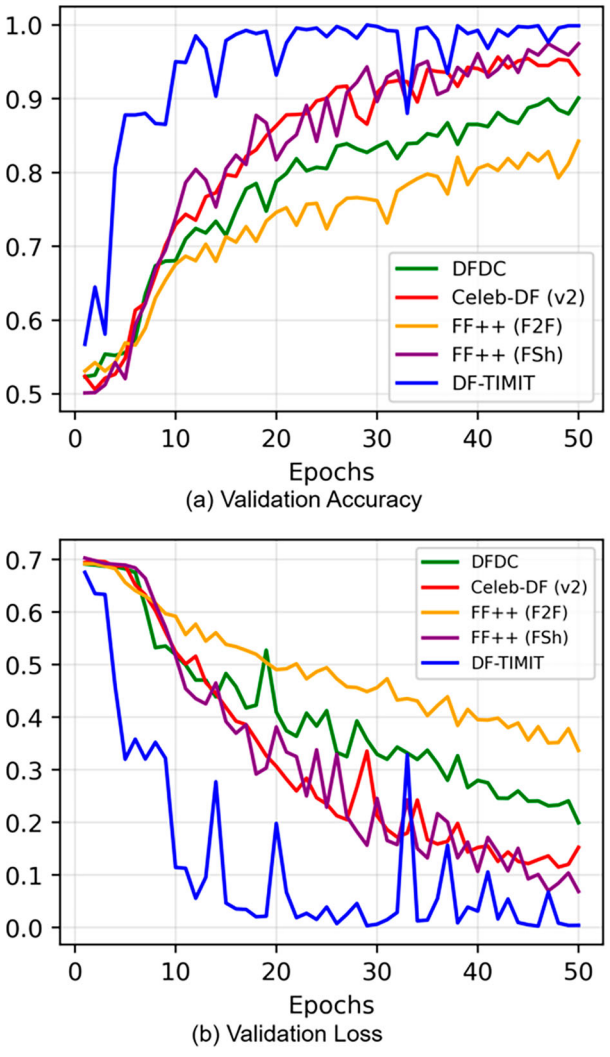


Figure 7: Validation accuracy and loss curve

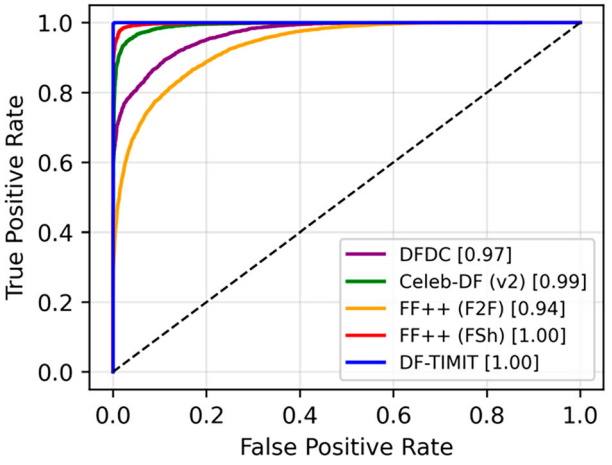


Figure 8: ROC curve of the proposed architecture on each dataset

Table 7: Results of ablation study

Method	FF++ (F2F)		Celeb-DF (v2)	
	Accuracy	F1-Score	Accuracy	F1-Score
Method 1	49.96	35.78	54.63	66.32
Method 2	50.17	18.53	56.89	63.00
Method 3	49.98	27.39	88.08	88.01
Method 4	50.02	66.67	94.85	94.86
Method 5	84.27	83.34	95.80	95.77

cross-domain test datasets. Experiments under compression and noise degradations are conducted using Gaussian blur with kernel sizes of 3 and 7, Gaussian noise with standard deviations of 5 and 10, and JPEG compression with quality factors of 23 and 40. The quantitative results from these evaluations are shown in Table 8. Figure 11 illustrates the comparative performance across these conditions with the original unaltered dataset. Cross-dataset generalization is also evaluated on each dataset pair by training on one dataset, such as Celeb-DF v2, and testing on another, like DFDC, using the same experimental setup. The inter-dataset generalization results are presented in Figure 9 with its limited performance. It is often seen in deepfake detection due to reliance on dataset-inherent artifacts like lighting changes, identity similarity, and pose variation. However, the proposed method

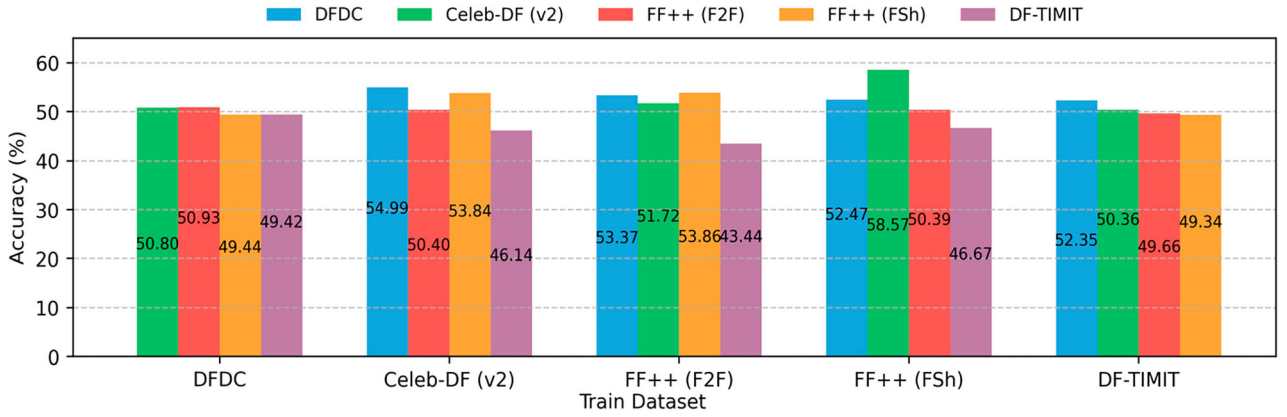


Figure 9: Inter-dataset generalization evaluation in terms of accuracy (%)

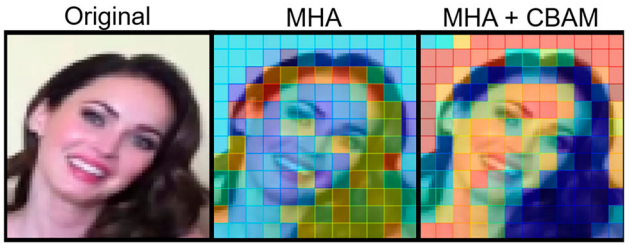


Figure 10: Attention heatmap visualization

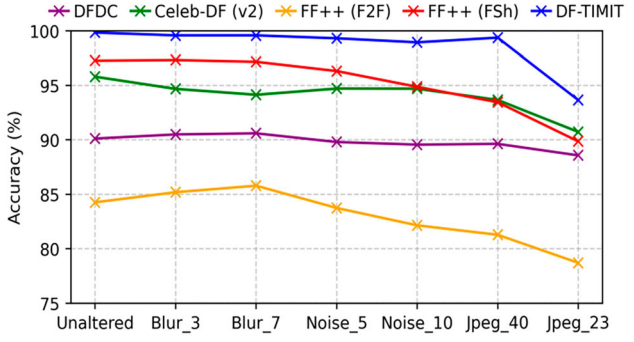


Figure 11: Performance comparison between unaltered and distorted image datasets

Table 8: Performance across different image distortions

	Gaussian blur		Gaussian noise		JPEG compress	
Dataset	3	7	5	10	40	23
DFDC	90.49	90.59	89.79	89.55	89.62	88.56
Celeb-DF (v2)	94.68	94.14	94.71	94.70	93.66	90.72
FF++ (F2F)	85.20	85.80	83.75	82.15	81.28	78.70
FF++ (FSh)	97.33	97.17	96.32	94.87	93.45	89.85
DF-TIMIT	99.58	99.58	99.31	98.94	99.37	93.65

exhibits strong robustness against various image degradations, including blur, noise, and JPEG compression.

To analyze the performance against existing baseline methods, experiments are conducted using two strong benchmark models: MesoInception-4 [37] and

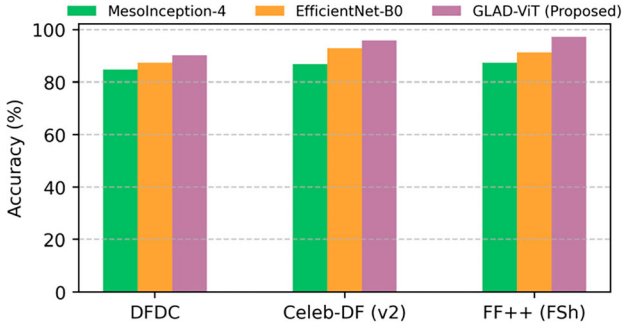


Figure 12: Accuracy comparison with baseline methods

Table 9: Comparison between GLAD-ViT and other baseline methods

Method	Dataset	Accuracy	F1-Score
MesoInception-4 [37]	DFDC	84.75	85.03
	Celeb-DF (v2)	86.82	87.32
	FF++ (FSh)	87.27	88.01
EfficientNet-B0 [38]	DFDC	87.32	87.17
	Celeb-DF (v2)	92.95	92.67
	FF++ (FSh)	91.31	91.72
GLAD-ViT	DFDC	90.11	89.98
	Celeb-DF (v2)	95.80	95.77
	FF++ (FSh)	97.27	97.22

EfficientNet-B0 [38]. The comparative analysis results are provided in Table 9. As we can see, GLAD-ViT achieves better performance in terms of both accuracy and F1-score. The accuracy comparison is also illustrated in Figure 12. These results indicate that GLAD-ViT detects deepfakes more accurately across these baseline methods.

A comparison is also performed between the GLAD-ViT and other state-of-the-art deepfake detection systems using Accuracy and AUC values (rounded to two decimal places), as shown in Table 10. "–" indicates that AUC

Table 10: Performance comparison of the proposed architecture with state-of-the-art methodologies

Method	Year	Dataset	Accuracy	AUC
LBPNet [7]	2022	DFDC	80.04	–
		Celeb-DF (v2)	92.38	–
DSLRFN [8]	2022	DFDC	72.74	0.84
		Celeb-DF (v2)	86.43	0.94
DFGNN [23]	2023	Celeb-DF (v2)	93.90	0.95
		FF++ (F2F)	62.49	0.69
BFLDL [24]	2024	DF-TIMIT	97.20	0.98
E-TAD [25]	2024	FF++ (FSh)	95.24	0.99
ViViT [12]	2024	DFDC	88.03	–
		Celeb-DF (v2)	87.17	–
		DF-TIMIT	94.14	–
GenConViT [10]	2025	Celeb-DF (v2)	90.94	0.98
		DF-TIMIT	98.28	1.00
BMNet [11]	2025	DFDC	84.72	0.89
		Celeb-DF (v2)	80.20	0.76
GLAD-ViT (Proposed)		DFDC	90.11	0.97
		Celeb-DF (v2)	95.80	0.99
		FF++ (F2F)	84.27	0.94
		FF++ (FSh)	97.27	1.00
		DF-TIMIT	99.84	1.00

was not reported in the source articles [7] [12]. Comparison results show that the proposed architecture demonstrates superior performance on the respective datasets compared to state-of-the-art methods.

6. CONCLUSION & FUTURE SCOPE

In this research, GLAD-ViT, a dual attention Vision Transformer leveraging global–local attention duality, is proposed and implemented for deepfake detection. The proposed method has shown impressive results in the experimental study on benchmarked deepfake datasets such as F2F (84.27%), FSh (97.27%) of FaceForensics++, DFDC (90.11%), Celeb-DF (95.80%), and DF-TIMIT (99.84%). The ablation study shows that GLAD-ViT gives better performance than other ViT variants incorporating various attention mechanisms. The main contribution of the work is not just using CBAM in Vision Transformer, but emphasizing the positioning of CBAM after MHA inside ViT encoder structure for high-quality deepfake detection. Despite limited cross-dataset capability, the method remains robust to blur, noise, and JPEG compression artifacts. It also outperforms several baseline and state-of-the-art detection methodologies. The demonstrated outcomes across a range of datasets showcase its flexibility and establish a robust foundation for multimedia authentication in forensic examination. Additionally, the proposed method can be extended for multi-modal analysis, incorporating facial movements with speech patterns to enhance accuracy.

DISCLOSURE STATEMENT

No potential conflict of interest was reported by the author(s).

REFERENCES

1. I. J. Goodfellow, *et al.* "Generative adversarial networks," 2014. Available: https://arxiv.org/abs/1406.2661.
2. Y. Mirsky, and W. Lee, "The creation and detection of deepfakes: a survey," *ACM Comput. Surv.*, Vol. 54, no. 1, Jan. 2021. DOI: 10.1145/3425780.
3. I. Goodfellow, Y. Bengio, and A. Courville. *Deep Learning*. Cambridge, MA: MIT Press, 2016.
4. J. Thies, M. Zollhöfer, M. Stamminger, C. Theobalt, and M. Nießner, "Face2Face," *Commun. ACM*, Vol. 62, no. 1, pp. 96–104, 2018. DOI: 10.1145/3292039.
5. A. Pumarola, A. Agudo, A. M. Martinez, A. Sanfeliu, and F. Moreno-Noguer, "GANimation: anatomically-aware facial animation from a single image," *Comput. Vis. – ECCV 2018*, 835–51, 2018. DOI: 10.1007/978-3-030-01249-6_50.
6. S. Liu, *et al.* "Face aging with contextual generative adversarial nets," in *Proc. 25th ACM Int. Conf. Multimed.*, 2017, pp. 82–90. DOI: 10.1145/3123266.3123431.
7. S. Kingra, N. Aggarwal, and N. Kaur, "LBPNet: exploiting texture descriptor for deepfake detection," *Forensic Sci. Int. Digit. Investig.*, Vol. 42-43, pp. 301452, 2022. DOI: 10.1016/j.fsidi.2022.301452.
8. A. H. Khalifa, N. A. Zaher, A. S. Abdallah, and M. W. Fakhr, "Convolutional neural network based on diverse gabor filters for deepfake recognition," *IEEE. Access.*, Vol. 10, pp. 22678–86, 2022. DOI: 10.1109/ACCESS.2022.3152029.
9. A. Dosovitskiy, *et al.* "An image is worth 16X16 words: transformers for image recognition at scale," in *ICLR 2021 - 9th Int. Conf. Learn. Represent.*, 2021. DOI: 10.48550/arXiv.2010.11929.
10. D. W. Deressa, H. Mareen, P. Lambert, S. Atnafu, Z. Akhtar, and G. Van Wallendael, "GenConViT: deepfake video detection using generative convolutional vision transformer," *Appl. Sci.*, Vol. 15, no. 12, pp. 6622, 2025. DOI: 10.3390/app15126622.
11. D. Xiong, Z. Wen, C. Zhang, D. Ren, and W. Li, "BMNet: enhancing deepfake detection through BiLSTM and multi-head self-attention mechanism," *IEEE. Access.*, Vol. 13, pp. 21547–56, 2025. DOI: 10.1109/ACCESS.2025.3533653.
12. K. N. Ramadhani, R. Munir, and N. P. Utama, "Improving video vision transformer for deepfake video detection using facial landmark, depthwise separable convolution and self attention," *IEEE. Access.*, Vol. 12, no. November 2023, pp. 8932–9, 2024. DOI: 10.1109/ACCESS.2024.3352890.
13. J. Hu, L. Shen, and G. Sun. "Squeeze-and-Excitation Networks," *in 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018, pp. 7132–7141. DOI: 10.1109/CVPR.2018.00745.

14. S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, *CBAM: Convolutional Block Attention Module*. Cham: Springer International Publishing, 2018. DOI: 10.1007/978-3-030-01234-2_1.

15. Y. Li, M. C. Chang, and S. Lyu. "In Ictu Oculi: exposing AI created fake videos by detecting eye blinking," in *10th IEEE Int. Work. Inf. Forensics Secur. WIFS 2018*, 2018, pp. 1–7. DOI: 10.1109/WIFS.2018.8630787.

16. F. Matern, C. Riess, and M. Stamminger. "Exploiting visual artifacts to expose deepfakes and face manipulations," in *Proc. – 2019 IEEE Winter Conf. Appl. Comput. Vis. Work. WACVW 2019*, 2019, pp. 83–92. DOI: 10.1109/WACVW.2019.00020.

17. X. Yang, Y. Li, and S. Lyu. "Exposing deep fakes using inconsistent head poses," in *ICASSP 2019 - 2019 IEEE Int. Conf. Acoust. Speech Signal Process.*, 2019, pp. 8261–8265. DOI: 10.1109/ICASSP.2019.8683164.

18. T. Jung, S. Kim, and K. Kim, "DeepVision: deepfakes detection using human eye blinking pattern," *IEEE. Access.*, Vol. 8, pp. 83144–54, 2020. DOI: 10.1109/ACCESS.2020.2988660.

19. B. Xu, J. Liu, J. Liang, W. Lu, and Y. Zhang, "DeepFake videos detection based on texture features," *Comput. Mater. Contin.*, Vol. 68, no. 1, pp. 1375–88, 2021. DOI: 10.32604/cmc.2021.016760.

20. Y.-K. Lin, and C.-Y. Chang. "Face forgery detection based on deep learning," in *Secur. with Intell. Comput. Big-Data Serv. 2019*, 2020, pp. 270–278. DOI: 10.1007/978-3-030-46828-6_23.

21. A. Singh, A. S. Saimbhi, N. Singh, and M. Mittal, "DeepFake video detection: a time-distributed approach," *SN Comput. Sci.*, Vol. 1, no. 4, pp. 212, Jun. 2020. DOI: 10.1007/s42979-020-00225-9.

22. J. Fei, Z. Xia, P. Yu, and F. Xiao, "Exposing AI-generated videos with motion magnification," *Multimed. Tools Appl.*, Vol. 80, no. 20, pp. 30789–802, 2020. DOI: 10.1007/s11042-020-09147-3.

23. F. Khalid, J. Ali, Q. Ain, H. Ilyas, and A. Irtaza, "DFGNN: an interpretable and generalized graph neural network for deepfakes detection," *Expert Syst. Appl.*, Vol. 222, pp. 119843, 2023. DOI: 10.1016/j.eswa.2023.119843.

24. A. Heidari, N. J. Navimipour, H. Dag, S. Talebi, and M. Unal, "A novel blockchain-based deepfake detection method using federated and deep learning models," *Cognit. Comput.*, Vol. 16, no. 3, pp. 1073–91, May 2024. DOI: 10.1007/s12559-024-10255-7.

25. J. Gao, *et al.*, "Texture and artifact decomposition for improving generalization in deep-learning-based deepfake detection," *Eng. Appl. Artif. Intell.*, Vol. 133, no. PC, pp. 108450, 2024. DOI: 10.1016/j.engappai.2024.108450.

26. D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," in *3rd Int. Conf. Learn. Represent.*, 2015, pp. 1–15. DOI: 10.48550/arXiv.1409.0473.

27. A. Vaswani, *et al.*, "Attention is all you need," *Adv. Neural Inf. Process. Syst.*, Vol. 2017-Decem, no. Nips, pp. 5999–6009, 2017. DOI: 10.48550/arXiv.1706.03762.

28. Y.-J. Heo, W.-H. Yeo, and B.-G. Kim, "DeepFake detection algorithm based on improved vision transformer," *Appl. Intell.*, Vol. 53, no. 7, pp. 7512–27, Apr. 2023. DOI: 10.1007/s10489-022-03867-9.

29. S. Kumar, and A. Solanki, "An abstractive text summarization technique using transformer model with self-attention mechanism," *Neural Comput. Appl.*, Vol. 35, no. 25, pp. 18603–22, Sep. 2023. DOI: 10.1007/s00521-023-08687-7.

30. D. Dagar, and D. K. Vishwakarma, "Tex-Net: texture-based parallel branch cross-attention generalized robust Deepfake detector," *Multimed. Syst.*, Vol. 30, no. 4, pp. 233, Aug. 2024. DOI: 10.1007/s00530-024-01424-7.

31. J. Xiang, and G. Zhu. "Joint face detection and facial expression recognition with MTCNN," in *2017 4th Int. Conf. Inf. Sci. Control Eng.*, 2017, pp. 424–427. DOI: 10.1109/ICISCE.2017.95.

32. P. Korshunov, and S. Marcel. "DeepFakes: a new threat to face recognition? Assessment and detection," 2018. DOI: 10.48550/arXiv.1812.08685.

33. B. Dolhansky, *et al.* "The DeepFake detection challenge (DFDC) dataset," 2020. DOI: 10.48550/arXiv.2006.07397.

34. Y. Li, X. Yang, P. Sun, H. Qi, and S. Lyu, "Celeb-DF: a large-scale challenging dataset for DeepFake forensics," in *2020 IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 3204–3213. DOI: 10.1109/CVPR42600.2020.00327.

35. A. Rössler, D. Cozzolino, L. Verdoliva, C. Riess, J. Thies, and M. Niessner, "FaceForensics++: learning to detect manipulated facial images," in *2019 IEEE/CVF Int. Conf. Comput. Vis.*, 2019, pp. 1–11. DOI: 10.1109/ICCV.2019.00009.

36. C. Sanderson, and B. C. Lovell. "Multi-region probabilistic histograms for robust and scalable identity inference," in *Lect. Notes Comput. Sci.* (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics), vol. 5558 LNCS, 2009, pp. 199–208. DOI: 10.1007/978-3-642-01793-3_21.

37. D. Afchar, V. Nozick, J. Yamagishi, and I. Echizen. "MesoNet: a compact facial video forgery detection network," in *10th IEEE Int. Work. Inf. Forensics Secur. WIFS 2018*, 2018, pp. 1–7. DOI: 10.1109/WIFS.2018.8630761.

38. M. Tan, and Q. V. Le. "EfficientNet: rethinking model scaling for convolutional neural networks," 2019. arXiv:1905.11946.

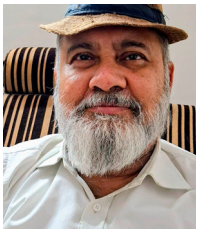
AUTHORS

Pawan Pandey is currently pursuing PhD in the Department of Computer Science & Engineering at Gautam Buddha University, Greater Noida, Uttar Pradesh, India. He received his MTech (CSE) from Amity University, Noida in 2013, and his BTech (CSE) from Raj Kumar Goel Institute of Technology, Ghaziabad (affiliated with Uttar Pradesh Technical University, Lucknow) in 2007. His research interests include artificial intelligence (AI), especially in the subfields of machine learning and deep learning techniques.

Email: pawank.fyi@gmail.com

Arun Solanki is currently working as an assistant professor in the Department of Computer Science and Engineering, Gautam Buddha University, Greater Noida, Uttar Pradesh, India. He received his PhD in Computer Science and Engineering from Gautam Buddha University in 2014 and MTech degree in Computer Science and Engineering from YMCA University, Faridabad, Haryana, India. His research interests span expert systems, artificial intelligence, machine learning, and search engines. he has published more than 70 research articles in SCI/Scopus-indexed International journals/ conferences and more than 10 Books with reputed publishers like IGI Global, CRC, Elsevier, and AAP. He also worked as associate editor in the International Journal of Web-Based Learning and Teaching Technologies (IJWLTT), IGI publisher (Scopus indexed), and Guest Editor for special issues in Recent Patents on Computer Science, Bentham Science Publishers.

Corresponding author. Email: ymca.arun@gmail.com



Sanjay Kumar Sharma is currently working as a professor in the Department of Computer Science and Engineering and Director (Indian Culture Study Center) at Gautam Buddha University, Greater Noida, Uttar Pradesh, India. He has published more than 165 articles in leading National/International Journals, Conferences, edited books, and also completed more than 10 research projects sponsored by ISRO, CSIR, DRDO, MHRD, IUAC & MOEFCC. He has worked with some of the premier public and private Institutions in different parts of the country, like Space Application Center, ISRO, National Institute of Technology, Kurukshetra, J. C. Bose University of Science and Technology, E-Max Group of Institutions, and M. M. University, Ambala, as a Research Scientist, Professor, Dean, and Director/Principal. As a professional in the area of Nuclear Physics, Material Science, Nanotechnology, and Artificial Intelligence, he has contributed to the Advisory Board of academic activities, member of PG and UG Board of studies of various universities, institutions, and colleges, Member of Executive Council of Nuclear Track Society of India, Society for Scientific Values (SSV), Indian Association for Physics Teachers (IAPT) and Secretary, VigyanBharti, Haryana.

Email: sanjay.sharma@gbu.ac.in