Journal of Theoretical and Applied Information Technology
15th March 2026. Vol.104. No.5

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

ENHANCED STATIC DEEPFAKE DETECTION THROUGH FACE SEGMENTATION AND TRANSFORMER-BASED MODELING

SAURABH KUMAR JAIN[1],MOHD AKBAR[2]

[1,2]Integral University, Department of Computer Science & Engineering, India

E-mail: [1]saurabh.jaincse@gmail.com, [2]akbar@iul.ac.in

ABSTRACT

The introduction of advanced generative models such as Generative Adversarial Networks (GANs) and diffusion architectures has made it easier to create hyper-realistic fake or tampered facial images. This advancement is very challenging to ensure the reliability of visual content in digital media. Detection of fake images is difficult and a major problem for issues related to privacy, security, and trust on the online platforms. The work proposed in this article demonstrates an integrated framework for real and fake face image detection that combines the advanced image preprocessing method with the Vision Transformer (ViT) based detection model. The image processing stage is applied to remove the backgrounds and isolate key facial regions. This process reduces unwanted objects and segments the discriminative facial features. The Vision Transformer restructures the face images into sequences of non-overlapping patches and applies global self-attention to capture subtle structural and textural inconsistencies that distinguish real and fake face images. Experimental results are presented that compare the performance of background removal, which helps to increase detection accuracy compared to unprocessed images. The proposed approach demonstrates strong performance on a very diverse dataset and successfully highlights the potential of transformer-based architectures for providing a scalable, interpretable, and robust deepfake detection method.

Keywords: *Deepfake, Vision Transformer, Face Detection, Image Segmentation, Data Mining*

1. INTRODUCTION

The rapid advancement of artificial intelligence and generative models such as Generative Adversarial Networks (GANs) and diffusion architectures has enabled the creation of highly realistic synthetic facial images that are often indistinguishable from genuine photographs. AI-generated deepfakes present significant challenges to digital security, privacy, and trust, as they can be exploited for misinformation, identity theft, and malicious social engineering. Fake image detection methods that are used in traditional approach are generally observed to be based on based on convolutional neural networks (CNNs) or designed to find abnormality in features value. These methods are generally failed to provide robust and homogeneous performance when exposed to new unknown data produced from generative models or manipulations in data at high-quality level. It may be stated that the subtle visual inconsistencies introduced by advanced generators are difficult to detect by methods depending only on local feature. It limits the effectiveness of CNN-based detection models for real-world data produced by rapidly evolving generative techniques. To address these challenges, this work has proposed an integrated framework that combines advanced image preprocessing with Vision Transformer (ViT)-based classification model for detecting real and fake face images. The preprocessing stage of face image data removes the complex structures and irrelevant background regions and focuses only the segmented face region to retain only the most informative object of interest such as eyes, nose, and mouth. After the segmentation of face are the transformation of image data is performed to get a sequence of non-overlapping patches and apply the mechanism of global self-attention. The Vision Transformer is used for this purpose to capture long-range correlation in pixel and subtle structural changes in patches that distinguish authentic faces from fake ones. This combination of image processing for accurate segmentation of face objects and vision transformer model helps to enhance the detection of imperceptible anomalies. In this way improvement in cross-dataset adaptability is

obtained that provides a scalable solution for Deepfake detection of face images.

The novelty of this work lies in the integration of an image processing method for background removal to segment the facial region with the Vision Transformer for the detection of real and fake face images. The conventional models are CNN, and their detection schemes primarily depend on local feature maps. The method proposed in this article utilizes the mechanism of the transformer model to analyze global contextual relationships between image patches while the preprocessing pipeline ensures that only the most relevant facial structures are provided to the transformer model. This two-stage approach, "segmentation + transformer model," helps to suppress the irrelevant background artifacts and enhances sensitivity to minute inconsistencies in texture, frequency, and spatial arrangement, allowing reliable detection of fake faces generated by anonymous generative models. This research work contributes a preprocessing and transformer-based detection framework that improves both accuracy and generalization in static facial deepfake identification. The preprocessing stage ensures that the model focuses exclusively on discriminative facial features, while the Vision Transformer provides patch-wise precise attention for capturing the minute anomalies across the segmented image. Integration of both stages helps to design a system that performs robust operation across diverse datasets and generative sources. It is also offering interpretability using attention visualizations that highlight the relevant segments, improving the classification accuracy.

Recent studies focusing on deepfake detection by application of CNNs, handcrafted feature extraction, and frequency-domain artifact analysis for identification of visual inconsistencies due to image manipulation [1]. Latest research works are introducing transformer-based models and hybrid architectures for the improvement of detection accuracy and generalization across diverse datasets and generative models [2].

The present work differs by integrating facial region segmentation with a Vision Transformer-based classification approach. This method works by reducing background and enhances extraction of feature of important face regions objects, it leads to improvement in detection accuracy, computational efficiency, and robustness as compared to existing methods [3].

2. LITERATURE REVIEW

The rapid advancement of generative models, like GANs and diffusion architectures, supported the field of AI-generated images. The images are very close to real data, and it is very difficult to distinguish them from the original data. Existing deep learning-based detection methods are generally developed using the CNN. These often fail to detect fake images generated from unknown generators or made from complex manual manipulations. In this way, accurate detection of fake images needs robust, generalizable, and interpretable detection frameworks that are capable of identifying images under diverse real-world scenarios. Deepfake detection has become a very important field in digital media applications due to the rapid growth in the use of AI-generated images produced by generative adversarial networks and diffusion models. This section covers the review of the latest high-quality literature works that have inspired this article on the application of fake face image detection by the fusion of image segmentation with a transformer model-based deep learning scheme. Solaiyappan and Wen [1] proposed machine learning techniques for medical images that extract the features and use neural networks to effectively identify localized manipulations in the images. This method often fails in generalized applications when applied to images produced using unseen generators or different image datasets. Chan et al. [4] utilized unsupervised domain adaptation to detect GAN-generated images from previously unseen models, achieving better performance than standard convolutional neural network-based detectors. Dong et al. [2] analyzed spectral artifacts, showing that inconsistencies in the frequency domain provide robust discriminative features. For improving the robustness against a variety of fake data generator sources, Mandelli et al. [3] designed an orthogonal training scheme that improves the feature learning method that is independent of specific types of GAN architectures. Guarnera et al. [5] highlighted the requirement for a multi-class discrimination method by introducing a method that categorizes images based on generator type and authenticity. Zhang et al. [6] presented an identity-based framework. It considers the reference images for quantifying the identity inconsistencies. It enables a straightforward detection method for facial image deepfakes. In the application of image deepfake detection, vision transformers and transformer-based architectures have recently become more popular. Their ability to model long-range dependencies across image patches allows the capture of subtle manipulations that

conventional convolutional networks may miss. Kavthekar et al. [8] developed MIM-ViT, which uses masked image modeling to detect localized manipulations effectively.

Table 1: Summarized literature review on latest approach applied for fake image data detection

Ref.	Author(s)	Approach	Pros	Cons
[1]	Solaiyappan &Wen(2021)	ML with handcrafted features for medical image fake detection	Simple and interpretable; effective on small data	Limited generalization to unseen generators/domains
[2]	Dong et al. (2022)	Spectral artifact analysis on GAN-generated images	Captures frequency inconsistencies	May fail on high-quality or post-processed images
[3]	Mandelli et al. (2022)	Orthogonal training to enforce generator-invariant features	Improves robustness across multiple GANs	Requires careful training; computational overhead
[4]	Chen et al. (2022)	Unsupervised domain adaptation for cross-generator detection	Generalizes to unseen generators	Complex setup; relies on source domain quality
[5]	Guarnera et al. (2023)	Multi-class discrimination of GAN vs. diffusion models	Handles multiple generator; flexible	Performance may drop for novel generators
[6]	Zhang et al. (2023)	Identity-based framework using reference images	Explainable results; high accuracy for fake faces	Needs reference images; less effective for non-face image
[7]	Kumar M et al. (2023)	GAN-based CNN model for social media images	Effective for social media datasets; fast	Limited generalization; may miss subtle manipulations
[8]	Kavthekar et al. (2024)	MIM-ViT (masked image modeling with ViT)	Captures long-range dependencies	Requires large datasets; high computational cost
[9]	Cozzolino et al. (2024)	Zero-shot detection of AI-generated images	Generalizes without training; practical	Slightly lower accuracy to supervised methods
[10]	Zhang J. et al. (2024)	Real images only training for detection	Avoids dependency on synthetic datasets; robust	Needs high-quality real datasets
[11]	Chen et al. (2024)	CLIP-ViT with multi-stage feature fusion	Strong generalization for global/local features	Complex architecture; requires GPU resources
[12]	Nguyen et al. (2024)	FakeFormer (focusing on artifact-prone patches)	High accuracy; robust to unseen generators	Computationally intensive; patch sensitive
[13]	Feuer et al. (2024)	Benchmarking and evaluation with curated image datasets	Provides standard evaluation	Dataset limited to specific generators
[14]	Say et al. (2025)	Mixed datasets with artifact analysis	Comprehensive evaluation; handles multiple generator	High complexity; dataset preparation effort
[15]	Kundu R. et al. (2025)	TruthLens: explainable detection	Provides interpretable outputs; good accuracy	More computationally expensive
[16]	Jiang Y. et al. (2025)	Loupe: adaptive framework for image forgery detection	Generalizable; adaptive to new manipulations	Needs tuning; may require large datasets
[17]	Anan K. et al. (2025)	Hybrid CNN + attention + frequency features	Captures local/global artifacts; robust scheme.	Complex model; higher training time
[18]	Huang et al. (2025)	SIDA: social media image detection and localization	Detects fake regions; suitable for social media	May struggle with unseen generator types
[19]	Chen B. (2025)	Deepfake image forensics	Ensures privacy protection; interpretable	Performance depends on dataset quality
[20]	Zhuang W. et al. (2025)	PV-ISM: Patch-based Vision Transformer	Patch-level analysis; high accuracy	Requires large datasets; computationally intensive
[21]	Wodajo &Atnafu(2021)	Combines CNN with ViT for spatial-temporal fake detection	Captures local and global facial inconsistencies	Computationally heavy; limited generalization
[22]	Arnab et al. (2021)	Pure Transformer architecture applied to video frames of space-time tokens	Achieves results on action recognition; avoids CNN dependency	Requires extensive computational resources; data-hungry model
[23]	Coccomini et al. (2022)	Extracts spatial features by EfficientNet, then use ViT for temporal and relation learning	Reduces training cost; improved generalization and interpretability	Architectural complexity; moderate latency in processing

Figure 1: Real faces

Figure 2: Fake faces

Chen et al. [11] proposed a combination of frozen CLIP-ViT features with multi-stage features. It demonstrated a strong generalization for application with unseen generators.Nguyen et al. [12] proposed FakeFormer as a transformer model that mainly focuses on attention towards artifact-vulnerable patches. It gave high performance results on image datasets. The latest ViT-based method, PV-ISM [20], applied at patch-level analysis for differentiating the real and fake media data at high accuracy. It highlights the relevance of transformer-based approaches. Anan K. et al. [17] integrated convolutional, attention, and frequency-domain features for capturing the local and global discrepancies to improve the robustness across the diversified datasets. Kundu et al. [15] introduced TruthLens to detect manipulated images by providing deep learning to forensic AI applications. Benchmarking studies like Deepfake-Eval-2024 [13] and SIDA [18] provide datasets and evaluation protocols; they are showing that traditional models generally face a reduction in performance on newly generated image datasets. It gave emphasis on the requirement for adaptive and generalizable detection frameworks. Overall, the literature survey has revealed that a clear progression exists from artifact-based CNN detectors to transformer-based and hybrid models. It is driven by the increasing popularity of AI-generated data. Analysis of current research work shows a focus on three fields: improving generalization to unseen generators, employing long-range contextual modeling using transformer models, and enhancement of interpretability.

Although several methods have been proposed for deepfake detection, most focus on specific generator types or video sequences, limiting their applicability to static images. Transformer-based approaches like Vision Transformers (ViTs) show promise in capturing long-range dependencies and subtle artifacts, but their adoption in image-only detection is still limited. Moreover, current approaches often lack interpretability and cross-dataset generalization, highlighting the need for hybrid and robust methods that perform reliably on fake images.

2.1 Research Questions

1. How effectively can the proposed segmentation and Vision Transformer-based framework detect deepfake images compared to existing detection methods?
2. Does facial region segmentation improve feature extraction and overall detection accuracy in deepfake identification?
3. How robust is the proposed model when tested on different deepfake datasets and manipulation techniques?
4. Can the proposed method reduce computational complexity while maintaining high detection performance?

2.2 Research Hypothesis

H1: The integration of facial region segmentation with Vision Transformer-based classification significantly improves deepfake detection accuracy compared to conventional methods.

H2: Removing background noise through facial segmentation enhances feature extraction and increases model robustness.

H3: The proposed framework provides improved computational efficiency and reliable performance across multiple deepfake datasets.

3. DATA PREPROCESSING

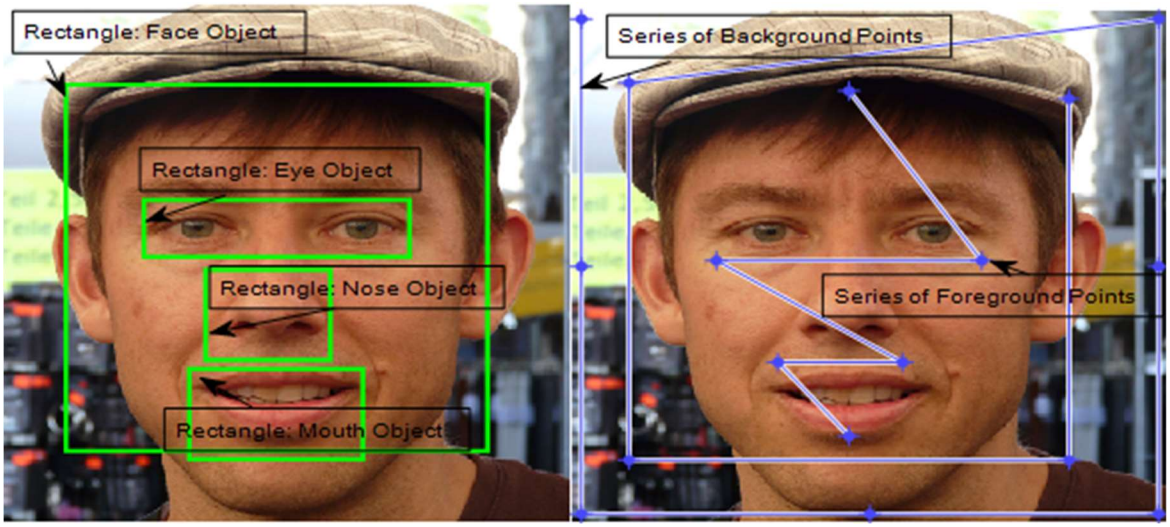

Figure 3: Identification of face, nose, eye and mouth objects (left). Indication the background area points for segmentation of face region of interest (right)

Figure 1 and 2 are showing some of the images of real and fake faces from the collected dataset. The dataset consists of two separate folders of 779 real face images and 668 fake face images. In the fake face images, generally the single or both eye, nose, mouth, eyes+nose, nose+mouth, etc. type combinations are observed as tampered or replaced by eye, nose, or mouth from an unknown face, as shown in the image data samples in figure 2. Both real and fake image folders have face images captured in front or side direction at random rotations of the face. The faces have beards, mustaches, different hairstyles, caps, etc. that hide the face information. The illumination level is also not uniform on the faces; hence, the database is a highly rough and uncorrelated type of variation and lacks any uniformity in each image.

Fig 4a: Border Lines extraction

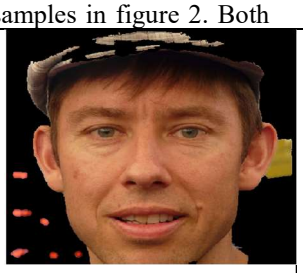

Fig 4b: Removal of Background

Fig 4c:Image with Eye, Nose, Mouth object

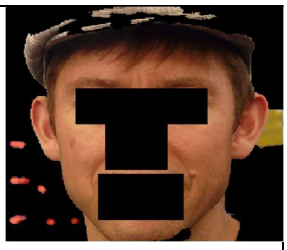

Fig 4d:Image without Eye, Nose, Mouth object

All the images from both folders of real and fake faces are passed through face object detection commands in three stages to find rectangular boundaries of the face object, eye object, nose object, and mouth object, as shown in figure 3 left. In detecting the eye, nose, or mouth, the algorithm detects a false eye, nose, or mouth. More than one region is pointed out as a nose or mouth object. To select only one object from multiple choices, the location-based logic is integrated. The rectangle whose center point is nearest to the midpoint of the lower line of the eye object rectangle is extracted out as the nose object. Similarly, if more than one mouth object is found, then the rectangle lower to the bottom line of the nose object rectangle and nearest to the nose object is taken as the mouth object. After detecting all the objects of the face, the front region and background region are pointed out automatically to eliminate the background as shown in Figure 3 (right). The data processing algorithm was developed on Matlab software for segmentation of the face image, excluding the background or other irrelevant information. It shows the automated localization of the front part as the top middle point of the face object, the corners of the eyes/nose, and the middle point of the lower point of the mouth object.

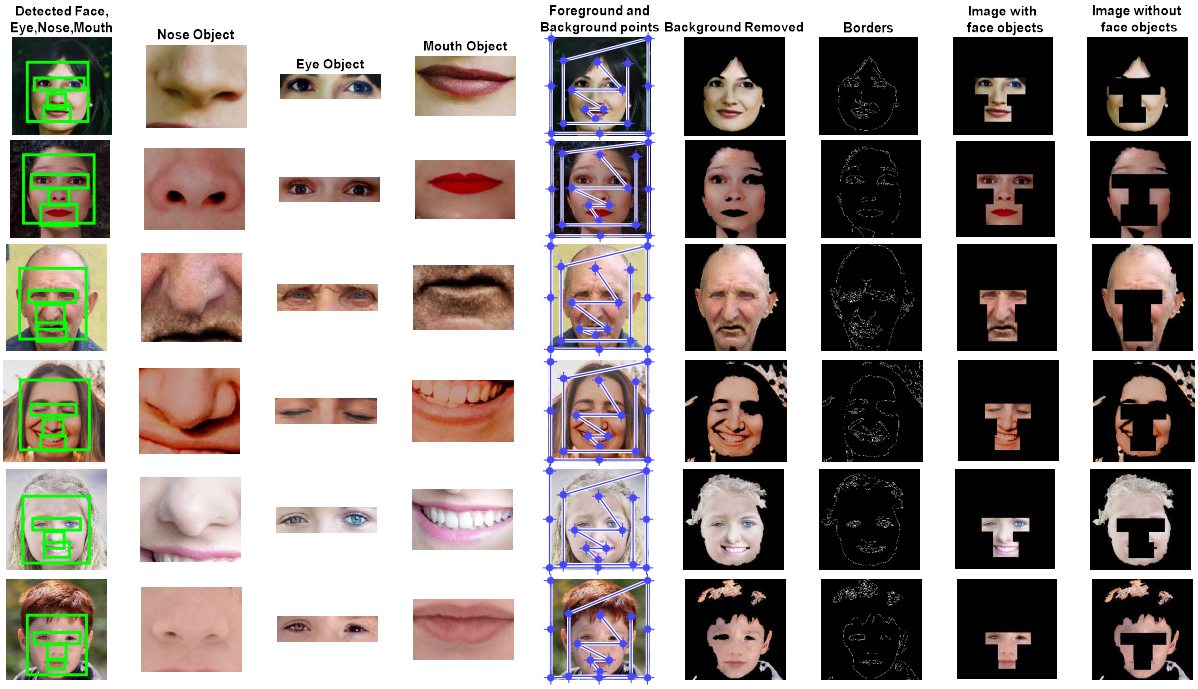

Figure 5: Data preprocessing and segmentation on real faces training data to obtain images with and without face

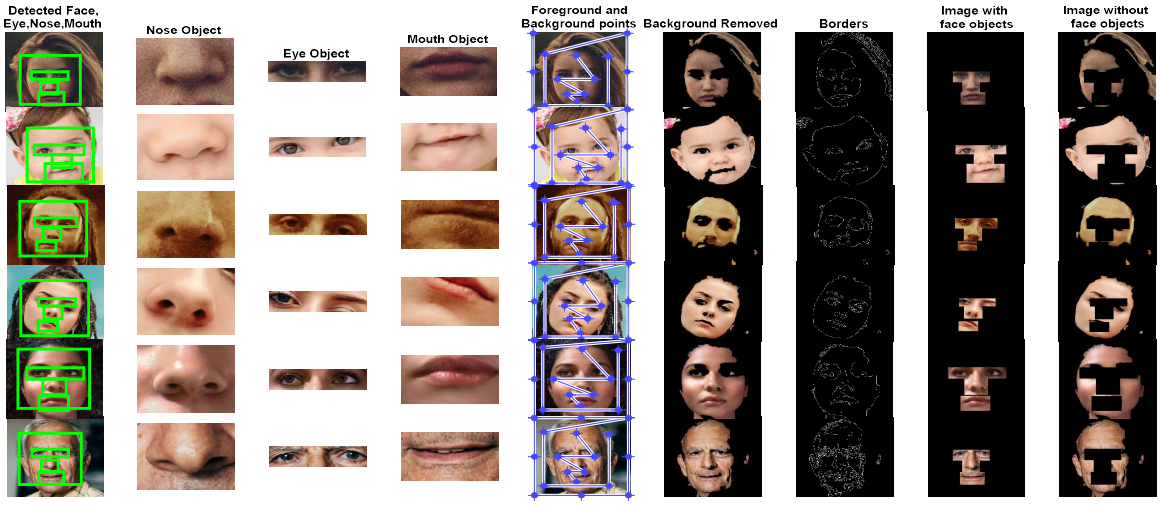

Figure 6. Data preprocessing and segmentation on fakefaces training data to obtain images with and without face.

The figure 5 and 6 are showing samples of some images from real and fake face data that are preprocessed and segmented to make them ready for learning of ViT. During the removal of the background, some eye/nose/mouth regions are also eliminated, which has created a distraction in the face image, so the segregated image is finally superimposed on the segmented face object to suppress the elimination of relevant facial pixel information that occurred during automated background removal.

3. METHODOLOGY

Figures The Vision Transformer (ViT) for image classification operates by reformulating images into sequential representations, enabling the direct application of Transformer architectures originally designed for natural language processing. The process begins with data preparation, where input images are resized to a fixed resolution, normalized, and labeled. Each image is divided into non-overlapping patches of fixed size, which are

subsequently flattened into vectors. These vectors are linearly projected into a latent embedding space, thereby forming a sequence of patch embeddings. A learnable classification token is prepended to the sequence, and positional encodings are added to preserve spatial information, as the Transformer architecture itself is permutation-invariant. The self-attention layer is the core computational block in the VT model. It learns the contextual relations between the different patches. It generates the queries, keys, and values by applying the linear projections of the features embedded in input data. The attention weights are computed using a scaled dot product followed by the softmax operation. These weights are used by the model to emphasize or suppress the information present in different patches depending on their relevance. The output of the attention layer is projected back into the embedding space to maintain consistency of data dimension. This mechanism captures long-range dependencies in the patches of the image.

In parallel, the Multi-Layer Perceptron (MLP) block works for refining the response. It consists of two fully connected layers separated by a non-linear activation such as GELU or ReLU. The first projection increases the representational capacity by mapping embedding to a higher-dimensional space. The second stage compresses them to the original embedding dimension. Self-attention and the MLP block run in parallel to design the functional blocks of the Vision Transformer. Under the classification process, only the class token, which consists of aggregated global information of all patches, is passed through a final classifier layer. A softmax activation function determines the probability distributions over all the classes, and cross-entropy loss is used for measuring the prediction error against the true labels. The training process tunes the model parameters that include projection matrices, positional embeddings, and classifier weights. These parameters are optimized using methods based on the value of the gradient. Gradients are calculated from differentiation of error signals calculated along backpropagation through the network layers. The process of patch embedding, self-attention computation, and MLP transformation is performed repeatedly across mini-batches of training images within each epoch. After every forward and backward pass, parameters are updated under a specific optimization process. After running training over a successive number of epochs, the model gradually learns the discriminative representations that improve the accuracy of classification. Performance is evaluated using standard metrics such as accuracy, precision, recall, and F1-score, with confusion matrices providing further insight into class-wise errors. Thus, the Vision Transformer framework effectively transfers the strengths of sequence modeling to image understanding by leveraging patch-based embeddings and global self-attention mechanisms.The integration of custom components—such as patch flattening, self-attention, and MLP sub-functions—ensures that local spatial information and global contextual relations are both captured, leading to robust and accurate classification performance.

Table 2: List of Abbreviations and Symbols

Abbreviations
ViT – Vision Transformer
CNN – Convolutional Neural Network
MHSA – Multi-Head Self-Attention
MLP – Multi-Layer Perceptron
FC – Fully Connected
CLS – Classification Token
ReLU – Rectified Linear Unit
GELU – Gaussian Error Linear Unit
SGD – Stochastic Gradient Descent
Adam – Adaptive Moment Estimation Optimizer
TP – True Positive, FP – False Positive
FN – False Negative, TN – True Negative
Symbols
X – Input image tensor
H – Image height
W – Image width
C – Number of image channels
N – Number of images in dataset
P – Patch size
np – Number of patches per image
d – Embedding dimension
m – Hidden dimension of MLP
Q – Query matrix in self-attention
K – Key matrix in self-attention
V – Value matrix in self-attention
WQ, WK, WV, WO – Projection weight matrices for attention
A – Attention score matrix
O – Output of self-attention layer
Z – Classifier output label before softmax
$\hat{y}$ – Predicted class probabilities
y – True class label
L – Cross-entropy loss
η – Learning rate
θ – Learnable parameters of the model
Acc – Accuracy metric
Prec – Precision metric
Rec – Recall metric
F1 – F1-score metric

Table 3: Algorithm: Vision Transformer for Fake Face Image Classification

Step 1: Data Preparation

1. Load dataset using image Datastore with subfolder names as labels.
2. Resize all images to fixed size, for example 32 by 32 with 3 channels.
3. Normalize pixel values to the range 0 to 1.
4. Store images in tensor X and labels in Y.
5. Extract class names using categories(Y).

Step 2: Model Initialization

1. Define hyper parameters such as patch size, embedding dimension, hidden dimension of MLP, learning rate, number of epochs, and mini-batch size.
2. Compute the number of patches per image.
3. Initialize learnable parameters: class token, positional embedding, fully connected classifier weights and bias.
4. Create custom layers: SelfAttentionLayer and MlpBlockLayer.

Step 3: Training Loop

For each epoch: a. Shuffle training data. b. For each mini-batch:
i. Extract batch images and labels.
ii. Convert batch to dlarray.
iii. Call modelGradientsFull to compute loss and gradients.
iv. Update learnable parameters using optimizer.
c. Print average loss for the epoch.

Step 4: Prediction-For each test image:
a. Call patchFlatten to divide the image into patches and embed them.
b. Add class token and positional embedding's.
c. Call SelfAttentionLayer.forward to apply attention.
d. Call MlpBlockLayer.forward to process the representation.
e. Extract class token output.
f. Pass through classifier fully connected layer.
g. Select class with maximum score as predicted label.

Step 5: Evaluation-

1. Compute accuracy as ratio of correct predictions to total samples.
2. Generate confusion matrix between predicted and true labels.
3. For each class compute precision, recall, and F1-score.
4. Present results.

◆ Sub Function Call: SelfAttentionLayer

Input: Representation X of size [embedding dimension × (number of patches + 1)]
Output: Attention-transformed representation

1. Compute queries Q by multiply X with WQ.
2. Compute keys K by multiply X with WK.
3. Compute values V by multiply X with WV.
4. Compute attention scores using softmax of (Q × K transpose divided by square root of dimension).
5. Multiply attention scores with V to obtain attended features.
6. Apply output projection using WO.
7. Return the transformed representation.

◆ Sub Function Call: MlpBlockLayer

Input: Representation X
Output: Processed representation

1. Multiply X with W1 and add bias b1.
2. Apply activation function such as ReLU or GELU.
3. Multiply with W2 and add bias b2.
4. Return the processed output.

◆ Sub Function Call: patchFlatten

Input: Image tensor of size [H × W × C]
Output: Flattened patch embedding's

1. Divide image non-overlapping patches P × P size.
2. Flatten each patch into a vector.
3. Project each patch into embedding dimension d.
4. Concatenate all patch embedding's into a sequence.
5. Return the sequence.

◆ Sub Function Call:modelGradientsFull

Input: Mini-batch of images, labels, and all learnable parameters
Output: Loss and gradients

1. Call patchFlatten to convert images into patch embeddings.
2. Add class token at beginning of sequence.
3. Add positional embeddings.
4. Call SelfAttentionLayer to apply self-attention.
5. Call MlpBlockLayer to further process O/P.
6. Extract class token output.
7. Pass through classifier fully connected layer.
8. Apply softmax to compute predicted probabilities.
9. Compute cross-entropy loss with true labels.
10. Use automatic differentiation to compute gradients with respect to all parameters.

Return loss and gradients.

3.1 Study Design

1. Problem Identification and Objective Formulation
Define the research problem related to deepfake image detection and establish study objectives.
2. Dataset Selection and Preparation
Collect benchmark datasets containing real and manipulated facial images and perform preprocessing.
3. Model Development
Design the deepfake detection framework by integrating facial segmentation and Vision Transformer-based classification.
4. Implementation and Training
Train the proposed model using prepared datasets and optimize model parameters.
5. Performance Evaluation
Analyze detection performance using evaluation metrics and compare results with existing approaches.

3.2 Research Protocol

1.Data Acquisition: Gather real and deepfake image datasets from standard public repositories.
2.Data Preprocessing: Perform image normalization, resizing, and facial region segmentation to remove background noise.
3.Feature Extraction and Classification:Apply Vision Transformer architecture toextract discriminative features and classify images.
4.Model Validation: Evaluate model performance using accuracy, precision, recall, and F1-score metrics.
5.Comparative and Statistical Analysis: Compare the obtained results with existing deepfake detection techniques to verify effectiveness and reliability.

5. RESULTS

The segmented face image and real and fake faces from the training data are used by the ViT model to learn for developing a fake face image detection system. The algorithm given in table 3 is followed up for training the ViT model. During the learning process the model is updated on the basis of targeting the objective of minimizing the loss function as the training epochs successively increase. Figure 7 is showing the convergence on the loss function over the 10 epochs of the learning process. This bar chart consists of the plot of the loss function when no face segmentation is applied in blue bars and the loss function variation with respect to epochs 1 to 10 when face region segmentation is applied. It may be observed that in each epoch the loss is lower when training is performed on segmented faces data. The values of the bar chart are shown in table 4; it clearly demonstrates that the loss function reduces successively as the training epochs move forward for both the without and with face region segmented cases. Loss function is the indication of how many predictions made by the model match the actual target values. It actually quantifies the error between the predicted output and the true output. If y is the true output and $\hat{y}$ is the predicted output for N number of input samples than loss function $L(y, \hat{y})$ is:

$$L(y, \hat{y}) = \frac{1}{N}\sum_{i-1}^{N}(y_i - \hat{y}_i)^2 \quad (1)$$

Table 4: Loss function at different epochs

	Loss Function	
Epoch	Without Image Segmentation	With Image Segmentation
1	0.1461	0.1293
2	0.145 0	0.1315
3	0.1397	0.1316
4	0.1396	0.129 0
5	0.1381	0.1312
6	0.1351	0.1328
7	0.134 0	0.1336
8	0.1332	0.127 0
9	0.1328	0.1315
10	0.1313	0.128 0

Table 5: Performance results

	Accuracy	Precision	Recall	F1 score
Without Image Segmentation	80.47%	0.7932	0.8126	0.8286
With Image Segmentation	85.55%	0.8571	0.8667	0.8619

Finally, the precision, recall, and F1-score are calculated by generating the true positive (TP), false positive (FP), true negative (TN), and false negative (FN) values from the confusion matrix for both the original images and the images with the face region segmented. The results are shown in table 5. The results for ViT-based detection using segmented face images are observed to be higher. Table 6 is showing the results for validating the proposed work with the latest state-of-the-art methods that are related to the proposed work in this article. The proposed work is named by method as “Segmentation+Vit” in the last row. The top

three rows are showing other authors' names and approaches that have used Convolutional ViT, ViVe, and Convolutional Cross ViT Efficient Net B0. In Convolutional ViT, the CNN-extracted features are utilized by ViT to detect fake face images (Face Forensic++ and DeepFake Detection Challenge (DFDC) dataset). It has the limitations of computational complexity and failed to see manipulation on face data. The Video Vision Transformer (ViViT) method [22] used the Kinetics dataset and is a pure use of only ViT, but it was more computationally complex and requires large data to give high accuracy. The third state-of-the-art method is Conv. Cross ViT EfficientNet [23], which extracts spatial features using EfficientNet and then applies ViT for temporal and relational learning that is applied to the DFDC and Celebrity DeepFake Dataset (CelebDF). It has drawbacks in that, due to architectural complexity, it introduces moderate latency in the process of detection. The proposed method in this work removes the background and objects other than the face images; hence, the data's size as well as redundancy in data is reduced to half. This results in a reduction of architectural and computational complexity during the learning and detection phases of the ViT model. The removal of unwanted background helps to produce a relevant feature; that is why the accuracy of the proposed work is observed to be higher, i.e., 85%, compared to Convolutional ViT at 67%, ViVe at 84%, and Conv. Cross ViT EfficientNet at 80% accuracy, respectively.

Table 6: Comparison to state of art methods

Author	Method	Accuracy (%)
Wodajo et al. [21]	Convolutional ViT	67%
A. Arnab et al. [22]	ViViT	84%
Davide Coccomini et al. [23]	Conv. Cross ViT EfficientNet	80%
Proposed Work	Segmentation+ ViT	85%

The proposed work is motivated by recent advancements under the concept of hybrid deep learning architectures that is combining CNN and transformer-based models for improvement of analysis of image. Hussain et al. [24] demonstrated that fusion-based CNN and Vision Transformer models enhance feature extraction capability and classification accuracy in medical image datasets. Similarly, Kumari and Saxena [25] and Rastogi et al. [26] performs the exploration of CNN for retinal image classification, that highlight the importance of deep feature learning but also reveal the performance limitations in standalone CNN models. Rahatwal et al. [27] further showed that integrating CNN and transformer architectures significantly improves deepfake detection performance using Celeb-DF datasets.Building upon these findings, the proposed work aims for development of an enhanced framework for deepfake detection by integration of facial objects segmentation hybrid with Vision Transformer architecture. Unlike existing methods that focus primarily on global features of image, the proposed method incorporates on extraction of discriminative facial objects features for improving detection accuracy, reducing background noise interference, and enhancing model generalization under diverse deepfake datasets.

Related studies on deepfake detection are widely published using CNN models, frameworks of hybrid deep learning, and architectures based on transformer for identifying the manipulation of artifacts and structural inconsistencies of the images [1], [2]. These studies report improvement in detection accuracy but highlights new challenges like limited generalization and dependency on characteristics of dataset.

The results presented in this proposed work align with previous studies by demonstration of enhanced detection performance by the integration of segmentation and transformer model-based classification [3]. However, the proposed approach depends on accurate extraction of face objects and has requirement of large training datasets. Additionally, its performance against the emerging deepfake methods requires further validation, indicates scope for future improvement [3].

5.1 Research Problems

Deepfake detection is becoming challenging due to rapid advancements of generative models that are producing highly realistic manipulated images. Existing detection methods often suffering from limited generalization when applying on unseen datasets or newly developed methods of deepfake generation. Many models rely heavily on complete image analysis, which introduces background noise and reduces detection accuracy. Additionally, deep learning-based detection systems generally require large labeled datasets and high computational resources, making real-time deployment difficult. These challenges create the need for more robust, efficient, and adaptable detection frameworks.

5.2 Open Research Issues

There is a need for developing deepfake detection models that effectively generalize on multiple datasets and emerging data manipulation scheme. Improvement in real-time detection with reduced computational complexity remaining an open

challenge. Integration of multimodal data like video, audio, and temporal features is another unexplored area that may be useful in enhancing detection reliability.

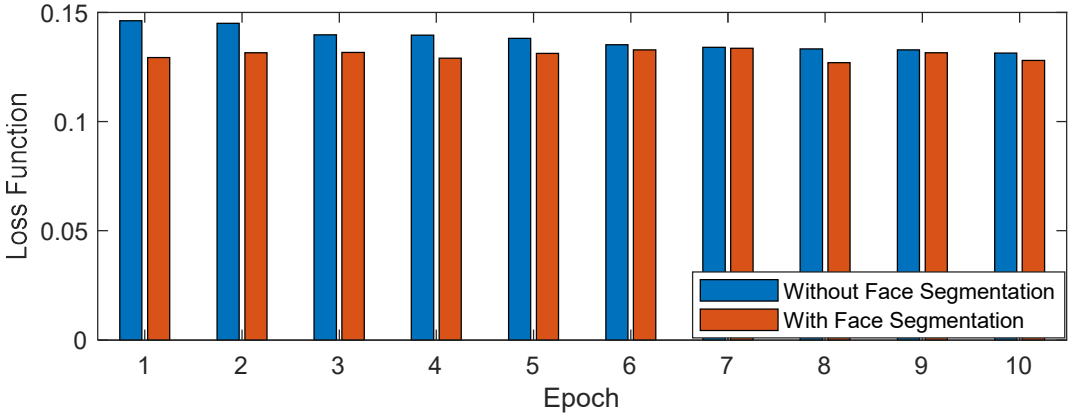

Figure 7: Loss Function During Training Process Using Face Data Without And With Face Segmentation.

Furthermore, improving explainability and interpretability of model is crucial for building trust and supporting practical deployment in security and forensic applications.

6. CONCLUSIONS

This work demonstrates the benefits of combining image preprocessing and Vision Transformer modeling. The integration provides an accurate solution for detecting fake face images. The image processing helps remove the irrelevant background content information and gives focused facial regions with minimum redundancy. The preprocessing stage of background removal and facial region segmentation ensures that only the key features are retained for analysis. Such an approach makes the ViT operation effective in performing patch-wise embeddings and global self-attention to detect subtle inconsistencies that conventional convolutional networks are unable to capture. The results generated in terms of precision, recall, F1 score, and accuracy were used to validate that face segmentation is useful to improve model performance. The proposed framework has established a reliable method to identify real and fake facial images. The approach has clearly enhanced the detection accuracy and improved the interpretability by using attention-based visualization, making it suitable for real-world applications in the field of forensics and media authentication. Future research may be extended by incorporating multimodal information like voice or video data for improving the robustness of deepfake detection applications. The integration of frequency-domain applications, self-supervised training, and few-shot learning strategies may have a chance to further enhance the quality of detection. Cross-domain evaluation on emerging diffusion models and hybrid datasets is still desirable in this work for ensuring long-term adaptability.Future research focusing on improvement of the ability of model for detection of deepfakes generated by advanced and emerging generative schemes. Further studies may use for exploration of multimodal detection by integration of image, video, and audio features for enhancement of robustness and real-world application. Additionally, optimizing the model for reducing computational complexity and real-time detection may be brought under consideration.

ACKNOWLEDGEMENT

The author would like to acknowledge IntegralUniversity for providing acknowledgement no. IU/R&D/2026-MCN0004257

REFERENCES:

[1] S. Solaiyappan and Y. Wen, “Machine Learning Based Medical Image Deepfake Detection: A Comparative Study”, *Machine Learning with Applications*, Vol. 8, 2022, pp. 100298.

[2] C. Dong, A. Kumar, and E. Liu, “Think Twice Before Detecting GAN-Generated Fake Images from Their Spectral Domain Imprints”, *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022, pp. 7865–7874.

[3] S. Mandelli, N. Bonettini, P. Bestagini, and S. Tubaro, “Detecting GAN-Generated Images by Orthogonal Training of Multiple CNNs”, *Proceedings of IEEE International Conference*

on Image Processing (ICIP), October 2022, pp. 3091–3095.

[4] B. Chen and S. Tan, "FeatureTransfer: Unsupervised Domain Adaptation for Cross-Domain Deepfake Detection", *Security and Communication Networks*, Vol. 2021, No. 1, 2021, pp. 9942754.

[5] L. Guarnera, O. Giudice, and S. Battiato, "Level up the Deepfake Detection: A Method to Effectively Discriminate Images Generated by GAN Architectures and Diffusion Models", *Proceedings of Intelligent Systems Conference*, Springer Nature Switzerland, July 2024, pp. 615–625.

[6] Y. Zhang, Z. Pang, S. Huang, C. Wang, and X. Zhou, "Unmasking AI-Created Visual Content: A Review of Generated Images and Deepfake Detection Technologies", 2025.

[7] M. Kumar and H. K. Sharma, "A GAN-Based Model of Deepfake Detection in Social Media", *Procedia Computer Science*, Vol. 218, 2023, pp. 2153–2162.

[8] S. Kavthekar, S. Vaidya, V. Pujari, and S. Mane, "MIM-ViT: Deepfake Detection Using Masked Image Modelling and Vision Transformer", *Proceedings of International Conference on Soft Computing for Problem Solving*, Springer Nature Singapore, August 2023, pp. 1–14.

[9] D. Cozzolino, G. Poggi, M. Nießner, and L. Verdoliva, "Zero-Shot Detection of AI-Generated Images", *Proceedings of European Conference on Computer Vision (ECCV)*, Springer Nature Switzerland, September 2024, pp. 54–72.

[10] J. Zhang, J. Ni, F. Nie, and J. Huang, "Domain-Invariant and Patch-Discriminative Feature Learning for General Deepfake Detection", *ACM Transactions on Multimedia Computing, Communications and Applications*, Vol. 21, No. 2, 2024, pp. 1–19.

[11] L. Chen, Y. Zhao, and X. Li, "Multi-Stage Feature Fusion Using CLIP-ViT for Robust Deepfake Detection", *Pattern Recognition Letters*, Vol. 179, 2024, pp. 23–34.

[12] D. Nguyen, M. Astrid, E. Ghorbel, and D. Aouada, "Fakeformer: Efficient Vulnerability-Driven Transformers for Generalisable Deepfake Detection", *arXiv Preprint*, arXiv:2410.21964, 2024.

[13] B. Feuer, J. Xu, N. Cohen, P. Yubeaton, G. Mittal, and C. Hegde, "SELECT: A Large-Scale Benchmark of Data Curation Strategies for Image Classification", *Advances in Neural Information Processing Systems*, Vol. 37, 2024, pp. 136620–136645.

[14] T. Say, M. Alkan, and A. Kocak, "Advancing GAN Deepfake Detection: Mixed Datasets and Comprehensive Artifact Analysis", *Applied Sciences*, Vol. 15, No. 2, 2025, pp. 923.

[15] R. Kundu, A. Balachandran, and A. K. Roy-Chowdhury, "TruthLens: Explainable DeepFake Detection for Face Manipulated and Fully Synthetic Data", *arXiv Preprint*, arXiv:2503.15867, 2025.

[16] Y. Jiang, J. Chu, J. Zhao, X. Zhang, X. Yang, L. Jin, and X. Li, "LOUPE: A Generalizable and Adaptive Framework for Image Forgery Detection", *arXiv Preprint*, arXiv:2506.16819, 2025.

[17] K. Anan, A. Bhattacharjee, A. Intesher, K. Islam, A. F. Fuad, U. Saha, and H. Imtiaz, "Hybrid Deepfake Image Detection: A Comprehensive Dataset-Driven Approach Integrating Convolutional and Attention Mechanisms with Frequency Domain Features", *arXiv e-prints*, arXiv:2502, 2025.

[18] Z. Huang, J. Hu, X. Li, Y. He, X. Zhao, B. Peng, and G. Cheng, "SIDA: Social Media Image Deepfake Detection, Localization and Explanation with Large Multimodal Model", *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025, pp. 28831–28841.

[19] B. Chen, X. Liu, Z. Xia, and G. Zhao, "Privacy-Preserving DeepFake Face Image Detection", *Digital Signal Processing*, Vol. 143, 2023, pp. 104233.

[20] W. Zhuang, Q. Chu, Z. Tan, Q. Liu, H. Yuan, C. Miao, and N. Yu, "UIA-ViT: Unsupervised Inconsistency-Aware Method Based on Vision Transformer for Face Forgery Detection", *Proceedings of European Conference on Computer Vision (ECCV)*, Springer Nature Switzerland, October 2022, pp. 391–407.

[21] D. Wodajo and S. Atnafu, "Deepfake Video Detection Using Convolutional Vision Transformer", *arXiv e-prints*, arXiv:2102, 2021.

[22] A. Arnab, M. Dehghani, G. Heigold, C. Sun, M. Lučić, and C. Schmid, "ViViT: A Video Vision Transformer", *Proceedings of IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021, pp. 6836–6846.

[23] D. A. Coccomini, N. Messina, C. Gennaro, and F. Falchi, "Combining EfficientNet and Vision Transformers for Video Deepfake Detection", *Proceedings of International Conference on Image Analysis and Processing*, Springer

International Publishing, May 2022, pp. 219–229.

[24] T. Hussain, H. Shouno, A. Hussain, D. Hussain, M. Ismail, T. H. Mir, F.-R. Hsu, T. Alam, and S. A. Akhy, "EFFResNet-ViT: A Fusion-Based Convolutional and Vision Transformer Model for Explainable Medical Image Classification", *IEEE Access*, Vol. 13, 2025, pp. 54040–54068.

[25] P. Kumari and P. Saxena, "Pathologic Myopia Diagnosis and Localization from Retinal Fundus Images Using Custom CNN", *Neural Computing and Applications*, Vol. 36, No. 23, 2024, pp. 14309–14325.

[26] P. Rastogi, J. Jain, P. Jain, K. Arjun, and R. Aggarwal, "Analysis of CNN Models for Eye Disease Classification Using Retinal Images", *Proceedings of 1st International Conference on Advanced Computing and Emerging Technologies (ACET)*, August 2024, pp. 1–6.

[27] K. P. Rahatwal, S. Pundir, M. Wazid, and V. Bhat, "A Novel Approach to Deepfake Detection: Leveraging Fused Facial and Body Dynamics with a CNN–Transformer Hybrid Network", *IEEE Access*, Vol. 13, 2025, pp. 197085–197108.