Abstract. The rise of deepfake technology has posed significant challenges for digital media authenticity. In this paper, we propose a novel approach for detecting manipulated videos using Vision Transformers (ViTs). Unlike traditional convolutional neural networks (CNNs), ViTs leverage self-attention mechanisms to capture global dependencies within video frames, making them particularly effective in identifying subtle inconsistencies typical of deepfakes. Our method is evaluated on benchmark datasets and shows improved accuracy and robustness compared to state-of-the-art models. This work advances the field of deepfake detection by providing an interpretable, reliable, and adversarial-resistant framework using ViTs, with potential real-world applications to mitigate the spread of deceptive content.

Keywords:

Keywords: Deepfake Detection, Vision Transformer (ViT), Binary Cross Entropy (BCE), Video Analysis, Computer Vision, Machine Learning, Adversarial Attacks, Self-Attention Mechanism, Video Manipulation Detection, Temporal Feature Extraction.

1 Introduction

Deepfakes represent manipulated media be it images, videos, or audio—that alter the original context, often creating misleading content that can deceive viewers. Using advanced AI techniques such as Generative Adversarial Networks (GANs), deepfake creators can modify facial expressions, voices, or even simulate entire videos. Deepfake technology, fueled by artificial intelligence (AI), deep learning, and machine learning algorithms, has revolutionized the way we edit and manipulate videos. There are several DeepFake Apps like **DeepFaceLab**, **Reface**, **Deepfakes Web**, with these advanced tools, we can seamlessly perform face swaps, create realistic deepfake videos, and generate high-quality visual content. Notably, these manipulations can be convincing, as demonstrated by viral videos, including fabricated celebrity appearances.

While deepfakes may serve benign purposes such as satire or entertainment, they also pose significant threats across various domains. From privacy concerns to political manipulation and corporate espionage, deepfakes have the potential to undermine public trust and damage reputations. The rapid improvement in deepfake quality requires equally advanced detection mechanisms to mitigate their spread. Traditional methods for detecting deepfakes rely heavily on convolutional neural networks (CNNs) and recurrent neural networks (RNNs), which focus on local pixel patterns. However, these models may struggle with identifying subtle or widespread inconsistencies across video frames. In contrast, Vision Transformers (ViTs) excel at capturing long-range dependencies by using a self-attention mechanism, making them a promising approach for detecting deepfakes. By processing entire video frames as sequences, ViTs can identify minor artifacts and manipulations more effectively.

Problem Statement: Current deepfake detection models, while effective in controlled settings, lack the robustness required to detect complex, high-quality deepfakes across diverse datasets. Their inability to generalize to unseen deepfake generation techniques often results in reduced detection accuracy and increased susceptibility to adversarial attacks. Thus, there is a need for more adaptive and scalable detection models that can address these shortcomings.

In this work, we propose a novel deepfake detection framework based on Vision Transformers (ViTs). ViTs have recently gained attention in the field of computer vision due to their ability to capture long-range dependencies and model global contextual information more effectively than CNNs. By leveraging the strengths of transformers, our model is designed to enhance feature extraction and improve detection accuracy, even in the face of subtle and sophisticated manipulations found in high-fidelity deepfakes.

The primary contributions of this paper are as follows:

- We propose a hybrid model using **Transformer** for local feature extraction and capturing long-range dependencies in deepfake videos.
- We did extensive experiments on the **FaceForensics++** and **Celeb-DF** datasets to evaluate the model's generalization across different deepfake generation techniques.
- We provide a comprehensive analysis of our model's performance, comparing it with state-of-the-art methods to demonstrate its effectiveness in improving both detection accuracy and robustness.

2 Related Work

Wang and et al. [**df16**] presented a Siamese network-based approach in order to make deepfake detection less sensitive to image quality variations. The network leverages both original and degraded images to effectively provide steady segmentation maps of the tampered regions. Besides, the model utilized a Mask-Guided Transformer to model the cooccurrences between manipulated areas and their neighbors. Their multi-task learning framework amalgamates segmentation, classification, and localization invariance losses. In comparison with state-of-the-art techniques, *LiSiam* accomplished better performance with 91.44% AUC on the FaceForensics++ dataset in low-quality settings, with strong generalization for cross-database evaluation.

A survey on deepfake video detection techniques using deep learning by Athirasree Das and et al. [**df15**] conducted a comparative study of various deep learning models for detecting deepfake videos. They reviewed techniques such as CNN (ResNet, VGG16, EfficientNet) and RNN (LSTM), highlighting how deep learning methods can capture features that distinguish between real and fake videos. The authors emphasized that hybrid models combining CNN and RNN solve issues like temporal inconsistencies and provide better detection results. They concluded that CNN models combined with SVM achieved the highest accuracy (98%) on the DFDC dataset, surpassing CNN-RNN hybrid approaches in some cases.

Aditya Devasthale et al.[**df13**] (2022) proposed an adversarially robust method for detecting deepfake videos. They utilized a VGG19-based deep network architecture, enhanced by adversarial training using the Iterative Fast Gradient Sign Method (I-FGSM), to improve the accuracy of deepfake detection. The study achieved notable results in adversarial robustness, demonstrating the model's ability to withstand various white-box adversarial attacks. Their experiments on the FaceForensic++ dataset showed significant improvements in accuracy compared to non-adversarial models, establishing the effectiveness of the proposed approach. 4. Comparative Analysis of Deepfake Video Detection Using Inception Net and Efficient Net Geetha Rani E et al. (2022) performed a comparative analysis of deepfake video detection techniques using Inception Net and Efficient Net. The paper emphasizes the effectiveness of convolutional neural networks (CNN) in detecting deepfakes. EfficientNet achieved significantly higher precision than Inception Net, demonstrating the benefit of combining local and global image processing for anomaly detection. The proposed EfficientNetB7 model obtained an accuracy of 83.4% on the ImageNet dataset, surpassing earlier models in terms of speed and accuracy. Extensive experiments confirmed that Efficient Net, in combination with Vision Transformers, outperformed traditional CNN models in deepfake detection.

Kaddar et al. (2021) proposed a deepfake detection method named HCiT, which combines Convolutional Neural Networks (CNN) with Vision Transformer (ViT). The hybrid model leverages CNN's ability to extract local information and ViT's self-attention mechanism for detecting deepfakes. The HCiT method was tested on the Faceforensics++ and DeepFake Detection Challenge preview datasets, significantly outperforming state-of-the-art methods. The model achieved high generalization across various deepfake generation techniques, showcasing competitive performance, with 96% accuracy on the Faceforensics++ dataset and 97.82% on DFDC.

Doshi et al. (2022) proposed a real-time deepfake detection system using Video Vision Transformer (ViViT). The system extracts spatio-temporal features from videos and applies a transformer-based architecture for deepfake detection. It was tested on the FaceForensics++ dataset and achieved a training accuracy of 95.8%, testing accuracy of 93.1%, and validation accuracy of 93.4%. Compared to traditional models like ResNet-50 and Xception, ViViT demonstrated superior performance with an inference time of 3.84 ms, making it effective for real-time detection of fake videos.

Yang Yu and et al. [**df09**] in their paper *MSVT: Multiple Spatiotemporal Views Transformer for DeepFake Video Detection*, the authors introduced a novel framework called Multiple Spatiotemporal Views Transformer (MSVT). This framework incorporates two components: the *Local Spatiotemporal View (LSV)* and *Global Spatiotemporal View (GSV)*, designed to capture dynamic inconsistencies in video sequences. For the LSV, consecutive frames are processed to mine local temporal inconsistencies, which are then passed through a temporal transformer and feature fusion module to generate group-level spatiotemporal features. The GSV takes into account the entire video by feeding frame-level features through another temporal transformer and feature fusion module to extract global temporal clues. Finally, a Global-Local Transformer (GLT) is used to integrate both local and global features, leading to the extraction of comprehensive and subtle spatiotemporal clues. Experiments on six large-scale datasets demonstrated significant

improvements in DeepFake detection accuracy, showcasing the effectiveness of the MSVT approach.

In the papaer A Novel Framework Based on a Hybrid Vision Transformer and Deep Neural Network for Deepfake Detection, Shahin et al. (2024) [**df02**] proposed a hybrid deepfake detection framework combining Vision Transformer (ViT) and Convolutional Autoencoders (CAE) to address the limitations of traditional forgery detection methods. The framework introduces two models: the first integrates ViT with CAE for enhanced image analysis and reconstruction, while the second uses CAE with classical machine learning algorithms to improve feature extraction and classification. The proposed models achieved an accuracy of approximately 87%, showcasing enhanced performance compared to state-of-the-art techniques.

Dagar et al. (2023) [**df05**] proposed a hybrid deepfake video detection model using Xception and LSTM with channel and spatial attention mechanisms (CBAM). The Xception model, employing depthwise separable convolution, captures spatial artifacts, while LSTM handles temporal discrepancies across frames in manipulated videos. The CBAM module refines the feature maps along both spatial and channel dimensions. The model was evaluated on the Div-DF dataset, consisting of face-swap, facial reenactment, and lip-sync manipulations, achieving an accuracy of 93% and an AUC of 0.98, outperforming several state-of-the-art deepfake detection models.

Sun et al. (2023)[**df06**] introduced a multi-scale Convolution-Transformer network, named ConTrans-Detect, for deepfake video detection. This model integrates a multi-scale CNN module for spatial feature extraction using 3D Inception blocks and a multi-branch Transformer module for capturing temporal dynamics. The approach effectively identifies subtle changes in video frames by learning low-level spatial features alongside temporal variations. The model achieved an AUC of 0.929 and an F1 score of 0.920 on the DeepFake Detection Challenge dataset, outperforming several state-of-the-art models while maintaining fewer parameters, highlighting its efficiency and effectiveness.

Mohsin Albazony et al. (2023)[**df08**] proposed a novel model for deepfake video detection by combining Convolutional Neural Networks (CNN) and Recurrent Neural Networks (RNN) alongside image preprocessing techniques. The model was evaluated using a dataset containing 135 real videos and 677 fake videos generated through various manipulation tools. Performance was assessed under two scenarios: different dimensions of training data and varying sizes of the training data. The results indicated that the proposed model outperformed previous models, achieving significant improvements in accuracy, precision, recall, and F-Measure, thus demonstrating its effectiveness in distinguishing between real and fake videos.

In the paper ISTVT: interpretable spatial-temporal video transformer for deepfake detection (ISTVT) (2023), C Zhao el at. [**df11**] designed to enhance both the performance and interpretability of deepfake detection models. The ISTVT leverages a decomposed spatial-temporal self-attention mechanism alongside a self-subtract mechanism, which improves detection accuracy and robustness over existing video-based deepfake detection approaches, including other video transformers. A key contribution of this work is a novel visualization technique that separates and displays the temporal and spatial features captured by the model, allowing for deeper insights into which features contribute to the prediction process. This interpretability method is generalizable and can be applied to other video transformers as well. Future research aims to optimize this visualization strategy to further enhance the model's interpretability and effectiveness in detecting deepfakes.

3 Research Objective and Approach

The objective of this research is to enhance the effectiveness of deepfake video detection using Vision Transformers (ViTs). Deepfakes pose significant societal risks, and while Convolutional Neural Networks (CNNs) have been widely employed in video forgery detection, they often struggle with the generalization of spatiotemporal information inherent in videos. To address this challenge, the study leverages the transformer architecture with its inherent ability to capture long-range dependencies and spatial relationships between frames, aiming to improve detection accuracy and robustness.

3.1 The proposed approach involves:

- **Utilizing Vision Transformers (ViT)**: Instead of relying solely on CNNs, the ViT is used for better feature extraction from video frames. This transformer-based architecture captures both spatial and temporal patterns from the frames.
- **Dataset Enhancement**: To prevent over-fitting, a large dataset of deepfake videos is employed (FaceForensics++), ensuring a diverse range of manipulations and challenges for the model.
- **Training and Optimization**: The model is trained using advanced optimization techniques to ensure high accuracy, precision, and generalizability across different types of deepfakes.

4 Algorithm

Vision Transformer (ViT) is a new structure that first successfully adapted the Transformer Model from natural language processing applications for visual tasks. This structure is described in the paper published by Dosovitskiy et al. [**df01**] in 2020 and quickly became a major trend in the deep learning world when working with images. The main difference between traditional computer vision approaches, which include convolutional neural networks (CNNs), is that ViT does not have any convolutional layers, using only attention arrays instead.

4.1 Architecture

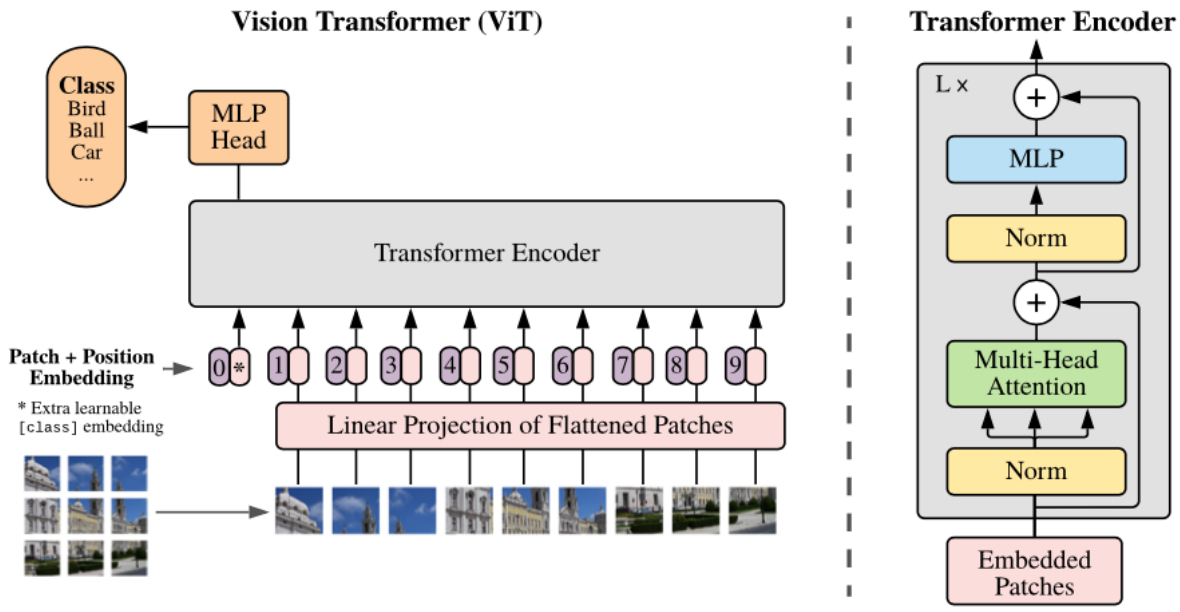

Fig. 1. ViT Architecture
[**df01**]

The Vision Transformer (**ViT**), as illustrated in the Fig 1 is designed to operate similarly to transformers used in natural language processing, but instead of processing words, it processes image patches. Below is a detailed breakdown of the architecture based on the diagram:

4.2 Patch and Position Embedding

Input : The Vision Transformer takes as input an image, which is a full rectangular surface and is first cut into square patches without overlap. Each patch is usually of dimension 16×16 pixels but may vary from this size in a given implementation.

Flattening : Each of these patches is then flattened into a vector. For example, for a patch of size $\mathbf{16 \times 16}$ in RGB coloration, this equates to a vector of size $\mathbf{16 \times 16 \times 3} = 768$.

Patch Embedding : The flat patches are mapped to a fixed-length vector using a linear projection that can be learned. This process is quite similar to how we have token embeddings in NLP.

Positional Embedding : Since transformers do not understand the position of patches with respect to others, it is necessary to append some positional information to every patch embedding. These embeddings specify in which area of the image each of the patches is situated. This step concludes with a sequence of embedded patches that consists of the patch content in addition to its respective geographical location.

4.3 Transformer Encoder

Input Sequence : Following the patch embedding and the position encoding, the input transforms into a series of vectors (or tokens), each of the token being a patch in the respective position in the image.

Class Token : The architecture incorporates a class token that is learnable at the start of the sequence, much in the same way left a "class" token in NLP transformers known as "[CLS]" which gathers up information of all the patches and is ultimately used for classification.

Transformer Layers The sequence of patch embeddings is joined with the class token and they are then transformed through a series of Transformer Encoder layers. Each layer is made up of:

- **Multi-Head Self-Attention:** The self-attention mechanism permits the model to make connections between various patches, hence it is able to comprehend the large image patches with the help of long-range dependencies.
- **Layer Normalization:** Normalization layers are used to maintain the training process volatile.
- **Feed-Forward Neural Network (MLP):** After the self-attention mechanism, a feed-forward neural network is applied to each token. This is typically a two-layer MLP with a ReLU activation function in between.
- **Residual Connections:** Residual connections are included in both the self-attention layer and the feed-forward neural network to make sure that the gradients flow smoothly during training.

This process is repeated for L layers, where L is the number of transformer encoder layers in the model.

4.4 MLP Head for Classification

- At the end of the transformer encoder layers, the original sequence of patch tokens and also the class token is used for classification by the class token.
- The class token's output is fed into an MLP Head composed of one or more fully connected layers shrinking the capacity and a softmax or another activation function is then used for classification.
- **Output**: The output of this MLP head is the class label (e.g., bird, ball, car), which tells what the image is.

5 Dataset

For training and testing our deepfake detection model, we used **FaceForensics++** dataset. This dataset is well-known in the deepfake detection community for their high-quality video manipulations and diverse forgery techniques, making them ideal for evaluating the model's performance under real-world conditions.

5.1 FaceForensics++

FaceForensics++ is a comprehensive dataset that contains manipulated facial videos and provides a benchmark for both video and image-based forgery detection. It is widely used for training and evaluating deepfake detection models, including various forgery types such as deepfakes, Face2Face, FaceSwap, and NeuralTextures. This dataset offers a comprehensive benchmark for both image and video forgery detection models and serves as a cornerstone for evaluating the robustness of deepfake detection methods.

6 Implementation

The implementation of the deepfake detection model involves several key stages, from data preprocessing to model evaluation. The overall process can be divided into the following steps:

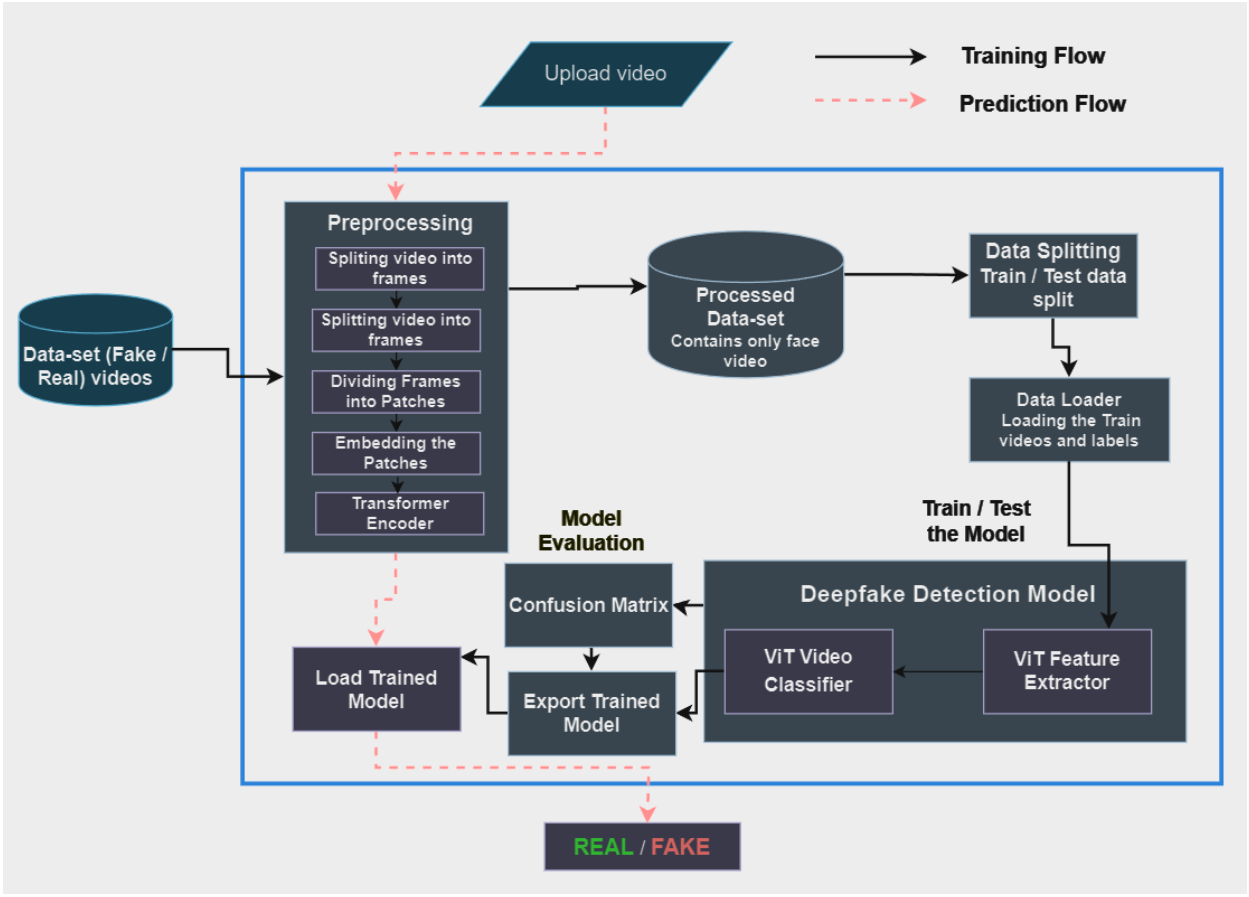

Fig. 2. Application of Implementation
[**df01**]

6.1 Data Preprocessing

Before training the model, the input videos from the FaceForensics++ dataset undergo preprocessing. The videos are split into individual frames, and face detection is applied to isolate only the facial regions. These cropped facial frames are then resized into consistent dimensions and saved for further use. This step ensures that the model focuses on the facial area, which is critical for detecting subtle manipulations in deepfake videos.

6.2 Dataset Splitting

The processed frames are divided into training and testing sets. This splitting allows the model to learn from the training set and generalize to unseen data from the testing set, which helps in evaluating its performance.

6.3 Loading Data with DataLoader

A DataLoader is utilized to efficiently load batches of data (videos and their corresponding labels) into the model during training. This process ensures that the GPU/CPU is optimally used by feeding data in manageable chunks, reducing memory overhead, and speeding up the training process.

6.4 Model Training

: The model is trained using a supervised learning approach, with the video labels (real or fake) serving as the ground truth. The training process involves optimizing the model's parameters using a loss function (e.g., cross-entropy loss) and an optimizer like Adam. During training, the model learns to minimize the error between its predictions and the actual labels by updating its internal weights.

6.5 Evaluation and Testing

After training, the model is evaluated on the test set using metrics such as accuracy, precision, recall, and F1-score. A confusion matrix is generated to assess the model's performance in distinguishing between real and fake videos. The trained model is then exported for deployment. **Deepfakes** represent manipulated media—be it images, videos, or audio—that alter the original context, often creating misleading content that can deceive viewers. Using advanced AI techniques such as Generative Adversarial Networks (GANs), deepfake creators can modify facial expressions, voices, or even simulate entire videos. Notably, these manipulations can be convincing, as demonstrated by viral videos, including fabricated celebrity appearances.

While deepfakes may serve benign purposes such as satire or entertainment, they also pose significant threats across various domains. From privacy concerns to political manipulation and corporate espionage, deepfakes have the potential to undermine public trust and damage reputations. The rapid improvement in deepfake quality requires equally advanced detection mechanisms to mitigate their spread.

Traditional methods for detecting deepfakes rely heavily on convolutional neural networks (CNNs) and recurrent neural networks (RNNs), which focus on local pixel patterns. However, these models may struggle with identifying subtle or widespread inconsistencies across video frames. In contrast, Vision Transformers (ViTs) excel at capturing long-range dependencies by using a self-attention mechanism, making them a promising approach for detecting deepfakes. By processing entire video frames as sequences, ViTs can identify minor artifacts and manipulations more effectively.

In this paper, we present a novel deepfake detection system using ViTs, which outperforms conventional CNN-based approaches. Our work contributes to improving the detection accuracy and robustness of models against adversarial manipulations. The following sections detail our methodology, experiments, and results, underscoring the potential of ViTs in enhancing digital content authenticity.
Deepfakes represent manipulated media—be it images, videos, or audio—that alter the original context, often creating misleading content that can deceive viewers. Using advanced AI techniques such as Generative Adversarial Networks (GANs), deepfake creators can modify facial expressions, voices, or even simulate entire videos. Notably, these manipulations can be convincing, as demonstrated by viral videos, including fabricated celebrity appearances.

While deepfakes may serve benign purposes such as satire or entertainment, they also pose significant threats across various domains. From privacy concerns to political manipulation and corporate espionage, deepfakes have the potential to undermine public trust and damage reputations. The rapid improvement in deepfake quality requires equally advanced detection mechanisms to mitigate their spread.

Traditional methods for detecting deepfakes rely heavily on convolutional neural networks (CNNs) and recurrent neural networks (RNNs), which focus on local pixel patterns. However, these models may struggle with identifying subtle or widespread inconsistencies across video frames. In contrast, Vision Transformers (ViTs) excel at capturing long-range dependencies by using a self-attention mechanism, making them a promising approach for detecting deepfakes. By processing entire video frames as sequences, ViTs can identify minor artifacts and manipulations more effectively.

In this paper, we present a novel deepfake detection system using ViTs, which outperforms conventional CNN-based approaches. Our work contributes to improving the detection accuracy and robustness of models against adversarial manipulations. The following sections detail our methodology, experiments, and results, underscoring the potential of ViTs in enhancing digital content authenticity.
Deepfakes represent manipulated media—be it images, videos, or audio—that alter the original context, often creating misleading content that can deceive viewers. Using advanced AI techniques such as Generative Adversarial Networks (GANs), deepfake creators can modify facial expressions, voices, or even simulate entire videos. Notably, these manipulations can be convincing, as demonstrated by viral videos, including fabricated celebrity appearances.

While deepfakes may serve benign purposes such as satire or entertainment, they also pose significant threats across various domains. From privacy concerns to political manipulation and corporate espionage, deepfakes have the potential to undermine public trust and damage reputations. The rapid improvement in deepfake quality requires equally advanced detection mechanisms to mitigate their spread.

Traditional methods for detecting deepfakes rely heavily on convolutional neural networks (CNNs) and recurrent neural networks (RNNs), which focus on local pixel patterns. However, these models may struggle with identifying subtle or widespread inconsistencies across video frames. In contrast, Vision Transformers (ViTs) excel at capturing long-range dependencies by using a self-attention mechanism, making them a promising approach for detecting deepfakes. By processing entire video frames as sequences, ViTs can identify minor artifacts and manipulations more effectively.

In this paper, we present a novel deepfake detection system using ViTs, which outperforms conventional CNN-based approaches. Our work contributes to improving the detection accuracy and robustness of models against adversarial manipulations. The following sections detail our methodology, experiments, and results, underscoring the potential of ViTs in enhancing digital content authenticity.