Transforming Detection: Vision Transformer-Based Deepfake Video Identification

Govinda Mandal[1] and Gautam Kumar[2]

[1] National Institute of Technology, Delhi, India
232211009@nitdelhi.ac.in
[2] National Institute of Technology, Delhi, India
gautam@nitdelhi.ac.in

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: Deepfake detection, Vision Transformers, Self-attention mechanism, Computer vision, Machine learning, Adversarial attacks.

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. 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.

2 Related Work

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 adversar-

ial manipulations. The following sections detail our methodology, experiments, and results, underscoring the potential of ViTs in enhancing digital content authenticity.

3 Objective

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.

4 Algorithm

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.

5 Dataset

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.

6 Implementation

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 au-

thenticity.

7 Results

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.

8 Conclusion

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.

Acknowledgment

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.