mathematics

MDPI

Article

Unmasking Deception: Empowering Deepfake Detection with Vision Transformer Network

Muhammad Asad Arshed [1,2,*], Ayed Alwadain [3], Rao Faizan Ali [2], Shahzad Mumtaz [4], Muhammad Ibrahim [1] and Amgad Muneer [5,6,*]

[1] Department of Computer Science, The Islamia University of Bahawalpur, Bahawalpur 63100, Pakistan; muhammad.ibrahim@iub.edu.pk
[2] School of Systems and Technology, University of Management and Technology, Lahore 54770, Pakistan; faizan.ali@umt.edu.pk
[3] Computer Science Department, Community College, King Saud University, Riyadh 145111, Saudi Arabia; aalwadain@ksu.edu.sa
[4] Department of Data Science, The Islamia University of Bahawalpur, Bahawalpur 63100, Pakistan; shahzad.mumtaz@iub.edu.pk
[5] Department of Imaging Physics, The University of Texas MD Anderson Cancer Center, Houston, TX 77030, USA
[6] Department of Computer and Information Sciences, Universiti Teknologi Petronas, Seri Iskandar 32160, Malaysia
* Correspondence: asad.arshed@umt.edu.pk (M.A.A.); muneeramgad@gmail.com (A.M.)

Abstract: With the development of image-generating technologies, significant progress has been made in the field of facial manipulation techniques. These techniques allow people to easily modify media information, such as videos and images, by substituting the identity or facial expression of one person with the face of another. This has significantly increased the availability and accessibility of such tools and manipulated content termed 'deepfakes'. Developing an accurate method for detecting fake images needs time to prevent their misuse and manipulation. This paper examines the capabilities of the Vision Transformer (ViT), i.e., extracting global features to detect deepfake images effectively. After conducting comprehensive experiments, our method demonstrates a high level of effectiveness, achieving a detection accuracy, precision, recall, and F1 rate of 99.5 to 100% for both the original and mixture data set. According to our existing understanding, this study is a research endeavor incorporating real-world applications, specifically examining Snapchat-filtered images.

Keywords: deepfake; identification; Vision Transformer; pretrained; fine tuning

MSC: 68T07

check for updates

Citation: Arshed, M.A.; Alwadain, A.; Faizan Ali, R.; Mumtaz, S.; Ibrahim, M.; Muneer, A. Unmasking Deception: Empowering Deepfake Detection with Vision Transformer Network. *Mathematics* **2023**, *11*, 3710. https://doi.org/10.3390/math11173710

Academic Editors: Sheng Li, Zhenjun Tang and Guorui Feng

Received: 28 July 2023
Revised: 18 August 2023
Accepted: 24 August 2023
Published: 29 August 2023

1. Introduction

The explosion of social media platforms and the widespread availability of affordable equipment such as cameras, cellphones, and laptops over the last decade led to an exponential increase in online material, particularly in the form of images and movies. These platforms have revolutionized the way individuals exchange and broadcast information, allowing for the quick distribution of content and simple access to a wide collection of media.

The human face is the most distinguishing feature of an individual, and it plays an important role in identity recognition and communication. Rapid advancements in face synthesis technology have created a significant security risk. In some circumstances, deepfake technology, a subdomain of artificial intelligence (AI), has raised concerns about the authenticity and integrity of facial photographs.

To generate high-resolution deepfake images, this technology requires complicated algorithms, typically based on deep learning (DL) models such as generative adversarial networks (GANs) [1]. The spread of deepfake technology brings several issues and potential hazards to numerous industries. For example, in the cybersecurity empire, the capacity to convincingly manipulate facial photos raises worries about identity theft, deception, and unauthorized access to critical information. Furthermore, the popularity of deepfakes poses a significant risk to public trust, as unscrupulous actors can use this technology to create deceptive visual indicators, spread misinformation, or harm individuals' reputations. To address these concerns, academics have concentrated their efforts on developing approaches for detecting and mitigating the impact of deepfakes [2]. This encompasses leveraging advancements in computer vision, machine learning, and forensic analysis to identify significant signs of image manipulation and distinguish between genuine and manipulated facial images. Understanding the underlying mechanisms of deepfake generation and developing robust detection methods are essential for the integration of visual content and maintaining trust in digital media.

Several methods were proposed to identify deepfakes and most of the methods are based on deep learning. The United States Defense Advanced Research Projects Agency (DARPA) recently launch media forensic research to develop methods for the detection of fake media [3]. Furthermore, Facebook with the collaboration of Microsoft launches the AI-based deepfake identification challenge [4].

The paper is structured as follows: In Section 2, we present the literature review and discuss related works addressing the deepfake problem using deep learning. Section 3 outlines the proposed research methodology and provides a detailed explanation of the validation dataset used for evaluating the proposed Vision Transformer (ViT) model. Moving to Section 4, we present the results and conduct an in-depth analysis. Finally, in Section 5, we conclude the study, emphasizing key takeaways and identifying potential directions for future research.

2. Literature Review

There are currently many well-known methods proposed for the identification of fake images, but the generalization capability of these models is significantly low. The performance of these models drops due to the frequent updating of deepfake or manipulation methods. Akhtar et al. [5] considered SqueezeNet [6], VGG16 [7], ResNet [8], DenseNet [9], and GoogleNet [10] in their study for the identification of face manipulation. They achieved effective accuracy for the same manipulation type of training and testing samples, but the performance decreased for the novel manipulation that was not considered during the training.

Z. Akhtar and D. Dasgupta [11] examined the possibility of local feature descriptors to recognize manipulated faces. This study reported on a comparative experimental analysis of ten local feature descriptors using the 'DeepfakeTIMIT' database.

In the study of Bekci et al. [12], a deepfake detection system utilizing metric learning and steganalysis-rich models is proposed to improve performance under unseen data and manipulations. To assess the generalization of the suggested approach, an empirical analysis was conducted on the FaceForensics++, DeepFakeTIMIT, and CelebDF data sets, which are all openly accessible. Their suggested framework achieved accuracy increases of 5% to 15% when subjected to hidden modifications.

Li et al. [13] conducted a study in which they observed distinctions between eye-blinking patterns in deepfake videos compared to those exhibited by humans. As a result, they developed a novel eye-blinking detection technique specifically designed for identifying deepfake videos. Gupta et al. considered EEG signal features and eye movement for the identification of deepfake videos.

In the study of Nguyen et al. [14], the eyebrow region was utilized as a set of features for detecting deepfake videos. Four deep learning methods, namely LightCNN, Resnet, DenseNet, and SqueezeNet, were employed for this purpose. Notably, the achieved highest AUC (Area Under Curve) values on the UADFV and Celeb-DF data sets were 0.984 and 0.712, respectively.

Trans-DF, a deepfake detection approach based on random forests, was proposed by Patel et al. [15]. The Trans-DF model's detection accuracy of 0.902 demonstrates how well it can spot deepfake videos. Yang et al. introduced an approach based on SVM classifiers for distinguishing deepfake images and videos. Their method leverages the differences in head poses as crucial features for discrimination. Utilizing this technique, they successfully developed a system that effectively detects and distinguishes deepfake content with an AUROC score of 0.890.

By using biological cues to analyze residuals, Ciftci et al. [16] introduced a novel method for locating the sources of deepfake content. This innovative study was the first to use biological clues in deepfake source detection. They used the Face Forensics++ data set for experimental evaluations, including several ablation tests, to validate their approach. Surprisingly, they were able to identify sources with an accuracy rate of 93.39% using four deepfake generators. These results demonstrate the effectiveness of their suggested approach and its potential for precisely locating the origins of deepfake content.

A deepfake detection model based on machine learning named MSTA_Net was introduced by Yang et al. [17] in 2022. This model focused on evaluating an image's texture properties to find abnormalities indicating deepfake manipulation. Notably, the MSTA_Net model considered the full image instead of only the facial regions. The model established relationships between the forged and unmanipulated areas of the image. The detection procedure includes looking for irregularities in the texture of the image. Any variations were flagged as fake if they were found. On the other hand, if no variations were discovered, the image was labeled as non-fake, indicating a higher probability of authenticity. Based on the overall texture characteristics, the identification of genuine and manipulated images is possible with their proposed model.

In the field of deepfake discrimination, the integration of multiple attention mechanisms and models has emerged as a crucial approach. In a recent study, Zhao et al. [18] introduced a multi-attentional deepfake detection method designed to identify subtle and partial features present in both real and fake images. The proposed technique consisted of three essential components. First, multiple spatial attention heads were employed to focus on distinct regions of the images, enabling the model to capture intricate details. Second, a textural-feature-enhancement block was incorporated to enhance the discriminative power of the detected features. Third, an aggregate module was utilized to consolidate the information gathered from the various attention heads and facilitate a comprehensive decision-making process. By leveraging these components, the multi-attentional deepfake detection method proposed by Zhao et al. [18] aimed to improve the accuracy and effectiveness of deepfake discrimination by effectively capturing subtle visual cues and enhancing textural features.

In their research, Wang et al. [19] presented a novel deepfake detection approach using a multi-modal, multi-scale transformer. The proposed model was designed to effectively identify deepfake images by analyzing various image patches of different sizes. By employing a multi-scale approach, the model addressed the need to capture local inconsistencies at different spatial levels within the image. This allowed for the detection of subtle manipulations or artifacts that may be present in different regions of the image. The multi-modal aspect of the proposed model indicates that it incorporates multiple sources of information or modalities to enhance the detection process. Overall, the multi-modal, multi-scale transformer model introduced by Wang et al. [19] offers a promising approach to deepfake detection, enabling the analysis of image patches at different spatial levels and leveraging multiple modalities for improved accuracy and robustness. Based on the localization and the utilization of the VGG16 model for the detection of various forgery types, as explored by Shelke et al. [20,21], Wang et al. [22] proposed an approach based on frequency domain analysis and residual networks for the purpose of identifying or detecting deepfake content.

CNNs have exhibited impressive accuracy in the realm of deepfake identification, underscoring their significance in this domain. Despite CNNs' ability to capture features of minute objects within images using deep architectures, pinpointing critical regions accurately can pose challenges. To address this limitation, our study incorporates the Vision Transformer model. In the ViT framework, the input image undergoes segmentation into blocks during the model's general training phase, treating each block as an independent entity. Through self-attention modules, the ViT model discerns relationships among these embedded patches. Notably, ViT has showcased exceptional performance in conventional classification tasks. The transformer's self-attention mechanism enhances the importance of pivotal features while mitigating the influence of noise-inducing features [23]. Inspired by this particular standpoint, the present investigation introduces a deepfake image recognition system utilizing the Vision Transformer (ViT) architecture. The results demonstrate that the suggested framework produces favorable results in the realm of deepfake image identification. This study brings forth notable contributions to the discipline in the subsequent manners:

- The fine-tuned ViT model presented in this study demonstrates superior performance compared to existing state-of-the-art models in the domain of deep-fake identification.
- A patch-wise self-attention module and global feature extraction technique considered in this study.
- Evaluating model for real-world in the deepfake detection task, with a focus on Snapchat.
- After conducting a thorough analysis of various standard data sets, our research substantiates the exceptional robustness and generalizability of the proposed method, surpassing numerous state-of-the-art techniques.

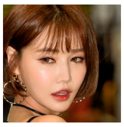
3. Materials and Methods

In this section, we have outlined and presented the methodologies that were employed and proposed to attain accurate identification of fake images.

3.1. Data Set

For our experimental investigation, we employed a data set procured from [24], see Table 1. However, it is essential to acknowledge that the data set size was constrained by the available resources. As a result, our study encompassed a selection of 100,000 images, evenly split between 50,000 authentic images and 50,000 images generated using GAN techniques, all sourced from the original data set. To ensure optimal model performance, we have trained the model with a balanced data set.

Table 1. Real and GAN Generated (Fake) Images.

					
Fake	Fake	Fake	Real	Real	Real

3.2. ViT Architecture

This section introduces the ViT framework, highlighting its key principles, structure, self-attention mechanism, multi-headed self-attention, and the mathematical foundations that motivate its design. Initially, the ViT was introduced in 2020 [25] as a deep neural network architecture specifically optimized for image recognition workloads. It extends the Transformer architecture, which was originally designed for natural language processing, and incorporates the innovative concept of considering images as sequences of tokens, which are commonly represented by image patches. ViT effectively handles these token sequences by leveraging the capabilities of the transformer design. Notably, the transformer architecture that underpins ViT has been successfully applied to a variety of tasks, including picture restoration and object detection [26,27], demonstrating its broad applicability and effectiveness [28].

Tokenization and embedding of the input picture are crucial stages in the ViT architecture. The image is divided into a grid of non-overlapping patches, flattened, and mapped to a higher-dimensional space using a linear transformation followed by normalization. The ViT model gains the capacity to capture both global and local information from the image by conducting tokenization and embedding, aiding comprehensive learning.

The Transformer architecture, while capable of processing sequences, does not explicitly take into account the positioning information of each token inside the sequence. The ViT design uses pre-defined positional embeddings to address this constraint. These embeddings are supplementary vectors that encode the position of each token in the sequence before being sent into the transformer layers. The model can now grasp the relative positions of tokens and extract spatial information from the input image thanks to this integration.

The Multi-head Self-Attention (MSA) mechanism is at the heart of the ViT design. This component enables the model to attend to many parts of the image at the same time. MSA is made up of distinct "heads," each of which computes attention independently. These attention heads can focus on different portions of the image, resulting in a variety of representations that are then concatenated to generate the final image representation. By attending to several sections continuously, the ViT can record complicated interactions between input elements. This enhancement, however, increases complexity and computational cost because it necessitates more attention to heads and more processing to aggregate the outputs from all heads. *MSA* can be stated mathematically as follows:

$$MSA(q,k,v) = Concat\ (h1, h2, \ldots, hn) \tag{1}$$

In Equation (1), the letters q, k, and v represent the query, key, and value inputs, respectively. The self-attention mechanism is the cornerstone of transformers, allowing for explicit modeling of interactions and linkages across all sequences in prediction tasks. Unlike CNNs, the self-attention layer gathers insights and features from the whole input sequence to collect both local and global information. This distinguishing feature of self-attention distinguishes it from CNNs since it promotes a more comprehensive interpretation and representation of the information.

The attention mechanism computes the dot product between the query and key vectors, normalizes the attention scores using SoftMax, and modulates the value vectors to provide enhanced output representation. A study was conducted by Cordonnier et al. [29] to investigate the link between self-attention and convolution operations. Their findings demonstrated that when endowed with a large number of characteristics, self-attention emerges as a highly flexible and versatile mechanism capable of extracting both local and global properties. This shows that self-attention is more versatile and adaptable than typical convolution procedures.

The abstract level ViT network diagram can be seen in Figure 1 and is based on the following main components of the ViT model.

Patch Embedding: In ViT, the input image is divided into fixed-size non-overlapping patches. Each patch is linearly projected to an embedding space using a learned linear transformation represented by the matrix. This step transforms the 2D spatial information of the image into a sequence of embeddings.

Positional Embedding: Since the transformer architecture does not inherently understand the spatial arrangement of the patches, positional information needs to be injected. Positional embeddings are added to the patch embeddings to provide information about their spatial positions within the image.

Transformer Encoder: The positional embeddings (E_POS) are passed through a transformer encoder. The encoder consists of multiple layers, and each layer contains self-attention mechanisms and feedforward neural networks. The self-attention mechanism allows each patch to attend to other patches, capturing global relationships in the image. The feedforward neural networks further process the attended representations. The output of the encoder is a set of contextualized embeddings for each patch, which captures both local and global image information.

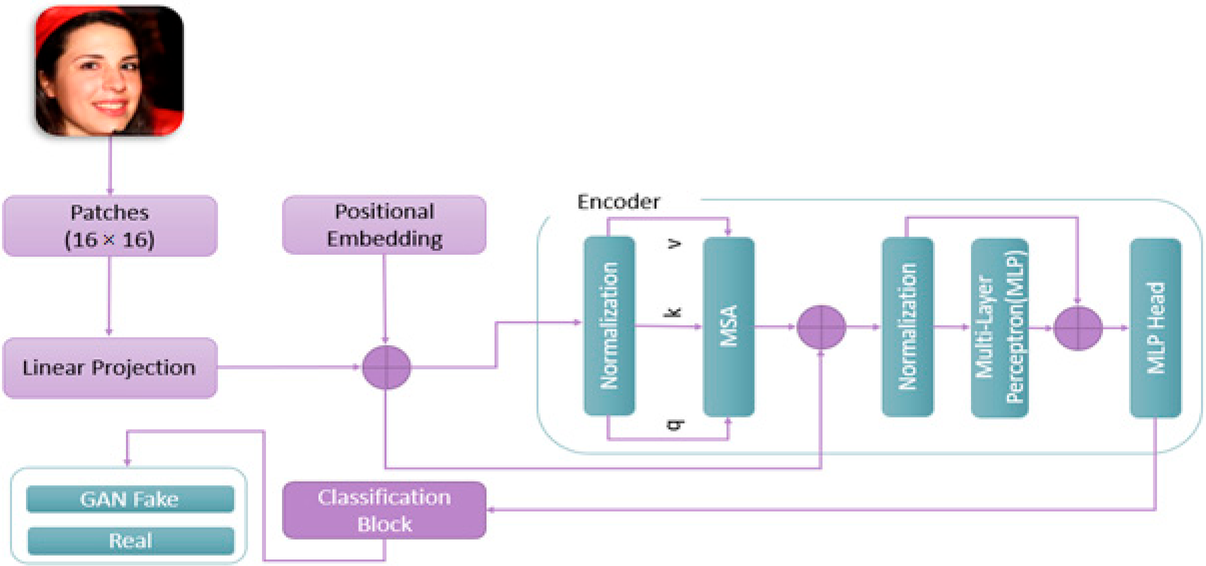


Figure 1. Abstract Level Architectural Diagram of the ViT [25].

Classification Head: The final contextualized embeddings from the transformer encoder are used for downstream tasks such as image classification. For classification tasks, the contextualized embeddings can be processed in different ways. One common approach is to take the average of all embeddings or a specific token's embedding (e.g., a classification token) and pass it through one or more fully connected layers to make class predictions.

3.3. Hyper-Parameters for ViT Pretrained Model

In this research study, the initial images are preprocessed and resized to 224 × 224 that further separated into patches of size 16 × 16 pixels. The technique of reducing the input image into smaller fixed-size patches comprises splitting the image into 16-pixel-wide and 16-pixel-tall pieces.

The model used in this work was trained on a large data set known as ImageNet-21k. This data set, which contains approximately 14 million photos classified into 21,841 different classes, is specifically tailored for large-scale image classification tasks. The model's structure consists of 12 transformer layers, each with 768 hidden components. The overall capacity of the model is reflected in its 85.8 million trainable parameters which is helpful in the learning process. The parameter values and configurations employed in the ViT model can be seen in Table 2.

Table 2. ViT Configurations.

Parameters	**Values**
Encoder and Pooling Layers Dimensionality	768
Transformer Encoder Hidden Layers	12
Feed-Forward Layer Dimensionality	3072
Hidden Layers Activation	Gelu
Hidden Layer Dropout	0.1
Image Size	224 × 224
Channels	3
Patches	16 × 16
Balanced	True

Figure 2 presents the abstract-level diagram illustrating the proposed methodology.

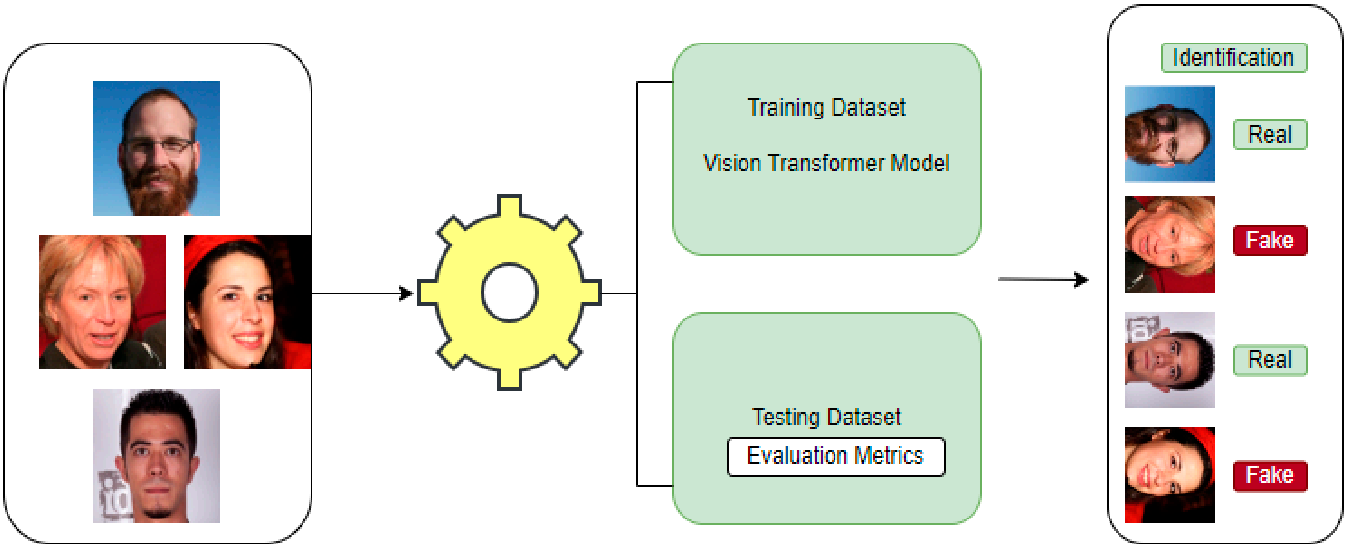


Figure 2. Abstract Level Diagram of the Proposed Methodology.

4. Experiments Results & Discussion

This section provides a comprehensive discussion of the evaluation measures, experimental details, and the results obtained through the proposed methodology.

4.1. Evaluation Metrics

Metrics for evaluating the performance of machine learning and deep learning models are essential. These measures are crucial in the fields of machine learning, deep learning, and statistical research. In this study, our attention was directed towards six essential assessment metrics to evaluate the effectiveness of our proposed model.

- Accuracy: The metric of accuracy assesses the comprehensive correctness of the model's predictions by calculating the ratio of accurately classified instances to the total samples. However, in scenarios involving imbalanced data sets or situations where distinct types of errors hold differing degrees of significance, relying solely on accuracy might not suffice for a thorough evaluation.

$$Accuracy = {}^{TP+TN}/_{TP+FP+TN+FN} \tag{2}$$

- Precision: Precision gauges a model's proficiency in recognizing positive samples within the set of actual positives. This metric quantifies the ratio of true positives to the sum of true positives and false positives.

$$P = {}^{TP}/_{TP + FP} \tag{3}$$

- Recall: The model's competence in precisely identifying positive samples from the pool of actual positives is measured using recall, which is alternatively known as sensitivity or the true positive rate. This metric is derived by calculating the ratio of true positives to the sum of true positives and false negatives. Essentially, recall provides an assessment of the thoroughness of positive predictions.

$$R = {}^{TP}/_{TP + FN} \tag{4}$$

- F1 Score: The F1 score, determined by the harmonic mean of precision and recall, serves as a singular metric that strikes a balance between these two measures. This becomes particularly advantageous in scenarios where there is an unequal distribution among classes or when there exists an equal emphasis on both types of errors. Ranging between 0 and 1, the $F1$ score attains its peak performance at 1.

$$F1 = {(2 * P * R)}\Big/{(P + R)} \tag{5}$$

4.2. Results & Discussion

We have trained the ViT model with different aspects of the data set. The classification report regarding the different aspects can be seen in the below sections.

4.2.1. Experiment 1: Results for the Kaggle Data Set

In experiment 1, we have initially considered the Kaggle [24] data set to train the ViT model from this data set due to the limited resources we have considered the 50K real images as well as 50K GAN-generated images that were labeled as fake images. From the total data set, we have used 20% data for testing purposes, and the remaining data are used for training purposes. We have trained the model for five epochs with a learning rate of 2×10^{-5}. The training loss start from 0.12000 and ended up at 0.00001 for the last epochs whereas the validation loss starts from 0.03010 and ended up on 0.00010 at the last epoch. We have achieved 100% accuracy with the fine-tuned ViT model; further, in Table 3, class-wise precision, recall, and f1 scores can be seen for 20K test images.

Table 3. ViT performance as Class Wise for Experiment 1.

Class Name	Precision	Recall	F1	Support
Real	1.0000	1.0000	1.0000	10,000
Fake	1.0000	1.0000	1.0000	10,000
Accuracy			1.0000	20,000
Macro Avg	1.0000	1.0000	1.0000	20,000
Weighted Avg	1.0000	1.0000	1.0000	20,000

To ensure the resilience of the fine-tuned model, we incorporated a website, accessible at https://thispersondoesnotexist.com/ (accessed on 1 July 2023). This website is based on the 'StyleGAN' algorithm [30] and generates distinct human faces with each visit. The samples extracted from this website are presented in Table 4. Through extensive testing using 50 images, our refined model consistently and accurately identified all of them as synthetic or "fake".

Table 4. StyleGAN [31] Testing Samples and Predicted Label.

Fake	Fake	Fake	Fake	Fake	Fake

4.2.2. Experiment 2: Real (Kaggle) + Fake (StyleGAN-Based) Data Set

To broaden the scope of our experiments, we acquired a substantial data set comprising 9451 images obtained from an online source [31], encompassing both fake/GAN images and authentic images. In addition, we collected an equal number of 9451 real images from the Kaggle source [24]. To address potential overfitting, we mitigated the issue of class imbalance by training the model on a balanced data set. To evaluate the model's performance, we allocated 33% of the data for testing purposes.

Utilizing the balanced approach, we trained the model on 6311 fake images and 6353 real images. Subsequently, we assessed its accuracy using a subset of 3140 fake images and 3098 real images. Remarkably, our model achieved an accuracy of 99.95%, as illustrated in Table 5.

Table 5. ViT Class Wise performance for Experiment 2.

Class Name	Precision	Recall	F1	Support
Real	1.0000	0.9990	0.9995	3140
Fake	0.9990	1.0000	0.9995	3098
Accuracy			0.9995	6238
Macro Avg	0.9995	0.9995	0.9995	6238
Weighted Avg	0.9995	0.9995	0.9995	6238

4.2.3. Experiment 3: Real (Kaggle) + Fake (Kaggle + StyleGAN-Based)

In order to demonstrate the effectiveness of the ViT model, we expanded our data set to include 30,000 real images sourced from Kaggle. For the fake images, we collected 15,000 from Kaggle and another 15,000 from the website https://thispersondoesnotexist.com (1 July 2023). It is important to note that this time the fake class consisted of images from two distinct sources.

To ensure the model's generalization capability, we trained it using the same parameters. We reserved 20% of the data, amounting to 12,000 images, for evaluation purposes. Impressively, our model achieved an accuracy of 99.66%, as shown in Table 6. During training, the initial training loss was 0.11620, gradually decreasing to 0.00000 by the last epoch. Similarly, the validation loss began at 0.02460 and reached 0.01270 by the final epoch. These results indicate the model's exceptional performance and strong convergence.

Table 6. ViT Class Wise performance for Experiment 3.

Class Name	Precision	Recall	F1	Support
Real	0.9951	0.9980	0.9965	5917
Fake	0.9980	0.9952	0.9966	6083
Accuracy			0.9966	12,000
Macro Avg	0.9966	0.9966	0.9966	12,000
Weighted Avg	0.9966	0.9966	0.9966	12,000

4.2.4. Model Robustness

To assess the robustness of our proposed fine-tuned ViT model, we conducted tests using filtered images. For this purpose, we utilized Snapchat [32], a widely used multimedia messaging application known for its diverse range of filters. These filters can be applied to images, enabling users to enhance or modify their appearance.

We gathered a data set comprising 24 images from Snapchat. These images depicted three distinct individuals and are presented in Table 7. Remarkably, our fine-tuned ViT model achieved a perfect accuracy of 100% when tested on these Snapchat filtered images. This result underscores the model's effectiveness in accurately identifying and classifying faces even in the presence of various filters applied to the images.

Table 7. Snapchat Images for Testing.

Real	Real	Real

The method put forward demonstrates superior performance when compared to state-of-the-art techniques, as indicated by the performance analysis presented in Table 8 of the study, while direct comparison proves challenging due to the inherent data set variations.

Table 8. Comparison between the proposed study and State-of-the-Art studies.

Authors	Method	Data Set	Evaluation Metric	Results
(Gandhi et al., 2020) [33]	ResNet Pretrained Architecture	10,000 Images	Accuracy	Test 94.75%
(Hu et al., 2021) [34]	Corneal Specular Highlights	1000 Images	Accuracy	Test 90.48%
(Yousaf et al., 2022) [35]	Two-Stream CNN	11,982 Images	Accuracy	Test Accuracy for StyleGAN 90.65%
Proposed (Experiment 1)	Vision Transformer	20,000 Test Images (Kaggle)	Accuracy Precision Recall F1	100.0% 100.0% 100.0% 100.0%
Proposed (Experiment 2)	Vision Transformer	6238 Test Images (Real Images from Kaggle and Fake from StyleGAN Based website)	Accuracy Precision Recall F1	99.95% 99.95% 99.95% 99.95%
Proposed (Experiment 3)	Vision Transformer	12,000 Test Images (Real Images from Kaggle and Fake from (Kaggle + StyleGAN Based website))	Accuracy Precision Recall F1	99.96% 99.96% 99.96% 99.96%

4.2.5. Theoretical and Practical Implications

From a theoretical standpoint, the study contributes to the field of deepfake detection by proposing the use of Vision Transformer Networks. This introduces a novel approach to tackle the escalating issue of deceptive media in the digital age.

The study’s findings have several practical implications. First, the implementation of Vision Transformer Networks enhances the accuracy and reliability of deepfake detection algorithms. This empowers individuals and organizations to identify and mitigate the harmful effects of manipulated media, such as misinformation, fraud, and privacy breaches.

Furthermore, the study’s outcomes hold promise for the development of more robust and efficient deepfake detection systems. This has wide-ranging applications, including the protection of individuals from reputation damage, the preservation of trust in digital content, and the prevention of cybercrimes.

Additionally, the research sheds light on the advancements and limitations of Vision Transformer Networks, i.e., huge data required, providing valuable insights for further refinement and optimization of these models. This contributes to the ongoing efforts in the field of computer vision and artificial intelligence, aiding in the continuous evolution of technologies to combat emerging threats posed by deepfakes.

5. Conclusions

Deepfaking has emerged as a novel method extensively utilized for propagating disinformation. Although not all deepfake content is inherently malicious, it is crucial to identify and address such creations, as certain instances pose a significant threat to society. The primary objective of this study was to assess the efficacy of ViT in detecting deepfake images. The utilization of global feature mapping and self-attention mechanisms inherent in Vision Transformer has proven to be highly effective. Through careful evaluation across multiple data sets, we have achieved exceptional accuracy rates ranging from 99.5% to 100% when considering three distinct perspectives. Additionally, within the scope of this investigation, we conducted an evaluation of filtered images from Snapchat. Remarkably, we achieved a perfect accuracy rate of 100% in accurately identifying and classifying such images. In future, our research endeavors aim to broaden the scope of our current work by incorporating additional data sets that have been specifically curated and released for deepfake research (i.e., diffusion-based deepfake image detection). This expansion is crucial to enhance the diversity, accuracy, and overall robustness of our methods and findings. Another research direction will be exploring the deepfake problem as a multiclass task.

Author Contributions: Conceptualization, M.A.A., M.I. and A.A., methodology, S.M., M.A.A., R.F.A. and M.I.; validation, M.A.A. and A.A.; investigation, M.A.A., A.M. and A.A.; data curation, M.I., S.M. and M.A.A.; writing—original draft preparation, M.A.A. and M.I., writing—review and editing, M.A.A., A.M. and M.I. All authors have read and agreed to the published version of the manuscript.

Funding: The authors extend their appreciation to the Deputyship for Research & Innovation, Ministry of Education in Saudi Arabia for funding this research work through the project no. (IFKSUOR3–057-3).

Data Availability Statement: The open-source data set used in this study for training purposes is available on Kaggle (https://www.kaggle.com/datasets/xhlulu/140k-real-and-fake-faces, accessed on 1 July 2023) and StyleGAN based website (https://thispersondoesnotexist.com/, accessed on 1 July 2023). Further, we have shared the cloab file for testing purposes based on experiment 1 of this study https://github.com/Muhammad-Asad-Arshed/Fake-Real-Images-Identification, accessed on 10 July 2023.

Conflicts of Interest: The authors declare no conflict of interest.

Abbreviations

TP	True Positive
FP	False Positive
TN	True Negative
FN	False Negative
ViT	Vision Transformer
(DARPA)	Defense Advanced Research Projects Agency
GAN	Generative Adversarial Networks
AI	Artificial Intelligence
ML	Machine Learning
DL	Deep Learning

References

1. Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; Bengio, Y. Generative adversarial nets. *IEEE Signal Process. Mag.* **2018**, *10*, 53–65. Available online: https://proceedings.neurips.cc/paper_files/paper/2014/hash/5ca3e9b122f61f8f06494c97b1afccf3-Abstract.html (accessed on 11 July 2023).
2. Nguyen, T.T.; Nguyen QV, H.; Nguyen, D.T.; Nguyen, D.T.; Huynh-The, T.; Nahavandi, S.; Nguyen, C.M. Deep learning for deepfakes creation and detection: A survey. *Comput. Vis. Image Underst.* **2018**, *223*, 103525. Available online: https://www.sciencedirect.com/science/article/pii/S1077314222001114 (accessed on 11 July 2023). [CrossRef]
3. Media Forensics. Available online: https://www.darpa.mil/program/media-forensics (accessed on 11 July 2023).
4. Deepfake Detection Challenge Results: An Open Initiative to Advance AI. Available online: https://ai.facebook.com/blog/deepfake-detection-challenge-results-an-open-initiative-to-advance-ai/ (accessed on 11 July 2023).
5. Akhtar, Z.; Mouree, M.R.; Dasgupta, D. Utility of Deep Learning Features for Facial Attributes Manipulation Detection. In Proceedings of the 2020 IEEE International Conference on Humanized Computing and Communication with Artificial Intelligence, HCCAI 2020, Irvine, CA, USA, 21–23 September 2020; pp. 55–60. [CrossRef]
6. Iandola, F.N.; Han, S.; Moskewicz, M.W.; Ashraf, K.; Dally, W.J.; Keutzer, K. SqueezeNet: AlexNet-level accuracy with 50x fewer parameters and <0.5 MB model size. *arXiv* **2016**, arXiv:1602.07360. Available online: https://arxiv.org/abs/1602.07360v4 (accessed on 12 July 2023).
7. Simonyan, K.; Zisserman, A. Very Deep Convolutional Networks for Large-Scale Image Recognition. In Proceedings of the 3rd International Conference on Learning Representations, ICLR 2015—Conference Track Proceedings, San Diego, CA, USA, 7–9 May 2015. Available online: https://arxiv.org/abs/1409.1556v6 (accessed on 12 July 2023).
8. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep Residual Learning for Image Recognition. In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778. [CrossRef]

9. Huang, G.; Liu, Z.; Van Der Maaten, L.; Weinberger, K.Q. Densely Connected Convolutional Networks. In Proceedings of the—30th IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2016; pp. 2261–2269. [CrossRef]
10. Szegedy, C.; Liu, W.; Jia, Y.; Sermanet, P.; Reed, S.; Anguelov, D.; Erhan, D.; Vanhoucke, V.; Rabinovich, A.; Liu, W.; et al. Going deeper with convolutions. In Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, USA, 7–12 June 2015; pp. 1–9. [CrossRef]
11. Akhtar, Z.; Dasgupta, D. A comparative evaluation of local feature descriptors for deepfakes detection. In Proceedings of the 2019 IEEE International Symposium on Technologies for Homeland Security (HST), Woburn, MA, USA, 5–6 November 2019. Available online: https://ieeexplore.ieee.org/abstract/document/9033005/ (accessed on 11 July 2023).
12. Bekci, B.; Akhtar, Z.; Ekenel, H.K. Cross-Dataset Face Manipulation Detection. In Proceedings of the 2020 28th Signal Processing and Communications Applications Conference, SIU 2020—Proceedings, Gaziantep, Turkey, 5–7 October 2020. [CrossRef]
13. Li, Y.; Chang, M.C.; Lyu, S. In Ictu Oculi: Exposing AI created fake videos by detecting eye blinking. In Proceedings of the 10th IEEE International Workshop on Information Forensics and Security, WIFS 2018, Delft, The Netherlands, 9–12 December 2019. [CrossRef]
14. Eyebrow Recognition for Identifying Deepfake Videos. IEEE Conference Publication. Available online: https://ieeexplore.ieee.org/document/9211068/authors#authors (accessed on 12 July 2023).
15. Patel, M.; Gupta, A.; Tanwar, S.; Obaidat, M.S. Trans-DF: A Transfer Learning-based end-to-end Deepfake Detector. In Proceedings of the 2020 IEEE 5th International Conference on Computing Communication and Automation, ICCCA 2020, Greater Noida, India, 30–31 October 2020; pp. 796–801. [CrossRef]
16. Ciftci, U.A.; Demir, I.; Yin, L. How do the hearts of deep fakes beat? deep fake source detection via interpreting residuals with biological signals. In Proceedings of the IJCB 2020—IEEE/IAPR International Joint Conference on Biometrics, Houston, TX, USA, 28 September–1 October 2020. [CrossRef]
17. Yang, J.; Xiao, S.; Li, A.; Lu, W.; Gao, X.; Li, Y. MSTA-Net: Forgery Detection by Generating Manipulation Trace Based on Multi-scale Self-texture Attention. *IEEE Trans. Circuits Syst. Video Technol.* **2021**, *32*, 4854–4866. Available online: https://ieeexplore.ieee.org/abstract/document/9643421/ (accessed on 12 July 2023). [CrossRef]
18. Zhao, H.; Zhou, W.; Chen, D.; Wei, T.; Zhang, W.; Yu, N. Multi-Attentional Deepfake Detection. *openaccess.thecvf.com.* Available online: https://openaccess.thecvf.com/content/CVPR2021/html/Zhao_Multi-Attentional_Deepfake_Detection_CVPR_2021_paper.html?ref=https://githubhelp.com (accessed on 12 July 2023).
19. Wang, J.; Wu, Z.; Ouyang, W.; Han, X.; Chen, J.; Jiang, Y.G.; Li, S.N. M2TR: Multi-modal Multi-scale Transformers for Deepfake Detection. In Proceedings of the 2022 International Conference on Multimedia Retrieval (ICMR 2022), Newark, NJ, USA, 27–30 June 2022; pp. 615–623. [CrossRef]
20. Shelke, N.A.; Kasana, S.S. Multiple forgery detection and localization technique for digital video using PCT and NBAP. *Multimed. Tools Appl.* **2022**, *81*, 22731–22759. [CrossRef]
21. Shelke, N.A.; Kasana, S.S. Multiple forgery detection in digital video with VGG-16-based deep neural network and KPCA. *Multimed. Tools Appl.* **2023**. [CrossRef]
22. Wang, B.; Wu, X.; Tang, Y.; Ma, Y.; Shan, Z.; Wei, F. Frequency Domain Filtered Residual Network for Deepfake Detection. *Mathematics* **2023**, *11*, 816. [CrossRef]
23. Zhang, D.; Zheng, Z.; Li, M.; Liu, R. CSART: Channel and spatial attention-guided residual learning for real-time object tracking. *Neurocomputing* **2020**, *436*, 260–272. [CrossRef]
24. 140 k Real and Fake Faces | Kaggle. Available online: https://www.kaggle.com/datasets/xhlulu/140k-real-and-fake-faces (accessed on 12 July 2023).
25. Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner THoulsby, N. An Image is Worth 16 × 16 Words: Transformers for Image Recognition at Scale. *arXiv* **2020**, arXiv:2010.11929. Available online: https://arxiv.org/abs/2010.11929v2 (accessed on 12 May 2023).
26. Carion, N.; Massa, F.; Synnaeve, G.; Usunier, N.; Kirillov, A.; Zagoruyko, S. End-to-End Object Detection with Transformers. In Proceedings of the European Conference on Computer Vision, Glasgow, UK, 23–28 August 2020; Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics). Volume 12346 LNCS, pp. 213–229.
27. Zhang, D.; Zheng, Z.; Wang, T.; He, Y. HROM: Learning High-Resolution Representation and Object-Aware Masks for Visual Object Tracking. *Sensors* **2020**, *20*, 4807. [CrossRef] [PubMed]
28. Devlin, J.; Chang, M.-W.; Lee, K.; Google, K.T.; Language, A.I. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In Proceedings of the 2019 Conference of the North, Minneapolis, MN, USA, 2–7 June 2019; pp. 4171–4186. [CrossRef]
29. Cordonnier, J.B.; Loukas, A.; Jaggi, M. On the relationship between self-attention and convolutional layers. In Proceedings of the 8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, 26–30 April 2020.
30. Karras, T.; Laine, S.; Aila, T. A Style-Based Generator Architecture for Generative Adversarial Networks. *IEEE Trans. Pattern Anal. Mach. Intell.* **2020**, *43*, 4217–4228. [CrossRef]
31. thispersondoesnotexist.com (1024 × 1024). Available online: https://thispersondoesnotexist.com/ (accessed on 12 July 2023).
32. Share the Moment | Snapchat. Available online: https://www.snapchat.com/ (accessed on 12 July 2023).

33. Gandhi, A.; Jain, S. Adversarial perturbations fool deepfake detectors. In Proceedings of the 2020 International Joint Conference on Neural Networks (IJCNN), Glasgow, UK, 19–24 July 2020. Available online: https://ieeexplore.ieee.org/abstract/document/9207034/ (accessed on 13 July 2023).
34. Hu, S.; Li, Y.; Lyu, S. Exposing GaN-generated faces using inconsistent corneal specular highlights. In Proceedings of the ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing—Proceedings, Addis Ababa, Ethiopia, 26–30 April 2021; pp. 2500–2504. [CrossRef]
35. Yousaf, B.; Usama, M.; Sultani, W.; Mahmood, A.; Qadir, J. Fake visual content detection using two-stream convolutional neural networks. *Neural Comput. Appl.* **2022**, *34*, 7991–8004. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.