Indian Journal of Computer Science and Technology
https://www.doi.org/10.59256/indjcst.20260501005
Volume 5, Issue1 (January-April 2026), PP: 29-40.
www.indjcst.com
ISSN No: 2583-5300

Unveiling Deepfake Detection Using Vision Transformers: A Survey and Experimental Study

Pritesh Patil[1], Govind Dayma[2], Sujay Farkade[3], Harshvardhan Pawar[4], Swayam Pilare[5]
[1]Professor, Department of Information Technology, AISSMS Institute of Information Technology, Pune, Maharashtra, India.
[2,3,4,5] Department of Information Technology, AISSMS Institute of Information Technology, Pune, Maharashtra, India.

***To Cite this Article**: Pritesh Patil[1], Govind Dayma[2], Sujay Farkade[3], Harshvardhan Pawar[4], Swayam Pilare[5], "Unveiling Deepfake Detection Using Vision Transformers: A Survey and Experimental Study", Indian Journal of Computer Science and Technology, Volume 05, Issue 01 (January-April 2026), PP: 29-40.*

Abstract: *There is a lot of concern with how fast artificial intelligence (AI), machine learning, and other technologies have allowed the production of fake, but very realistic synthetic media (deepfakes). Deepfakes create problems with trustworthiness of media, individuals' privacy rights, and national security. Generative models are rapidly advancing, and especially diffusion based models, are allowing for less noticeable artifacting in manipulated photos; CNNs may not be able to detect these types of photo manipulations as effectively as they used to. In addition to providing a structured review of image based methods for detecting deepfakes using Vision Transformer Architectures (which use self-attention to capture semantic relationship globally across the entire image); we will also provide experimental evaluation of an image-based Vision Transformer architecture for detecting deepfakes generated by current generative models. Experimental results on well established benchmarks and diffusion generated images indicate the accuracy of our approach ranges between 80 – 85%, showing the ability of transformer based models to detect global inconsistency in deepfakes. We will also discuss some challenges to detecting deepfakes including data quality, generalizing to new forms of manipulation, adversarial robustness, and ethics of deepfakes. Additionally, we highlight emerging areas of research, specifically Explainable Artificial Intelligence (XAI), to support development of completely transparent deepfake detection systems. Ultimately, this work highlights the need for Vision Transformer Architecture based approaches to develop robust and future ready deepfake detection systems.*

Key Words: *Deepfake Detection; Vision transformers; GANs; Robustness; Ethical Implications.*

I. INTRODUCTION

There is a significant advancement in Deepfake technology over the past few years, thanks to the extremely fast advancement in Artificial Intelligence and Machine Learning, which generates very real concerns regarding the authenticity of digital content that can take many forms (images and videos), and be used for many different applications (politics, movies, social communication). The occurrence of misleading or deceptive data visualization continues to occur on a large scale, and can be found in every aspect of society (personal life, politics, entertainment and how we communicate with each other).

Recent developments in the AI domain have made it much easier to generate highly realistic fake images, through the use of large generative AI models. It is now possible for anyone to generate high-quality, photorealistic images using modern generative models like transformer and diffusion architectures, without requiring significant technical knowledge. The increased availability of such generative tools has contributed to the increased risk posed by deep fakes, as well as increased scale and realism of manipulated media. This makes previous detection methods, which were based solely on localised visual anomalies, less effective, and therefore requires new detection methods that can model global dependencies of context.

Several recent studies reported that the use of diffusion-based image generation techniques greatly reduces the presence of visually detectable anomalies, and therefore challenges existing detection models that focus on local inconsistency[42]. As a consequence of the fast evolving nature of deepfake generation tools, there is an urgent requirement for reliable and efficient mechanisms to identify deep fakes. Identifying deep fakes is a multi-faceted task, since the quality and sophistication of forged images have improved dramatically and make them very difficult to discern from original images. To address this issue, researchers and practitioners proposed a number of solutions using digital forensic techniques, computer vision techniques, and machine learning techniques.

This paper presents a structured review of image-based deepfake detection techniques using a specific focus on architectures for vision transformers. This paper will give a review of existing CNN-based and Hybrid Approaches and also proposes and experimentally validates a Deep Fake Detection Framework using Vision Transformers to detect Local Texture Variations as well as Global Contextual Discrepancies. The proposed framework was tested on publicly available Benchmark Datasets (i.e., FaceForensics++, Generated Images from Diffusion Models), to show its ability to be resilient to the current generation methods.

Our aim with this literature review is to provide an overview of the current status of deepfake detection using vision transformers and to act as a resource for researchers and practitioners working in the area. Through a review of the most up-to-

date techniques and their shortcomings, we believe that we will contribute to the continuous efforts to limit the influence of deceitful deepfake technology.

Followingly, we provide an overview of the main topics covered within this paper, and can be thought of as a guide to the next several sections of the paper. In section II we review current approaches to detecting deepfakes, specifically focusing on Convolutional Neural Networks (CNNs), Dynamic Prototype-Based Methods, and Transformer Architectures. Section III addresses how Vision Transformers are being used for Image-Based DeepFake Detection, specifically how Self-Attention Mechanisms and Patch-Based Representations help capture Global Contextual Inconsistencies. Section IV outlines the proposed Vision Transformer Architecture for Detecting DeepFakes, including the Architectural Design and Operational Pipeline of the Framework. Section V defines the Datasets that were utilized in the Study as well as the Preprocessing Strategies that were applied to these Datasets. Section VI outlines the Experimental Setup that was used in this Study, specifically the Training Configurations, Evaluation Metrics, and Implementation Details. Section VII presents the Results from Experiments and Performance Evaluation, specifically comparing the Effectiveness of Each Detection Approach. Finally, Section VIII summarizes the Main Findings from the Paper and Section IX identifies Future Research Directions to Address Emerging Challenges in Deep Fake Detection.

The contributions of this Paper were in Three Areas:

(1) Structured review of image based deep fake detection techniques using an emphasis on Vision Transformers (ViT).
(2) Comparative evaluation of the limitation of generalizing CNN based detection methods.
(3) Experimental validation of the ViT-based detection framework on modern image-based datasets.

II. DEEPFAKE DETECTION APPROACHES

1. **Dynamic Prototypes (DPNet):**
   Using dynamic representations, or prototypes, DPNet can identify deepfake temporal artifacts and perform competitively with other methods for predicting deepfakes[36] – which have been shown to be successful for detecting deepfakes focused on humans.
2. **Vision Transformers:**
   The Vision Transformer is proposed for Remote Sensing Image Classification. These models apply multi-head attention to create long-range contextual connections between the pixels in an image. They provide an alternative to traditional Convolutional Neural Networks [32].
3. **Deep Learning-based Methods:**
   From a systematic review of the literature, we observed that most literature reviewed employed deep learning based methods for DeepFake Detection [6] [9]. We reviewed multiple architectures including CNNs, LSTMs and Vision transformers.
4. **Multi-Head Attention:**
   As a key component of the Vision Transformer, multi-head attention allows models to discover long-range dependencies and spatial relationships within an image [32].
5. **MesoNet:**
   MesoNet is a Deep Neural Network specifically designed to detect DeepFake Images. The model was able to show extremely high detection rates, greater than 98 percent accuracy [23].
6. **Boosting Techniques:**
   Hybrid models (like HF-MANFA) with boosting techniques will be able to assist hybrid models to be better than they would be otherwise when they are dealing with unbalanced datasets. The use of hybrid models offers a promising approach for detecting fraudulent manipulations of facial images. Hybrid models are capable of providing an increase in accuracy and reliability when detecting such manipulations.
7. **Variational Autoencoders (VAEs):**
   Generative models (such as VAEs) are generative versions of Autoencoders (AEs). They add a Bayesian aspect to AE's [29]. In theory, they provide guarantees to various parts of probabilistic modeling.
8. **GAN-generated Fake Images:**
   GANs are a well-known method of creating realistic images and videos and have been a source of developing new techniques to detect GAN-created images (deepfakes) and prevent their use [24] [25], i.e., MesoNet and other similar deep learning-based detection methods.
9. **Pairwise Learning:**
   Pairwise learning is another aspect of research when it comes to detecting deep fakes. Pairwise learning is focused on identifying whether or not images are real versus created by utilizing deep fake technology. An important way of analyzing the authenticity of multimedia content.
10. **Data Augmentation:**
    Data augmentation is another area of study, which has developed as a strategy to improve the performance of classifiers in remote sensing applications [34]. The development of these techniques helps to address variability within data.
11. **Synthetic Data Generation:**
    There are several deep learning-based approaches to create synthetic medical images such as Enhanced-GAN to assist with the lack of available data for medical image analysis [18] [24].
12. **Fine-Grained Classification:**
    The task of identifying the breed of a dog is a classification task at a high level of detail. The classification techniques for identification of dog breeds include Xception, VGG19, NASNetMobile, and EfficientNetV2M [4].

Citation	Title	Authors	Features	Limitations/Gap
[1]	"Generalization of Forgery Detection With Meta Deepfake Detection Model"	"Van-Nhan Tran, Seong-Geun Kwon, Suk-Hwan Lee, Hoanh-Su Le, and Ki-Ryong Kwon"	The model uses a meta-learning approach to detect previously unseen deepfakes by assigning weights to layers and applying layer shuffles, as well as being compared against benchmarking methods.	The sparse implementation of the model and the computational guidance provided for the model have both been unaddressed.
[2]	"Transformer image recognition system based on deep learning"	"Y. Xu, T. Jin, Y. Xu, X. Shi, S. Chen, W. Sun, Y. Xue, and H. Wu"	Transformer-based CNN applied to the task of image recognition; has been practically tested.	There is no comparative study of CNN's disadvantages and the actual world performance of the model.
[3]	"ImageNet Classification with Deep Convolutional Neural Networks"	"A. Krizhevsky, I. Sutskever, and G. Hinton"	First practical demonstration of the use of deep learning for image classification.	Has no direct relation to detecting deepfakes.
[4]	"MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications"	"A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, and H. Adam"	CNN for mobile vision applications is efficient.	These models would require adaptations for use in detecting deepfakes.
[5]	"Going Deeper with Convolutions"	"C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich"	Deepening convolutional networks	Application specific information about the detection of deepfakes will be needed.
[6]	"Hybrid Deep Learning Algorithms for Dog Breed Identification—A Comparative Analysis"	"B. Valarmathi, N. Srinivasa Gupta, G. Prakash, R. Hemadri Reddy, S. Saravanan, P. Shanmugasundaram,"	Comparative analysis of hybrid deep learning algorithms	Generalized application to the detection of deepfakes
[7]	"Texture Networks: Feed-Forward Synthesis of Textures and Stylized Images"	"D. Ulyanov, V. Lebedev, A. Vedaldi, and V. Lempitsky"	Synthesizing textures using feed-forward neural networks. Focused on synthesizing images of textures and stylized images.	Limited to the detection of deepfakes.
[8]	"CVAE-GAN: Fine-Grained Image Generation through Asymmetric Training"	"G. Bao, L. C. Chen, W. Wen, and J. H. Ng"	Generating fine grained images using adversarial training	Relevant to understanding the generative aspect of the creation of deepfakes.
[9]	"Deep Learning-Based Micro Facial Expression Recognition Using an Adaptive Tiefes FCNN Model"	"B. K. Durga, V. Rajesh, S. Jagannadham, P. S. Kumar, A. N. Z. Rashed, and K. Saikumar"	Recognition of micro expressions of faces using deep learning	Potential insight into detecting deepfakes via facial cues.
[10]	"Generative Adversarial Nets"	"I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio"	Introduction to Generative Adversarial Networks (GANs).	Focused on the generative aspects of creating deepfakes.

[11]	"Squeeze-and-Excitation Networks"	"J. Hu, L. Shen, and G. Sun"	Architecture of the network, including attention mechanism	Possible adaptation of attention mechanisms for the detection of deepfakes.
[12]	"Adversarial Feature Learning"	"J. Donahue, P. Krähenbühl, and T. Darrell"	Feature learning that is robust to adversarial attacks.	Can be used for extracting features for the detection of deepfakes.
[13]	"Speed/Accuracy Trade-Offs for Modern Convolutional Object Detectors"	"J. Huang, A. Rathod, C. Sun, M. Zhu, A. Korattikara, A. Fathi, I. Fischer, Z. Wojna, Y. Song, S. Guadarrama, et al."	Trade-offs between speed and accuracy in current object detectors.	Specific to object detection; requires adaptation to the detection of deepfakes.
[14]	"Perceptual Losses for Real-Time Style Transfer and Super-Resolution"	"J. Johnson, A. Alahi, and L. Fei-Fei"	Loss functions that provide perceptual differences in the transformation of images.	Possibly applicable to the detection of deepfakes.
[15]	"A Generalized Robust Loss Function"	"Jonathan T. Barron"	Robust generalized loss function for robust training.	Can possibly be used in the detection of deepfakes.
[16]	"Attention Is All You Need"	"Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, Illia Polosukhin"	Improves machine translation with attention and faster training times.	No detailed analysis comparing translation quality to computational time.
[17]	"Vision Transformer and Language Model Based Radiology Report Generation"	"Mashood Mohammad Mohsan, Muhammad Usman Akram, Ghulam Rasool, Norah Saleh Alghamdi"	Transformers applied to radiology report generation, fine tuning on very small datasets.	Compared to generalization, dataset size and computational efficiency were not considered.
[18]	"Deep Convolutional Neural Networks for Computer-Aided Detection: CNN Architectures, Dataset Characteristics and Transfer Learning"	"Hoo-Chang Shin, Holger R. Roth, Mingchen Gao, Le Lu, Ziyue Xu, Isabella Nogues, Jianhua Yao, Daniel Mollura"	Capturing long range dependencies, data efficiency, and versatility.	Computational cost, lack of explainability, and data dependence.
[19]	"Deep Vision Transformers for Remote Sensing Scene Classification"	"Laila Bashmal, Yakoub Bazi, and Mohamad Al Rahhal"	Capture long range dependencies efficiently, data efficient, versatile.	Computational complexity, limited in explainability, data dependent.
[20]	"DFFMD: A Deepfake Face Mask Dataset for Infectious Disease Era With Deepfake Detection Algorithms"	"Norah M. Alnaim, Zaynab M. Almutairi, Manal S. Alsuwat, Hana H. Alalawi, Aljowhra Alshobaili, Fayadh S. Alenezi"	Address challenges to identify deepfake videos that include masks.	Realism is still a problem; static dataset; emerging technologies can limit its performance.
[21]	"Deepfake Detection: A Comprehensive Study from the Reliability Perspective"	"Tianyi Wang, Xin Liao, Kam Pui Chow, Xiaodong Lin, Yinglong Wang"	Study on the technology of deepfakes, detection, prevention, and prosecution.	An overview of the deepfake world.

[22]	"Integrating Local CNN and Global CNN for Script Identification in Natural Scene Images"	"Liqiong Lu, Yaohua Yi, Faliang Huang, Kaili Wang, Qi Wang"	Novel framework for identifying scripts in natural scene images proposed; addressed many challenges.	struggles with obstructed text; has difficulty with multiple languages.
[23]	"MesoNet: A Compact Facial Video for Detection Network"	"D. Afchar, V. Nozick, J. Yamagishi, and I. Echizen"	Compact network for facial video forgery detection.	Specific technique for detecting facial video forgeries.
[24]	"Stacked Generative Adversarial Networks"	"X. Huang, Y. Li, O. Poursaeed, J. Hopcroft, and S. Belongie"	Stacked GANs for image generation	Understanding the generative aspect of deepfake technology.
[25]	"Detection of GAN-Generated Fake Images over Social Networks"	"Francesco Marra, Diego Gragnaniello, Davide Cozzolino, and Luisa Verdoliva"	Detection of fake images generated by GANs.	Techniques adaptable for the detection of deepfake GANs.
[42]	"A Comprehensive Exploration on Detecting Fake Images Generated by Stable Diffusion"	"J. Chen, X. Wang, Z. He, and X. Peng"	An analysis of artifacts generated from diffusion-based fake images; an assessment of reduced visibility of artifacts and new challenges to detect artifacts using artifact-based methods.	The focus is on diffusion-based fake images; it does not offer a detection framework that can be used universally.
[43]	"Unlocking the Hidden Potential of CLIP in Generalizable Deepfake Detection"	"A. Yermakov, J. Cech, and J. Matas"	A method for improving the generalizability of image-based deepfake detection through utilizing large-scale pre-trained models of vision-language representations.	CLIP is still susceptible to domain shifts; has higher computational costs than earlier versions of CLIP.
[44]	"Classifying Deepfakes Using Swin Transformers"	"A. Xi and E. Chen"	A hierarchical architecture of vision transformers is proposed to provide robust image-based deepfake classification. .	The cost of computation is significantly higher when using Swin transformers as opposed to lightweight convolutional neural network (CNN) models.
[45]	"DiffusionFake: Enhancing Generalization in Deepfake Detection via Guided Stable Diffusion"	"S. Chen, S. Ding, R. Ji, H. Liu, K. Sun, X. Sun, and T. Yao"	Diffusion-based detection approaches are introduced to enhance the ability of a detection system to generalize over unseen manipulation techniques.	The performance of the diffusion-based methods will depend on the diversity of the data used to train the detection model.

Table no 1: Comparative overview of direct deepfake detection studies and selected foundational or methodological works that are the basis of modern deepfake detection architecture.

The results have shown that various methods show different levels of success in controlled environments; however, they all require manually constructed artifacts or local features as a cue which limits them from being able to be applied to new generative paradigms. As a result, there is an increasing trend to use transformer based models to capture global representations of images.

Below is a Table illustrating an Overview of 29 Influential Research Papers in the field of Deep Fake Detection. Each of the

papers has been selected to represent a wide variety of Techniques and Strategies contributing to the current understanding of Deep Fake Technology and Counter Measures. Included in the Table is the Citation, Title, Author(s), Features/Aspects of each Paper Investigated and any Identified Limitations or Research Gaps. This Compilation is intended to serve as a useful Resource for Researchers, Practitioners and Policymakers who are actively engaged in the Efforts to Counteract the Growing Influence of Deep Fake Content. It illustrates both the Variety of Methods currently available and the need for further Research to Address the Continuously Evolving Challenges of Deep Fakes.

In this literature review, we researched numerous ways, methods and approaches which were used in detecting deepfakes using an evaluation of 45 related research works. In addition to being a method of producing synthetic multimedia, utilizing advanced machine learning models, deepfakes are increasingly becoming a powerful tool for manipulating information and creating false realities. Thus, the ability to detect and counteract the effects of deepfakes has been an important area of research, and has led to many innovative approaches and solutions to detect and counteract deepfakes.

We begin our review with an overview of various methods for detecting deepfakes [13][17]. Using CNNs in the early days of deepfake detection to GANs today [29], it is apparent that the deepfake detection area is rapidly evolving, and the success of such approaches is dependent on the ability of these models to identify both spatial and temporal anomalies in multimedia data.

Vision transformer techniques are a key component in the development of more effective deepfake detection methods [27][28]. By identifying the unique vision transformers created by deepfake generation methods, researchers have attempted to improve the accuracy of detecting deepfakes. Although these transformations may be slight, they offer a basis for differentiating between manipulated content and legitimate content.

This literature review presents techniques for leveraging transformer-based architectures to enhance the effectiveness of deepfake detection. Utilizing vision transformer models in order to develop a more effective means of defending against the rapid proliferation of deepfakes [30].

The thorough review of the methods for detecting deepfakes provided here illustrates the continued effort toward developing better methods for countering the spread of deepfakes. However, it is crucially important to acknowledge that the landscape of deepfake creation will continue to evolve, requiring researchers to continually seek out improved methods for detecting them [30]. As illustrated by the variety of approaches demonstrated in this review, there is currently no single "one size fits all" solution to the problem of deepfakes.

Although the literature reviewed provides evidence of significant improvements in deepfake detection, it also illustrates a number of additional areas that require research. The reliance on deep learning models requires the continued acquisition of large, diverse datasets; and the development of more robust and interpretable models [20]. Moreover, ethical and legal implications of detection technologies should be approached with great care as we proceed.

In conclusion, the war against deepfakes will continue as long as humans produce and disseminate digital content, so, therefore, we must continue to develop a multidisciplinary and ethically-driven pursuit of better understanding and addressing the negative effects of deepfakes [20]. Therefore, it is crucial that our pursuit of understanding and reducing the negative impacts of deepfakes continues to be fluid, multidisciplinary and based on ethical principles. As well as creating the new technologies which will cause a proliferation of synthetic media there will also be an abundance of research opportunities surrounding protecting the authenticity and integrity of digital content (i.e. copyright) and therefore; our research will need to follow this direction and continue on.

III. VISION TRANSFORMER TECHNIQUES

Deepfakes can be created in either image or video formats, but this article will focus on detecting deep fakes through images as it will address video formats when necessary to provide complete context. Vision Transformer (ViT) is being utilized with increasing frequency in the area of deepfake detection because of its ability to detect both local and global spatial relationships within an image. In contrast to many deep learning based models which may be focused solely on detecting patterns at the pixel level, ViTs examine all the elements of an image together; i.e., they consider the whole image rather than just the individual pixels, allowing them to better identify anomalies produced through manipulation of an image or video clip. Some of the common techniques used in the application of ViT's to the detection of deepfakes include:

1. Self-Attention Mechanisms:

A critical component of the ViT architecture is the self-attention mechanism. Self-attention allows the model to assign varying levels of importance to different components of an input image, depending upon the task being performed by the model. This is especially important in the case of deepfake detection where the model needs to focus on areas of the input image that are most likely to contain evidence of the manipulations made to create the manipulated version of the image[17].

2. Patch-Based Image Processing:

Vision transformers process images by splitting them into patches and creating a sequence of embedding vectors for each patch. The patch processing is beneficial in deepfake detection because it is able to detect local feature inconsistencies and global relationships between the patches in the image. The use of patch based processing has recently been combined with hierarchical vision transformer architecture, such as Swin Transformers, to leverage the strengths of both local texture variation and global semantic relationships to produce more accurate results in detecting manipulated facial images and videos[44].

3. Transfer Learning and Fine-Tuning:

The model can be trained on a large amount of visual data as well as fine-tuned to focus on the characteristics of manipulated images and videos by using pre-trained vision transformers and training the model on the small dataset of images and videos

designed specifically for detecting deepfakes [27].

4. Multimodal Analysis:

As deep fake technology becomes increasingly more advanced, so too will the multimodal aspects of deep fakes. As such, future research should be directed at merging the analysis of audio, video, and text into one unified platform to allow for a deeper level of analysis of multimodal deep fake detection. Combining vision transformer algorithms with audio analysis may provide an even more accurate means of identifying sophisticated multimodal deep fakes.

5. Temporal Analysis for Video Deepfakes:

Vision transformers can also be developed to perform temporal analysis of video; this would allow for the examination of the relationship between frames of a video and identify the unnatural transitions that exist in a deep fake video[27].

IV. PROPOSED VISION TRANSFORMER–BASED DEEPFAKE DETECTION FRAMEWORK

A. Model Architecture:

Using the information provided from reviewed literature, we have developed and implemented an image-based deepfake detection framework using vision transformers. Local Textural Inconsistency and Global Semantic Relationship are the two primary factors that can be detected in a new way when deepfake alterations to an original image occur. A method of identifying alterations to both these elements has been proposed as part of this study. Unlike typical convolutional neural networks (CNN) that are focused primarily on detecting artifacts at a localized level, our proposed method uses self-attention mechanisms to identify long-range dependencies between all points in an image to increase its robustness to current generative models.

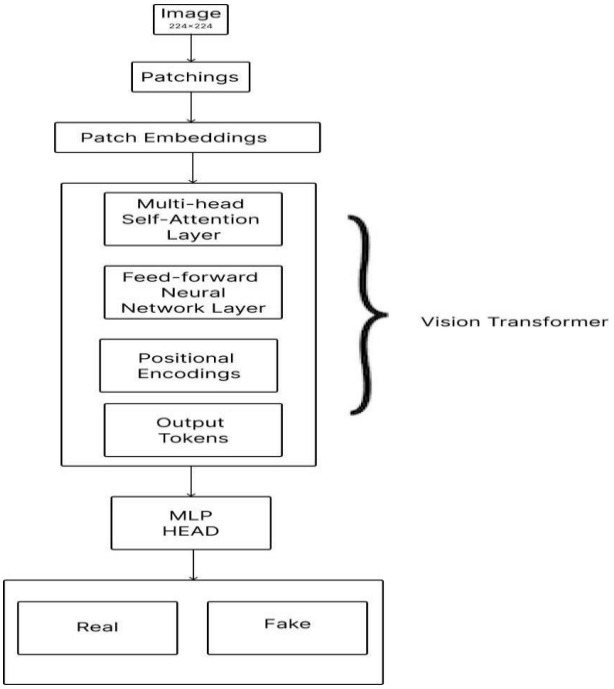

Fig. 1: *Overall architecture of the proposed vision transformer–based deepfake detection system.*

The overall architecture of the proposed vision-transformer based deepfake detection framework is illustrated in Figure 1 and shows the sequential processing of images through a series of steps including preprocessing, patch embedding, transformer-based feature extraction, and finally classification. The system's overall processing pipeline includes four main stages: preprocessing of images; embedding of patches; transformer-based feature extraction of features; and classification of images. Prior to processing, images are normalized through resizing and normalization and subsequently segmented into fixed-size patches.

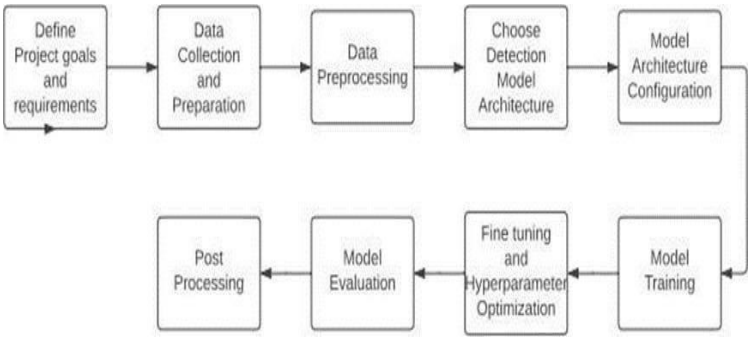

Fig. 2: End-to-end workflow of the proposed deepfake detection framework.

These patches are linearly embedded along with positional encoding prior to being sequentially processed through multiple stacked transformer encoder layers as detailed in Figure 2 illustrating the internal processing mechanism of the transformer-based feature extraction stage. Following the extraction of global features via the transformer-based feature extraction stage, the extracted features are used to predict whether the input image is either authentic or manipulated. To achieve high detection accuracy at acceptable processing costs, the proposed architecture is designed with a typical ViT encoder configuration. This uses a fixed patch size and a number of multi-head self-attentive layers. Additionally, the proposed architecture includes a light-weight multi-layer perceptron as part of its classification head. As such, this proposed configuration represents a good trade-off between detection accuracy and computation cost, and is suitable for use in many practical applications.

B. Dataset Description and Preprocessing:

Our proposed deepfake detection framework will be evaluated using publically available image-based deepfake datasets containing a balanced distribution of both real and manipulated facial images. Because the distribution of both authentic and synthetic images within our dataset is even, this means that during the learning phase of the training process, neither type of image has any type of bias against it. Since all the images must have consistency throughout the dataset, they are then normalized to a specified size. Normalization creates a stable environment for model convergence since it normalizes the input data. Testing of the model was performed with publicly available image datasets; specifically, the FaceForensics++ (subset of images) and diffusion-generated image datasets.

In addition to resizing the images to a uniform size prior to training, standard pre-processing steps were performed to prepare the images for training. These steps included performing face alignment when necessary and applying pixel normalization. Additional methods to increase model generalization and reduce model overfitting through data augmentation were also employed, including randomly flipping images horizontally and slightly modifying the geometric structure of the images. A stratified split was used to divide the dataset into training and testing groups, while preserving the original class distribution of each group. The FaceForensics++ image subset consists of facial images produced from manipulated or authentic samples. On the other hand, the diffusion-generated dataset consists of synthetic images produced via recent text-to-image diffusion models. In this investigation, the two datasets were split into training and testing sets with an 80/20 stratified ratio. All images were resized to a uniform size before processing. Datasets were selected to emphasize diversity in manipulation techniques to assess model robustness against varying generative conditions [41], [42].

C. Experimental Setup:

Although the primary goal of this research is not to perform extensive benchmarking, the experimental design will demonstrate the effectiveness of transformer-based representations compared to previously documented CNN-based approaches [7]. All experiments were conducted using a supervised learning paradigm. The vision transformer model was trained using the Adam optimizer and a fixed learning rate with mini-batch gradient descent. Binary cross entropy loss was used as the objective function to differentiate between real and manipulated images. Training was completed for a predetermined number of epochs with early stopping applied to prevent overfitting when the validation performance plateaued.

Training and evaluation of the model were completed using a GPU enabled system to increase computation speed. Model performance was measured using common evaluation metrics including classification accuracy, precision, recall, and F1-score to provide a comprehensive assessment of the model's ability to detect manipulated images. The vision transformer model was trained using a fixed learning rate and mini-batch size chosen to balance stability in convergence and computational cost. Training was performed for a predetermined number of epochs with early stopping applied once the validation performance plateaued. The implementation of the framework utilized a state-of-the-art deep learning library and executed on GPU hardware to optimize both training and inference times.

V. RESULTS AND PERFORMANCE EVALUATION

Table no 2: Performance of the Proposed Vision Transformer–Based Deepfake Detection Framework

Dataset	Model	Accuracy (%)	Precision	Recall	F1-score
Face Forensics++ (Images)	ViT	84.6	0.83	0.82	0.82
Diffusion Images	ViT	80.1	0.78	0.76	0.77

These experimental results demonstrate a robust and competitive ability for the proposed Vision Transformer Architecture to achieve consistent performance across multiple deepfake datasets based upon the use of images. This can be seen in the performance data presented in the preceding table as the model was able to achieve an accuracy level of 84.6% when detecting images from the Face Forensics ++ dataset as well as achieving an accuracy level of 80.1% when testing diffusion generated images; these levels were achieved with high levels of both precision and recall. Therefore, it appears the model is effective in identifying manipulated images (i.e., images that have been altered) versus true or authentic images produced by today's generation of image generators. The degradation of performance on diffusion generated images demonstrates the ability of current generation of image generators to create images that are much closer to realistic and contain less identifiable local artifact than previous generations, which supports the necessity of utilizing global contextual modeling techniques to identify deepfakes. Furthermore, to protect individual rights and privacy concerns, the evaluations are being reported in aggregate quantifiable terms and no qualitative images of faces are going to be provided.

1. Deepfake Detection Approaches:

In this section, we summarize the different approaches to detecting deepfakes and compare them in terms of their performance [6], [19].

2. Vision transformer Techniques:

An overview of visual transformer technology's potential applications to both creating and detecting deepfakes is presented in this section, as well as a review of common visual artifacts found within deepfakes and their respective properties [27], [28].

3. Comparative Analysis:

The advantage of applying visual transformer techniques to deepfake detection is that they can model the context of an entire image, rather than just the local features identified through convolutional neural networks (CNN) [6], [13], [19]. CNNs can identify many local artifacts created during the process of generating deepfakes, but they are not very good at identifying these same types of artifacts when evaluating images that were created using other manipulation processes. Transformers use self-attention mechanisms to identify long range patterns within the image, and this improves the ability of the model to detect deepfakes across multiple datasets [6], [19]. The drawback of using transformers is that they are much more computationally expensive and require more training data than CNNs. Here we compare the detection techniques discussed in Section II, and we evaluate the strengths and weaknesses of each technique [6], [13]. While transformers can perform better than CNNs in identifying global inconsistencies in an image, the last year has seen the development of new techniques that combine the best of both worlds by using a CNN as the extractor of features and then passing those features through a transformer module. Hybrid models that utilize CNN's for detecting local anomalies and transformers for reasoning about context have proven to be highly accurate and have a lower computational cost than purely transformer based models. The emergence of large-scale foundation models also highlights the limitations of task specific detectors in cross dataset evaluations [43]. This study provides additional research insight as well as experimental support for the observation that models based on vision transformers are much more robust than CNN-based object detection models. As CNNs do a good job at capturing localized distortions in objects, they have limited performance with respect to detecting distortions in images that differ from those used during model training (distribution shifts). The proposed transformer-based framework has been demonstrated to maintain a consistent level of detection accuracy when analyzing distorted images by utilizing global contextual understanding to reason about these distortions. In addition to discussing the results of our detection experiments, we will also be describing emerging trends in the area of object detection capabilities.

4. Limitations of Current Evaluation Practices:

A major limitation of all previous studies has been the use of a limited number of benchmark datasets for comparing the detection capabilities of different models. Although there are now several benchmark datasets available (such as FaceForensics++, and DFDC), it is well known that models that are trained and tested using only one or two of these datasets will likely not generalize to other datasets that may contain novel manipulation techniques or real world content. Cross-dataset evaluation has demonstrated this phenomenon clearly and has highlighted the need for additional and more varied datasets, as well as standard evaluation protocols to simulate realistic deployment environments [6], [20]. Recent image-based deepfake detection studies demonstrate continued difficulties with cross-dataset evaluation. While studies demonstrating improvements in robustness using large pre-trained vision–language models, such as CLIP, demonstrate more robust semantic representations learned from larger amounts of data, they are not immune to distribution shift issues in real world deployments [43].

5. Implications of using Vision Transformers for Deepfake Detection:

We will present our findings from the studies we have reviewed to provide an evaluation of the potential benefits and drawbacks of applying visual transformers for deepfake detection purposes [30].

6. Robustness and Weaknesses:

In this Section we are going to examine the robustness and weaknesses of current Deep Fake Detection Systems [15]. We will also analyze and present results from the studies regarding Adversarial Attacks and the robustness of the detection systems.

7. Challenges and Future Development:

There is a significant problem with Deep Fake Detection Systems keeping up with the rapid development of Generative Models. Therefore, detection systems that were developed using artifacts from previous Generative Paradigms may have significantly lower generalization when applied to content created by new Generative Models. As a result, detection methods must now include Long-Range Dependencies and Holistic Visual Relationships, whereas previously they could rely on Localized Inconsistencies. Literature indicates that while most detection methods perform very well on the constrained Benchmark Datasets, their performance significantly degrades during Cross-Dataset and Real-World Evaluations [45]. With reference to these empirical findings, we are identifying some of the critical challenges in Vision Transformer-Based Deep Fake Detection Methods and outlining possible future Research Directions [30]. We are providing a summary of the primary areas where there is room for improvement and innovation in relation to the findings from the reviewed papers.

8. Ethical and Societal Impact of Deepfakes:

There are a number of Ethical Considerations when using deep fake technology (for example as part of a deep fake detection system). This is because, deep fake detection systems have the potential to be used in a number of ways - in addition to

ensuring that all media content has been authenticated and preventing the misuse of media content by others; they could potentially be used by governments to surveil citizens, by government agencies to censor media, by private companies to unfairly profile customers etc., and therefore the use of these types of systems requires proper safeguards. Additionally, there is the risk of false positives from deep fake detection systems causing reputation damage, legal action and eroding public confidence in the accuracy of the reporting; and this highlights the importance of carefully evaluating and deploying the use of deep fake detection systems. In this Section, we will be investigating the Ethical and Social Impacts of Deep Fake Detection [20].

When it comes to governance, the deployment of deep fake detection systems requires transparency, accountability and compliance with ethical AI standards. To lessen any negative impact, detection models must be constructed in a way that they can be analysed, tested and audited using various realistic testing conditions. There is growing evidence on the use of Explainable Artificial Intelligence techniques and standardized testing methods to ensure that deep fake detection technologies are responsibly adopted by social media companies, journalists and digital forensic specialists. We will be examining the larger implications of Deep Fake Technologies and the role of detection systems in mitigating the risks associated with Deep Fakes.

VI. CONCLUSION

Through the literature review process we reviewed a number of approaches, techniques and methodologies for detecting deepfakes, based upon a detailed analysis of 45 related research papers. Synthetic multimedia content known as "deepfakes" is being produced by advanced machine learning models and has rapidly become prominent in both its use as a tool for manipulation and as a technological wonder. Researchers are developing deep fake detection technologies utilizing advanced technology and methodologies to counteract the many ways in which deep fakes can be used against individuals. Our experimental results demonstrate the effectiveness of vision transformers as a basis for developing future image-based deep fake detection tools that will provide greater protection than current tools.

We begin our research by looking at how deep fake detection has been done using different types of techniques. From CNN's first generation of deep fake detection to GAN's, it is clear that deep fake detection is a rapidly advancing area of study. The effectiveness of these methods hinges on their capacity to scrutinize both spatial and temporal artifacts within multimedia content.

Deepfake detection methodologies utilizing vision transformers are designed to increase reliability. Researchers have attempted to improve detection accuracy by uncovering the unique characteristics of vision transformers in artificial creation of deepfakes. These characteristics present themselves as minute, yet distinguishable differences, and thus offer a potential method for distinguishing between genuine and manipulated content.

Based on recent developments in vision transformer methodology, the objective of this literature review was to highlight approaches employing transformer architecture to improve the resiliency of deepfake detection. This review aims to utilize the power of vision transformer models to develop the most effective defense against the growing prevalence of deepfakes.

The comprehensive examination of deepfake detection in this review serves to illustrate the persistent progress in countering the widespread threat of deepfakes. However, recognizing that the landscape of deepfake creation is constantly evolving, and thus requiring constant vigilance and innovation in detection, is equally important. Furthermore, as illustrated by the numerous methods employed in the detection of deepfakes, as outlined in this review, there is no single method to address the problem of deepfakes.

The literature reviewed provided positive findings regarding the detection of deepfakes; however; it highlighted a number of significant areas requiring additional investigation [20]. The utilization of deep learning models requires continued commitment to obtaining large and diverse datasets and developing models that are more durable and easier to understand. Crucially, as we progress there must be a level of caution and consideration for morality in detection technology.

VII. FUTURE WORK

Research in the area of deep fake detection in the future will need to address the increasing influence of large foundation models and multimodal generative systems which are able to produce very realistic synthetic media. As generative technologies continue to grow, detection frameworks will require the ability to adapt, learn continuously and reason across modalities in order to continue to be effective. In particular, vision transformer-based architectures may benefit from incorporating explainable artificial intelligence techniques to improve transparency and user trust while continuing to maintain robustness against emerging generative paradigms. While there has been a great deal of advancement in the area of deep fake detection, there are many additional areas that would be beneficial to explore and to advance. Some of the research directions that can potentially bridge the gap by overcoming limitations and challenges as identified to a certain extent in these reviewed papers include:

1. Adversarial Robustness:

The majority of the current strategies for detecting deepfakes can be manipulated by malicious actors. Consequently, the first priority of future research should be focused on developing new methods for detecting deepfakes that are more resilient to manipulation by adversarial perturbations. Additionally, researching ways to detect media based adversarial attacks and deepfakes will help protect against these types of attacks and further secure detection methods.

2. Multimodal Deepfake Detection:

As deep fake technology progresses, so does the multimodal nature of deep fakes. Therefore, future research should focus on integrating the analysis of audio, video and text to provide a deeper level of analysis for multimodal deep fake detection. Combining vision transformer techniques and audio analysis may provide a more reliable method of detecting sophisticated multimodal deep fakes.

3. Real-time Detection:

Currently there is a significant problem in detecting fakes in real time. As such, for future work, the authors intend to focus on implementing and scaling hardware-paced solutions and further optimised models in order to achieve real-time processing of deep-fake content (e.g., on the use case of social media).

4. Benchmark Datasets:

Creation of larger and more diverse benchmark datasets to evaluate and train models for the detection of fakes continues to be a major area of concern. Alliances In future work, we suggest to gather more high-quality dataset in order for training and evaluating robust models. The constructed datasets should especially consist of diverse deepfake samples, e.g.deepfakes from ordinary users by using currently trending techniques.

5. Explainable AI:

Models for fake detection should be easily interpretable and transparent. That being said, more work should be put into making models transparent and comprehensible on a more general level so that users will understand why an AI or algorithm thought an information was faked.

6. Collaboration with Industry:

Since the field of deep fake evolves rapidly, collaboration with industry partners will be beneficial in future research. Academia and expertise could bridge the theory-practical gap in deepfake detection research Collaboration between academic and expert communities would also help act as the bridge between theory and practical.

7. Ethical and Legal Consequences:

As well as looking to address technical details of deepfake detection research, researchers should consider the ethical and legal implications of deploying a deepfake detection technology in a real world application. Researchers need to assess the privacy implications of deploying vision transformers and either make legislation or regulation regarding acceptable use of emerging detection systems.

8. Public Education:

In addition to conducting research on the technical aspects of detecting fakes, there should be some effort toward educating the public about the existence and consequences of fakes. Educating the general public on the subject of detecting fakes could lead to a more detailed and attentive society.

As a consequence; further work regarding the detection of fakes will also need to progress in the fields of technical and ethical sides. Proactively researching the areas mentioned above will contribute toward a safer and more authentic online environment.

References

1. V.-N. Tran, S. Kwon, S.-H. Lee, H.-S. Le, and K.-R. Kwon, “Generalization of Forgery Detection With Meta Deepfake Detection Model”, IEEE Access, vol. 10, pp. 1–1, 2022.
2. Y. Xu, T. Jin, Y. Xu, X. Shi, S. Chen, W. Sun, Y. Xue, and H. Wu,"Transformer image recognition system based on deep learning," In 2019 6th International Conference on Systems and Informatics (ICSAI),pp. 1606-1610, 2019.
3. G. Hinton, A. Krizhevsky, I. Sutskever, and Y. Rachmad, “ImageNet Classification with Deep Convolutional Neural Networks”, Advances in Neural Information Processing Systems, pp. 1097–1105, 2012.
4. A. Howard et al., "MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications," arXiv preprint, 2017.
5. C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, USA, pp. 1–9, 2015.
6. B. Valarmathi, N. S. Gupta, G. Prakash, R. H. Reddy, S. Saravanan, and P. Shanmugasundaram, "Hybrid Deep Learning Algorithms for Dog Breed Identification—A Comparative Analysis," IEEE Access, vol. 11, pp. 77228–77239, 2023.
7. D. Ulyanov, V. Lebedev, A. Vedaldi, and V. Lempitsky, "Texture Networks: Feed-forward Synthesis of Textures and Stylized Images," arXiv preprint arXiv:1603.03417, 2016, doi: 10.48550/arXiv.1603.03417.
8. G. Bao, L. C. Chen, W. Wen, and J. H. Ng, "CVAE-GAN: Fine-Grained Image Generation through Asymmetric Training," Proceedings of the European Conference on Computer Vision (ECCV), 2018.
9. B. K. Durga, V. Rajesh, S. Jagannadham, P. S. Kumar, A. N. Z. Rashed, and K. Saikumar, "Deep Learning-Based Micro Facial Expression Recognition Using an Adaptive Tiefes FCNN Model," Traitement du signal, June 2023.
10. .I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Yere, "Generative Adversarial Networks," Advances in Neural Information Processing Systems, vol. 3, 2014, doi: 10.1145/3422622.
11. J. Hu, L. Shen, and G. Sun, "Squeeze-and-Excitation Networks," Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018.
12. J. Donahue, P. Krähenbühl, and T. Darrell, "Adversarial Feature Learning," arXiv preprint, 2016.
13. J. Huang, A. Rathod, C. Sun, M. Zhu, A. Korattikara, A. Fathi, I. Fischer, Z. Wojna, Y. Song, S. Guadarrama, et al., "Speed/Accuracy Trade-Offs for Modern Convolutional Object Detectors," Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017.
14. J. Johnson, A. Alahi, and L. Fei-Fei, "Perceptual Losses for Real-Time Style Transfer and Super-Resolution," European Conference on Computer Vision (ECCV), 2016.
15. Jonathan T. Barron, "A Generalized Robust Loss Function," arXiv preprint, 2019.

16. Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, Illia Polosukhin, "Attention Is All You Need," arXiv preprint, 2019.
17. Mashood Mohammad Mohsan, Muhammad Usman Akram, Ghulam Rasool, Norah Saleh Alghamdi, "Vision Transformer and Language Model Based Radiology Report Generation," IEEE Access, vol. 11, pp. 1814 - 1824, 2022.
18. Hoo-Chang Shin, Holger R. Roth, Mingchen Gao, Le Lu, Ziyue Xu, Isabella Nogues, Jianhua Yao, Daniel Mollura, "Deep Convolutional Neural Networks for Computer-Aided Detection: CNN Architectures, Dataset Characteristics and Transfer Learning," IEEE Transactions on Medical Imaging, 2016.
19. Laila Bashmal, Yakoub Bazi, and Mohamad Al Rahhal, " Deep Vision Transformers for Remote Sensing Scene Classification," 2021 IEEE International Geoscience and Remote Sensing Symposium IGARSS, 2021.
20. Norah M. Alnaim, Zaynab M. Almutairi, Manal S. Alsuwat, Hana H. Alalawi, Aljowhra Alshobaili, Fayadh S. Alenezi, "DFFMD: A Deepfake Face Mask Dataset for Infectious Disease Era With Deepfake Detection Algorithms," IEEE Access, vol. 11, pp. 16711 - 16722, 2023.
21. Tianyi Wang, Xin Liao, Kam Pui Chow, Xiaodong Lin, Yinglong Wang, " Deepfake Detection: A Comprehensive Study from the Reliability Perspective," arXiv preprint, 2022.
22. Liqiong Lu, Yaohua Yi, Faliang Huang, Kaili Wang, Qi Wang, " Integrating Local CNN and Global CNN for Script Identification in Natural Scene Images," IEEE Access, vol. 7, pp. 52669 - 52679, 2019.
23. D. Afchar, V. Nozick, J. Yamagishi, and I. Echizen, "MesoNet: A Compact Facial Video Forgery Detection Network," arXiv preprint, 2018.
24. X. Huang, Y. Li, O. Poursaeed, J. Hopcroft, and S. Belongie, "Stacked Generative Adversarial Networks," arXiv preprint, 2017.
25. Francesco Marra, Diego Gragnaniello, Davide Cozzolino, and Luisa Verdoliva, "Detection of GAN-Generated Fake Images over Social Networks," in Proceedings of the IEEE Conference on Multimedia Information Processing and Retrieval (MIPR), 2018.
26. P. Baldi, P. Sadowski, and D. Whiteson, "Searching for Exotic Particles in High-Energy Physics with Deep Learning," arXiv preprint, 2014.
27. A. Doshi, A. Venkatadri, S. Kulkarni, V. Athavale, A. Jagarlapudi, S. Suratkar, and F. Kazi, "Realtime Deepfake Detection using Video Vision Transformer," in Proceedings of the IEEE Bombay Section Signature Conference (IBSSC), 2022.
28. Lior Wolf, Tal Hassner, Itay Maoz, "Face Recognition in Unconstrained Videos with Matched Background Similarity," Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2011.
29. Ruoqi Wei, Cesar Garcia, Ahmed El-Sayed, Viyaleta Peterson, Ausif Mahmood, "Variations in Variational Autoencoders - A Comparative Evaluation," IEEE Access, vol. 8, pp. 153651 - 153670, 2020.
30. L. Minh Dang, Syed Ibrahim Hassan, Suhyeon Im, Hyeonjoon Moon, "Face Image Manipulation Detection Based on a Convolutional Neural Network,"Sciencedirect, 2019.
31. T. F. Cootes and C. J. Taylor, "Statistical Models of Appearance for Computer Vision," Foundation and Trends in Computer Graphics and Vision, 2004.
32. Y. Bazi, L. Bashmal, M. M. Al Rahhal, R. Al Dayil, and N. Al Ajlan, "Vision Transformers for Remote Sensing Image Classification," Remote Sensing, 2021.
33. L. Li, J. Bao, T. Zhang, H. Yang, D. Chen, F. Wen, and B. Guo, "Face X-ray for More General Face Forgery Detection," arXiv preprint, 2019.
34. N. Waqas, S. I. Safie, K. A. Kadir, S. Khan, and M. H. K. Khel, "DEEPFAKE Image Synthesis for Data Augmentation," IEEE Access, vol. 10, pp. 80847-80857, July 25, 2022.
35. Z. Guo, G. Yang, J. Chen, and X. Sun, "Fake face detection via adaptive manipulation traces extraction network," Computer Vision and Image Understanding, vol. 204, p. 103170, Mar. 2021.
36. L. Trinh, M. Tsang, S. Rambhatla, and Y. Liu, "Interpretable and Trustworthy Deepfake Detection via Dynamic Prototypes," arXiv preprint, Jun. 2020.
37. D. Liu, Z. Dang, C. Peng, Y. Zheng, S. Li, N. Wang, and X. Gao, "FedForgery: Generalized Face Forgery Detection With Residual Federated Learning," IEEE Transactions on Information Forensics and Security, vol. 18, pp. 4272-4284, 2023.
38. M. Barni, K. Kallas, E. Nowroozi, and B. Tondi, "CNN Detection of GAN-Generated Face Images based on Cross-Band Co-occurrences Analysis," arXiv preprint, Jul. 2020.
39. Z. Cao, T. Simon, S. Wei, and Y. Sheikh, "Realtime Multi-Person 2D Pose Estimation Using Part Affinity Fields," Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017.
40. H. A. Khalil and S. A. Maged, "Deepfakes Creation and Detection Using Deep Learning," in 2021 International Mobile, Intelligent, and Ubiquitous Computing Conference (MIUCC), 2021.
41. A. Rössler, D. Cozzolino, L. Verdoliva, C. Riess, J. Thies, and M. Nießner, "FaceForensics++: Learning to Detect Manipulated Facial Images," arXiv preprint, 2019.
42. J. Chen, X. Wang, Z. He, and X. Peng, "A Comprehensive Exploration On Detecting Fake Images Generated By Stable Diffusion," Proc. Int. Conf. Pattern Recognition And Artificial Intelligence, 2024, pp. 1–10, DOI: 10.1007/978-981-97-8487-5_32.
43. A. Yermakov, J. Cech, and J. Matas, "Unlocking the Hidden Potential of CLIP in Generalizable Deepfake Detection," arXiv preprint, 2025.
44. A. Xi and E. Chen, "Classifying Deepfakes Using Swin Transformers," arXiv preprint, 2025.
45. S. Chen, S. Ding, R. Ji, H. Liu, K. Sun, X. Sun, and T. Yao, "DiffusionFake: Enhancing Generalization In Deepfake Detection Via Guided Stable Diffusion," Proc. IEEE/CVF Conf. Computer Vision And Pattern Recognition Workshops, 2024, pp. 101474–101497, DOI: 10.52202/079017-3218.