Expert Systems With Applications 267 (2025) 126150

Contents lists available at ScienceDirect

Expert Systems With Applications

journal homepage: www.elsevier.com/locate/eswa

Enhanced deepfake detection with DenseNet and Cross-ViT

Fazeela Siddiqui, Jiachen Yang, Shuai Xiao *, Muhammad Fahad

School of Electrical and Information Engineering, Tianjin University, 92 Weijin Road, Nankai District, Tianjin, 300072, China

ARTICLE INFO

Keywords:
Deep learning
Vision transformer
Convolutional networks
DenseNet
DeepFake video forensics

ABSTRACT

Artificial intelligence technologies have transformed multiple domains by utilizing sophisticated methods like generative adversarial networks (GANs) to produce authentic-looking counterfeit samples and deepfake movies. These technologies can provide substantial hazards, inducing apprehension and instability by empowering individuals to produce and distribute false information that appears genuine. Thus, it is imperative to create resilient systems that can effectively differentiate between authentic and altered videos, especially in the always changing realm of social media. Convolutional neural networks (CNNs), namely DenseNet121, have become extensively employed in recent years for detecting deepfake videos, owing to their high efficacy. As deepfake techniques advance, the demand for enhanced detection methods becomes increasingly crucial. The primary objective of our research is to recognize video deepfakes, with a particular focus on the facial forgeries that are becoming more and more realistic, making them difficult to detect. Our concept involves integrating Vision Transformers with a neural feature extractor based on DenseNet to produce a novel technique. This approach yields results equivalent to the most recent Vision Transformer techniques without depending on intricate tactics like distillation or ensemble methods. In addition, we have devised a simple yet effective inference technique that employs a voting mechanism to address the identification of numerous faces in a single video frame. The highest-performing model has exhibited outstanding performance, with an Area Under the Curve (AUC) of 99.99% and a mean F1-score of 99.0% on the DeepForensics 1.0 dataset. Additionally, our model demonstrated a remarkable Area Under the Curve (AUC) of 97.4% and an F1-score of 95.1% on the CelebDF dataset, showcasing its effectiveness.

1. Introduction

The vast quantity of readily available content on social media, advanced tools, and affordable computing resources have made it remarkably easy for individuals to produce Deepfakes that spread falsehoods and misinformation. The convenience with which anyone can utilize these technologies to craft propaganda may provoke anxiety and chaos due to their rapid advancement (Rafique et al., 2023). Image, video, and audio editing technologies are advancing at a swift pace (Chesney & Citron, 2019; Zheng, Zhang, & Thing, 2019). The rise of artificial intelligence has significantly simplified the process of creating counterfeit videos that are difficult to distinguish from authentic ones (Suratkar & Kazi, 2023). Deep Fake is a technique that allows users to overlay their facial image onto another person's video, effectively creating a replica of the original individual. This enables the user to imitate the gestures and speech of the person being mimicked. Within this realm lies a particular form of Deepfake technology referred to as face-swapping (Ahmed, Sonuç, Ahmed, & Duru, 2022). In early 2021, a successful deepfake spoofing assault on facial recognition was discovered in China. Individuals engaged in tax fraud utilized pilfered facial pictures to generate deepfake movies and employed a modified phone equipped with a tampered camera to deceive the tax invoicing structure into validating these pre-generated deepfake individuals, resulting in a fraudulent acquisition of $76.2 million (Borak, 2021).

This study examines the technological viability of a spoofing attack on facial recognition systems. Initially, we thoroughly examine potential risks to comprehend which facial recognition scenarios enable the implementation of deepfake spoofing assaults. Using this approach, we establish the attacker model for the facial recognition system attacks. This study displays the capacity of deepfakes to deceive two commercial facial recognition systems. The researcher will explore potential methods to mitigate these spoofing assaults (Salko, Firc, & Malinka, 2024).

The advancement of deepfake creation techniques and their growing availability compels the research community to develop efficient methods for differentiating between modified videos and authentic ones. Simultaneously, there is a growing trend in the field of Computer Vision

* Corresponding author.
E-mail addresses: sidd_fazeela@tju.edu.cn (F. Siddiqui), yangjiachen@tju.edu.cn (J. Yang), xs611@tju.edu.cn (S. Xiao), mfahadgull77@tju.edu.cn (M. Fahad).

https://doi.org/10.1016/j.eswa.2024.126150
Received 29 July 2024; Received in revised form 31 October 2024; Accepted 10 December 2024
Available online 17 December 2024

towards using more models based on Transformers. These models have been proven to deliver outstanding results in various applications, such as image processing (Khan et al., 2022; Pang et al., 2020), document retrieval (MacAvaney et al., 2020), and efficient visual-textual matching (Messina, Amato, Esuli et al., 2021; Messina, Falchi, Esuli and Amato, 2021). They are particularly useful in large-scale multi-modal retrieval systems (Amato et al., 2023; Messina, Amato, Falchi, Gennaro and Marchand-Maillet, 2021).

Deepfakes employ techniques that allow for the creation of human images with unprecedented levels of editing, rendering it challenging to distinguish between authentic and fabricated visuals. Those deepfakes that pose a threat to society are generated using a technique called generative adversarial neural networks (Suganthi et al., 2022). By utilizing the FSGAN framework, films and pictures featuring face-swapping and reenactment can also be generated (Nirkin, Hassner, & Keller, 2022; Nirkin, Keller, & Hassner, 2019). Deepfake videos can be classified into categories such as face-swapping, synthesis, and manipulation of facial features. Face-swap Deepfakes involve substituting an individual's face with that of another to create a video that falsely implicates them in offenses they did not commit (Liu & Liu, 2022), which can severely harm their public image (Harwell, 2018).

A multitude of research articles have been published in this area, encompassing facial landmark detection techniques (Vezzetti, Marcolin, Tornincasa, & Maroso, 2016; Zhang, Zhang, Liu, & Tang, 2014), the Viola–Jones face detector (Viola & Jones, 2001), and lip detection methods (Bazarevsky, Kartynnik, Vakunov, Raveendran, & Grundmann, 2019). The authors explored a deep learning-based approach for detecting deepfakes in videos using XGBoost (Ismail, Elpeltagy, S. Zaki, & Eldahshan, 2021). The main challenges with existing methods are their excessive processing time and precision issues. In this research, we employ deep learning algorithms for analyzing images and detecting deepfake faces in videos. When it comes to human interactions, the implementation of biometrics-based human authentication and identity services significantly emphasizes the importance of facial recognition. Hence, alterations in facial features can potentially undermine trust in digital communication and security systems (Akhtar, Dasgupta, & Banerjee, 2019). Differentiating between authentic and manipulated videos has become a significant challenge because of their devastating global repercussions (Wodajo & Atnafu, 2021).

As the technology for creating AI-generated content advances, it will become increasingly challenging to distinguish it from authentic content. The ability to detect and analyze faces in photos or videos is essential for recognizing bogus material. Even if Deepfake information is not intentionally harmful, it still needs to be detected because some of it can be detrimental to society by causing skepticism about what is true while also serving malicious purposes (Liu & Liu, 2022). The main goal of this study is to ascertain methods for identifying Deepfake images and videos. We provide an innovative approach for detecting and classifying deep fake videos using the Vision Transformers with a convolutional feature extractor based on DenseNet. The DL-based and vision transformer method is utilized on a dataset consisting of videos created by a GAN architecture. An analysis was conducted on the accuracy and losses of the model. The empirical findings demonstrate the efficacy of our suggested methodology.

- **Utilization of Diverse Deepfake Datasets:** We employ a wide range of sub-datasets, including FaceForensics++ (face swap, Face- 2Face, DeepFakes, neural textures), Celeb-DF, and DeeperForensics 1.0, to comprehensively assess the performance and generalize ability of our proposed model.
- **Hybrid Architecture for Feature Extraction:** Our approach integrates Vision Transformers (Cross-ViT) with a DenseNet-based convolutional feature extractor. This combination enhances the detection of subtle manipulations by leveraging both local and global feature analysis.
- **Accurate Classification of Manipulated Frames:** The model predicts and classifies video frames as either real or fake by analyzing the features extracted through the hybrid architecture, offering a robust solution to deepfake detection.
- **Performance Evaluation:** The effectiveness of our approach is demonstrated using standard metrics, including accuracy and F1-score, where the model consistently achieves superior results compared to existing techniques.

2. Related work

2.1. Deep detect

Deep learning has become renowned because it can accurately represent complex and multidimensional data (Fahad et al., 2024; Pavan Kumar & Jayagopal, 2021; Shad et al., 2021). The classification approach based on neural networks has been used by researchers to classify genuine images and Deepfakes accurately (Afchar, Nozick, Yamagishi, & Echizen, 2018; Güera & Delp, 2018; Korshunov & Marcel, 2018). The paper addresses data challenges, training resources, reliability issues, and emerging manipulation approaches, emphasizing deep learning-based methods' dominance in detecting deepfakes despite limitations (Kaur, Noori Hoshyar, Saikrishna, Firmin, & Xia, 2024). This paper evaluates deepfake detection strategies using deep learning algorithms, categorizing them based on multimedia applications, highlighting new discoveries, security flaws, and areas requiring further investigation (Heidari, Jafari Navimipour, Dag, & Unal, 2024). The framework uses contrastive learning to represent dynamic style latent vectors, and a style attention module detects visual and temporal artifacts, performing well in cross-dataset and cross-manipulation settings (Choi, Kim, Jeong, Baek, & Choi, 2024). The authors have classified Deepfakes using a Multilayered perceptron from the Face Forensics dataset, resulting in an AUC score of 0.85 (Matern, Riess, & Stamminger, 2019). The authors offer a deep fake detection and classification technique. Our suggested system analyses the picture at the fault level, and a Dense Network is used to extract deep features. The characteristics of the detected face frames were extracted using the VGG-19 network as in Nguyen, Yamagishi and Echizen (2019) from Nguyen et al. while the reference study (Nguyen, Fang, Yamagishi and Echizen, 2019) programmed an autoencoder with a convolutional neural network to detect altered videos from the FaceForensics++ dataset (c23) dataset edited videos. In addition, the study locates the actual facial areas in the movie where the manipulation was detected. The authors in Li, Yang, Sun, Qi, and Lyu (2020) utilize various convolutional neural network (CNN) architectures, including InceptionV3, MesoInception4, ResNet-50, Meso4, XceptionNet, FWA-based Dual Spatial Pyramid (DPS), and a CNN constructed based on a multilayer feed-forward network. Subsequently, these diverse architectures undergo training using distinct datasets and are evaluated using the Celeb-DF (Li et al., 2020) dataset. Mohammad Farukh Hashmi et al. (2020) introduced a deep fake detection approach based on CNN and LSTM. However, this method is unsuitable for scaling up systems with greater resources.

2.2. Face detection from images

The most frequent method for face identity fraud involves swapping facial features (Zhu, Li, Wang, Xu, & Sun, 2021). It involves primarily two methods to producing realistic faces: Generative adversarial Networks (GANs) (Goodfellow et al., 2020) and variational AutoEncoders (VAEs) (Kingma & Welling, 2013). GANs utilize two separate networks. Discriminator is responsible for distinguishing between real and fraudulent videos, while generator in network convincingly alters the video to trick the discriminator. GANs have yielded highly plausible and lifelike outcomes, and several techniques have been developed throughout time, including MCS-GAN (Xiao et al., 2023) and GAN-CNN (Sharma, Kumar, & Sharma, 2024). The most impressive achievements in this domain have been accomplished using StyleGAN-V2 (Karras et al., 2020). Latent Forensics, based on StyleGAN has been proposed for high quality images (Delmas, Kacete, Paquelet, Leglaive, & Seguier, 2023) VAE-based systems need two encoder–decoder pairs. Each team learns to dissect and recreate one of the exchange faces. The decoding procedure is then adjusted to rebuild the desired person's visage. This approach is used in DeepFaceLab (Perov et al., 2020).

Several methods have been proposed for detecting deepfake video (Hu, Liao, Wang, & Qin, 2021)[, which used a frame-temporality two-stream convolutional network to detect compressed deepfakes in social networks. Additionally, Liao, Wang, Wang, Hu, and Wu (2023) introduced FAMM, leveraging facial muscle motions, while Zhang et al. (2024) proposed a multi-feature fusion and local enhancement technique. In contrast to these approaches, our method employs DenseNet for spatial feature extraction and CrossViT for global context analysis, improving the detection of compressed and low-resolution deepfakes across various datasets.

The detection of faces can be done according to race and gender (Ju, Hu, Jia, Chen, & Lyu, 2024) The look of the face is a highly conspicuous characteristic for identifying people. The swift progress of face synthesis technology presents a progressively substantial threat to national security. Due to the rapid progress of CNNs (Alom et al., 2019; Haridas & Jyothi, 2019), GANs (Goodfellow et al., 2014), and their alternatives (Hong, Hwang, Yoo, & Yoon, 2019), it is now feasible to generate hyper-realistic images (Ledig et al., 2017), videos (Zakharov, Shysheya, Burkov, & Lempitsky, 2019) and audio signals (Donahue, McAuley, & Puckette, 2018; Suwajanakorn, Seitz, & Kemelmacher-Shlizerman, 2017) that are considerably difficult to distinguish from genuine, unmodified audiovisual content. Deepfakes, although being a quite recent technical development, have drawn plenty of research attention. In the latter half of 2020, there has been a substantial increase in the development of deep fake articles. Using ML and DL algorithms, many scholars have automated deep fake detection in audiovisual content. These technologies make original and counterfeit content easier to spot.

2.3. Vision transformers

The Transformer (Vaswani et al., 2017), a self-attention architecture first used in natural language processing (NLP) and showed outstanding performance. In 2020, the Google team introduced ViT (Dosovitskiy et al., 2020), a computer vision version that gained popularity. Attention is often utilized in vision to preserve structure in convolutional networks. The ViT model shows that relying on Convolutional Neural Networks (CNNs) is unnecessary. The initial stage in applying the transformer to images is partitioning them into patches, which are then mapped to a linear embedding. As a classifier, it passes through numerous layers of self-attention modules to generate the final discriminant vector. The self-attention modules blend the properties of each patch via the self-attention process. The self-attention mechanism is outstanding at detecting global relationships and gathering global data.

The ViT model is ideal for tasks that require developing long-range partnerships. In the task of temporal action localization, such as in Refs. Tirupattur, Duarte, Rawat, and Shah (2021), Yang, Peng, Zhang, Fu, and Han (2020) and Zhao, Han, Yang, Wang, and Zhang (2021), gathering information from the surrounding context of an action is crucial. The dependencies in Tirupattur et al. (2021) are represented using a modified version of ViT. It is a potentially effective solution for detecting deepfake videos, since it considers the global perspective and focuses on the crucial locations. Nevertheless, we have determined that directly employing ViT for deepfake detection is inefficient. Thus, we get knowledge from ViT and introduce an innovative method. It directs the main network to prioritize important areas by consolidating worldwide data.

3. Proposed framework

3.1. Overview and context

The overall proposed model explained in Fig. 1, which consists of two parallel branches processing consecutive frames of the entire face, learning complementary dynamics at multiple levels, and integrating their predictions for final detection. We introduce a hybrid architecture that combines convolutional and transformer models. This architecture is designed to process pre-extracted facial data and generate a probability score indicating whether the face has been edited. Both architectures are trained using a supervised approach to distinguish between genuine and manipulated cases. Thus, researchers address the recognition objective by seeing the problem as a binary categorization issue.

While most public datasets lack identifiable defect areas, in the true deepfake scenario, the loss function is quite significant. Even humans struggle to identify valuable questionable spots when dealing with realistic forgery. We use our experience in fine-grained categorization for weakly guided learning. This also provides certain advantages. The suggested module can be readily applied to several models as a pluggable module, allowing for a more accurate comparison of performance impact without relying on additional benefits from annotation. The loss function for the model is a cross-entropy loss of both the network output and the input label.

We specifically employed the Densenet121 and the Convolutional Cross ViT models, which will be discussed further in the subsequent paragraphs.

3.1.1. Cross vit

The Convolutional Cross ViT extends the Efficient ViT and multi-scale Transformer architecture, as mentioned in Ref. Chen, Fan, and Panda (2021). It takes a step further by splitting into two paths: a small patch S-branch and a large patch L-branch to widen the receptive field. Cross attention combines the visual tokens outputted from Transformer Encoders from both branches, permitting through interface amid the dual directions. The CLS elements that correlate with the outcomes from both forks are subsequently utilized to produce two distinct logits. The logits are summed and then subjected to a sigmoid function to get the final probability result. This process is depicted in Fig. 1. The Convolutional Cross ViT uses a different model backbones.

3.1.2. DenseNet

Densenet121 (Huang, Liu, Van Der Maaten, & Weinberger, 2017) a popular extension of ResNet, the Residual CNN architecture. Unlike ResNet and other convolutional neural networks, DenseNet directly connects each layer to all future layers. We wanted to keep Keras' DenseNet121 model correct while adding a dense layer as the last layer. Model consisted of four thick blocks with interconnected layers, including Batch Standardization (BN) and 3 × 3 turnaround. The design has a transition layer between dense blocks, with an average pooling layer of 2 × 2 and a concentration of 1 × 1. After the last dense block, we added a sigmoid-activated dense layer.

3.2. Netwrok architecture

The model consists of two primary components: Densenet121 and the Cross-ViT. The facial recognition system extracts identifiable facial traits from the images, which will utilized for the purpose of learning. The DenseNet 121 takes the Feature Maps from images for a feed and then transforms them into a series of image pixels to be used in the subsequent detecting process.

The dropout layers, with a rate of 0.05, help in regularizing the model, reducing the risk of overfitting. The Attention class implements the multi-head self-attention mechanism. Using linear layers, it transforms the input into queries, keys, and values.This attention mechanism

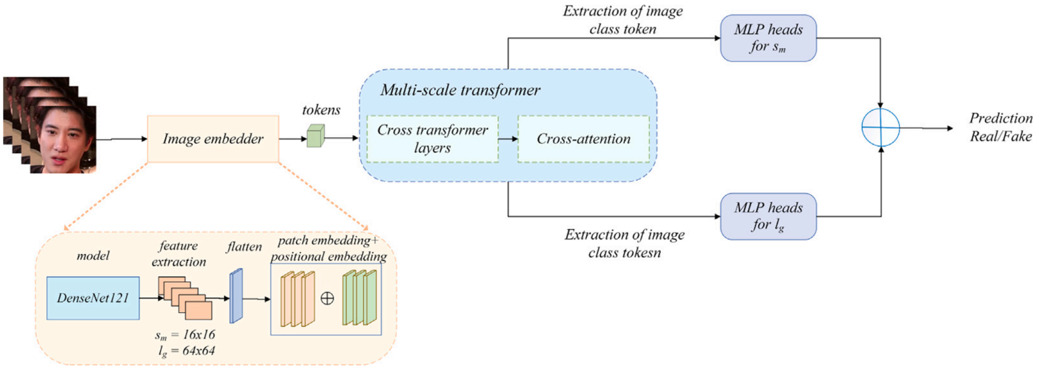

Fig. 1. Architectural design of the proposed network model.

enables the model to focus on different parts of the input sequence, capturing intricate dependencies. The output of the attention process is transformed back to the original dimension using another linear layer followed by dropout. This mechanism is essential for capturing long-range dependencies and interactions within the input data.

The CrossTransformer class alternates attention between small and large tokens to integrate multi-scale features through cross-attention mechanisms. This enhances the model's ability to effectively combine detailed and coarse-grained features, allowing it to leverage information from multiple scales to handle complex tasks. The model consists of four multi-scale encoding blocks, which stack layers of small and large-scale transformers and cross-transformers. In each layer, the small and large tokens are processed separately using transformer encoders, followed by cross-attention to integrate the features from both scales.

This approach captures a wide range of spatial information and dependencies within the input images by leveraging the strengths of different resolutions. The dropout rate for embedding is 0.15. The model uses DenseNet121 to extract features, producing maps with 1024 channels. Adaptive pooling ensures these feature maps are of a fixed size, which is then flattened and linearly projected to generate patch embedding. The patch sizes are 16 × 16 for small images and 32 × 32 for large images. Multi-head attention mechanisms within the transformers focus on different parts of these patches, capturing spatial relationships effectively.

The Cross-ViT class combines the components to form a complete model. It initializes image embedders for small and large scales and uses a multi-scale encoder to process these embeddings. The final class tokens are processed through MLP heads, providing the logits combined to generate the final output.

3.2.1. Cross transformer layers

As shown in Fig. 2, in the cross-transform layers there are two tokens: one is for small images and the other is for large images. So the small image tokens, i.e., s_{m_tokens} split into $s_{m\_cls} \in \mathbb{R}^{B\times 1\times D_{s_m}}$ and $s_{m\_patch\_tokens} \in \mathbb{R}^{B\times N_{s_{m\_patches}}\times D_{s_m}}$.

Similarly, l_{g_tokens} splits into $l_{g\_cls} \in \mathbb{R}^{B\times 1\times D_{l_g}}$ and $l_{g\_patch\_tokens} \in \mathbb{R}^{B\times N_{l_{g\_patches}}\times D_{l_g}}$. The small image's class token (s_{m_cls}) interacts with the $l_{g\_patch\_tokens}$ to update the s_{m_cls}. Similarly, the large image's class token (l_{g_cls}) updates the $s_{m\_patch\_tokens}$. To the token sequences, we concatenate the updated class tokens with corresponding patch tokens, yielding the updated token, which is $B\times(1+N_{s_{m\_patches}})\times D_{s_m}$ for s_{m_tokens} and $B \times (1 + N_{l_{g\_patches}}) \times D_{l_g}$ for l_{g_tokens}. These updated tokens can interact with cross-attention layers to compute the information between tokens.

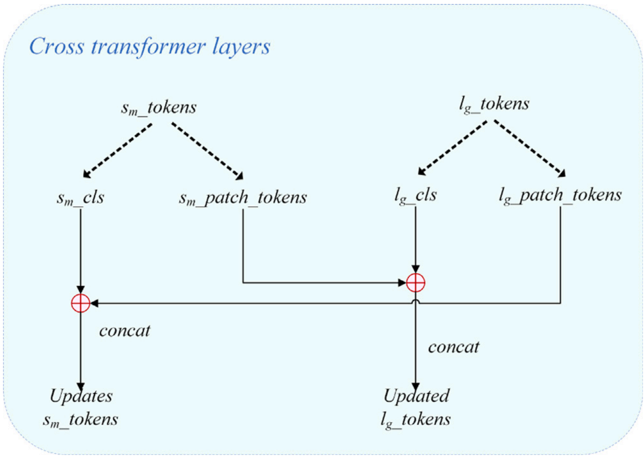

Fig. 2. Detailed operation of cross-transformer layers for *sm* and *lg* image tokens.

3.2.2. Cross-attention map

It is discussed to improve the model further. As shown in Fig. 3, input characteristics are denoted by s_m-branch and l_g-branch as $X_s \in \mathbb{R}^{B\times N\times D}$ and $X_l \in \mathbb{R}^{B\times N\times D}$. We first transform the representation $X_s \in \mathbb{R}^{B\times N\times D}$ using spatially average adaptive pooling. To get a Query $X_q^t \in \mathbb{R}^{B\times h\times N\times d_k}$, a trainable projection and shape transformation are used. X_q^t and Key $X_k^t \in \mathbb{R}^{B\times h\times N\times d_k}$, and Value $X_v^t \in \mathbb{R}^{B\times h\times N}$ are acquired by a trainable projection with reshaping. Similarly, X_s is converted to obtain $X_q^t \in \mathbb{R}^{B\times h\times N\times d_k}$, $X_k^t \in \mathbb{R}^{B\times h\times N\times d_k}$, and $X_v^t \in \mathbb{R}^{B\times h\times N}$.

After, we created the attention map $A = \text{Softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right) V$ we get the refined feature $B \times h \times N \times d_k$ where X_s output is $\in \mathbb{R}^{B\times h\times N\times d_k}$. The original embedding dim is re-aligned using a linear projection. $X_{\text{out}} \in \mathbb{R}^{B\times N\times (h\cdot d_k)}$. The same operation is executed on l_g images, where $A = \text{softmax(dots)} \in \mathbb{R}^{B\times h\times N\times N}$, and $\text{dots} = \text{einsum}(q, k) \cdot \text{scale} \in \mathbb{R}^{B\times h\times N\times N}$, after applying the attention then, and $X_{\text{out}} = \text{einsum}(A, v) \in \mathbb{R}^{B\times h\times N\times d_k}$, after shape transformation is done then $X_{\text{out}} = \mathbb{R}^{B\times N\times (h\cdot d_k)}$ then a Linear projection applied to obtain $X_{\text{out}} \in \mathbb{R}^{B\times N\times D}$. Finally, the attention heads are concatenated back together to form the output tensor $X \in \mathbb{R}^{B\times N\times D}$.

4. Experimental discussion

4.1. Settings

4.1.1. Datasets

Initially, we performed testing on FaceForensics++(FF++). The collection consists of authentic and counterfeit videos created using various deepfake-generating methods. During our assessment, we considered the modified movies generated in the Deepfakes (DF), Face2Face(F2F), FaceSwap (FS), and NeuralTextures(NT) sub-datasets. Furthermore, we utilize the Celeb-DF (C-DF) (Li et al., 2020) dataset, including 590 authentic and 5639 counterfeit videos, encompassing

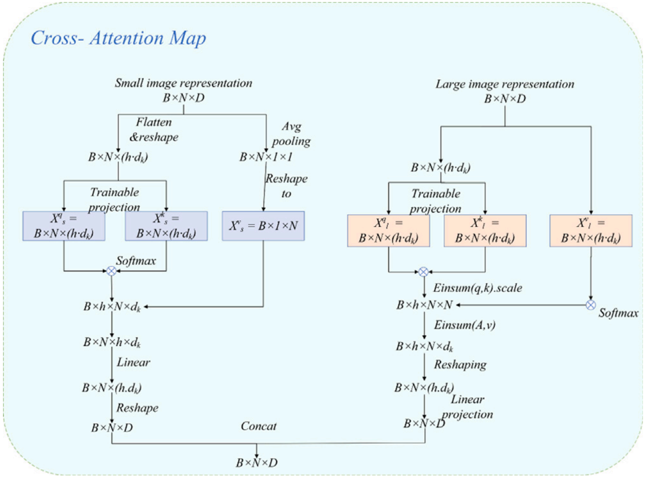
Cross- Attention Map
Small image representation
$B\times N\times D$
Flatten &reshape
Avg pooling
$B\times N\times(h\cdot d_k)$
$B\times N\times 1\times 1$
Reshape to
Trainable projection
$X^q_s = B\times N\times(h\cdot d_k)$
$X^k_s = B\times N\times(h\cdot d_k)$
$X^v_s = B\times 1\times N$
Softmax
$B\times h\times N\times d_k$
$B\times N\times h\times d_k$
Linear
$B\times N\times(h.d_k)$
Reshape
$B\times N\times D$
Concat
$B\times N\times D$
Large image representation
$B\times N\times D$
$B\times N\times(h\cdot d_k)$
Trainable projection
$X^q_l = B\times N\times(h\cdot d_k)$
$X^k_l = B\times N\times(h\cdot d_k)$
$X^v_l = B\times N\times(h\cdot d_k)$
Einsum(q,k).scale
$B\times h\times N\times N$
Softmax
Einsum(A,v)
$B\times h\times N\times d_k$
Reshaping
$B\times N\times(h.d_k)$
Linear projection
$B\times N\times D$

Fig. 3. Comprehensive functionality of cross-attention.

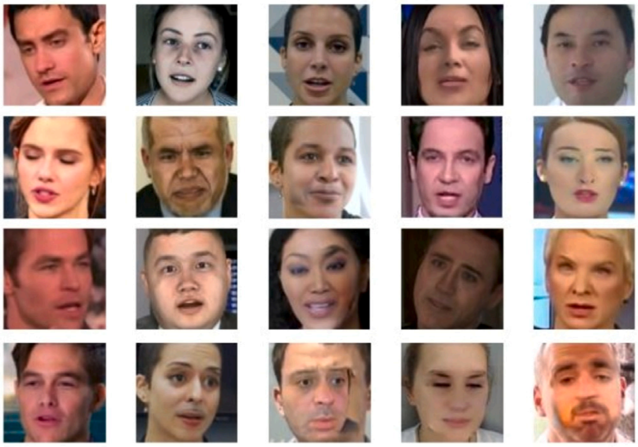

Fig. 4. Sample of fake extracted images.

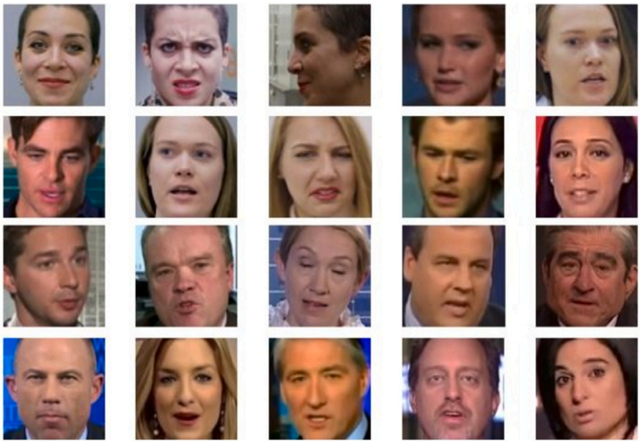

Fig. 5. Sample of real extracted images.

more than 2 million video frames. The model was trained using a training directory containing training videos from the Celeb-DF (Li et al., 2020), DeeperForensics 1.0 (DF 1.0) (Jiang, Li, Wu, Qian, & Loy, 2020), UADFV (Yang, Li, & Lyu, 2019) dataset, as well as forged videos from each of the FaceForensics++ sub-datasets (Rossler et al., 2019) The trained model was then used to generate the accuracy metrics, which have been defined distinctly. To assess the proposed approach on the Celeb-DF, DeeperForensics, UADFV, and FaceForensics++ test set, we performed experiments using the Convolutional Vision Transformer (Wodajo & Atnafu, 2021) on these videos. We were then able to gather the necessary AUC and F1-score values for comparison. During training, we utilized the Augmentations library (Buslaev et al., 2020) to provide common transformations like as blur, Gaussian noise, transposition, rotation, and various isotropic resizes.

4.1.2. Computational cost

The model was trained and tested using a Linux server equipped with an NVIDIA GeForce RTX 2080 GPU and Python 3.9 with the required libraries. The training time for the full model (DenseNet + CrossViT) on the datasets used in our study was approximately 75-80 min per epoch on one Dataset (it could be faster if we use a powerful GPU). The model required a minimum of 1.6 GB of GPU memory during training. Given the complexity of CrossViT's multi-scale attention mechanism and DenseNet's deep feature extraction, the computational cost remains manageable, with efficient performance on the 2080 GPU.

4.1.3. Implementation details

In each video frame, MTCNN was used to extract and align face areas, which were then scaled to 224 × 224. The framework was constructed using open-source PyTorch (Paszke et al., 2017) DenseNet 121 as our backbone with weights initialized using the pre-trained model. L has 12 blocks and D has 256 embedding dimensions. Setting the cross attention dim head D_k to 64. The framework was trained with a 0.01 learning rate and 1e−7 weight decay using the SGD optimizer. Using 32 batches, 30 epochs were trained. Two equally weighted cross entropy losses during inference comprise the loss function during training, we take a 32-frame sample. When more than one face is found in a single frame, the classifier is applied to each face, and the frame's predicted confidence is based on the highest fakeness confidence. After we have all of the frame projections, we can average them to determine the video's prediction. We set the confidence to 0.5 for videos when no face is detected in all frames so that we may compare all test sets fairly. All the dataset contains the raw data of videos that can be used to extract faces from the videos from different datasets. Fig. 4, shows the fake images extracted, while Fig. 5, contains the different images extracted from the manipulated videos. Take notice that all these images contain 224 × 224 RGB format.

4.2. Base

4.2.1. Effectiveness of the backbone network

Several tests were carried out on various datasets to thoroughly examine the efficiency of our network architecture. For the FF++, VIT and Cross ViT were used for comparison. The comparative findings are shown in Table 1. It is observed that our proposed method with GeLU activation achieves the best detection results within the dataset, but Cross ViT with DenseNet outperforms all competitors in terms of robustness as well as detection performance within the dataset, demonstrating that it can learn a stronger dynamic representation.

Various video-level Transformer topologies were tested to verify their efficacy in characterizing. For comparison, we investigated widely-used models in video interpretation and analysis, such as Conv. ViT (Wodajo & Atnafu, 2021) and Cross-ViT (Coccomini et al., 2022). Table 2, demonstrates that Cross-ViT delivers superior classification results on F2F, but, it lacks robustness when applied to DeepFake techniques. The approach we present has exceptional detection accuracy when applied to the FF++ dataset, and it also displays much-improved robustness capabilities.

Table 1
Models accuracy based on %.

Model	FaceForensics++					DFD	UADFV	C-DF	DF 1.0
	Mean	DF	FS	F2F	NT				
Conv. ViT (Wodajo & Atnafu, 2021)	85.24	**99.68**	82.55	83.46	75.28	93.9	88.33	93.52	94.85
Conv. Cross ViT Eff.net B0 (Coccomini, Messina, Gennaro, & Falchi, 2022)	88.52	92.8	87.37	84.58	89.95	**98.2**	90.78	94.78	95.92
Conv. Cross ViT GeLU (our)	94.05	93.67	94.26	93.66	**94.62**	91.75	82.66	92.66	90.58
Conv. Cross ViT DenseNet121 (our)	**97.84**	99.57	**98.88**	**98.96**	93.96	97.96	**98.29**	**99.93**	**99.66**

Table 2
Test results on FaceForensics++ (Rossler et al., 2019).

Models	DF		FS		F2F		NT	
	AUC	F1-score	AUC	F1-score	AUC	F1-score	AUC	F1-score
Conv. ViT (Wodajo & Atnafu, 2021)	0.8433	77.82%	0.8257	78.79%	0.9236	87.48%	0.7405	78.52%
Conv. Cross ViT (Coccomini et al., 2022)	0.9512	88.05%	0.8593	83.33%	0.9998	99.15%	0.8864	85.99%
XceptionNet (Rossler et al., 2019)	0.9636	–	0.9029	–	0.8686	–	0.8067	–
Conv. Cross ViT GeLU (our)	0.8675	81.39%	0.8251	78.93%	0.8695	95.26%	0.7962	75.64%
Conv. Cross ViT DenseNet121 (our)	0.9988	99.34%	0.9999	99.15%	0.9996	98.90%	0.9893	89.44%

Table 3
Compression robustness.

Models	Video-wise AUC		
	Unprocessed	High-res.	Low-res.
Conv. ViT (Wodajo & Atnafu, 2021)	0.998	**0.993**	0.92
Conv. Cross ViT Eff.net B0 (Coccomini et al., 2022)	**0.999**	0.978	0.777
Conv. cross ViT GeLU (our)	0.997	0.990	0.958
Conv. Cross ViT DenseNet121 (our)	**0.999**	0.985	**0.969**

4.2.2. Robustness under compression

Mostly compressed videos are widely distributed on social media networks, Because compressed videos are commonly transmitted on social media networks, we ran additional tests using FF++ at various compression levels to assess the resilience of our method. Table 4's findings reveal that all models perform nearly flawlessly on raw videos. However, their robustness differs when trained at different compression settings. Frame-based detection approaches, such as Cross-ViT, which depends on detecting face borders, suffer the most under compression circumstances due to the eradication of intra-frame abnormalities, and their performance significantly decreases on low-resolution movies. Compared to state-of-the-art techniques, our model performs similarly on unprocessed and low-resolution versions, suggesting that the proposed characteristics are unaffected by resolution or noise (see Table 3).

4.3. Robustness across datasets

To ensure a comprehensive evaluation of the created models, we conducted additional experiments on different datasets. Based on the data presented in Table 4, our models have demonstrated superior performance compared to the original Convolutional ViT (Wodajo & Atnafu, 2021) and Convolutional Cross ViT across various datasets. It is crucial to emphasize that the accuracy results obtained from different datasets support the findings in Wodajo and Atnafu (2021) specific deepfake methodologies, Nevertheless, the average performance of our models surpasses that of the Convolutional ViT. The Convolutional Cross ViT shows improved accuracy results across different datasets, particularly when combined with a dense network to enhance accuracy. Conversely, the accuracy of our newly introduced model (Cross ViT with GeLU activation) is comparatively lower when benchmarked against other models. Thus, our proposed model exhibits notable enhancements when compared to models introduced in Coccomini et al. (2022) and Wodajo and Atnafu (2021) If we assess the accuracy on DFD, C-DF, UADFV, DF 1.0, it demonstrates an impressive average accuracy of 97.84%, while the best results are obtained from DeeperForensics 1.0, with an outstanding accuracy of 99.99%, equivalent to the current state-of-the-art. To ensure that our model produces the best results, we evaluated it on several datasets. Tables 4 and 5 indicates that all models using Densenet121 on various datasets get greater accuracy and F1 scores than the convolutional ViT (Coccomini et al., 2022; Wodajo & Atnafu, 2021). It should also be noted that Cross-efficient DenseNet121 produces the best results on Deeper Forensics 1.0. To be more sure, in Fig. 6, we have plotted the detailed ROC for different architectures on the deeper forensics 1.0 dataset.

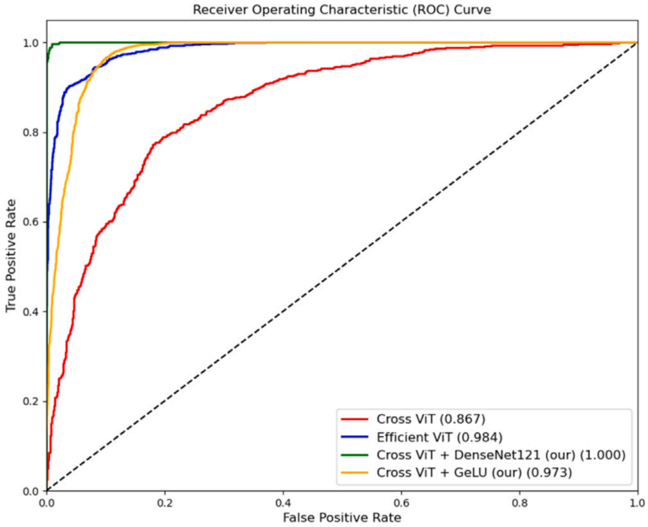

Fig. 6. ROC curve comparison between our model and others on the Deeper Forensics 1.0 dataset.

4.4. Ablation study

We evaluated four configurations: the DenseNet backbone independently to evaluate its feature extraction capabilities, the Vision Transformer (ViT) independently to assess the efficacy of attention mechanisms, DenseNet in conjunction with a standard ViT to investigate the synergy between convolutional feature extraction and transformer-based attention, and ultimately, the proposed DenseNet + CrossViT model as our comprehensive approach. The results, provided in Table 5, demonstrate that each component contributes to the overall performance. However, the DenseNet + CrossViT combination significantly outperforms the other configurations, showing substantial improvements in terms of AUC and F1-scores across all datasets, including FaceForensics++, DFD, UADFV, and DeeperForensics 1.0. This indicates that the CrossViT module is critical for achieving the highest levels of

Table 4
Test results on different datasets.

Models	DFD		UADFV		C-DF		DF 1.0	
	AUC	F1-score	AUC	F1-score	AUC	F1-score	AUC	F1-score
Conv. ViT (Wodajo & Atnafu, 2021)	0.8988	84.55%	0.843	80.32%	0.8695	82.52%	0.9298	91.68%
Conv. Cross ViT Eff.net B0 (Coccomini et al., 2022)	0.9239	89.49%	0.8272	78.66%	0.8824	84.96%	0.9659	92.62%
Conv. Cross ViT GeLU (our)	0.8752	85.55%	0.8982	82.98%	0.9856	96.86%	0.9128	86.95%
Conv. Cross ViT DenseNet121 (our)	**0.9996**	**99.19%**	**0.9689**	**92.62%**	**0.9746**	**95.60%**	**0.9999**	**99.60%**

Table 5
Performance comparison of various backbone models across deepfake detection datasets.

Models	DFD		UADFV		C-DF		DF 1.0	
	AUC	F1-score	AUC	F1-score	AUC	F1-score	AUC	F1-score
DenseNet121 backbone	0.8612	79.35%	0.8102	87.12%	0.8823	89.01%	0.7432	77.10%
ViT backbone	0.8255	75.80%	0.8041	74.86%	0.8415	82.95%	0.6952	72.00%
DenseNet121 with ViT	0.8971	84.02%	0.8357	79.98%	0.9054	90.23%	0.7684	80.22%
Conv. Cross ViT DenseNet121 (our)	**0.9996**	**99.19%**	**0.9689**	**92.62%**	**0.9746**	**95.60%**	**0.9999**	**99.60%**

accuracy and robustness, particularly in handling compressed video artifacts. We believe this ablation study further validates the architecture and demonstrates the effectiveness of our proposed approach.

5. Conclusion

In this research, we focused on the important problem of identifying fake video frames. This concern affects many people because deepfake technology is becoming more common. Our model was designed to improve existing methodologies, improve accuracy, and reduce error rates. The model performed exceptionally well predicting fake videos, achieving an accuracy score of 99.99%. This score is much higher than the accuracy scores of previous models, demonstrating its superiority. We used a combination of different architectures to improve our model. We used convolutional networks like DenseNet121 to extract visual features and Vision Transformers to get a complete global description for other tasks. This combination has proven effective, allowing for state-of-the-art results without the need for distillation techniques used in other models based on convolutional or ensemble networks. More specifically, the patch extractor performed better than generic convolutional networks trained from the beginning even in its smallest network configuration. This was observed in comparison with the work of Wodajo et al. Although we have made significant progress, our proposed model cannot still detect unknown fake frames effectively. In the future, we will work on integrating a more extensive feature dictionary to make the model more flexible and improve its ability to detect things. In short, this research highlights the potential of mixed convolutional-transformer networks in advancing the field of deepfake detection and lays the groundwork for future enhancements.

CRediT authorship contribution statement

Fazeela Siddiqui: Conceptualization, Methodology, Software, Investigation, Writing – original draft. **Jiachen Yang:** Formal analysis, Investigation, Resources, Supervision. **Shuai Xiao:** Conceptualization Methodology, Software, Investigation, Writing – original draft. **Muhammad Fahad:** Methodology, Software, Writing – review & editing.

Funding

This work was supported by the National Natural Science Foundation of China under Grant 62271345, 62301356.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgment

The authors especially acknowledge the Artificial Intelligence and Marine Information Processing Lab (AIMIP) at Tianjin University, Tianjin, China. The authors wish to acknowledge the editor and reviewers for their insightful comments, which have improved the quality of this publication.

Data availability

Data will be made available on request.

References

Afchar, D., Nozick, V., Yamagishi, J., & Echizen, I. (2018). Mesonet: a compact facial video forgery detection network. In *2018 IEEE international workshop on information forensics and security* (pp. 1–7). IEEE.

Ahmed, S. R., Sonuç, E., Ahmed, M. R., & Duru, A. D. (2022). Analysis survey on deepfake detection and recognition with convolutional neural networks. In *2022 international congress on human-computer interaction, optimization and robotic applications* (pp. 1–7). IEEE.

Akhtar, Z., Dasgupta, D., & Banerjee, B. (2019). Face authenticity: An overview of face manipulation generation. *Detection and Recognition*, *5*.

Alom, M. Z., Taha, T. M., Yakopcic, C., Westberg, S., Sidike, P., Nasrin, M. S., et al. (2019). A state-of-the-art survey on deep learning theory and architectures. *Electronics*, *8*(3), 292.

Amato, G., Bolettieri, P., Carrara, F., Falchi, F., Gennaro, C., Messina, N., et al. (2023). VISIONE at video browser showdown 2023. In *International conference on multimedia modeling* (pp. 615–621). Springer.

Bazarevsky, V., Kartynnik, Y., Vakunov, A., Raveendran, K., & Grundmann, M. (2019). Blazeface: Sub-millisecond neural face detection on mobile gpus. arXiv preprint arXiv:1907.05047.

Borak, M. (2021). Chinese government-run facial recognition system hacked by tax fraudsters: report. *South China Morning Post*.

Buslaev, A., Iglovikov, V. I., Khvedchenya, E., Parinov, A., Druzhinin, M., & Kalinin, A. A. (2020). Albumentations: fast and flexible image augmentations. *Information*, *11*(2), 125.

Chen, C.-F. R., Fan, Q., & Panda, R. (2021). Crossvit: Cross-attention multi-scale vision transformer for image classification. In *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 357–366).

Chesney, B., & Citron, D. (2019). Deep fakes: A looming challenge for privacy, democracy, and national security. *California Law Review*, *107*, 1753.

Choi, J., Kim, T., Jeong, Y., Baek, S., & Choi, J. (2024). Exploiting style latent flows for generalizing deepfake video detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 1133–1143).

Coccomini, D. A., Messina, N., Gennaro, C., & Falchi, F. (2022). Combining efficientnet and vision transformers for video deepfake detection. In *International conference on image analysis and processing* (pp. 219–229). Springer.

Delmas, M., Kacete, A., Paquelet, S., Leglaive, S., & Seguier, R. (2023). LatentForensics: Towards lighter deepfake detection in the StyleGAN latent space. arXiv preprint arXiv:2303.17222.

Donahue, C., McAuley, J., & Puckette, M. (2018). Adversarial audio synthesis. arXiv preprint arXiv:1802.04208.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., et al. (2020). An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929.

Fahad, M., Zhang, T., Iqbal, Y., Ikram, A., Siddiqui, F., Abdullah, B. Y., et al. (2024). Advanced deepfake detection with enhanced resnet-18 and multilayer CNN max pooling. *Visual Computer*, 1–14.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., et al. (2014). Generative adversarial nets. *Advances in Neural Information Processing Systems*, *27*.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., et al. (2020). Generative adversarial networks. *Communications of the ACM*, *63*(11), 139–144.

Güera, D., & Delp, E. J. (2018). Deepfake video detection using recurrent neural networks. In *2018 15th IEEE international conference on advanced video and signal based surveillance* (pp. 1–6). IEEE.

Haridas, R., & Jyothi, R. (2019). Convolutional neural networks: A comprehensive survey. *International Journal of Applied Engineering Research*, *14*(3), 780–789.

Harwell, D. (2018). Scarlett Johansson on fake AI-generated sex videos: 'Nothing can stop someone from cutting and pasting my image'. *Washington Post*, *31*, 12.

Hashmi, M. F., Ashish, B. K. K., Keskar, A. G., Bokde, N. D., Yoon, J. H., & Geem, Z. W. (2020). An exploratory analysis on visual counterfeits using conv-lstm hybrid architecture. *IEEE Access*, *8*, 101293–101308.

Heidari, A., Jafari Navimipour, N., Dag, H., & Unal, M. (2024). Deepfake detection using deep learning methods: A systematic and comprehensive review. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, *14*(2), Article e1520.

Hong, Y., Hwang, U., Yoo, J., & Yoon, S. (2019). How generative adversarial networks and their variants work: An overview. *ACM Computing Surveys*, *52*(1), 1–43.

Hu, J., Liao, X., Wang, W., & Qin, Z. (2021). Detecting compressed deepfake videos in social networks using frame-temporality two-stream convolutional network. *IEEE Transactions on Circuits and Systems for Video Technology*, *32*(3), 1089–1102.

Huang, G., Liu, Z., Van Der Maaten, L., & Weinberger, K. Q. (2017). Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 4700–4708).

Ismail, A., Elpeltagy, M., S. Zaki, M., & Eldahshan, K. (2021). A new deep learning-based methodology for video deepfake detection using XGBoost. *Sensors*, *21*(16), 5413.

Jiang, L., Li, R., Wu, W., Qian, C., & Loy, C. C. (2020). Deeperforensics-1.0: A large-scale dataset for real-world face forgery detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 2889–2898).

Ju, Y., Hu, S., Jia, S., Chen, G. H., & Lyu, S. (2024). Improving fairness in deepfake detection. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision* (pp. 4655–4665).

Karras, T., Laine, S., Aittala, M., Hellsten, J., Lehtinen, J., & Aila, T. (2020). Analyzing and improving the image quality of stylegan. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 8110–8119).

Kaur, A., Noori Hoshyar, A., Saikrishna, V., Firmin, S., & Xia, F. (2024). Deepfake video detection: challenges and opportunities. *Artificial Intelligence Review*, *57*(6), 1–47.

Khan, S., Naseer, M., Hayat, M., Zamir, S. W., Khan, F. S., & Shah, M. (2022). Transformers in vision: A survey. *ACM Computing Surveys (CSUR)*, *54*(10s), 1–41.

Kingma, D. P., & Welling, M. (2013). Auto-encoding variational bayes. arXiv preprint arXiv:1312.6114.

Korshunov, P., & Marcel, S. (2018). Deepfakes: a new threat to face recognition? assessment and detection. arXiv preprint arXiv:1812.08685.

Ledig, C., Theis, L., Huszár, F., Caballero, J., Cunningham, A., Acosta, A., et al. (2017). Photo-realistic single image super-resolution using a generative adversarial network. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 4681–4690).

Li, Y., Yang, X., Sun, P., Qi, H., & Lyu, S. (2020). Celeb-df: A large-scale challenging dataset for deepfake forensics. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 3207–3216).

Liao, X., Wang, Y., Wang, T., Hu, J., & Wu, X. (2023). FAMM: facial muscle motions for detecting compressed deepfake videos over social networks. *IEEE Transactions on Circuits and Systems for Video Technology*, *33*(12), 7236–7251.

Liu, Y., & Liu, X. (2022). Spoof trace disentanglement for generic face anti-spoofing. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *45*(3), 3813–3830.

MacAvaney, S., Nardini, F. M., Perego, R., Tonellotto, N., Goharian, N., & Frieder, O. (2020). Efficient document re-ranking for transformers by precomputing term representations. In *Proceedings of the 43rd international ACM SIGIR conference on research and development in information retrieval* (pp. 49–58).

Matern, F., Riess, C., & Stamminger, M. (2019). Exploiting visual artifacts to expose deepfakes and face manipulations. In *2019 IEEE winter applications of computer vision workshops* (pp. 83–92). IEEE.

Messina, N., Amato, G., Esuli, A., Falchi, F., Gennaro, C., & Marchand-Maillet, S. (2021). Fine-grained visual textual alignment for cross-modal retrieval using transformer encoders. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, *17*(4), 1–23.

Messina, N., Amato, G., Falchi, F., Gennaro, C., & Marchand-Maillet, S. (2021). Towards efficient cross-modal visual textual retrieval using transformer-encoder deep features. In *2021 international conference on content-based multimedia indexing* (pp. 1–6). IEEE.

Messina, N., Falchi, F., Esuli, A., & Amato, G. (2021). Transformer reasoning network for image-text matching and retrieval. In *2020 25th international conference on pattern recognition* (pp. 5222–5229). IEEE.

Nguyen, H. H., Fang, F., Yamagishi, J., & Echizen, I. (2019). Multi-task learning for detecting and segmenting manipulated facial images and videos. In *2019 IEEE 10th international conference on biometrics theory, applications and systems* (pp. 1–8). IEEE.

Nguyen, H. H., Yamagishi, J., & Echizen, I. (2019). Capsule-forensics: Using capsule networks to detect forged images and videos. In *ICASSP 2019-2019 IEEE international conference on acoustics, speech and signal processing* (pp. 2307–2311). IEEE.

Nirkin, Y., Hassner, T., & Keller, Y. (2022). FSGANv2: Better subject agnostic face swapping and reenactment. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1.

Nirkin, Y., Keller, Y., & Hassner, T. (2019). Fsgan: Subject agnostic face swapping and reenactment. In *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 7184–7193).

Pang, M., He, L., Xiong, F., Yang, X., He, Z., & Han, X. (2020). Developing an image-based 3D model editing method. *IEEE Access*, *8*, 167950–167964.

Paszke, A., Gross, S., Chintala, S., Chanan, G., Yang, E., DeVito, Z., et al. (2017). Automatic differentiation in pytorch. *Openreview*.

Pavan Kumar, M., & Jayagopal, P. (2021). Generative adversarial networks: a survey on applications and challenges. *International Journal of Multimedia Information Retrieval*, *10*(1), 1–24.

Perov, I., Gao, D., Chervoniy, N., Liu, K., Marangonda, S., Umé, C., et al. (2020). DeepFaceLab: Integrated, flexible and extensible face-swapping framework. arXiv preprint arXiv:2005.05535.

Rafique, R., Gantassi, R., Amin, R., Frnda, J., Mustapha, A., & Alshehri, A. H. (2023). Deep fake detection and classification using error-level analysis and deep learning. *Scientific Reports*, *13*(1), 7422.

Rossler, A., Cozzolino, D., Verdoliva, L., Riess, C., Thies, J., & Nieß ner, M. (2019). Faceforensics++: Learning to detect manipulated facial images. In *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 1–11).

Salko, M., Firc, A., & Malinka, K. (2024). Security implications of deepfakes in face authentication. In *Proceedings of the 39th ACM/SIGAPP symposium on applied computing* (pp. 1376–1384).

Shad, H. S., Rizvee, M. M., Roza, N. T., Hoq, S. A., Monirujjaman Khan, M., Singh, A., et al. (2021). [Retracted] comparative analysis of deepfake image detection method using convolutional neural network. *Computational Intelligence and Neuroscience*, *2021*(1), Article 3111676.

Sharma, P., Kumar, M., & Sharma, H. K. (2024). GAN-CNN ensemble: A robust deepfake detection model of social media images using minimized catastrophic forgetting and generative replay technique. *Procedia Computer Science*, *235*, 948–960.

Suganthi, S., Ayoobkhan, M. U. A., Bacanin, N., Venkatachalam, K., Štěpán, H., Pavel, T., et al. (2022). Deep learning model for deep fake face recognition and detection. *PeerJ Computer Science*, *8*, Article e881.

Suratkar, S., & Kazi, F. (2023). Deep fake video detection using transfer learning approach. *Arabian Journal for Science and Engineering*, *48*(8), 9727–9737.

Suwajanakorn, S., Seitz, S. M., & Kemelmacher-Shlizerman, I. (2017). Synthesizing obama: learning lip sync from audio. *ACM Transactions on Graphics (ToG)*, *36*(4), 1–13.

Tirupattur, P., Duarte, K., Rawat, Y. S., & Shah, M. (2021). Modeling multi-label action dependencies for temporal action localization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 1460–1470).

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., et al. (2017). Attention is all you need. *Advances in Neural Information Processing Systems*, *30*.

Vezzetti, E., Marcolin, F., Tornincasa, S., & Maroso, P. (2016). Application of geometry to rgb images for facial landmark localisation-a preliminary approach. *International Journal of Biometrics*, *8*(3–4), 216–236.

Viola, P., & Jones, M. (2001). Rapid object detection using a boosted cascade of simple features. *vol. 1*, In *Proceedings of the 2001 IEEE computer society conference on computer vision and pattern recognition*. Ieee, I–I.

Wodajo, D., & Atnafu, S. (2021). Deepfake video detection using convolutional vision transformer. arXiv preprint arXiv:2102.11126.

Xiao, S., Lan, G., Yang, J., Lu, W., Meng, Q., & Gao, X. (2023). MCS-GAN: A different understanding for generalization of deep forgery detection. *IEEE Transactions on Multimedia, 26*, 1333–1345.

Yang, X., Li, Y., & Lyu, S. (2019). Exposing deep fakes using inconsistent head poses. In *ICASSP 2019-2019 IEEE international conference on acoustics, speech and signal processing* (pp. 8261–8265). IEEE.

Yang, L., Peng, H., Zhang, D., Fu, J., & Han, J. (2020). Revisiting anchor mechanisms for temporal action localization. *IEEE Transactions on Image Processing*, *29*, 8535–8548.

Zakharov, E., Shysheya, A., Burkov, E., & Lempitsky, V. (2019). Few-shot adversarial learning of realistic neural talking head models. In *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 9459–9468).

Zhang, D., Chen, J., Liao, X., Li, F., Chen, J., & Yang, G. (2024). Face forgery detection via multi-feature fusion and local enhancement. *IEEE Transactions on Circuits and Systems for Video Technology*.

Zhang, Z., Zhang, W., Liu, J., & Tang, X. (2014). Multiview facial landmark localization in RGB-D images via hierarchical regression with binary patterns. *IEEE Transactions on Circuits and Systems for Video Technology*, *24*(9), 1475–1485.

Zhao, T., Han, J., Yang, L., Wang, B., & Zhang, D. (2021). SODA: Weakly supervised temporal action localization based on astute background response and self-distillation learning. *International Journal of Computer Vision*, *129*(8), 2474–2498.

Zheng, L., Zhang, Y., & Thing, V. L. (2019). A survey on image tampering and its detection in real-world photos. *Journal of Visual Communication and Image Representation*, *58*, 380–399.

Zhu, Y., Li, Q., Wang, J., Xu, C.-Z., & Sun, Z. (2021). One shot face swapping on megapixels. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 4834–4844).