Multimedia Tools and Applications
https://doi.org/10.1007/s11042-024-20548-6

Video deepfake detection using a hybrid CNN-LSTM-Transformer model for identity verification

Georgios Petmezas[1] · **Vazgken Vanian**[1] · **Konstantinos Konstantoudakis**[1] · **Elena E. I. Almaloglou**[1] · **Dimitris Zarpalas**[1]

Received: 21 June 2024 / Revised: 14 October 2024 / Accepted: 19 December 2024

Abstract
The proliferation of deepfake technology poses significant challenges due to its potential for misuse in creating highly convincing manipulated videos. Deep learning (DL) techniques have emerged as powerful tools for analyzing and identifying subtle inconsistencies that distinguish genuine content from deepfakes. This paper introduces a novel approach for video deepfake detection that integrates 3D Morphable Models (3DMMs) with a hybrid CNN-LSTM-Transformer model, aimed at enhancing detection accuracy and efficiency. Our model leverages 3DMMs for detailed facial feature extraction, a CNN for fine-grained spatial analysis, an LSTM for short-term temporal dynamics, and a Transformer for capturing long-term dependencies in sequential data. This architecture effectively addresses critical challenges in current detection systems by handling both local and global temporal information. The proposed model employs an identity verification approach, comparing test videos with reference videos containing genuine footage of the individuals. Trained and validated on the VoxCeleb2 dataset, with further testing on three additional datasets, our model demonstrates superior performance to existing state-of-the-art methods, maintaining robustness across different video qualities, compression levels and manipulation types. Additionally, it operates efficiently in time-sensitive scenarios, significantly outperforming existing methods in inference speed. By relying solely on pristine, unmanipulated data for training, our approach enhances adaptability to new and sophisticated manipulations, setting a new benchmark for video deepfake detection technologies. This study not only advances the framework for detecting deepfakes but also underscores its potential for practical deployment in areas critical for digital forensics and media integrity.

Keywords Video deepfake detection · 3D Morphable Models (3DMMs) · Transformer networks · Video forensics · Biometric authentication · Identity verification

✉ Georgios Petmezas
petmezgs@iti.gr

[1] Centre for Research and Technology Hellas, 57001 Thessaloniki, Greece

1 Introduction

In recent years, advancements in artificial intelligence (AI) and machine learning (ML) have led to the proliferation of deepfake technology, which has the potential to both amaze and concern the world [1]. Video deepfakes are highly realistic manipulated videos generated using advanced AI algorithms [2]. These digital forgeries can convincingly replace a person's likeness with another, create false scenarios, and even make individuals appear to say or do things they never did. As video deepfakes become increasingly sophisticated and accessible, the risk of their malicious use grows significantly, posing severe threats to privacy, security and trust in media content [3]. The implications are vast and varied, touching on legal, ethical and social aspects. For instance, the potential use of video deepfakes in disinformation campaigns can undermine democratic processes, influence elections and incite social unrest. Similarly, in the personal domain, unauthorized video deepfakes can lead to severe violations of privacy and harm to individual reputations.

Detecting video deepfakes has emerged as a critical research area to safeguard against misinformation and manipulation. While early video deepfakes were relatively easy to spot due to their obvious flaws, such as unnatural facial movements or inconsistent lighting, the rapid progress of DL, a subset of ML, has made it increasingly challenging to distinguish deepfake videos from pristine ones with the naked eye alone [4]. As a result, researchers have had to turn to advanced AI-based approaches to develop robust and accurate video deepfake detection systems [5, 6], including convolutional neural networks (CNNs) and recurrent neural networks (RNNs), among others [7]. These techniques excel in processing and analyzing complex visual and temporal data, thereby identifying subtle discrepancies that indicate manipulation.

Although numerous methods for detecting video deepfakes have been proposed, many of them often struggle with problems such as overfitting, high computational demands and a lack of generalizability across various datasets and deepfake creation techniques [8]. Additionally, the typical training approach for these systems involves using both authentic and manipulated data [9]. This requirement poses further challenges, particularly due to the imbalance in available datasets, where genuine videos far outnumber the deepfakes. This disparity can lead to biased outcomes in detection capabilities, favoring the detection of real content over fake.

Given these challenges, the present study proposes a novel approach to video deepfake detection that leverages the capabilities of CNNs, LSTMs and Transformers, aiming to significantly enhance detection accuracy and efficiency in identifying manipulated content. The proposed methodology employs 3DMMs to extract detailed facial biometrics from the input videos. This biometric data serves as a basis for our hybrid CNN-LSTM-Transformer model, which is specifically trained to extract features from a target identity and learn unique characteristics that distinctly differentiate it from any other identity or deepfake. These learned features serve as a benchmark against which characteristics computed from potentially manipulated test videos are compared. This enables the detection of discrepancies that signify manipulation, enhancing the model's ability to accurately distinguish authentic content from forgeries. Remarkably, the proposed model is trained exclusively with pristine data, ensuring it focuses on authentic facial features and behaviors. This approach makes the model agnostic to specific manipulation techniques, thereby improving its ability to accurately distinguish authentic content from forgeries, regardless of the techniques used to generate them.

Overall, the key contributions of this work are:

- It employs 3D Morphable Models (3DMMs) for detailed facial biometrics extraction and a hybrid CNN-LSTM-Transformer model that combines fine-grained spatial analysis with short- and long-term temporal dynamics, thereby enhancing the detection of subtle manipulations in video content.
- The proposed model is designed to process large volumes of video data efficiently, enabling optimized deepfake detection capabilities.
- The model demonstrates strong performance across different levels of video compression, quality settings and types of manipulation, ensuring its utility in diverse real-world applications.
- By training exclusively with unmanipulated data, the proposed model avoids the common pitfalls of overfitting to specific artifacts, and remains adaptable to new and sophisticated deepfake techniques.

The remainder of this paper is structured as follows: In the 2 section, background knowledge for video deepfake detection is presented and related state-of-the-art DL methods are reviewed. In the 3 section, the research methodology is presented in detail, while in the 4 section, the performance of the proposed model is evaluated and discussed. Finally, the conclusions of the study are summarized in the 5 section.

2 Related work

The task of video deepfake detection usually involves leveraging various computer vision and ML methods to analyze and classify images or videos as either genuine or manipulated [10]. During the last few years, several researchers have leveraged massive datasets of pristine and deepfake images or videos to train DL or ML models capable of identifying subtle artefacts and patterns indicative of manipulation. In particular, Li and Lyu [11] developed a CNN-based approach that detects AI-generated deepfake videos by capturing face-warping artefacts, while Güera and Delp [12] proposed a temporal-aware system that is composed of a CNN for frame feature extraction and a long short-term memory (LSTM) network for temporal sequence analysis to detect whether a video has been subject to manipulation or not. Afchar et al. [13] suggested two DL approaches, a CNN and a residual network (ResNet), that focus on the mesoscopic properties of images to detect face tampering in videos.

At the same time, Agarwal et al. [14] introduced a biometric-based forensic method that combines static facial recognition and temporal behavioral traits, which are learned through a CNN, for the detection of face-swap deepfakes, while a study by Wang and Dantcheva [15] compared three different CNN-based models, namely 3D ResNet, 3D ResNeXt and I3D, for detecting video manipulations, such as face-swap, facial reenactment [16], and neural textures [17]. Another work by Wodajo and Atnafu [18] utilized a Convolutional Vision Transformer (CvT) to identify videos with evidence of manipulation, where the CNN extracts learnable features, and the Vision Transformer (ViT) [19] receives the learned features as input and classify them using an attention mechanism.

Moreover, Mo et al. [20] employed a CNN to separate fake from real images using two different image datasets for the experiments, one including both the original face and background, and one containing cropped images that represent only the facial region of the person. A study by de Rezende et al. [21] applied a ResNet-50 on raw images to detect computer-generated images. Şengür et al. [22] combined a pre-trained CNN with a support

vector machine (SVM) to detect face liveness using raw images as input to the hybrid model, while another study by Hsu et al. [23] trained a DenseNet using both real and GAN-generated fake images.

On the other hand, Dong et al. [24] proposed the Identity Consistency Transformer, which receives images as input and learns a pair of identity vectors, one for the inner face and one for the outer face, in order to detect identity inconsistency in inner and outer facial regions. The model is trained both on real facial images and images that resulted from swapping the inner face of two real faces belonging to different identities. Also, Giudice et al. [25] employed the discrete cosine transform (DCT) in order to detect anomalies in GAN-generated deepfakes. Moreover, Kosarkar et al. [26] used a simple CNN structure to separate real and manipulated videos on a frame-by-frame basis, whereas Wodajo et al. [27] trained a generative CvT on several well-known benchmark deepfake datasets including the Deepfake Detection Challenge (DFDC) dataset [28] and the FaceForensics++(FF++) [29] dataset, among others, to classify real and fake videos.

All of the approaches presented above are of great interest; however, they all use raw video frames as input for the proposed models, which may limit the ability of the models to exploit hidden information that is present in the frame. For this reason, several researchers have tried to exploit additional features that can be derived from a manipulated video by applying some feature extraction before feeding the model. More specifically, D'Avino et al. [30] presented a hybrid approach that combines an autoencoder (AE) and an LSTM network for video forgery detection, utilizing the network's ability to learn temporal dependencies and detect manipulated video frames. The authors computed the raw image residuals via high-pass filtering, then quantized them, and finally used them to extract a histogram of co-occurrences which was fed as input to the DL model. On the other hand, a study by Amerini et al. [31] proposed the use of optical flow fields as input to two semi-trained models, namely VGG16 and ResNet-50, to detect deepfake videos by exploiting possible inter-frame dissimilarities.

Furthermore, Yang et al. [32] proposed an SVM for image deepfake detection that receives as input a feature vector containing the difference in estimated head poses. This difference is derived by comparing head poses estimated using all facial landmarks and head poses estimated using only the central facial region. Agarwal et al. [33] suggested using the frequency spectrum of raw images as input to a hybrid model that consists of a CNN and a capsule layer in order to detect GAN-generated fake images, while Frank et al. [34] trained a CNN on discrete cosine transform (DCT) frequency spectra of raw images for the same purpose. Another study by Tan et al. [35] converted facial videos into graphs to capture both facial structural information and facial action dependencies, and trained a graph convolutional network on them to classify videos as real or fake.

Nonetheless, all the above studies, both the ones that use raw frames as input and those that perform some feature extraction before feeding the DL/ML models, have a common feature that they follow a global model training approach; this means that the models are trained using a mixture of real and fake images from multiple individuals in order to learn how to answer to a specific binary question; that is, "is the given image/video real or not?", no matter the depicted individual. This method, although time efficient, limits the model's generalizability, and can lead to overfitting and, thus, low performance in predicting new unseen data.

On the contrary, a more personalized approach would require more time for the training stage, since the model has to be fine-tuned for each individual input; however, in this way, a more accurate and targeted prediction could be achieved. Such a study is the one by Cozzolino et al. [36], who introduced ID-Reveal, a novel identity-aware video deepfake

detection method that learns temporal facial features using adversarial training. The authors propose a novel scheme in which the DL model is trained only on real videos containing many different subjects and investigates whether the face under test preserves all the biometric traits of the involved subject (both structural and motional), instead of answering the binary question "real or fake" as is usually done.

Nevertheless, in order to increase the predictive capacity of their model and ensure that its predictions are based also on behavioral instead of just visual information, a generative adversarial network (GAN) is applied as follows: the generator creates manipulated videos by combining the appearance of the involved subject with the expressions of a second one, while the discriminator, which is the already trained DL model, tries to predict whether the input video is pristine or fake. The above process is repeated every time a different individual needs to be identified, which increases its predictive ability on the targeted dataset and provides generalization to different manipulation methods.

Also, another study by Cozzolino et al. [37] trained both an audio and a video DL network using segments of real talking-face videos in order that the two models learn how to extract features that are close to ones extracted from segments of the same identity, while maintaining a considerable distance from segments associated with different identities. To achieve this, the authors proposed four different similarity indices, namely an audio, a video, an audio–video, and a fusion similarity index, which combines the previous three.

Both the last two approaches transform the "real or fake" classification problem into a "is this really the person of interest?" question, which creates the conditions for training models with a higher generalizability, since they are focused on extracting distinctive features of the depicted subject and not on identifying specific forgery techniques, which may be the case for the rest of the aforementioned studies. This is also supported by the fact that they use only pristine videos for training their models. However, these methods could be said to have a potentially important drawback, which is that they presuppose the existence of a set of reference videos in addition to the test one in order to make the required comparison.

3 Materials and methods

3.1 Datasets

3.1.1 Training

The primary dataset used in this study is VoxCeleb2 [38], a large-scale audio-visual dataset derived from videos uploaded to YouTube. It features over 1 million utterances by 6,112 unique identities across 150,480 videos. VoxCeleb2's diversity is a critical asset; it includes a near gender-balanced selection of speakers (61% male) from a broad array of ethnicities, accents, professions and ages, which supports the development of robust deepfake detection models that are effective across diverse demographic groups.

This dataset is noted for its challenging recording conditions, encompassing a wide range of visual and auditory environments, including interviews from red carpets, speeches in large stadiums and other noisy outdoor settings, as well as quiet indoor studio talks, professionally shot multimedia and low-quality handheld video recordings. These environments introduce natural variances in background noise, such as chatter, laughter, overlapping speech and room acoustics, which are critical for training models to recognize

authentic audio cues in real-world scenarios. The dataset also provides comprehensive visual data, with video segments captured "in the wild", characterized by variations in pose, lighting, image quality and motion blur.

By training exclusively on this pristine dataset, our model learns to accurately identify genuine biometric and behavioral patterns. This foundational knowledge is crucial when the models are subsequently tasked with detecting deviations from these established norms in manipulated media not present during training. Therefore, while VoxCeleb2 does not contain manipulated videos, the depth and realism of its data are invaluable for preparing models to handle a variety of deepfake scenarios, enhancing their predictive accuracy and generalizability in real-world applications.

It is also worth mentioning that, although VoxCeleb2 contains both visual and audio information, this study focuses solely on the visual components. Audio data is not consistently available in video deepfakes, making reliance on visual cues essential. This strategic decision ensures that our detection methodology remains applicable across all deepfake scenarios, including those where audio is absent or remains unaltered. Leveraging the "in the wild" nature of the visual data, we aim to develop a detection system that is both effective and reliable, enhancing the robustness of our model even in the absence of audio.

3.1.2 Testing

To rigorously evaluate the performance and generalizability of the model trained and validated on the VoxCeleb2 dataset, three additional datasets specifically designed for testing deepfake detection capabilities are utilized. These datasets are chosen to complement the training data by presenting new challenges and scenarios that help assess the robustness of the proposed model under diverse conditions.

The first of these, the DeepFakeDetection (DFD) [39] dataset from Google AI lab, comprises 363 original sequences featuring 28 paid actors across 16 different scenes, along with 3,068 manipulated videos produced using a face-swapping algorithm known as DeepFakes. The dataset provides two distinct video quality settings: high quality (HQ), where videos are compressed using a constant rate quantization parameter of 23 with H.264 encoding, and low quality (LQ), where a higher quantization parameter of 40 is applied. These settings allow for the evaluation of the performance of the proposed model under varying degrees of compression and potential video degradation.

The second testing dataset, Celeb-DF [40], includes 590 original videos sourced from publicly available YouTube clips featuring 59 celebrities from diverse genders, ages and ethnic groups. Alongside these are 5,639 DeepFake videos, totaling more than 2 million frames. These DeepFake videos are generated using an improved synthesis method that significantly enhances the overall visual quality compared to earlier datasets. The high quality and realism of the manipulations in the Celeb-DF dataset offer a solid platform for validating the effectiveness of our detection system under conditions that closely mimic real-world scenarios. The diverse demographic characteristics of the subjects in Celeb-DF further ensure that our model is tested across a broad spectrum of populations, enhancing its generalizability and robustness.

The third dataset used for testing is FF++[29], a large-scale dataset specifically designed for evaluating facial image manipulation detection. FF++comprises 1,000 pristine videos sourced from YouTube, featuring over 500,000 images. The dataset also includes 8,000 manipulated videos generated using four state-of-the-art face manipulation methods: Face2Face, FaceSwap, DeepFakes and NeuralTextures. These methods

encompass both face-swapping and facial reenactment techniques, enabling a comprehensive evaluation of the model's performance across the most prevalent types of video manipulations. To reflect real-world conditions, the manipulated videos are provided in two quality levels – HQ and LQ – achieved through H.264 compression with quantization parameters of 23 and 40, respectively. The diversity in manipulation techniques and compression levels in FF + + allows for a thorough assessment of the model's ability to detect facial forgeries under varying conditions, further testing its generalizability and robustness.

3.2 Preprocessing

The preprocessing of the raw video data into a format suitable for DL analysis is a critical step in the proposed methodology. This process involves several key stages, each designed to systematically transform raw videos into sequences of 3DMM vectors, which serve as input for the DL network. The first step in our preprocessing pipeline is frame extraction, where each video is systematically segmented into individual frames. This involves parsing the video file to extract still images at a consistent frame rate, ensuring that each moment captured in the video is available for further analysis. Frame extraction is crucial as it isolates each visual instance for detailed examination in subsequent steps. Following frame extraction, the RetinaFace [41] face detector is employed. Renowned for its accuracy and efficiency, particularly in detecting faces with high precision under various conditions, RetinaFace identifies and locates facial regions within the frames. Once detected, the faces are cropped and then undergo alignment to ensure that key facial features such as eyes, nose and mouth are positioned consistently across all images. This alignment step adjusts the orientation and scale of the faces, standardizing the input and improving the consistency of subsequent processing steps.

Subsequently, each isolated facial image undergoes a process to generate a 3DMM representation. 3DMMs [42] are generative models that utilize a set of low-dimensional parameters to create statistically meaningful representations of the input face. They employ advanced statistical techniques, including principal component analysis (PCA), to distill the inherent shape and texture variations of the face. These models have a profound capability to understand and represent the intricate variations in facial features, making them highly effective for transforming a 2D image into a detailed 3D representation. A neutral face without expressions serves as the basis for this transformation, with any new face being represented as a linear combination of shape and texture components. Specifically, the principal components obtained through PCA form the core basis of the 3DMM. These components allow for the synthesis of new instances by adjusting the coefficients/weights associated with each component, with each coefficient influencing specific facial variations such as the width of the nose or the intensity of skin pigmentation.

Given a new input – such as a 2D image of a face – the 3DMM is employed to reconstruct the underlying 3D shape and texture. This is achieved by iteratively adjusting the weights of the principal components to best match the input image, thereby minimizing the discrepancy between the reconstructed model and the original data. Recent advancements in the extraction of 3DMMs have seen the development of sophisticated techniques such as FLAME [43], RingNet [44] and DECA [45], which further enhance the accuracy and fidelity of the models. In the present study, a fast and lightweight 3D dense face alignment approach, named 3DDFA_V2 [46], is implemented. This framework predicts a vector of 62 coefficients for each input frame, comprising 40 parameters for the shape, 10 for the expression, and 12 for the rigid pose of the depicted face. Finally, these vectors are stacked

to form a sequential array of 3DMM vectors for all frames of the video, which serves as the input for subsequent DL analysis. This method enhances our model's ability to capture and analyze facial dynamics accurately throughout the video sequence.

3.3 Proposed approach

The proposed approach to video deepfake detection integrates a hybrid CNN-LSTM-Transformer model with 3D Morphable Models (3DMMs) to enhance both accuracy and efficiency in identifying manipulated content. This methodology is inspired by the techniques outlined in the ID-Reveal [36] study, which has been previously discussed in the *Related Works* section. The ID-Reveal methodology provides a robust foundation for understanding and addressing the challenges associated with video deepfake detection, particularly in terms of leveraging biometric data and temporal features for more accurate identification of manipulated videos. Also, by employing an identity verification strategy, where test videos are compared with a set of reference videos containing genuine footage of the individuals, authentic content can be discerned from manipulated content with high precision.

Building on this foundation, the present study has developed a more comprehensive detection system by integrating CNNs, LSTMs and Transformer networks, each contributing uniquely to the model's effectiveness. CNNs [47] are highly effective at capturing spatial hierarchies and detailed features from input data, making them ideal for extracting fine-grained information from 3DMM vectors. On the other hand, LSTM networks [48] excel in modeling temporal dynamics, enabling the capture of short-term dependencies and variations across consecutive frames. This ability to understand temporal changes over time is essential for accurately assessing the evolution of features in video sequences. The integration of Transformer networks further enhances the model by addressing long-range dependencies and capturing complex sequential patterns. Transformers [49] represent a significant breakthrough in neural network architecture, perfectly suited for applications that require the analysis of complex sequential data [50]. Unlike traditional models that relied on recurrent or convolutional layers, Transformers leverage a fully attention-driven approach [51]. This architecture eliminates the need for recurrence, facilitating substantial improvements in parallel processing capabilities and, thus, enabling the efficient handling of large volumes of video data [52]. Key features of this architecture include multi-headed attention mechanisms, positional embeddings, residual connections, layer normalization and feedforward networks, all of which work in concert to capture global relationships between input and output elements without the sequential processing constraints of earlier architectures. These advancements allow for enhanced processing of sequential tasks like video analysis, where each frame can influence the interpretation of every other frame in the sequence, thereby improving the detection and identification of subtle inconsistencies indicative of deepfake manipulation. The proposed CNN-LSTM-Transformer model effectively combines local detail extraction, short-term temporal modeling, and long-term dependency analysis, providing a powerful framework for video deepfake detection. The architecture capitalizes on the strengths of each component: the CNN for spatial detail, the LSTM for short-term temporal consistency, and the Transformer for long-term temporal relationships.

The detailed model architecture is also presented in Fig. 1. Similar to approaches from other fields that deploy Transformer-based models, such as ViT, a tokenization process is employed to convert input data into sequences. This is achieved by the CNN component of the network, which includes a 1D convolutional layer with a kernel size

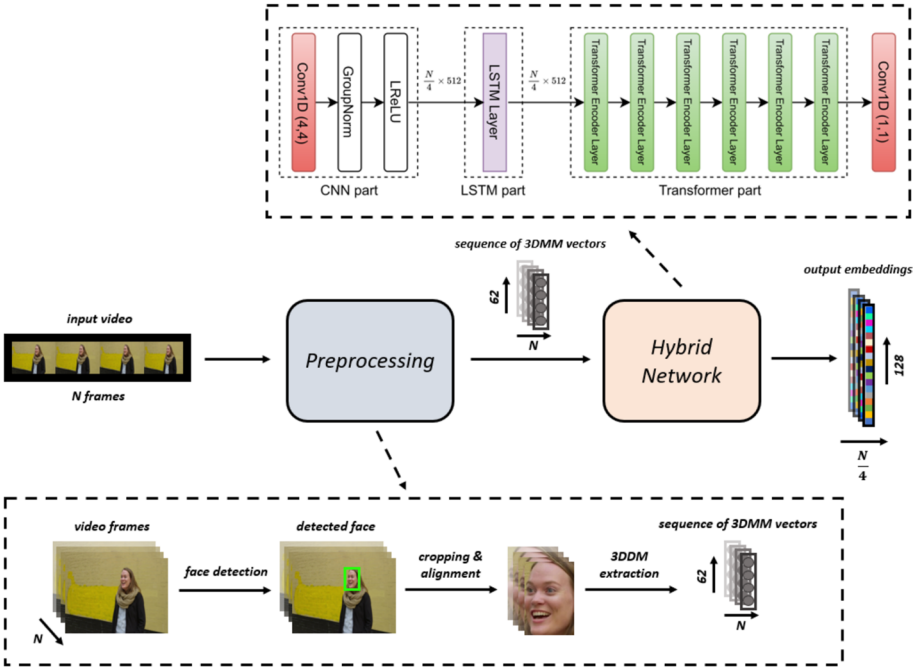

Fig. 1 The proposed approach: Raw video data (N frames) undergoes preprocessing through frame extraction, face detection, facial alignment and 3DMM vector generation. The resulting sequences are first processed by a CNN, which extracts detailed spatial features from the 3DMM vectors that act as tokens. These features are then passed to an LSTM module, which captures short-term temporal dynamics across consecutive frames. The output from the LSTM is subsequently fed into a Transformer network, where six Transformer encoder layers capture long-range dependencies within the video sequence. Finally, the encoded features are projected into a 128-dimensional space for each token, using a linear transformation, to produce the final output embedding

of 4 and a stride factor of 4, followed by GroupNormalization and LeakyReLU activation. The sequence of tokens is then passed to an LSTM module, which captures short-term temporal dynamics across consecutive frames, and the resultant output is then fed into a series of Transformer encoder layers. Each Transformer encoder layer consists of a multi-headed attention and a feedforward network, with each component followed by a LayerNormalization layer, as detailed in [49]. Finally, for each token within the sequence, the resulting features are linearly projected into a 128-dimensional vector, which represents the final output embedding, using a 1D convolutional layer with a kernel size of 1 and a stride factor of 1.

For model training, the official training set of VoxCeleb2 dataset consisting of 5,994 identities is used, while model validation is performed on the remaining 118 identities. During this process, a set of 8 random identities is selected and for each identity a total of 8 videos are chosen, resulting in a batch comprising 64 videos. From each video, a segment of 96 consecutive frames is sampled to ensure the continuity of the frames and maintain their sequential order. Similar to [36], a contrastive learning approach is deployed and each video of the batch the following loss function is calculated:

$$L_{\text{net}} = \frac{1}{N}\sum_{i,t}\left(-\log\left(\frac{\sum_{j\neq i} e^{sim_{+}(i,t,j)}}{\sum_{j\neq i} e^{sim_{+}(i,t,j)} + \sum_{j\neq i} e^{sim\_(i,t,j)}}\right)\right) \tag{1}$$

where N is the total number of segments, $sim+$ is the similarity of positive pairs – pairs of segments depicting the same identity –, and $sim-$ is the similarity of negative pairs – pairs of segments depicting different identities – given a pivot segment i within the current batch. The similarity function is calculated using the squared Euclidean distance as follows:

$$\text{sim}(i,t,j) = -\min_{t'}\|y_{i,t} - y_{j,t'}\|^2 \tag{2}$$

where i and j represent two segments, t and t' are the individual frames of the two segments, respectively, and y is the output embedding representing the encoded information extracted from the segments. This objective function encourages the network to learn similar representations for segments of the same identity and dissimilar representations for segments of different identities. The network is optimized using the Adam optimizer, configured with an initial learning rate of 10^{-5}, which is gradually reduced by a cosine learning rate scheduler [53], and a weight decay of 10^{-4}.

During model testing, to determine the authenticity of a video representing a specific identity, the following steps are taken: Initially, a set of pristine videos depicting the identity is required. These videos undergo processing by the proposed network, resulting in a sequence of embeddings. Subsequently, the video in question is inputted into the network, generating its own series of embeddings. By measuring the Euclidean distance between these embeddings and the reference embeddings, one can ascertain the legitimacy of the test video.

Specific enhancements have been implemented to refine the ID-Reveal principles in the present study, focusing on optimizing model performance and ensuring scalability across diverse datasets and real-world scenarios. These enhancements include more effective training procedures and the integration of an efficient CNN-LSTM-Transformer model that can better capture and analyze the temporal dynamics of facial expressions in videos. Another important remark is that the proposed approach diverges significantly from ID-Reveal by eschewing the use of GANs to generate manipulated content during training. Instead, we train our model exclusively on pristine, unmanipulated data. This strategic decision prevents the model from learning to detect only the particular characteristics of GAN-manipulated faces, which might not be representative of all possible manipulations. In this way, the generalization ability of the proposed model is enhanced, enabling it to more effectively identify and adapt to new, previously unseen types of manipulations. By not confining the training to the characteristics of GAN-generated deepfakes, our model is better equipped to handle a broader spectrum of deepfake technologies and techniques that might emerge, which is crucial for maintaining high detection accuracy in a landscape where manipulation methods are continually evolving and becoming more sophisticated.

4 Results and discussion

The development and testing of the DL model for video deepfake detection were conducted within a Python 3.9 environment, utilizing the PyTorch library (version 2.0.1) as the primary tool for DL model implementation. All computational experiments were performed on a workstation equipped with a NVIDIA GeForce RTX 3090 GPU, supported by CUDA

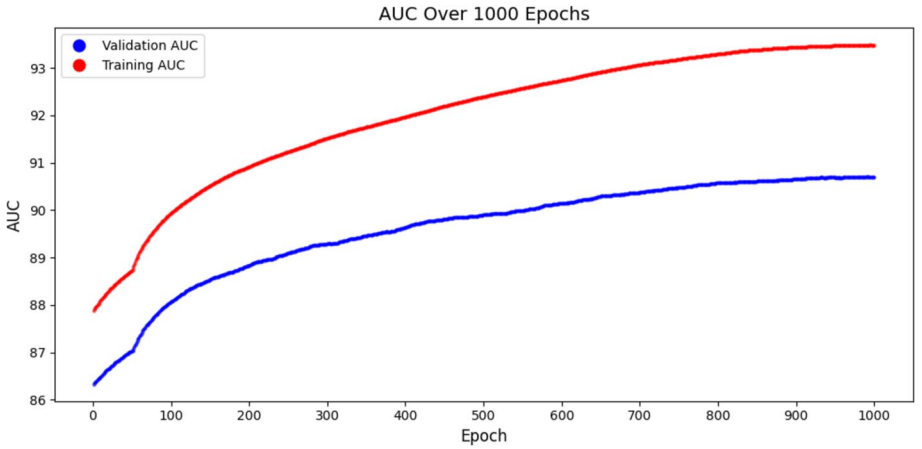

Fig. 2 Model performance (AUC) on the training and validation sets across different training epochs

Table 1 Model performance (AUC) on the validation set relative to the number of Transformer encoder layers

Layers	1	2	4	6	8
AUC	88.58%	89.99%	90.64%	90.82%	90.88%

version 11.8, which provided the necessary computational power to handle intensive training and testing processes. The choice of PyTorch was motivated by its flexibility and efficiency in handling tensor computations and its dynamic computational graph that facilitates rapid testing and iteration of different model architectures.

The proposed model underwent extensive training to identify the optimal balance between performance and computational efficiency. Training and validation were systematically conducted using separate subsets of the VoxCeleb2 dataset, with training occurring on the designated training set and evaluations on the validation set. This approach helps to prevent overfitting and ensures that the model's generalizability is accurately assessed. Initial tests aimed at ascertaining the point of diminishing returns for performance gains covered a range from 100 to 1,000 epochs. Figure 2 presents the area under the curve (AUC) measurements during training both on the training and validation sets at various epoch intervals.

The data in Fig. 2 shows that the model's performance generally improved as the number of training epochs increased, with a notable plateau in improvement beyond 500 epochs. The highest AUC reached was 90.82% at 1,000 epochs, after which we observed that further increases in epochs did not result in substantial gains in performance. This finding indicated a convergence point at 1,000 epochs, leading to the decision to standardize training at this duration for the final model. This decision was informed by the diminishing returns on further training and the need to balance computational resources with practical deployment considerations. Following the assessment of training duration, we conducted experiments to evaluate how variations in the number of Transformer encoder layers affect the model's performance, in particular measured by the AUC on the validation set. These findings are summarized in Table 1.

The results show a noticeable improvement in AUC as the number of layers increases from 1 to 6, suggesting that deeper models capture more complex patterns essential for accurate video deepfake detection. However, beyond six layers, the increases in AUC become marginal. Specifically, the AUC improvement from 6 to 8 layers was only 0.06%, a slight increment considering the additional complexity and computational overhead associated with larger models. Based on these results, we decided to adopt the 6-layer Transformer module for our system. This decision balances performance with computational efficiency, ensuring substantial detection capabilities without undue resource demands, suitable for practical deployment scenarios where both accuracy and processing speed are critical. Following the assessment of the number of Transformer encoder layers, we conducted an ablation study to investigate the impact of varying the number of LSTM layers on the model's performance, particularly measured by the AUC on the validation set. The results of this study are summarized in Table 2.

The results in Table 2 indicate that the model's performance, measured by the AUC, is highest with 1 LSTM layer (90.82%), with a very slight decrease when using 2 LSTM layers (90.80%). However, further increasing the number of LSTM layers to 4 or 6 leads to a noticeable decline in performance, with the AUC dropping to 90.20% and 87.53%, respectively. This suggests that while a minimal number of LSTM layers can effectively capture the temporal dependencies needed for accurate deepfake detection, adding more layers introduces unnecessary complexity, leading to overfitting and a subsequent reduction in performance. Based on these findings, we opted to use a single LSTM layer in our final model configuration. This choice ensures optimal performance while maintaining computational efficiency, aligning with our goal of developing a robust and practical deepfake detection method suitable for real-world applications.

After validating the proposed model's performance on the VoxCeleb2 validation set, further evaluations were conducted on the additional external testing datasets, DFD, Celeb-DF and FF++, to assess its effectiveness in more diverse and challenging environments. These datasets were tested under both HQ and LQ conditions to better simulate the range of real-world scenarios the model might encounter. In addition to assessing the model's standalone performance, we conducted a comparative analysis to further validate its effectiveness by comparing its results against other well-regarded models in the field. This comparison involves a variety of detection approaches that were also reviewed in the ID-Reveal study, encompassing frame-based methods like MesoNet [54], Xception [55], EfficientB7 [56] and FFD [57], ensemble methods such as ISPL [58] and the solution of the DFDC winner, Selim Seferbekov, temporal-based methods including Eff.B1+LSTM and ResNet+LSTM [12], as well as identity-based approaches like A&B [14], ID-Reveal and our own model. For detailed descriptions of these methods, the reader is referred to the supplemental document provided by the ID-Reveal study [36]. Additionally, for a more comprehensive comparison, we also evaluated two other models: a CNN-LSTM and a CNN-Transformer model.

To ensure a fair and comprehensive evaluation, all supervised models – spanning frame-based, ensemble and temporal approaches – were trained on specific datasets tailored to the testing scenarios, as described in [36]. For the DFD dataset, these models were trained on the DFDC dataset, which comprises approximately 100,000 fake and

Table 2 Model performance (AUC) on the validation set relative to the number of LSTM layers

Layers	1	2	4	6
AUC	90.82%	90.80%	90.20%	87.53%

20,000 real videos. For the Celeb-DF dataset, the training utilized the FF++ [29] dataset. This approach ensures that each model is challenged with data that mirror the conditions and manipulations they are expected to detect.

Conversely, identity-based methods, including A&B, ID-Reveal and our proposal, were consistently trained on the VoxCeleb2 dataset for both testing scenarios. This consistent training strategy helps to focus on biometric and behavioral consistencies rather than specific forgery signatures, enhancing their capability to generalize across different forms of manipulations. For testing on the FF++ dataset, it is important to note that the dataset does not provide multiple videos of the same subject. Therefore, for identity-based approaches, we adopted a similar approach to the ID-Reveal study by using only videos of at least 14 s in duration. The first 6 s of each pristine video were used as the reference dataset, while the last 6 s were used for performance evaluation. Testing was conducted on the DFD, Celeb-DF and FF++ datasets under both HQ and LQ video conditions, providing a comprehensive assessment of model performance across various data sources and manipulation techniques.

As depicted in Table 3, the proposed CNN-LSTM-Transformer model achieves superior performance outperforming all other models listed in almost all cases. Notably, our model demonstrates an AUC of 97% for both HQ and LQ conditions in the DFD dataset, and an AUC of 86% and 83% for HQ and LQ conditions in the Celeb-DF dataset, respectively. The FF++ dataset further underscores the superior performance of our model, with AUCs of 98% and 97% for HQ and LQ conditions in face swapping videos, and 99% and 98% in facial reenactment videos. These results highlight a dramatic improvement over existing approaches, whose AUCs are significantly lower. These results confirm the robustness and reliability of the CNN-LSTM-Transformer model against various forms of video manipulation, standing out particularly in environments where video quality may be degraded – a common challenge in real-world applications.

Table 3 Performance comparison (AUC) between the proposed model and other relevant studies

Dataset / Method	DFD		Celeb-DF		FF++			
					FS		FR	
	HQ	LQ	HQ	LQ	HQ	LQ	HQ	LQ
MesoNet	57%	53%	75%	67%	61%	62%	58%	57%
Xception	93%	63%	**88%**	58%	89%	79%	58%	57%
Efficient-B7	97%	64%	80%	56%	93%	80%	59%	54%
FFD	83%	69%	76%	59%	75%	70%	56%	56%
ISPL	93%	72%	83%	61%	83%	76%	59%	55%
Seferbekov	**98%**	67%	86%	62%	97%	87%	62%	55%
ResNet+LSTM	65%	64%	72%	60%	63%	66%	58%	58%
Eff.B1+LSTM	95%	76%	84%	58%	90%	78%	62%	58%
A&B	77%	61%	56%	55%	97%	65%	79%	53%
ID-Reveal	96%	94%	84%	80%	**99%**	**97%**	87%	83%
CNN-LSTM (Ours)	93%	93%	81%	78%	97%	96%	**99%**	**98%**
CNN-Transformer (Ours)	97%	**97%**	86%	82%	97%	**97%**	**99%**	97%
CNN-LSTM-Transformer (Ours)	97%	**97%**	86%	**83%**	98%	**97%**	**99%**	98%

The superior performance of the proposed model is significantly attributed to its training on an exclusively pristine dataset. This approach prevents the model from overfitting to specific artifacts of deepfake generation methods, which are often present in datasets containing manipulated media. By understanding the subtleties of genuine human expressions and interactions captured in unaltered videos, the model forms a more accurate baseline for identifying discrepancies. Consequently, this training strategy enhances the model's generalizability across unknown manipulations, where typical deepfake detection models might fail.

Furthermore, the integration of the CNN-LSTM-Transformer architecture significantly advances our approach beyond the CNN-based methodologies used in models like ID-Reveal, which employs a ResNet format. In our model, the CNN component initially extracts detailed spatial features from the input data, capturing fine-grained local details essential for identifying manipulations. The LSTM module then processes these features to model short-term temporal dynamics across consecutive frames, addressing the evolution of features over time. Finally, the Transformer network analyzes the entire sequence globally, capturing long-range dependencies and complex patterns that may span across the entire video. This combined approach allows our model to integrate both spatial and temporal information more comprehensively. By processing sequences in this integrated manner, our model excels in identifying subtle manipulative discrepancies that may not be consistently apparent across frames, offering enhanced performance in both HQ and LQ video content.

This comparative analysis not only establishes the efficacy of the proposed model in detecting deepfake videos but also illustrates its potential to be a leading solution in the fight against digital video manipulations. The performance advantage in LQ video conditions is especially significant, highlighting the model's advanced capability to handle the kind of noisy, compressed and less-than-ideal video data that is often encountered in practical scenarios.

The efficiency of the proposed model during inference is another critical aspect of its real-world applicability, particularly when considering deployment in environments where processing speed is crucial. To demonstrate the advancements our model offers in terms of inference speed, we conducted a comparative analysis with the ID-Reveal study across various operational metrics. This analysis was performed on the same workstation equipped with a NVIDIA GeForce RTX 3090 GPU and supported by CUDA version 11.8, as previously described at the beginning of this section, ensuring fairness and consistency in our comparative evaluation.

The results of this analysis are summarized in Table 4, which shows the time taken for different subtasks during the final stages of inference on the testing datasets. It's important to note that these times reflect the model's processing after initial preprocessing steps such as frame extraction, face detection and 3DMM extraction have been completed. These tasks involve: (i) *DFD ref inf*, the time required to process the DFD videos used as the reference set, (ii) *DFD test inf*, the time to process the DFD videos used as the test set, (iii) *DFD dist_calculation_ref*, the time to compute the distances between the encoded representations of the DFD reference set of videos, (iv) *DFD dist_calculation_test*, the time to calculate the distances between the encoded representations of the DFD test set of videos, (v) *celebDF ref inf*, the processing time for the Celeb-DF reference set of videos, (vi) *celebDF test inf*, the processing time for the Celeb-DF test set of videos, (vii) *celebDF dist_calculation_ref*, the time needed to compute the distances between the encoded representations of the Celeb-DF reference set of videos, (viii) *celebDF_dist_calculation_test*, the time required to calculate the distances between the encoded representations of the

Table 4 Inference time comparison between ID-Reveal and the proposed model

Subtask	Time	
	ID-Reveal	Ours
DFD ref inf (363vids)	38 s	**4 s**
DFD test inf (3068vids)	3 min 29 s	**11 s**
DFD dist_calculation_ref	16 s	**1 s**
DFD dist_calculation_test	1 min 33 s	**9 s**
celebDF ref inf (590vids)	15 s	**2 s**
celebDF test inf (5639vids)	2 min 11 s	**18 s**
celebDF dist_calculation_ref	1 s	**0 s (< 1 s)**
celebDF_dist_calculation_test	8 s	**2 s**
FF + + FS ref inf (694vids)	14 s	**4 s**
FF + + FS test inf (687vids)	13 s	**4 s**
FF + + FS dist_calculation_ref	0 s (< 1 s)	0 s (< 1 s)
FF + + _FS dist_calculation_test	0 s (< 1 s)	0 s (< 1 s)
FF + + FR ref inf (694vids)	13 s	**4 s**
FF + + FR test inf (687vids)	12 s	**4 s**
FF + + FR dist_calculation_ref	0 s (< 1 s)	0 s (< 1 s)
FF + + _FR dist_calculation_test	0 s (< 1 s)	0 s (< 1 s)

Celeb-DF test set of videos, (ix) *FF + + FS ref inf*, the processing time for the FF + + face swapping videos used as the reference set, (x) *FF + + FS test inf*, the processing time for the FF + + face swapping videos used as the test set, (xi) *FF + + FS dist_calculation_ref*, the time to compute the distances between the encoded representations of the FF + + face swapping reference set of videos, (xii) *FF + + FS dist_calculation_test*, the time to calculate the distances between the encoded representations of the FF + + face swapping test set of videos, (xiii) *FF + + FR ref inf*, the processing time for the FF + + facial reenactment videos used as the reference set, (xiv) *FF + + FR test inf*, the processing time for the FF + + facial reenactment videos used as the test set, (xv) *FF + + FR dist_calculation_ref*, the time to compute the distances between the encoded representations of the FF + + facial reenactment reference set of videos, and (xvi) *FF + + FR dist_calculation_test*, the time to calculate the distances between the encoded representations of the FF + + facial reenactment test set of videos.

These results demonstrate a significant reduction in inference time across all metrics when compared to the ID-Reveal model. Notably, the proposed model reduces the inference time dramatically for all three testing datasets, with the largest time savings observed in the inference for the test videos of the DFD dataset, where the proposed model completes the task in just 10 s compared to 3 min and 29 s for the ID-Reveal model. The substantial decrease in time required for distance calculations, which is particularly attributed to the advanced architecture of the CNN-LSTM-Transformer model, where each component contributes to faster processing., further highlights the efficiency of our approach, especially in real-time or near-real-time applications where rapid processing is essential. These improvements not only enhance the usability of our model in practical settings but also ensure that it can be deployed effectively in systems requiring high throughput and minimal latency.

Overall, the achievements of the present study are the following: (i) the integration of 3DMMs with CNNs, LSTMs and Transformer networks represents a significant

advancement in the field of deepfake detection offering a powerful tool for identifying subtle deepfake indicators, (ii) the proposed model addresses one of the critical barriers in deepfake detection – high computational demands – by combining efficient operation with advanced processing capabilities, (iii) the model demonstrates strong performance across different video qualities, compression levels and manipulation types., ensuring its utility in real-world applications where such conditions are variable and unpredictable, (iv) by training exclusively on unmanipulated data, the model avoids the pitfalls of overfitting to specific deepfake artifacts and remains adaptable to new and evolving deepfake techniques, enhancing its long-term applicability, and (v) the proposed model effectively manages temporal dynamics, providing robust and accurate detection across sequences, which is crucial for capturing subtle manipulations in video content.

On the other hand, the limitations of the present study could be summarized into three parts. Firstly, while our extensive testing confirms the model's effectiveness on multiple datasets, its generalizability to additional datasets or unseen manipulation techniques beyond those covered in this study remains an area for future exploration. Secondly, although the model is designed to adapt to new manipulation techniques, the rapidly evolving nature of deepfake technology may eventually require further adaptations or updates to maintain its effectiveness. Thirdly, we exclusively used Euclidean distance for identity verification due to its simplicity and effectiveness in related studies [36, 37], but alternative distance metrics could potentially offer additional insights that were not explored in this work.

To address these limitations and build on the current achievements, future work could explore the incorporation of more diverse datasets during training to enhance generalizability. In this regard, we have identified a critical gap in the availability of open-access datasets that include both manipulated and pristine videos of the same identities – datasets essential for applying identity verification approaches like ours. To overcome this limitation, we plan to create an extended deepfake dataset that supports both traditional supervised deepfake detection models and identity verification models. This dataset will be designed to include a wide range of manipulation techniques and pristine videos, facilitating the development of more robust and generalizable deepfake detection methods.

Additionally, we plan to adjust our model design to accommodate alternative distance metrics, such as cosine similarity and density-based probability, to further refine and enhance its performance. Finally, although this study focuses exclusively on video data due to the inconsistent availability of audio and the absence of textual information in the datasets used, future research will explore integrating multimodal data, such as audio and text, wherever available. Such an approach will aim to provide a more comprehensive deepfake detection framework. Continuous monitoring of emerging deepfake technologies and periodic model updates will also be essential to keep pace with the advancing manipulation techniques.

5 Conclusions

In this study, we introduced a novel and comprehensive integration of 3DMMs, CNNs, LSTMs and Transformer networks to enhance both the accuracy and efficiency of video deepfake detection. This approach leverages the strengths of 3DMMs in capturing intricate facial geometries, the spatial analysis capabilities of CNNs, the short-term temporal dynamics of LSTMs, and the long-term dependencies management of Transformers. To

the best of our knowledge, this is the first study to combine these advanced technologies for detecting manipulated videos, setting a new benchmark in the field. The proposed model excels in environments with varied video qualities, compression levels and manipulation types, maintaining high performance under challenging conditions. Its efficiency is underscored by rapid processing speeds that facilitate optimized deepfake detection capabilities – a significant improvement over existing methods, as detailed in our comparative analysis. Additionally, by training exclusively on pristine, unmanipulated data, the model avoids the common pitfalls of overfitting and remains effective against evolving, sophisticated manipulation techniques. This approach not only advances the technological framework for deepfake detection but also enhances the practical applicability of these systems in critical areas such as digital forensics and content authentication.

Acknowledgements This research has been supported by the European Commission funded program EITHOS, under Horizon Europe Grant Agreement 101073928.

Data availability The datasets used and analyzed during the present study are open-access and can be found as follows: the VoxCeleb2 dataset is available through https://www.robots.ox.ac.uk/~vgg/data/voxceleb/vox2.html, the DeepFakeDetection dataset is available through https://github.com/ondyari/FaceForensics/blob/master/dataset/README.md, the Celeb-DF dataset is available through https://github.com/yuezunli/celeb-deepfakeforensics, and the FaceForensics++dataset is available through https://github.com/ondyari/FaceForensics.

Declarations

Competing interest The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Verdoliva L (2020) Media Forensics and DeepFakes: An Overview. IEEE J Select Topics Signal Process 14:910–932
2. Yu P, Xia Z, Fei J, Lu Y (2021) A survey on deepfake video detection. IET Biometrics 10(6):607–624. https://doi.org/10.1049/bme2.12031
3. Mustak M, Salminen J, Mäntymäki M, Rahman A, Dwivedi YK (2023) Deepfakes: Deceptions, mitigations, and opportunities. J Bus Res 154:113368. https://doi.org/10.1016/j.jbusres.2022.113368
4. Nguyen TT, Nguyen QVH, Nguyen DT, Nguyen DT, Huynh-The T, Nahavandi S, Nguyên TT, Pham Q, Nguyen CM (2022) Deep learning for deepfakes creation and detection: A survey. Comput Vis Image Underst 223:103525. https://doi.org/10.1016/j.cviu.2022.103525
5. Rana MS, Nobi MN, Murali B, Sung AH (2022) Deepfake Detection: A Systematic Literature Review. IEEE Access 10:25494–25513. https://doi.org/10.1109/access.2022.3154404
6. Srivastava, A., Pandey, M. K., & Sahu, S. K. (2022). A review on deepfakes detection using machine learning techniques. Lect Notes Elect Eng, 641–651. https://doi.org/10.1007/978-981-19-5037-7_46
7. Malik A, Kuribayashi M, Abdullahi SM, Khan AN (2022) DeepFake detection for human face images and Videos: a survey. IEEE Access 10:18757–18775. https://doi.org/10.1109/access.2022.3151186

8. Naitali A, Ridouani M, Salahdine F, Kaabouch N (2023) Deepfake Attacks: generation, detection, datasets, challenges, and research directions. Computers 12(10):216. https://doi.org/10.3390/computers12100216
9. Almars AM (2021) DeepFakes Detection Techniques Using Deep Learning: A survey. J Comput Commun 09(05):20–35. https://doi.org/10.4236/jcc.2021.95003
10. Heidari A, Navimipour NJ, Dağ H, & Ünal M (2023) Deepfake detection using deep learning methods: A systematic and comprehensive review. Wiley Interdisciplinary Reviews. Data Mining and Knowledge Discovery/Wiley Interdisciplinary Reviews. Data Min Knowl Dis, 14(2). https://doi.org/10.1002/widm.1520
11. Li Y, Lyu S (2018) Exposing deepfake videos by detecting face warping artifacts. arXiv (Cornell University), pp 46–52. https://arxiv.org/pdf/1811.00656.pdf
12. Güera D, & Delp EJ (2018) Deepfake video detection using recurrent neural networks. 2018 15th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS). https://doi.org/10.1109/avss.2018.8639163
13. Afchar D, Nozick V, Yamagishi J, Echizen I (2018) MesoNet: a Compact Facial Video Forgery Detection Network. IEEE Int Workshop Inform Forensic Secur (WIFS) 2018:1–7
14. Agarwal S, Farid H, El-Gaaly T, & Lim SN (2020) Detecting deep-fake videos from appearance and behavior. 2020 IEEE Int Workshop Inform Forensic Secur (WIFS). https://doi.org/10.1109/wifs49906.2020.9360904
15. Wang Y, & Dantcheva A (2020) A video is worth more than 1000 lies. comparing 3DCNN approaches for detecting deepfakes. 2020 15th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2020). https://doi.org/10.1109/fg47880.2020.00089
16. Thies J, Zollhofer M, Stamminger M, Theobalt C, & Niessner M (2016) Face2Face: Real-time face capture and reenactment of RGB videos. 2016 IEEE Conf Comput Vis Patt Recog (CVPR). https://doi.org/10.1109/cvpr.2016.262
17. Thies J, Zollhöfer M, Nießner M (2019) Deferred neural rendering. ACM Transact Graph (TOG) 38:1–12
18. Wodajo D, Atnafu S (2021) Deepfake video detection using convolutional vision transformer. ArXiv, abs/2102.11126
19. Dosovitskiy A, Beyer L, Kolesnikov A, Weissenborn D, Zhai X, Unterthiner T, Dehghani M, Minderer M, Heigold G, Gelly S, Uszkoreit J, Houlsby N (2020) An image is worth 16x16 words: transformers for image recognition at scale. ArXiv, abs/2010.11929
20. Mo H, Chen B, & Luo W (2018) Fake faces identification via Convolutional Neural Network. Proceed 6th ACM Workshop Inform Hiding Multimedia Secur. https://doi.org/10.1145/3206004.3206009
21. de Rezende ERS, Ruppert GCS, & Carvalho T (2017) Detecting computer generated images with deep convolutional neural networks. 2017 30th SIBGRAPI Conference on Graphics, Patterns and Images (SIBGRAPI). https://doi.org/10.1109/sibgrapi.2017.16
22. Şengür A, Akhtar Z, Akbulut Y, Ekici S, & Budak U (2018) Deep feature extraction for face liveness detection. 2018 Int Conf Artif Intell Data Process (IDAP). https://doi.org/10.1109/idap.2018.8620804
23. Hsu CC, Zhuang YX, Lee CY (2020) Deep fake image detection based on pairwise learning. Appl Sci 10(1):370. https://doi.org/10.3390/app10010370
24. Dong X, Bao J, Chen D, Zhang T, Zhang W, Yu N, Chen D, Wen F, & Guo B (2022) Protecting Celebrities from DeepFake with Identity Consistency Transformer. 2022 IEEE/CVF Conf Comput Vis Patt Recog (CVPR). https://doi.org/10.1109/cvpr52688.2022.00925
25. Giudice O, Guarnera L, Battiato S (2021) Fighting deepfakes by detecting GAN DCT anomalies. J Imaging 7(8):128. https://doi.org/10.3390/jimaging7080128
26. Kosarkar U, Sarkarkar G, Gedam S (2023) Revealing and Classification of Deepfakes Video's Images using a Customize Convolution Neural Network Model. Procedia Comput Sci 218:2636–2652. https://doi.org/10.1016/j.procs.2023.01.237
27. Wodajo D, Atnafu S, & Akhtar Z (2023) Deepfake Video Detection Using Generative Convolutional Vision Transformer. ArXiv, abs/2307.07036
28. Dolhansky B, Bitton J, Pflaum B, Lu J, Howes R, Wang M, Canton-Ferrer C (2020) The deepfake detection challenge dataset. ArXiv, abs/2006.07397
29. Rössler A, Cozzolino D, Verdoliva L, Rieß C, Thies J, & Nießner M (2019) FaceForensics++: Learning to Detect Manipulated Facial Images. 2019 IEEE/CVF International Conference on Computer Vision (ICCV). https://doi.org/10.1109/iccv.2019.00009

30. D'Avino D, Cozzolino D, Poggi G, Verdoliva L (2017) Autoencoder with recurrent neural networks for video forgery detection. Media Watermarking, Security, and Forensics
31. Amerini I, Galteri, L, Caldelli R, & Del Bimbo A (2019) Deepfake video detection through optical flow based CNN. 2019 IEEE/CVF Int Conf Comput Vis Workshop (ICCVW). https://doi.org/10.1109/iccvw.2019.00152
32. Yang X, Li Y, & Lyu S (2019) Exposing deep fakes using inconsistent head poses. ICASSP 2019 - 2019 IEEE Int Conf Acoust, Speech Signal Process (ICASSP). https://doi.org/10.1109/icassp.2019.8683164
33. Agarwal, Samaksh, Girdhar N, & Raghav H (2021) A novel neural model based framework for detection of gan generated fake images. 2021 11th International Conference on Cloud Computing, Data Science & Engineering (Confluence). https://doi.org/10.1109/confluence51648.2021.9377150
34. Frank JC, Eisenhofer T, Schönherr L, Fischer A, Kolossa D, & Holz T (2020) Leveraging Frequency Analysis for Deep Fake Image Recognition. ArXiv, abs/2003.08685
35. Tan L, Wang Y, Wang J, Yang L, Chen X, Guo Y (2023) Deepfake video detection via facial action dependencies estimation. Proceed AAAI Conf Artif Intell 37(4):5276–5284. https://doi.org/10.1609/aaai.v37i4.25658
36. Cozzolino D, Rossler A, Thies J, Niesner M, & Verdoliva L (2021) Id-reveal: Identity-aware Deepfake Video detection. 2021 IEEE/CVF Int Conf Comput Vis (ICCV). https://doi.org/10.1109/iccv48922.2021.01483
37. Cozzolino D, Pianese A, Nießner M, & Verdoliva L (2023) Audio-visual person-of-interest deepfake detection. 2023 IEEE/CVF Conf Comput Vis Patt Recog Workshops (CVPRW). https://doi.org/10.1109/cvprw59228.2023.00101
38. Chung JS, Nagrani A, & Zisserman A (2018) VoxCeleb2: Deep Speaker Recognition. Proc Interspeech 2018. https://doi.org/10.21437/interspeech.2018-1929
39. Dufour N, Gully A, Karlsson P, Vorbyov AV, Leung T, Childs J, Bregler C (2019) Deepfakes detection dataset
40. Li Y, Yang X, Sun P, Qi H, & Lyu S (2020) Celeb-DF: A Large-Scale Challenging Dataset for DeepFake Forensics. 2020 IEEE/CVF Conf Comput Vis Patt Recog (CVPR). https://doi.org/10.1109/cvpr42600.2020.00327
41. Deng J, Guo J, Ververas E, Kotsia I, & Zafeiriou S (2020) RetinaFace: Single-Shot Multi-Level Face Localisation in the Wild. 2020 IEEE/CVF Conf Comput Vis Patt Recog (CVPR). https://doi.org/10.1109/cvpr42600.2020.00525
42. Blanz V, & Vetter T (1999) A morphable model for the synthesis of 3D faces. SIGGRAPH '99: Proceedings of the 26th Annual Conference on Computer Graphics and Interactive Techniques. https://doi.org/10.1145/311535.311556
43. Li T, Bolkart T, Black MJ, Li H, Romero J (2017) Learning a model of facial shape and expression from 4D scans. ACM Trans Graph 36(6):1–17. https://doi.org/10.1145/3130800.3130813
44. Sanyal S, Bolkart T, Feng H, Black MJ (2019) Learning to Regress 3D Face Shape and Expression From an Image Without 3D Supervision. IEEE/CVF Conf Comput Vis Patt Recog (CVPR) 2019:7755–7764. https://doi.org/10.1109/cvpr.2019.00795
45. Feng Y, Feng H, Black MJ, Bolkart T (2021) Learning an animatable detailed 3D face model from in-the-wild images. ACM Trans Graph 40(4):1–13. https://doi.org/10.1145/3450626.3459936
46. Guo J, Zhu X, Yang Y, Yang F, Lei Z, & Li SZ (2020) Towards fast, accurate and stable 3D dense face alignment. In Lecture notes in computer science (pp. 152–168). https://doi.org/10.1007/978-3-030-58529-7_10
47. Vaillant R, Monrocq C, Cun YL (1994) Original approach for the localisation of objects in images. IEE Proceed Vis Image Signal Process 141(4):245. https://doi.org/10.1049/ip-vis:19941301
48. Hochreiter S, Schmidhuber J (1997) Long Short-Term memory. Neural Comput 9(8):1735–1780. https://doi.org/10.1162/neco.1997.9.8.1735
49. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser Ł, & Polosukhin I (2017) Attention is all you need. arXiv (Cornell University). https://doi.org/10.48550/arxiv.1706.03762
50. Lin T, Wang Y, Li X, Qiu X (2022) A survey of transformers. AI Open 3:111–132. https://doi.org/10.1016/j.aiopen.2022.10.001
51. Zhao C, Wang C, Hu G, Chen H, Liu C, Tang J (2023) ISTVT: Interpretable Spatial-Temporal Video Transformer for Deepfake Detection. IEEE Trans Inf Forensics Secur 18:1335–1348. https://doi.org/10.1109/tifs.2023.3239223
52. Selva J, Johansen AS, Escalera S, Nasrollahi K, Moeslund TB, & Clapés A (2023) Video Transformers: a survey. IEEE Transactions on Pattern Analysis and Machine Intelligence, 1–20. https://doi.org/10.1109/tpami.2023.3243465

53. Loshchilov I, Hutter F (2016) SGDR: stochastic gradient descent with warm restarts. In: 5th International Conference on Learning Representations (ILCR 2017)
54. Afchar D, Nozick V, Yamagishi J, & Echizen I (2018) MesoNet: a Compact Facial Video Forgery Detection Network. 2018 IEEE Int Workshop Inform Forensic Secur (WIFS). https://doi.org/10.1109/wifs.2018.8630761
55. Chollet F (2017) Xception: Deep Learning with Depthwise Separable Convolutions. 2017 IEEE Confer Comput Vis Patt Recog (CVPR). https://doi.org/10.1109/cvpr.2017.195
56. Tan M, Le Q (2019) EfficientNet: Rethinking model scaling for convolutional neural networks. Int Conf Machine Learn 2019:6105–6114
57. Dang H, Liu F, Stehouwer J, Liu X, & Jain AK (2020) On the Detection of Digital Face Manipulation. 2020 IEEE/CVF Conf Comput Vis Patt Recog (CVPR). https://doi.org/10.1109/cvpr42600.2020.00582
58. Bonettini N, Cannas ED, Mandelli S, Bondi L, Bestagini P, & Tubaro S (2021) Video Face Manipulation Detection Through Ensemble of CNNs. 2020 25th Int Conf Patt Recog (ICPR). https://doi.org/10.1109/icpr48806.2021.9412711