Electronic Research Archive

http://www.aimspress.com/journal/era

ERA, 32(4): 2621–2641.
DOI: 10.3934/era.2024119
Received: 03 January 2024
Revised: 01 March 2024
Accepted: 20 March 2024
Published: 29 March 2024

Research article

Analyzing temporal coherence for deepfake video detection

Muhammad Ahmad Amin[1], **Yongjian Hu**[1,*] **and Jiankun Hu**[2]

[1] School of Electronic and Information Engineering, South China University of Technology, Guangzhou 510641, China

[2] School of Engineering and Information Technology, The University of New South Wales, Australian Defence Force Academy Canberra, ACT 2610, Australia

* **Correspondence:** Email: eeyjhu@scut.edu.cn; Tel: +8613642667090.

Abstract: Current facial image manipulation techniques have caused public concerns while achieving impressive quality. However, these techniques are mostly bound to a single frame for synthesized videos and pay little attention to the most discriminatory temporal frequency artifacts between various frames. Detecting deepfake videos using temporal modeling still poses a challenge. To address this issue, we present a novel deepfake video detection framework in this paper that consists of two levels: temporal modeling and coherence analysis. At the first level, to fully capture temporal coherence over the entire video, we devise an efficient temporal facial pattern (TFP) mechanism that explores the color variations of forgery-sensitive facial areas by providing global and local-successive temporal views. The second level presents a temporal coherence analyzing network (TCAN), which consists of novel global temporal self-attention characteristics, high-resolution fine and low-resolution coarse feature extraction, and aggregation mechanisms, with the aims of long-range relationship modeling from a local-successive temporal perspective within a TFP and capturing the vital dynamic incoherence for robust detection. Thorough experiments on large-scale datasets, including FaceForensics++, DeepFakeDetection, DeepFake Detection Challenge, CelebDF-V2, and DeeperForensics, reveal that our paradigm surpasses current approaches and stays effective when detecting unseen sorts of deepfake videos.

Keywords: deepfake video detection; temporal modeling; self-attention; coherence interpretability; temporal transformer

1. Introduction

The emergence of deepfake videos created by facial image manipulation methods poses a potential threat to our privacy and social safety. These deepfake videos can be effortlessly fabricated using

publicly available software like Deepfakes [1] and DeepFaceLab [2]. Thus, developing effective and robust deepfake detection methods has become crucial. Fortunately, many deepfake detectors have been developed, which have shown promising results on various deepfake databases.

Current methods for detecting deepfake videos can be split into frame-level and video-level categories. Frame-level approaches [3–8] use a single video frame as intake and concentrate on identifying spatial forgery patterns such as RGB information, frequency statistics, auxiliary mask or blending boundary information, etc. Although these methods have achieved a great success, they have two critical drawbacks. First, the facial video created through frame-by-frame manipulation has two main types of artifacts. The first type is spatially associated, including checkboard, blur and blending boundary cues. The second type is temporal incoherence. Even cutting-edge forgery methods may create remarkably unpretentious facial images, they cannot eradicate unnatural and minute image transitions [4] among frames or temporal incoherence like facial parts jittering [5]. Unfortunately, the frame-level approaches fail to detect these temporal discriminatory traits and have capped performance. Second, the spatially associated cues are more dominant than temporal incoherence, and frame-level approaches tend to overfit the spatial cues. As a result, without a straightforward temporal design, current methods count more on spatial features than temporal incoherence to identify real or deepfake videos.

On the other hand, to address the limitations of frame-based methods, video-based deepfake detection techniques analyze a sequence of frames to identify inconsistencies in texture and structure within the video, making the detection process more resilient [9–11]. The early research employed general video analysis techniques such as 3-dimensional convolutional neural network (3D CNN) [12], recurrent neural network (RNN) [13], and long short-term memory (LSTM) [14]. However, these approaches are computationally expensive and have limitations in mining meaningful temporal features. This is because they are not designed explicitly for facial deepfake video detection, making it challenging to identify temporal inconsistencies. Furthermore, these models are less effective than frame-based methods, which can compute the video-level decision by averaging frame-level scores. Recently, researchers have proposed methods [15–19] that uniformly sample single frames at gaps to assemble the conclusive input sequence, resulting in better temporal modeling of deepfake videos. However, these approaches are more computationally intensive. Concurrently, while vision transformer [20] (ViT) methods have shown encouraging achievement in other computer vision studies, such as activity identification [21, 22] and video object detection [23, 24], their application to deepfake video detection has shown promising results as well. Not surprisingly, some researchers have tried to employ video transformers for detecting deepfake videos, but they suffer from high computational complexity, limited temporal modeling, and a lack of interpretability.

Our research is motivated by two observed facts. First, current facial image manipulation methods are explicitly designed at the image level, which requires applying the methods to each frame individually to generate a fake video. Regardless unpretentious shifts in appearance, like lighting, noise, and motion, often lead to temporal incoherent consequences, as shown in Figure 3. Second, detecting temporal incoherence is challenging since the annotated areas of the incoherence in the video are unavailable. Adopting complex temporal networks [25] to discern real from deepfake by temporal incoherence leads to high computational complexity, limited temporal modeling, and a lack of interpretability. Thus, detecting temporal incoherence requires more careful study and an interpretable approach.

To address these challenges for generalized deepfake video detection, in this paper, we devise a

novel paradigm to mine and analyze comprehensive and unpretentious temporal coherence in two steps. In the first step, to leverage temporal coherence, we extract the spatial pixel variation of each red-green-blue (RGB) color channel from multiple facial areas that are forgery-sensitive, such as the forehead, eyes, nose, mouth, and chin areas. These areas form a set of multivariate sequences called the temporal facial pattern (TFP), which contains both local-successive temporal and global temporal perspectives. As the dynamic inconsistency occurs in local-successive temporal regions, these clues can characterize and interpret the distribution of forgery-sensitive areas in space and time, and they can be robustly used to expose deepfake videos. The second step involves interpreting the TFP significantly to understand the coherence of real and deepfake videos. Inspired by the affluent global attention attributes and self-attention mechanism of transformers in text and sequence processing, we propose a refined transformer framework, anointed temporal coherence analyzing network (TCAN), and employ its powerful mining ability to implicitly find the global intrinsic representation to expose the difference in coherence between real and synthetic videos. Specifically, in contrast to ViT, the TCAN framework redesigns the encoder and the classification head, eliminating some of its limitations of low-resolution features with fixed resolutions and high computation costs. In the first modification, the TCAN adapts position-free encoding, which helps adapt to any arbitrary resolution without impacting performance. Second, we introduce a hierarchical function that facilitates the encoder to generate high-resolution fine and low-resolution coarse features. Third, the classification head benefits from the transformer-induced features by aggregating information from different layers, combining global and local attention. Thus, we get explicit yet decisive representations for real and deepfake video classification. The proposed framework is depicted in Figure 1.

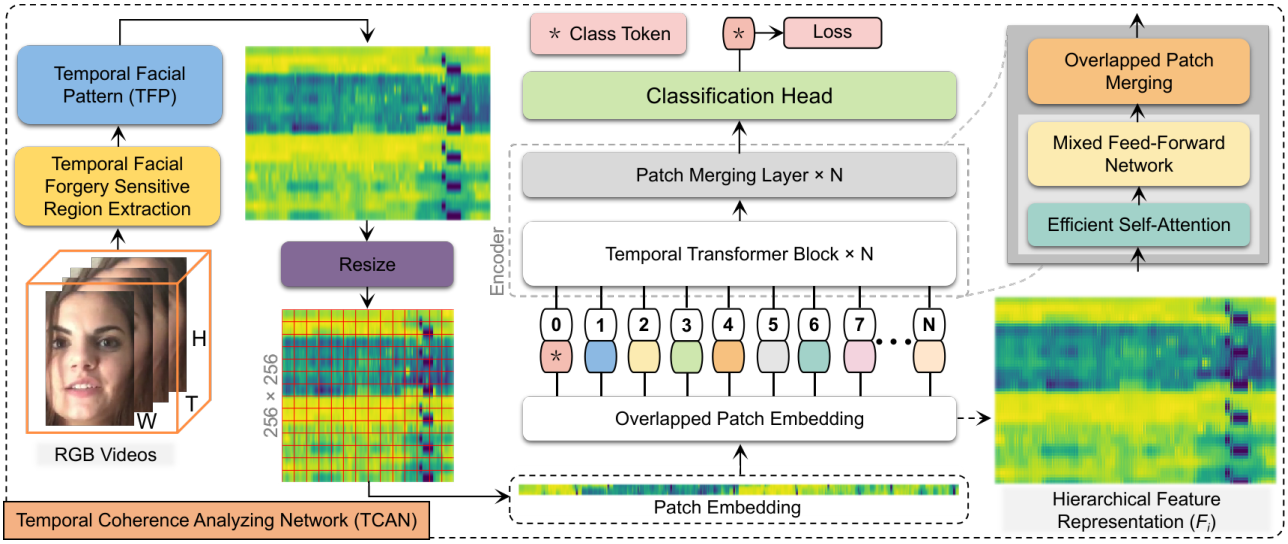

Figure 1. An illustration of our proposed deepfake video detection framework based on temporal facial pattern (TFP) and temporal coherence analyzing network (TCAN). TCAN has two main components: a hierarchical transformer encoder for feature extraction and a classification head for feature fusion and output prediction.

The key contribution of our approach includes:

- We propose a TCAN that explicitly analyzes the temporal clues to detect incoherence and exposes

synthetic videos.

- We introduce a novel TFP based on the pixel variation of each RGB color channel from multiple regions in the possible forgery-sensitive facial areas to represent the local-successive temporal and global temporal views of real and deepfake videos.
- Thorough experiments on diverse datasets exhibit the dominance of our proposed approach concerning the generalization ability against unseen forgeries.

This paper is arranged as follows. We will begin with a summary of prior works in Section 2, followed by an explanation of our presented methodology in Section 3. Section 4 will present a detailed account of the experimental setup, and Section 5 will present our evaluation, analysis, and ablation study. Finally, in Section 6, we will conclude the paper by our conclusions and future work.

2. Related work

In deepfake synthesis, facial video manipulations are performed frame-by-frame, leading to inconsistencies between frames. Therefore, spatio-temporal incoherence can be used to detect deepfake videos more broadly. Some studies have focused on using temporal clues to detect fake face videos. In [12], 3D CNN is employed to illustrate spatio-temporal features. Some studies initially perform frame-level CNN feature extraction and then employ RNN [13, 26] or LSTM [14, 27] to learn the temporal artifacts. Nevertheless, the temporal clues extracted by these approaches are relatively crude. Gu et al. [10] present the spatial-temporal inconsistency learning (STIL) block to grasp temporal incoherence along two orthogonal paths. LipForensics was presented by Haliassos et al. [15], founded on high-level semantic abnormalities in mouth movements, to accomplish robust and generalizable deepfake detection. Recent studies [16–19] suggest constructing a local-global temporal standpoint to catch spatio-temporal cues.

There are specific methods based on transformers that are designed to detect deepfake videos. To capture long-scope dependencies along the temporal dimension and achieve generalizable deepfake detection, Zheng et al. [16] propose fully temporal convolutional network (FTCN), which uses transformers to mine spatio-temporal clues. Shao et al. [19] proposed SeqFakeFormer, which aims to accurately predict a sequential vector of face manipulation procedures instead of merely giving a binary label output. Additionally, Yu et al. [28] developed multiple spatiotemporal views transformer (MSVT), a new transformer model that incorporates local spatio-temporal views to capture dynamic inconsistency and global spatio-temporal views to extract video-wide features. Then, a global-local transformer integrates multi-level cues, providing a thorough spatio-temporal feature representation to enhance detection performance.

Similarly, many multi-modal detection approaches are proposed [29–33], which include methods based on the consistency among visual and audio modalities. The complementarity between the two modalities allows one modality to enhance the feature representation of the other. Cheng et al. [29] introduced a method based on voice-face conformity to detect manipulated deepfake video (VFD), which examines the uniformity between facial identity and audio in raw videos. Regardless, VFD cannot catch fiddled content without identity change, such as while glasses are removed. In another effort, Yang et al. proposed audio-visual-based learning for detecting deepfake (AVoiD-DF) [30], which includes a two-stream spatial-temporal encoder and a multi-modal coordinated decoder for collective intrinsic relationship learning. Despite only utilizing visual temporal clues, our proposed method aims

to capture nuanced and thorough temporal features from local and global standpoints in an interpretable way with the least computation complexity, which is an improvement.

3. Proposed method

In this section, we foremost provide details regarding the multi-region TFP modeling process. Secondly, we present the transformer-based temporal coherence analyzing network for deepfake video detection, as illustrated in Figure 1. The TFP is modeled based on six facial areas, which are more susceptible to forgery artifacts, as shown in Figure 2. The areas from the forehead, eyes, nose, mouth, and chin are typically in constant motion. Hence, these regions will most likely contain temporal forgery cues while generating fake facial videos. A TFP is generated by placing these six regions over the length of a video, from top to bottom, in order from forehead to chin. This placement is designed with the patch embedding mechanism in mind, as used in a conventional transformer model. Every TFP row represents the varying color information extracted from each region in the pristine or deepfake video.

Further, the TCAN model relies on the hierarchical feature representation of the temporal hierarchical transformer encoder to analyze the temporal coherence within TFP for real and deepfake video identification. If we adopt a typical ViT and input the TFP directly for further learning, the original patch segmentation method cuts the images into nonoverlapping patches. This process will destroy the time information hidden in each row as it follows the specific patch size information and will discard discriminative information of some local neighboring informative regions. Therefore, to sort out this issue, we introduce the hierarchical representation through overlapping patches merging method to generate multi-level features similar to CNNs to focus on local information and modified ViT global attention mechanism based on sparse global connection, as illustrated in Figure 4, because generally ViT only generates fixed resolution feature map. Finally, all the features from the four encoder blocks within the TCAN are aggregated to obtain a final high-resolution fine and low-resolution coarse feature representation. Further, the self-attention and data-driven positional encoding mechanisms are employed to incorporate location information within TFP for temporal coherence analysis, and the classification head predicts the outcome based on this analysis.

3.1. Temporal facial pattern

The proposed multi-region TFP is a novel representation of the facial forgery-sensitive areas in terms of color variations that incorporate local and global cues. The approach is detailed in Figure 2 and further elaborated below.

3.1.1. Facial forgery sensitive areas detection and segmentation

In our approach, we employ an 81-point Dlib [34] face detector to catch the face landmarks. On the basis of detected landmarks, we determine the most instructive regions in the face that are susceptible to forgery, including the forehead (landmark points from 68 to 77), left eye (from 36 to 39), right eye (from 42 to 45), nose (from 27 to 35), mouth (from 48 to 67), and chin (from 4 to 12). As the detector can operate at a frame rate of 30 plus frames per second, we do face and landmark detection on each video sequence frame to obtain precise and uniform region facial areas.

We also must regard instances where some facial landmarks are not detected for rare frames, which can occur when the head of a person rotates or moves too quickly. This results in missing information

on signals that vary in color over time. To address this problem, we artificially simulate the missing data cases by randomly masking a miniature portion of the temporal pattern along the temporal dimension. We then use this somewhat masked temporal pattern as augmented information to improve the performance of the TCAN model.

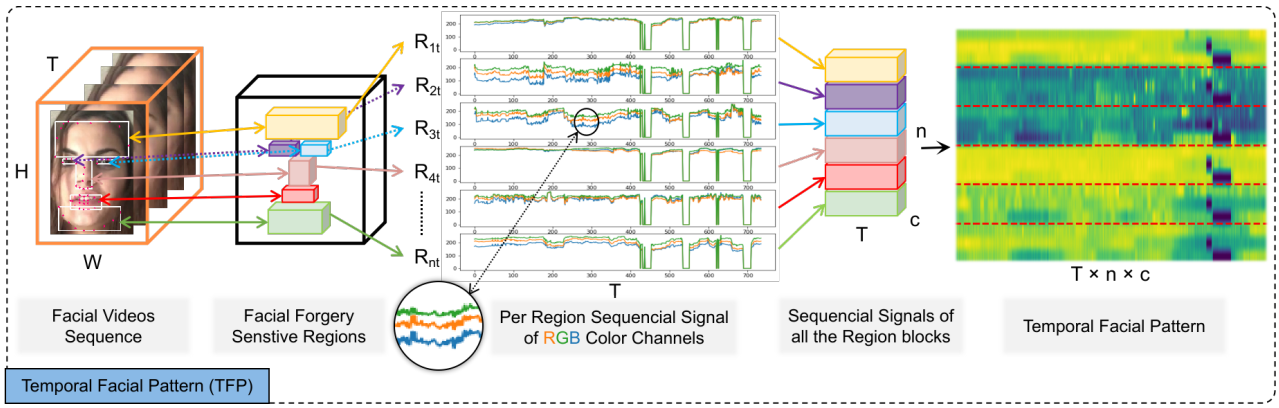

Figure 2. An illustration of TFP generation from a facial video.

3.1.2. Temporal pattern generation and modeling

In real or deepfake video processing, each video sequence with c color space dimensions and T frames undergoes a face and landmark detection process. If a face and its landmarks are detected, the facial regions per frame are defined and split into n blocks, and the mean color intensity of each block is calculated over time. This helps analyze and understand the changes in color intensity of different facial regions over time.

Further, to enhance the signal-to-noise ratio (SNR) and reduce noise within TFP, our method relies on the variation in pixel values of the RGB channels of the entire facial area. This type of average pooling presentation offers greater robustness than analyzing individual pixels. For each region block, three temporal sequences are generated based on the various channels of the RGB color space. Moreover, a min-max normalization per sequence is applied. The n temporal sequences for each color dimension are arranged in rows to create a TFP of the areas that are susceptible to forgery in the video clip.

In the context of a video with c color space channels and T frames, we extract facial regions in each frame from the forehead, mouth, nose, and eye areas. Further, we divide the face region of each frame into n block regions $R_t = \{R_{1t}, R_{2t}, \ldots, R_{nt}\}$. Let $C(x, y, t)$ represent the value at point (x, y) of the t^{th} frame using the various color space dimensions. We determine the average pooling of every channel for the t^{th} block region of the t^{th} frame, where $|R_i|$ represents the area of a block region (in terms of the number of pixels).

$$\overline{C_i}(t) = \frac{\sum_{x,y \in R_i} \cdot C(x, y, t)}{|R_i|} \tag{3.1}$$

So, for every video sequence, we can derive a set of temporal sequences $3 \times n$ for different channels of the RGB color space with the length of T, denoted as $C_i = \{C_i(1), C_i(2), \ldots, C_i(T)\}$, where C refers to one of the color channel within c and i refers to the region block index. After applying a min-max normalization to every temporal sequences and scaling the data into the [0, 255] range, we arrange the n

temporal signals into rows, which yields a TFP of the natural facial video sequence with $T \times n \times c$ size. The TFP samples of real and deepfake videos from the FF++ [35] dataset with various compression levels are shown in Figure 3.

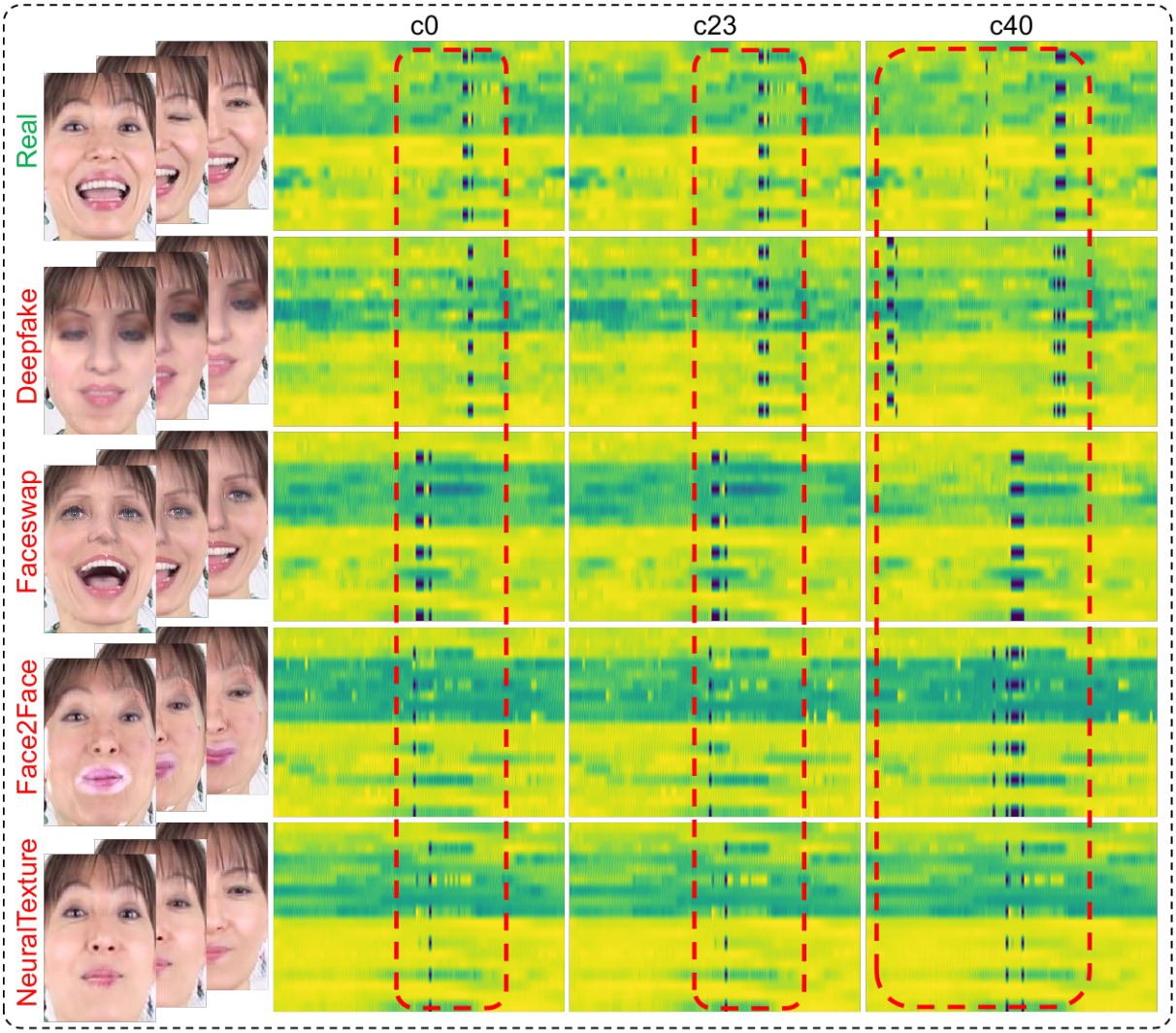

Figure 3. A comparison of TFPs of real and deepfake videos with three different compression levels (raw (c0), high quality (c23), and low quality (c40)) of FF++ [35] dataset. Deepfake videos are generated with four different manipulation methods.

3.2. Temporal coherence analyzing network

This section provides a brief overview of our proposed TCAN design. TCAN comprises two primary components: (1) a temporal hierarchical transformer encoder (THTE) that generates low-resolution fine and high-resolution coarse characteristics, and (2) a classification head, as illustrated in Figure 4. To begin, we divide a TFP of size $H \times W \times 3$ into patches of size 7×7. In contrast to ViT [20], which employs size 16×16 patches, smaller patches are more favorable for fine-grain temporal coherence analysis. We then input these patches to the THTE, which generates multi-level

features at $\left\{\frac{1}{4}, \frac{1}{8}, \frac{1}{16}, \frac{1}{32}\right\}$ that preserve the actual TFP resolution. Finally, the classification head intakes these multi-level features to foresee the output class.

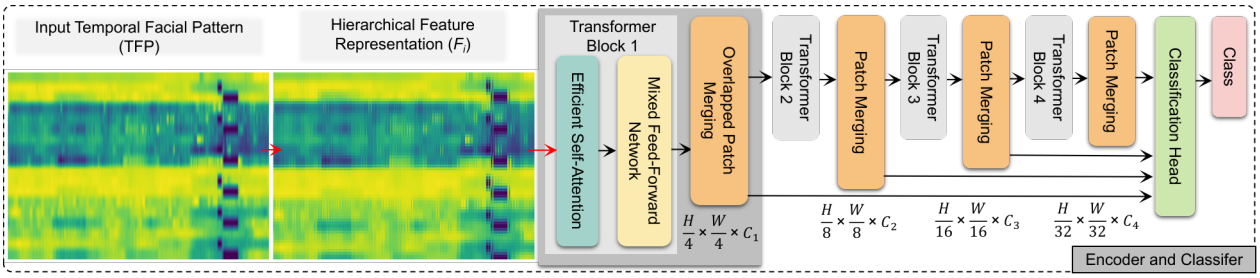

Figure 4. THTE design illustration.

3.2.1. Temporal hierarchical transformer encoder

The THTE design is partly inspired by ViT [20, 36] but tailored for deepfake video detection. It benefits from four key contributions, such as hierarchical feature representation, overlapped patch merging, efficient self-attention, and mixed feed-forward network. These modules are detailed below.

3.2.1.1. Hierarchical feature representation

The representation module is designed to generate multi-level features similar to CNNs, unlike ViT, which produces a uni-resolution cue map. The high-resolution fine-grained and low-resolution coarse features extraction from an input TFP image can enhance the performance of temporal classification tasks. To achieve this, we use patch merging on an input TFP image of resolution $H \times W \times 3$ to get a hierarchical feature map F_i with $\frac{H}{2^{i+1}} \times \frac{W}{2^{i+1}} \times C_i$ resolution, where $i \in \{1, 2, 3, 4\}$. The resolution of each level C_{i+1} is greater than the previous level C_i, as demonstrated in Figure 4.

3.2.1.2. Overlapped patch merging

The process of merging patches in a TFP patch involves unifying an $N \times N \times 3$ patch into a $1 \times 1 \times C$ vector, which is used in baseline ViT. We can extend this process to confine a $2 \times 2 \times C_i$ feature course into a $2 \times 2 \times C_{i+1}$ vector, thus obtaining hierarchical feature maps. Shrinking hierarchical features from $F_1\left(\frac{H}{4} \times \frac{W}{4} \times C_1\right)$ to $F_2\left(\frac{H}{8} \times \frac{W}{8} \times C_2\right)$ and iterating for additional feature maps in the hierarchy can be achieved using this method. However, this technique was originally devised to merge feature patches or non-overlapping images, resulting in the failure to keep the local continuity information about those patches, which is not suggestible for temporal information analysis. Thus, an overlapping patch merging method is employed instead, requiring us to define P, K, and S, where P is the padding size, K denotes patch size, and the stride between two adjacent patches is S. During experiments, we fixed $P = 3$, $K = 7$, and $S = 4$ and $P = 1$, $K = 3$, and $S = 2$ to achieve overlapping patch blending to produce features the exact size as the non-overlapping strategy.

3.2.1.3. Efficient self-attention

Self-attention layer is the primary computational bottleneck for the encoder blocks. In the standard multi-head self-attention procedure, every head (Q, K, V) has the exact dimensions $N \times C$ as the length

of the sequence ($N = H \times W$). The self-attention is calculated as follows:

$$Attention(Q, K, V) = Softmax\left(\frac{QK_T}{\sqrt{d_{head}}}\right) \cdot V. \tag{3.2}$$

The computational sophistication of the current procedure is $O\left(N^2\right)$, which can be a challenge for processing large image resolutions. However, we have an alternative approach that utilizes the sequence-lessening technique described in [37]. This technique involves using a reduction ratio R to effectively shorten the sequence length without compromising its quality as follows:

$$\hat{K} = Reshape\left(\frac{N}{R}, C \cdot R\right)(K), \tag{3.3}$$

$$K = Linear\left(C \cdot R, C\right) \cdot \left(\hat{K}\right). \tag{3.4}$$

In the technical context, while K represents the sequence that needs to be reduced, $\left(\frac{N}{R}, C \cdot R\right)(K)$ refers to the reshaping of K to match the shape of $\frac{N}{R} \times (C \cdot R)$. The $Linear\left(C_{in} \cdot C_{out}\right)(\cdot)$, in this context, directs to a linear layer that takes a C_{in}-dimensional tensor as intake and generates a C_{out}-dimensional tensor as outcome. The resulting tensor K has dimensions $\frac{N}{R} \times C$, which reduces the self-attention mechanism complexity from $O\left(N^2\right)$ to $O\left(\frac{N^2}{R}\right)$. We set R value to [64, 16, 4, 1] from block 1 to 4.

3.2.1.4. Mixed feed-forward network

The ViT [20] introduces positional encoding (PE) to incorporate location information. Nevertheless, the PE resolution is fixed, which causes accuracy drops when the inference resolution is dissimilar from the training resolution and the interpolation of positional code is required. To deal with this issue, the CPVT [38] model uses a 3×3 convolutional layer with the PE to create a data-controlled PE. We use a new mixed feed-forward network (mixed-FFN) approach that employs a 3×3 convolutional layer ($Conv$) straight in the FFN to consider the zero padding effect and minimize location information leakage, as in [39]. The mixed-FFN approach streamlines as in Eq (3.5).

$$x_{out} = MLP\left(GELU\left(Conv_{3\times3}\left(MLP\left(x_{in}\right)\right)\right)\right) + x_{in} \tag{3.5}$$

The self-attention module uses the feature x_{in}. Each FFN in mixed-FFN combines a multilayer perceptron (MLP) and a 3×3 convolution. A 3×3 convolution is enough for our temporal transformer to get positional information required for temporal coherence analysis. In respective, we employ depth-wise convolutions to reduce parameters and enhance performance.

3.2.2. Classification head

The TCAN model uses a classification head that consists of a single MLP layer. This layer allows us to fuse multi-level features from our THTE and achieve a more considerably effective receptive field (ERF) compared to conventional CNN encoders. Specifically, there are two main phases in the classification head. The first phase unifies the multi-level features F_i from THTE in the channel dimension. In the second phase, the concatenated feature embeddings are used to predict the output class categories. These steps can be formulated as:

$$y = Linear\left(4C_i, N_{cls}\right)\left(Concat\left(F_i\right)\right), \forall_i, \tag{3.6}$$

where y denotes the predicted class, $Linear\left(C_i, N_{cls}\right)(\cdot)$ and represents a linear layer with C_i and N_{cls} as intake and outcome vector dimensions, respectively. Table 1 provides parameters of TCAN model.

Table 1. TCAN model parameters and configuration details.

Model modules	Information
No. of encoder blocks	4
No. of attention heads	[1, 2, 5, 8]
Patch sizes	[7, 3, 3, 3]
Strides	[4, 2, 2, 2]
Classifier hidden size	256
Parameters	3.3 million
FLOPs	8.4 G

4. Experimental details

This section outlines the datasets, the evaluation metric, hyperparameters, and the loss function.

4.1. Datasets

The efficacy of the TCAN model was evaluated through experiments conducted on five renowned benchmark datasets, namely FaceForensics++ [35] (FF++), DeepFakeDetection (DFD) [40], Deepfake Detection Challenge [41] (DFDC), Celeb-DF-V2 [42] (CDF-V2), and DeeperForensics-1.0 [43] (Deeper). FF++ is comprised of subsets such as DeepFakes (DF), FaceSwap (FS), Face2Face (F2F), and NeuralTextures (NT), each having three compression ratios (raw (c0), high quality (c23), and low quality (c40)). For each dataset, 1000 videos were chosen and divided into separate subsets for training, validation, and testing as in [35]. A more detailed summary of the datasets is provided in Table 2.

Table 2. A summary of the datasets that were evaluated in this experimental study.

Dataset	Manipulation methods	Real/Fake
FF++ [35]	DF, F2F, FS, NT	1000/4000
DFD [40]	Improved DF	363/3068
DFDC [41]	DF-256, MM/NN, NTH, Audio swaps, DF-128, FSGAN, StyleGAN, refinement	23,654/104,500
CDF-V2 [42]	Enhanced FaceSwap DF	590/5639
Deeper [43]	DF-VAE with seven perturbations	1000/9000

4.2. Evaluation metric

We employ the area under the receiver operating characteristic curve [44] (AUC) as an evaluation metric following the previous works [16, 28], which is standard practice in most studies. The AUC

values are calculated based on video-level evaluation.

4.3. Hyperparameters settings

In the following subsection, we provide information regarding key hyperparameters and preprocessing techniques used in the training and testing of the TCAN model. We employ the learning rate (*lr*) of $5e^{-5}$ and the AdamW optimizer with a cosine annealing learning rate scheduler. All models are trained with a maximum of 15 epochs. All the TCAN models are trained on a NVIDIA RTX 3090Ti with batch size 16, using mix-precision training [45] (bfloat-16). The input TFP is resized to 256×256 during the training process. We apply a normalization transformation to preprocess the data using dataset-specific mean and standard deviation values. These values (i.e., approximately [0.2292, 0.5373, 0.4988] for means and [0.1439, 0.1881, 0.0986] for standard deviations) ensure the dataset is standardized for all the experimentation purposes.

4.4. Loss function

The cross-entropy loss function L_{CE} [46] is employed in the training of the TCAN model.

$$L_{CE} = -(y \cdot \log(p) + (1 - y) \cdot \log(1 - p)), \tag{4.1}$$

where y represents the binary label, and p in [0, 1] are the predicted probability for the positive class.

5. Results and analysis

In this section, we evaluate our proposed deepfake video detection framework through experiments that included intra-, cross-dataset, and robustness scenarios. These experiments are designed to simulate the detection of previously unseen forgeries, cross-domain, and robustness settings against compression ratios. To train our framework, we used the FF++ [35] subsets at the c0 compression level, and for testing, we utilized the c0, c23, and c40 sets (see Tables 3 and 5) to perform intra- and robustness evaluations. We trained our framework on the complete FF++ [35] (c0) dataset for cross-dataset evaluations and tested it on the DFD [40], DFDC [41], CDF-V2 [42], and Deeper [43] datasets (refer to Table 4). We also compared our results in intra- and cross-dataset settings against state-of-the-art methods. Additionally, we performed an ablation study by replacing the TCAN with different state-of-the-art transformer methods. Using publicly available codes, we implemented the comparison methods in the same environment as ours.

5.1. Intra-dataset evaluation against unseen forgeries

The differences among facial forgery videos largely stem from variations in reference videos and facial alteration techniques. To assess the cross-manipulation generalization ability of our method against various deepfake detectors and eliminate potential biases oriented by distinct reference videos, we performed experimentations on the FF++ [35] dataset, as this dataset supplies fabricated videos generated by four facial forgery techniques for the same reference videos. We utilized the leave-one-out approach to evaluate the efficacy of our face forgery detectors, as presented in the study by [28]. The evaluation process entailed using each of the four categories of deepfake videos in the FF++ [35] dataset as a test dataset. Conversely, the remaining three kinds formed the training dataset. The training and evaluation

phases are performed on the FF++ [35] dataset c0 variant. The outcomes are summarized in Table 3, while the learned feature spaces are visualized in Figure 5 using t-SNE [47] plots.

Table 3 demonstrates the impressive detection results of our approach, with a generalization AUC score of 98.82% for unique forgeries. Our approach surpasses most existent methods by a significant margin on all four types of manipulations in FF++ [35] (DF, FS, F2F, and NT) that use different methods and have distinct artifacts. Our method can learn generalized discriminative features on three manipulations and generalize to the remaining one. Compared to contemporary state-of-the-art approaches such as MSVT [28], FTCN [16], SeqFakeFormer [19], and LipForensics [15], our approach achieves a 1.48%, 2.56%, 3.09%, and 3.0% increase in video-level AUC, respectively. Moreover, our approach accomplishes the best performance with the tiniest number of parameters (**M** mean millions), without any external training data or pre-training, which demonstrates superiority of our framework.

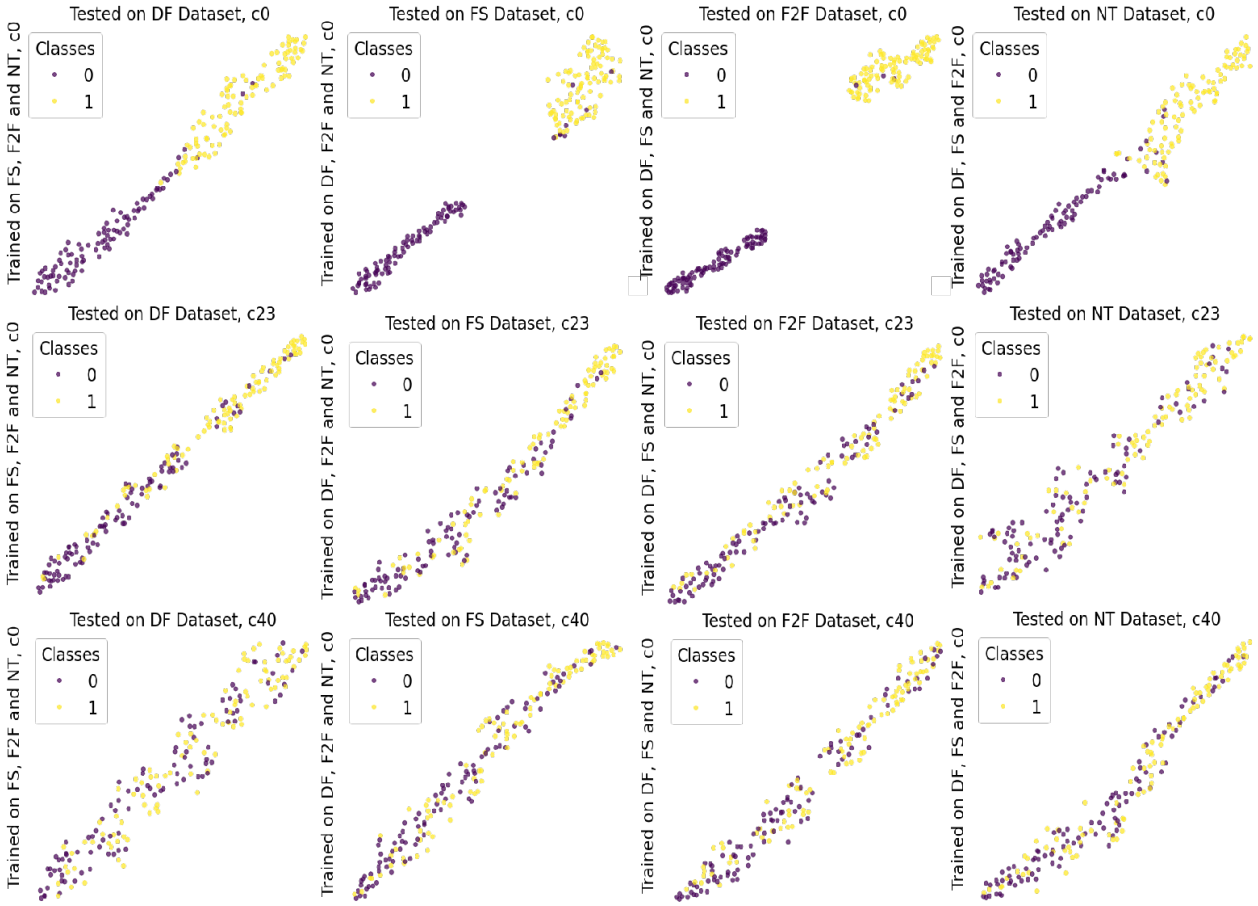

Figure 5. The t-SNE [47] visualization depicts the learned feature space of classes predicted by our proposed framework on cross-manipulation evaluation across three compression levels. Golden dots indicate real class, and purple dots signify manipulated class samples.

Table 3. Our proposed method is evaluated for intra-dataset settings on the FF++ [35]. Results are compared in terms of video-level AUC (%) with the state-of-the-art methods.

Type	Methods	Trained on remaining three				Avg.	#params.
		DF	FS	F2F	NT		
Frame-level	LRLNet [7]	91.37	89.44	90.22	85.83	89.22	46.3 M
	PEL [8]	89.78	87.67	88.91	87.78	88.54	33.7 M
Video-level	S-MIL [9]	87.68	85.26	85.85	84.49	85.82	42.2 M
	STIL [10]	90.71	88.73	91.73	88.78	89.99	32.5 M
	LipForensics [15]	97.83	90.51	98.02	96.90	95.82	36.0 M
	FTCN [16]	96.71	96.02	96.07	96.24	96.26	26.6 M
	DIL [17]	94.36	94.77	94.81	93.24	94.3	41.4 M
	ECGL [18]	92.49	90.96	89.59	89.19	90.56	37.3 M
	SeqFakeFormer [19]	96.82	95.08	95.75	95.27	95.73	104 M
	Two-branch [27]	89.81	87.28	90.06	86.92	88.52	67.1 M
	MSVT [28]	98.06	97.43	96.68	97.19	97.34	26.5 M
	VDF [29]	96.20	86.31	89.60	94.25	91.62	40 M
	Audio-DF [30]	97.31	94.74	94.50	95.10	95.41	89 M
	TCAN (Ours)	**99.16**	**98.93**	**99.03**	**98.15**	**98.82**	**3.3 M**

5.2. Cross-dataset evaluation against unseen cross-domain

Current deepfake videos are generally generated using a variety of synthesis sources. Therefore, it is crucial to test their cross-domain generalization capability. To test the generalization capacity of TCAN, we trained our model on the FF++ [35] dataset and tested it on DFD [40], DFDC [41], CDF-V2 [42], and Deeper [43] datasets. We achieved auspicious detection results corresponding to the aforementioned approaches in Table 4, indicating robustness improvement of our method. Our approach achieved promising results in comparison to both video-level and frame-level approaches. Specifically, our method achieved an outstanding AUC score of 98.80% on the Deeper [43] dataset, which was generated using advanced synthesis techniques on real videos from the FF++ [35] dataset. Likewise, our method yielded an AUC score of 91.09% on the DFD [40] dataset but slightly underperformed in comparison to the likes of MSVT [28] (91.36% AUC) and Audio-DF [30] (91.12% AUC), generated by Google using an improved deepfake generation pipeline. However, generalization was more challenging for unseen datasets, such as CDF-V2 [42] and DFDC [41]. We found that video-based methods, such as MSVT [28] (76.79% AUC), VDF [29] (85.10% AUC), Audio-DF [30] (82.31% AUC), three state-of-the-art transformer-based approaches, and our TCAN with 86.94% AUC, outperformed the frame-level state-of-the-art approaches on the DFDC [41] dataset, such as LRLNet [7] (76.61% AUC) and PEL [8] (65.87% AUC), which suggests that temporal information is crucial in enhancing the cross-domain generalization ability of deepfake video detection methods. Similar trends are also shown on the CDF-V2 [42] dataset in Table 4. Figure 6 shows the feature spaces learned by our method across different datasets.

Table 4. Our proposed method is evaluated for cross-dataset settings, where it is trained on the FF++ [35] and tested on DFD [40], DFDC [41], CDF-V2 [42], and Deeper [43] datasets. Results are compared in terms of video-level AUC (%) with state-of-the-art methods.

Type	Methods	Trained on all four subsets					
		DFD	DFDC	CDF-V2	Deeper	Avg.	#params.
Frame-level	LRLNet [7]	77.61	76.61	78.37	85.86	79.61	46.3 M
	PEL [8]	89.29	65.87	70.29	70.41	73.97	33.7 M
Video-level	S-MIL [9]	86.39	66.79	75.86	83.07	78.03	42.2 M
	STIL [10]	87.21	67.22	76.17	83.94	78.64	32.5 M
	LipForensics [15]	84.25	73.62	83.60	97.00	84.62	36.0 M
	FTCN [16]	90.46	74.17	86.95	97.95	87.38	26.6 M
	DIL [17]	88.15	72.52	82.82	86.09	82.40	41.4 M
	ECGL [18]	90.18	73.09	85.86	96.62	86.94	37.3 M
	SeqFakeFormer [19]	90.22	74.32	84.26	95.96	86.19	104 M
	Two-branch [27]	86.19	70.18	75.79	81.34	78.38	67.1 M
	MSVT [28]	91.36	76.79	88.81	98.42	88.85	26.5 M
	VDF [29]	90.50	85.10	80.70	84.30	85.15	40 M
	Audio-DF [30]	91.12	82.31	86.20	96.50	89.03	89 M
	TCAN (Ours)	**91.09**	**86.94**	**88.99**	**98.80**	**91.46**	**3.3 M**

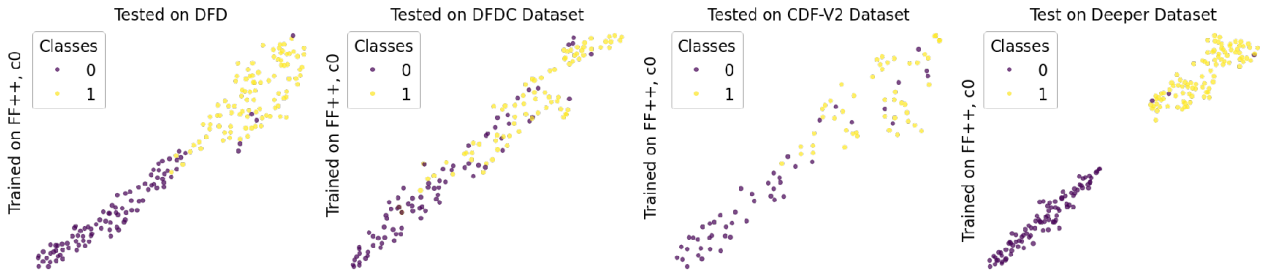

Figure 6. The t-SNE [47] visualization shows the learned feature space of predicted classes by our proposed deepfake video detection framework on cross-dataset evaluation. The golden dots represent the real class, and the purple dots indicate the altered class samples.

Further, we have observed that LipForensics [15], FTCN [16], MSVT [28], VDF [29], and Audio-DF [30], all attempt to utilize the temporal incoherence in deepfake videos to enhance the generalization detection capability. In contrast, our approach differs because we rely on two separate paradigms: temporal modeling to extract facial color variation effectively and coherence analysis through self-attention to comprehend the short-term inter-frame incoherence within TFP. On the other hand, the approaches mentioned employ complex networks or a temporal transformer for temporal information extraction to enhance long-term relationship learning, which increases computational complexity. Nevertheless, MSVT [28] has more extended input video frames and may function better on datasets like DFD [40] and CDF-V2 [42], where the head poses and lighting circumstances are even among frames.

Regardless, our approach offers several edges. Firstly, we achieve better cross-dataset performance on challenging datasets compared with MSVT [28], VDF [29], and Audio-DF [30], such as DFDC [41], where it is challenging to distinguish whether the incoherence is caused by deepfake cues, head pose shifts or lighting circumstances in elongate sequences. Secondly, our approach provides more reasonable interpretability as we individually oversee the temporal modeling and detection task in TCAN. Lastly, our model displays better intra-dataset performance against unseen manipulation sources and cross-dataset performance against unseen dataset source domains due to the fine-grain temporal incoherence learning ability of the transformer model based on the high-resolution fine and low-resolution coarse feature representation, resulting in better detection results on most datasets.

5.3. Ablation study

In the presented ablations, we scrutinize the robustness of our method against compression on FF++ [35] subsets and test the interoperability of the proposed TFP mechanism by adopting other state-of-the-art backbone transformer models.

5.3.1. Robustness analysis: impact of compression on the TCAN performance

The transmission of multimedia content through social media networks often goes through JPEG compression. Hence, robustness evaluation is a crucial scale to evaluate the performance of our proposed deepfake detector. By adjusting the quality factor of JPEG compression, we are able to control the level of perturbation in our robustness analysis. Our approach, TCAN, has demonstrated robustness in the results presented in Table 5. We assess the robustness of TCAN using the leave-one-out approach, following the methodology proposed by [28]. We employ the FF++ [35] dataset to train the TCAN model at compression level c0 and then test it at compression levels of c23 and c40.

At compression levels c23 and c40, our proposed TCAN framework performance drops by an approximate margin of 19% in comparison to the results stated in Table 3. The primary reason for the reduced detection performance is the lossy compression of test data, which discards some high-frequency image details and textures that contain forgery artifacts. Our TFP mechanism relies on the spatial information of facial videos, but compression makes it unable to extract the features as sensitively as those from the data at compression level c0. However, our proposed method, which includes temporal self-attention and feature fusion mechanisms, has been demonstrated to be positively influential in enhancing the robustness of the TCAN model.

Table 5. Ablation study: Robustness performance comparison of TCAN in terms of video-level AUC (%) across two compression levels—trained on the three subsets of FF++ [35] and tested on the remaining one.

Compression level	Trained on remaining three				Avg.
	DF	FS	F2F	NT	
c23	85.23	79.67	76.84	74.14	78.97
c40	70.21	73.16	70.59	69.28	70.81

5.3.2. Impact of others state of the art backbone model adaptation on performance

We conduct an ablation study to assess the performance impact of integrating our proposed TFP mechanism with various state-of-the-art backbone models, including ViT [20], RegNet [48], SwinV2 [49], BEiT [50], and PoolFormer (PF) [51]. The evaluation results are compared against our TCAN method in Table 6, where the FF++ [35] dataset is used for training at compression level c0 with two combinations for intra- and cross-dataset evaluation.

Table 6. Ablation study: Our proposed TFP mechanism adapts to other state-of-the-art backbone models, and its performance impact is analyzed. Results are compared with TCAN in terms of video-level AUC (%).

Methods	Trained on remaining three				Trained on all four			Avg.	#params.
	DF	FS	F2F	NT	DFDC	CDF-V2	Deeper		
ViT [20]	96.30	98.34	98.31	95.25	86.28	82.15	97.61	93.46	85.8 M
RegNet [48]	97.71	97.30	98.55	97.98	82.51	78.69	96.50	92.74	43.9 M
SwinV2 [49]	97.99	96.53	96.91	93.09	77.76	78.89	93.53	90.67	27.5 M
BEiT [50]	96.21	98.21	97.20	96.82	77.19	87.69	96.88	92.88	85.7 M
PF [51]	98.98	98.78	98.02	98.08	84.04	85.18	98.77	94.55	11.4 M
TCAN (Our)	99.16	98.93	99.03	98.15	86.94	88.99	98.80	95.71	**3.3 M**

Integrating our TFP mechanism with ViT [20] demonstrates a competitive AUC of 93.46%, showcasing its adaptability and effectiveness. Our TFP mechanism performs well when unified with RegNet [48], achieving a video-level AUC of 92.74%, indicating its compatibility and positive impact on diverse backbone architectures. Similarly, the SwinV2 [49] backbone, coupled with our TFP, achieves an AUC of 90.67%, affirming the versatility of our approach across different model architectures. Integrating our TFP with BEiT [50] results in a 92.88% AUC score, demonstrating its ability to learn TFP among various models. In one of the most promising second choices as a backbone network for coherence analysis, PF [51] showcases effective detection performance with an overall average AUC score of 94.55% in all categories of intra- and cross-dataset evaluations. In comparison, our TCAN method serves as the baseline, achieving the best and most competitive video-level AUC scores across all evaluated backbones, as shown in Table 6, emphasizing its effectiveness compared to existing methods. In short, our ablation study illustrates that the proposed TFP mechanism within our framework can be adapted to other backbone networks as well. Additionally, without any tailored network design for deepfake detection tasks, it consistently maintains its performance across various state-of-the-art

backbone models, emphasizing its adaptability and effectiveness in learning TFP representation for deepfake detection.

6. Conclusions and future work

In this paper, we have presented a novel bi-level paradigm for deepfake video detection, which consists of temporal modeling and coherence analysis. In level one, we have employed the temporal modeling mechanism for generalized detection, which decomposes real or deep fake videos into TFP containing subtle and dynamic local-global temporal clues from the facial forgery-sensitive areas based on color variations. The second level is based on the TCAN to mine and capture vital dynamic inconsistencies. TCAN first employs dynamic feature extraction and aggregation mechanisms to integrate multi-level temporal features effectively, and then analyzes the local-global temporal perspectives within TFP by utilizing global temporal self-attention for robust incoherence identification. Our proposed approach dramatically enhances robustness and generalized performance corresponding to prior deepfake video detection techniques, including current video transformers. The empirical validation of our approach is conducted on a range of large-scale deepfake datasets, such as intra- and cross-dataset evaluations on FF++ subsets, DFD, DFDC, CDF-V2, and Deeper. TCAN outperformed all its competitors in the intra- and cross-evaluations with comprehensive margins on the FF++ subsets (DF by 1.1%, FS by 1.5%, F2F by 1.01%, NT by 0.96%), DFDC by 1.84%, CDF-V2 by 0.18%, and Deeper by 0.38%, which further provide evidence of our proposed method's effectiveness. In future research, we will explore different color spaces to refine the temporal modeling design and incorporate audio modality for enhanced deepfake video detection. We will also devise a visualization strategy that can help us better comprehend the mechanics behind deepfake video detection by transformer models.

Use of AI tools declaration

The authors declare they have not used Artificial Intelligence (AI) tools in the creation of this article.

Acknowledgments

The Science and Technology Foundation of Guangzhou Huangpu Development District provided funding for this research work under Grant No. 2022GH15.

Conflict of interest

The authors declare there is no conflict of interest.

References

1. M. Kowalski, Deepfakes. Available from: https://www.github.com/MarekKowalski/FaceSwap/.
2. K. Liu, I. Perov, D. Gao, N. Chervoniy, W. Zhou, W. Zhang, Deepfacelab: integrated, flexible and extensible face-swapping framework, *Pattern Recognit.*, **141** (2023), 109628. https://doi.org/10.1016/j.patcog.2023.109628

3. D. Afchar, V. Nozick, J. Yamagishi, I. Echizen, MesoNet: a compact facial video forgery detection network, in *2018 IEEE International Workshop on Information Forensics and Security (WIFS)*, (2018), 1–7. https://doi.org/10.1109/WIFS.2018.8630761
4. F. Matern, C. Riess, M. Stamminger, Exploiting visual artifacts to expose deepfakes and face manipulations, in *2019 IEEE Winter Applications of Computer Vision Workshops (WACVW)*, (2019), 83–92. https://doi.org/10.1109/WACVW.2019.00020
5. Y. Qian, G. Yin, L. Sheng, Z. Chen, J. Shao, Thinking in frequency: face forgery detection by mining frequency-aware clues, in *ECCV 2020: Computer Vision – ECCV 2020*, Springer-Verlag, (2020), 86–103. https://doi.org/10.1007/978-3-030-58610-2_6
6. H. Liu, X. Li, W. Zhou, Y. Chen, Y. He, H. Xue, et al., Spatial-phase shallow learning: rethinking face forgery detection in bfrequency domain, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, (2021), 772–781.
7. S. Chen, T. Yao, Y. Chen, S. Ding, J. Li, R. Ji, Local relation learning for face forgery detection, in *Proceedings of the AAAI Conference on Artificial Intelligence*, **35** (2021), 1081–1088. https://doi.org/10.1609/aaai.v35i2.16193
8. Q. Gu, S. Chen, T. Yao, Y. Chen, S. Ding, R. Yi, Exploiting fine-grained face forgery clues via progressive enhancement learning, in *Proceedings of the AAAI Conference on Artificial Intelligence*, **36** (2022), 735–743. https://doi.org/10.1609/aaai.v36i1.19954
9. X. Li, Y. Lang, Y. Chen, X. Mao, Y. He, S. Wang, et al., Sharp multiple instance learning for DeepFake video detection, in *Proceedings of the 28th ACM International Conference on Multimedia*, (2020), 1864–1872. https://doi.org/10.1145/3394171.3414034
10. Z. Gu, Y. Chen, T. Yao, S. Ding, J. Li, F. Huang, et al., Spatiotemporal inconsistency learning for DeepFake video detection, in *Proceedings of the 29th ACM International Conference on Multimedia*, (2021), 3473–3481. https://doi.org/10.1145/3474085.3475508
11. S. A. Khan, H. Dai, Video transformer for deepfake detection with incremental learning, in *Proceedings of the 29th ACM International Conference on Multimedia*, (2021), 1821–1828. https://doi.org/10.1145/3474085.3475332
12. D. H. Choi, H. J. Lee, S. Lee, J. U. Kim, Y. M. Ro, Fake video detection with certainty-based attention network, in *2020 IEEE International Conference on Image Processing (ICIP)*, (2020), 823–827. https://doi.org/10.1109/ICIP40778.2020.9190655
13. E. Sabir, J. Cheng, A. Jaiswal, W. Abdalmageed, I. Masi, P. Natarajan, Recurrent convolutional strategies for face manipulation detection in videos, *Interfaces (GUI)*, (2019), 80–87.
14. A. Chintha, B. Thai, S. J. Sohrawardi, K. Bhatt, A. Hickerson, M. Wright, et al., Recurrent convolutional structures for audio spoof and video deepfake detection, *IEEE J. Sel. Top. Signal Process.*, **14** (2020), 1024–1037. https://doi.org/10.1109/JSTSP.2020.2999185
15. A. Haliassos, K. Vougioukas, S. Petridis, M. Pantic, Lips don't lie: a generalisable and robust approach to face forgery detection, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, (2021), 5039–5049.

16. Y. Zheng, J. Bao, D. Chen, M. Zeng, F. Wen, Exploring temporal coherence for more general video face forgery detection, in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, (2021), 15044–15054.

17. Z. Gu, Y. Chen, T. Yao, S. Ding, J. Li, L. Ma, Delving into the local: dynamic inconsistency learning for DeepFake video detection, in *Proceedings of the AAAI Conference on Artificial Intelligence*, **36** (2022), 744–752. https://doi.org/10.1609/aaai.v36i1.19955

18. X. Zhao, Y. Yu, R. Ni, Y. Zhao, Exploring complementarity of global and local spatiotemporal information for fake face video detection, in *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, (2022), 2884–2888. https://doi.org/10.1109/ICASSP43922.2022.9746061

19. R. Shao, T. Wu, Z. Liu, Detecting and recovering sequential DeepFake manipulation, in *ECCV 2022: Computer Vision – ECCV 2022*, Springer-Verlag, (2022), 712–728. https://doi.org/10.1007/978-3-031-19778-9_41

20. A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, et al., An image is worth 16 x 16 words: transformers for image recognition at scale, preprint, arXiv:2010.11929.

21. A. Arnab, M. Dehghani, G. Heigold, C. Sun, M. Lučić, C. Schmid, ViViT: a video vision transformer, in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, (2021), 6836–6846.

22. Y. Zhang, X. Li, C. Liu, B. Shuai, Y. Zhu, B. Brattoli, et al., VidTr: Video transformer without convolutions, in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, (2021), 13577–13587.

23. L. He, Q. Zhou, X. Li, L. Niu, G. Cheng, X. Li, et al., End-to-end video object detection with spatial-temporal transformers, in *Proceedings of the 29th ACM International Conference on Multimedia*, (2021), 1507–1516. https://doi.org/10.1145/3474085.3475285

24. Z. Xu, D. Chen, K. Wei, C. Deng, H. Xue, HiSA: Hierarchically semantic associating for video temporal grounding, *IEEE Trans. Image Process.*, **31** (2022), 5178–5188. https://doi.org/10.1109/TIP.2022.3191841

25. O. de Lima, S. Franklin, S. Basu, B. Karwoski, A. George, Deepfake detection using spatiotemporal convolutional networks, preprint, arXiv:2006.14749.

26. D. Güera, E. J. Delp, Deepfake video detection using recurrent neural networks, in *2018 15th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*, (2018), 1–6. https://doi.org/10.1109/AVSS.2018.8639163

27. I. Masi, A. Killekar, R. M. Mascarenhas, S. P. Gurudatt, W. AbdAlmageed, Two-branch recurrent network for isolating deepfakes in videos, in *ECCV 2020: Computer Vision – ECCV 2020*, Springer-Verlag, (2020), 667–684. https://doi.org/10.1007/978-3-030-58571-6_39

28. Y. Yu, R. Ni, Y. Zhao, S. Yang, F. Xia, N. Jiang, et al., MSVT: Multiple spatiotemporal views transformer for DeepFake video detection, *IEEE Trans. Circuits Syst. Video Technol.*, **33** (2023), 4462–4471. https://doi.org/10.1109/TCSVT.2023.3281448

29. H. Cheng, Y. Guo, T. Wang, Q. Li, X. Chang, L. Nie, Voice-face homogeneity tells deepfake, *ACM Trans. Multimedia Comput. Commun. Appl.*, **20** (2023), 1–22. https://doi.org/10.1145/3625231

30. W. Yang, X. Zhou, Z. Chen, B. Guo, Z. Ba, Z. Xia, et al., AVoiD-DF: Audio-visual joint learning for detecting deepfake, *IEEE Trans. Inf. Forensics Secur.*, **18** (2023), 2015–2029. https://doi.org/10.1109/TIFS.2023.3262148

31. M. Liu, J. Wang, X. Qian, H. Li, Audio-visual temporal forgery detection using embedding-level fusion and multi-dimensional contrastive loss, *IEEE Trans. Circuits Syst. Video Technol.*, 2023. https://doi.org/10.1109/TCSVT.2023.3326694

32. Q. Yin, W. Lu, B. Li, J. Huang, Dynamic difference learning with spatio–temporal correlation for deepfake video detection, *IEEE Trans. Inf. Forensics Secur.*, **18** (2023), 4046–4058. https://doi.org/10.1109/TIFS.2023.3290752

33. Y. Wang, C. Peng, D. Liu, N. Wang, X. Gao, Spatial-temporal frequency forgery clue for video forgery detection in VIS and NIR scenario, *IEEE Trans. Circuits Syst. Video Technol.*, **33** (2023), 7943–7956. https://doi.org/10.1109/TCSVT.2023.3281475

34. D. E. King, Dlib-ml: A machine learning toolkit, *J. Mach. Learn. Res.*, **10** (2009), 1755–1758. Available from: https://www.jmlr.org/papers/volume10/king09a/king09a.pdf.

35. A. Rossler, D. Cozzolino, L. Verdoliva, C. Riess, J. Thies, M. Niessner, FaceForensics++: Learning to detect manipulated facial images, in *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, (2019), 1–11. https://doi.org/10.1109/ICCV.2019.00009

36. E. Xie, W. Wang, Z. Yu, A. Anandkumar, J. M. Alvarez, P. Luo, SegFormer: Simple and efficient design for semantic segmentation with transformers, in *Advances in Neural Information Processing Systems*, **34** (2021), 12077–12090.

37. W. Wang, E. Xie, X. Li, D. P. Fan, K. Song, D. Liang, et al., Pyramid vision transformer: a versatile backbone for dense prediction without convolutions, in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, (2021), 568–578.

38. X. Chu, Z. Tian, B. Zhang, X. Wan, C. Shen, Conditional positional encodings for vision transformers, preprint, arXiv:2102.10882.

39. M. A. Islam, S. Jia, N. D. B. Bruce, How much position information do convolutional neural networks encode? preprint, arXiv:2001.08248.

40. N. Dufour, A. Gully, P. Karlsson, A. V. Vorbyov, T. Leung, J. Childs, et al., Contributing data to Deepfake detection research by Google Research & Jigsaw, 2019. Available from: https://blog.research.google/2019/09/contributing-data-to-deepfake-detection.html.

41. B. Dolhansky, J. Bitton, B. Pflaum, J. Lu, R. Howes, M. Wang, et al., The DeepFake detection challenge (DFDC) dataset, preprint, arXiv:2006.07397.

42. Y. Li, X. Yang, P. Sun, H. Qi, S. Lyu, Celeb-DF: A large-scale challenging dataset for DeepFake forensics, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, (2020), 3207–3216.

43. L. Jiang, R. Li, W. Wu, C. Qian, C. C. Loy, 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 (CVPR)*, (2020), 2889–2898.

44. A. P. Bradley, The use of the area under the ROC curve in the evaluation of machine learning algorithms, *Pattern Recognit.*, **30** (1997), 1145–1159. https://doi.org/10.1016/S0031-3203(96)00142-2

45. P. Micikevicius, S. Narang, J. Alben, G. F. Diamos, E. Elsen, D. Garcia, et al., Mixed precision training, preprint, arXiv:1710.03740.

46. Z. Zhang, M. R. Sabuncu, Generalized cross entropy loss for training deep neural networks with noisy labels, in *Advances in Neural Information Processing Systems*, **31** (2018).

47. L. van der Maaten, G. Hinton, Visualizing data using t-SNE, *J. Mach. Learn. Res.*, **9** (2008), 2579–2605.

48. I. Radosavovic, R. P. Kosaraju, R. Girshick, K. He, P. Dollár, Designing network design spaces, in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, (2020), 10425–10433. https://doi.org/10.1109/CVPR42600.2020.01044

49. Z. Liu, H. Hu, Y. Lin, Z. Yao, Z. Xie, Y. Wei, et al., Swin transformer v2: Scaling up capacity and resolution, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, (2022), 12009–12019.

50. H. Bao, L. Dong, S. Piao, F. Wei, BEiT: BERT pre-training of image transformers, preprint, arXiv:2106.08254.

51. W. Yu, M. Luo, P. Zhou, C. Si, Y. Zhou, X. Wang, et al., Metaformer is actually what you need for vision, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, (2022), 10819–10829.

AIMS Press