HolisticDFD: Infusing Spatiotemporal Transformer Embeddings for Deepfake Detection

Muhammad Anas Raza, Khalid Mahmood Malik*, Ijaz Ul Haq

[a] *Department of Computer Science, Oakland University, Rochester, 48309, MI, USA*

Abstract

Deepfakes, or synthetic audiovisual media developed with the intent to deceive, are growing increasingly prevalent. Existing methods, employed independently as images/patches or jointly as tubelets, have, up to this point, typically focused on spatial or spatiotemporal inconsistencies. However, the evolving nature of deepfakes demands a holistic approach. Inspection of a given multimedia sample with the intent to validate its authenticity, without adding significant computational overhead has, to date, not been fully explored in the literature. In addition, no work has been done on the impact of different inconsistency dimensions in a single framework. This paper tackles the deepfake detection problem holistically. HolisticDFD, a novel, transformer-based, deepfake detection method which is both lightweight and compact, intelligently combines embeddings from the spatial, temporal and spatiotemporal dimensions to separate deepfakes from bonafide videos. The proposed system achieves 0.926 AUC on the DFDC dataset using just 3% of the parameters used by state-of-the-art detectors. An evaluation against other datasets shows the efficacy of the proposed framework, and an ablation study shows that the performance of the system gradually improves as embeddings with different data representations are combined. An implementation of the proposed model is available at: https://github.com/smileslab/deepfake-detection/.

Keywords: Deepfake Detection, Intermediate Fusion, Multimedia Forensics, Transformers

1. Introduction

The rapid evolution of generative AI algorithms has led directly to an increase in cyber threats in the form of synthetic media, which may take various forms, including deepfakes, as shown in Figure 1, Javed et al. (2021). Deepfake videos, created to spread outright lies, may damage public perception of the target of

*Corresponding author

Email address: mahmood@oakland.edu (Khalid Mahmood Malik)

such an attack, e.g., a political leader. Taken to its logical extent, they may be used to undermine and destabilize governments. Moreover, malicious actors have been known to use deepfakes, distributed under false profiles, to disseminate disinformation on social media. Belief in convincing deepfake-created content, fostered by the development of easily available deepfake generation tools,has jeopardized the reputations of celebrities and world leaders, who are often the targets of such attacks. Deepfakes have also been used to finance phishing schemes, fund fake charities, and foment credit card fraud. More recently, bad-actors have combined deepfakes and shallow-fakes into complex forgeries to evade existing tools. A marked increase in the availability of open-source implementations for deepfake creation and tremendous improvements to generative algorithms, e.g., autocoders (AE), generative adversarial networks (GAN), and diffusion models, for deepfake generation, has made it possible for users with no knowledge of machine learning to generate exceptionally believable deepfakes. These are growing increasingly difficult for an average person to spot on social media because modern-day, sophisticated techniques are good enough to fool an uninformed public. It is therefore imperative that the potential damage caused by this new generation of deepfakes be curbed.

The research community is in active pursuit of tools and techniques to counteract the threat of media falsification and the mass spread of disinformation Khan et al. (2022); Khalid et al. (2023). However, detecting whether a video, audio or image is original or forged is a continuously evolving task. Initial efforts used hand-crafted features which effectively detected early versions of deepfakes and used discrepancies in head pose, eye-blinking, and face-warping artifacts. Generative algorithms, including recent developments such as the advent of image/video diffusion models, have greatly improved image/video synthesis Dhariwal and Nichol (2021). The resultant advances have increased the quality of deepfakes over time, and have rendered previously effective methods useless.

Performant algorithms for video deepfake detection may be classified into two categories, defined by the irregularities they focus on: spatial, or image-based feature exploitation; and spatiotemporal, or video-based features. Image-based methods focus only on spatial cues in individual frames and ignore temporal oddities. Detectors based on spatial anomalies analyze each frame in order to classify real and fake images. However, recent generative approaches are capable of synthesizing highly photo-realistic frames that do not have spatial inconsistencies. This causes image-based detection approaches, though previously effective, to perform poorly on modern deepfakes. In contrast, video-based feature extraction methods focus on sequence patterns and explore spatiotemporal inconsistencies to detect deepfakes. These techniques, however, do not detect inconsistencies that are distributed dynamically in multiple local regions within frames. Though both approaches have had some success, a deepfake detection technique that dynamically learns and automatically adjusts

Figure 1: Samples from the FaceForensics++ (top row), DFDC (middle row), and Celeb-Df (Bottom row) datasets. Forgery techniques in each dataset are different i.e., FaceForensics++ has DeepFakes, Face2Face, FaceSwap and NeuralTextures. DFDC and Celeb-DF chiefly utilizes Faceswap. As evident in the samples, these forgeries introduce spatial (incomplete glasses, beard) and temporal artifacts (missing beard between consecutive frames).

the weights of spatial or temporal features, while analyzing the inconsistency patterns in synthetic video, has not been attempted. We argue that variation in deepfake generation methods has a key impact on the performance of detection algorithms. In Figure 1, the artifacts in the deepfake samples generated with various forgery techniques can be grouped into those with spatial and temporal anomalies. These can then be exploited for detecting suspected deepfakes. As generative algorithms evolve to produce more realistic results, it is imperative to define a holistic approach that exploits spatial, spatiotemporal, and temporal features to detect deepfakes.

On the algorithmic side, transformers, well-known for their dominance in natural language processing, are increasingly being applied to vision tasks, and deepfake detection is no exception. For instance, Coccomini et al. (2022); Heo et al. (2021) use a transformer architecture for this problem. Similar approaches use large complex structures and a multitude of parameters. They focus either on the spatial or temporal inconsistencies extant in deepfakes. A deepfake may have all spatial artifacts, e.g., in one frame there may be a discrepancy in eye color, temporal artifacts, e.g., an unnatural transition between frames, or a joint spatiotemporal anomaly between frames, and any of these may be used to classify the sample as fake. However, a synthesis of spatial and spatiotemporal pattern mining for deepfake detection has not been extensively explored. In

addition, existing methods use models pre-trained for other tasks, or a computationally expensive, pure transformer-based architecture which requires a large dataset. In Hassani et al. (2021), the authors suggest that combining the benefits of convolution and transformers in a single network architecture may result in a more robust application. To address the challenges discussed above, this work proposes a novel framework that combines spatial, temporal, and spatiotemporal features. It accomplishes detection using a smaller parameter set than existing methods, and offers the following major contributions:

1. A novel multi-dimensional Model Infused Deepfake Detection method (HolisticDFD) is proposed that fuse embeddings independently as well as jointly from frozen models on the spatial and temporal dimensions of a video sample. The individual models are pre-trained to independently learn a single deepfake inconsistency dimension, and the proposed method combines different views of the same video sequence in a joint space.

2. Unlike existing methods that analyze a specific region of the subject (e.g., eyebrow movement), the proposed framework takes a holistic view, using sequence pooling technique to fuse the embeddings from spatial, temporal, and spatiotemporal data representations of a suspected deepfake. This method focuses on patterns extracted from potential forgery regions compared to other areas and weights them accordingly.

3. Unlike existing methods which either employ transfer learning or knowledge distillation, a compact transformer-based deepfake detection method is proposed which uses just 3% of the parameters required by state-of-the-art (SoTA) models.

The rest of the paper is organized as follows: Current deepfake generation and detection methods are reviewed in Section 2. Section 3 describes the research questions we attempt to answer and formalizes the problem of deepfake detection. The technical details of the proposed deepfake detection framework are described in Section 4. Section 5 is dedicated to the experimental results and analysis, followed by conclusions and future work in Section 6.

2. Related Work

A number of algorithms have been developed for deepfake generation and detection. This section provides a brief overview of the techniques currently in use for deepfake video.

2.1. Deepfake Generation

Deepfakes have a long history, dating to 1997 when a program was developed that could alter video footage. It could create new content, making it appear as if the individual in the video said words that were in the source clip. However, the

first “well-known ‘deepfake’” appeared in September 2017 when a forged pornographic content of known actresses was released. In the following years, two approaches to create genuine faces in a deepfake: Variational Auto Encoders, commonly called VAEs, and Generative Adversarial trained Networks (GANs) gained popularity for deepfake generation. VAEs employ two encoder- decoder sets, individually trained to encode and decode the faces to be exchanged. Faces are then encoded in a latent distribution by the encoder. In the last step, the decoder synthesizes the target face.

In contrast, GANs utilize a different technique, also composed of two steps. The first, a discriminator, which distinguishes real and synthetic data. The other component, the generator, modifies the input sample in an attempt to deceive the discriminator. Convincing output requires multiple iterations of this process, but very realistic forgeries have been attained with GANs. Several GAN-based methods have been invented, e.g., StarGAN Choi et al. (2018), and top results have been achieved through StyleGAN-V2 Karras et al. (2020). GANs support manipulation, such as image restoration and style transfer. Neither of these can be achieved by classical forgery generation methods. The next generation of diffusion models are being developed which outperform these individual approaches. GANs have significant computational requirements and require large datasets for training, which has limited their availability to the general public.

Regardless of the technique used to carry out the manipulation, deepfake generation approaches may be categorized by the specific way in which the image is modified Masood et al. (2022). Deepfakes may be loosely grouped into the following categories:

- **Facial Synthesis** uses latent representations of the facial datasets for generating a hyper-realistic “person.” The resulting image is not a representation of a real person, in whole or in part, as it is synthesized without a target subject.
- **Facial Transfer** transfers both identity-aware and identity-agnostic content (e.g., expression and pose) from a source face to the target face.
- **Facial Swapping** transfers the identity of the source face to the target face while preserving identity-agnostic content.
- **Facial Stacked Manipulation (FSM)** is a set of methodologies that transfer both the identity and the attributes of the target to the source, while others alter the attributes of the swapped target after the transfer of the identity.
- **Facial Reenactment** preserves the identity of the source subject but manipulates intrinsic attributes such as mouth or expression.
- **Facial Editing** modifies external attributes such as age, gender, or ethnicity.

Recently, audiovisual deepfake datasets, where the audio and images have both been manipulated, have been presented to the research community Khalid et al. (2021). Thus, it may be inferred that deepfake generation technologies are evolving faster than detection technologies and even more realistic content is expected to be seen in the future.

2.2. Deepfake Detection

A deepfake detection method has recently gained widespread attention across the globe that finds traces of forgeries by exploiting the contextual and semantic understanding of the data. Researchers in the area of multimedia forensics have provided a wide range indicators to be used when spotting fake media, including but not limited to: face wobble, distortion and shimmer, waviness in movement, inconsistency in facial movement and speech, anomalies in the movement of an object between frames, inconsistency in lighting, shadows and reflections, blurred edges, abnormal facial angles and feature blurring, breathing patterns, eye direction inconsistencies, missing facial features, e.g., a known cheek mole, weight and softness of hair and clothing, over smoothness of skin, missing teeth and hair details, lack of alignment in facial symmetry, inconsistency in pixel contours, and strange or implausible behavior Masood et al. (2022).

• **Frame-based detection methods:** Deepfake detection using frame-based models is less complex and focuses on blending defects. For instance, Jia et al. (2021) developed a two-branch multi-task learning framework based on within-image and between-frame inconsistencies for classifying single frame. Zhao et al. (2021) regarded deepfake detection as a fine-grained classification task, and constructed a multi-attentional network to focus on local discriminative features from multiple face attentive regions. The method in Coccomini et al. (2022) achieved competitive results by joining Transformers to a pre-trained convolutional network. Heo et al. (2021), improved the performance of the network by using distillation in Vision Transformer from a pre-trained EfficientNet-B7. Xia et al. (2022) leveraged textural disparities in facial images from multi-color channels for detecting forged multimedia. Tian et al. (2023) developed a frequency-aware contrastive approach to differentiate deepfakes from real, and Liang et al. (2023) used facial geometry analysis along with a CNN-LSTM network for the same purpose. Lastly, Wang et al. (2022) developed a fast landmark-based method employing feature point defects.

• **Sequence-based detection:** Sequence based methods use a stack of frames for detecting a suspected deepfake. Nguyen et al. (2021) propose a 3D CNN that learns spatiotemporal features from sequence of video. In Hu et al. (2021), a two-stream detector detects fakes by analyzing the frame-level and temporal artifacts of compressed videos. A more recent study, Gu et al. (2021), proposes spatial and temporal inconsistency learning using separate spatial and temporal modules for deepfake detection. Another method, Chen et al. (2022), uses spatiotemporal attention and a convolutional LSTM for tackling deepfakes. Yang et al. (2023) proposes a graph relation-based approach used for spatiotemporal deepfake detection, and finally, Zhao et al. (2023) uses a spatiotemporal video vision transformer for performant deepfake detection.

Patch-based solutions: The arrival of transformers in vision has seen increased use of patches for the task of recognizing fabricated faces. Li et al. (2020b) crop facial regions and no-face regions into different patches and use a dual-branch learning framework to distinguish between bonafide and forged facial patches. This method also detects inconsistencies in the facial and background regions. Similarly, Zhao et al. (2020) hypothesizes that source artifacts are still there after the original image is forged. Using real-video features and monitoring consistency, the forged images can be found. In Chai et al. (2020), a convolutional patch-based classifier is developed which gives predictions (bonafide or forged) over moving patches of a target image. Schwarcz and Chellappa (2021) employ facial parsing and develop separate patch-based detectors on truncated outputs, while Heo et al. (2023) use concatenated CNN features, along with patch-based positioning, to specifying the artifact regions in the face.

Summarizing existing work, frame-, sequence-, and patch- based models each have pros and cons. Frame-based approaches for deepfake detection have received more attention in the literature compared to spatial and temporal artifact- based methods Gu et al. (2021); Zhao et al. (2023); Haiwei et al. (2022). However, a framework that combines the strengths of all three approaches in a lightweight manner has not been explored. Therefore, the proposed framework integrates the strengths of each approach using convolutions, vision transformers, and a unique sequence-pooling technique. Our framework uses transformer encoders with patches extracted in frames and sequences. Differences between consecutive frames are exploited, allowing a holistic analysis of deepfake video. By doing this, the proposed framework is able to capture both the spatial and temporal features of the video, improving the accuracy of deepfake detection in a lightweight manner.

3. Problem Definition and Formulation

The tremendous success and continuously evolving deepfake generation technology can create content that is indistinguishable from reality to the human eye. Several approaches exist that describe deepfake detection as classification problem and employ either spatial or spatiotemporal artifacts for classing. However, no work has been done which studies the effect of spatial, temporal, and spatiotemporal artifacts together. Therefore, this paper attempts to answer the following research questions:

- Does the holistic approach, i.e., using spatial, temporal, and spatiotemporal features together, give better detection accuracy than these features individually?
- How does the holistic detection approach perform in terms of area under curve (AUC) and accuracy on multiple datasets and across corpora?
- How does holistic approach perform when it processes the spatial, temporal, or spatiotemporal information of a given video individually vs. as one entity to classify it as real or deepfake?

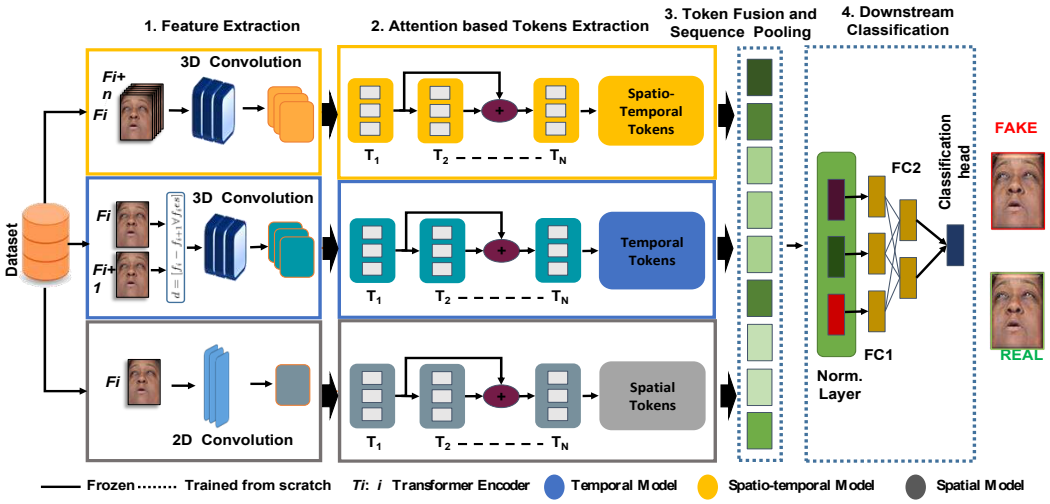

Figure 2: A graphical representation of the proposed method. A suspected deepfake is fed to the proposed framework. Spatial, Temporal and Spatiotemporal Features, extracted from the images and sequences in step 1, are fed to the attention-based token extraction. This is then fused and sequence-pooled in step 3. Finally, the fused embeddings from suspected deepfakes are fed to a classification head in step 4 to give the final prediction.

To answer the above questions, we formulate deepfake detection as a binary classification problem of real and fake sequences. A given video V is divided into n sequences $\{S_1, S_2, ..., S_n\}$ and each sequence composed of 30 frames of dimensions 224 × 224. Each sequence S_i is fed to the HolisticDFD framework fn as follows:

$$prob = fn(S_i) \tag{1}$$

The proposed framework returns a probability, prob, in $[0, 1]$. The determination if a video is real or a deepfake is done with the following decision function:

$$f(prob) = \begin{cases} fake, & if\ prob \geq 0.5 \\ real, & otherwise \end{cases}$$

4. Proposed Method

This section provides a detailed overview of the proposed HolisticDFD framework, which is divided into four stages. The first stage extracts feature maps from frozen spatial, temporal, and spatiotemporal models. In step two, frozen attention-based tokenization is performed in order to extract self-attended embeddings. Step three concatenates the tokens and performs pooling, and finally, a downstream classifier is trained to identify bonafide and forged video in Step 4. A pictorial representation of the proposed framework is shown in Figure 2 and the pseudo-code of the complete framework is presented in Algorithm 1. A detailed discussion of each step is presented in the following subsections.

4.1. Feature Extraction

In this section, we provide the details of feature extraction for the spatial, temporal, and spatiotemporal features using the Compact Convolutional Transformer

Algorithm 1 Model Infusion for Deepfake Detection

Input: Sequence of frames *f*, spatial network *sn*, temporal network *t*, spatiotemporal network k, sequence pooling layer s_pool, classifier MLP

Output: Probability fakeness *pred*

images ← flattened sequence S
diff ← s[:, 1:,] - s[:, :-1] ▷ *s is of shape [batch, time, height, width, channels]*
k_features ← k(seq)
s_f = sn(images)
t_f ← t(diff)
c_f ← conc.(s_f, t_f, k_f)
pool ← s pool(c_f)
pred ← MLP(pool)

method described by Hassani et al. (2021). The spatial features make use of 2D Convolution and 2D Max Pooling and the temporal and spatiotemporal features use 3D Convolution and 3D Max Pooling layers. The convolution layers, followed by pooling, significantly reduce the number of required parameters for the model and the number of features extracted from images and sequences making the entire architecture extremely lightweight.

4.2. Feature Extraction

In this section, we provide the details of feature extraction for the spatial, temporal, and spatiotemporal features using the Compact Convolutional Transformer method described by Hassani et al. (2021). The spatial features make use of 2D Convolution and 2D Max Pooling and the temporal and spatiotemporal features use 3D Convolution and 3D Max Pooling layers. The convolution layers, followed by pooling, significantly reduce the number of required parameters for the model and the number of features extracted from images and sequences making the entire architecture extremely lightweight.

4.2.1. Spatial Feature Extraction

Spatial features are extracted from video frames in order to detect image level discrepancies, e.g., irregularities in eye-color, skin tone, or texture. When extracting features which focus on spatial inconsistencies in the input images, a 2D Convolution operation is performed on a frame x_i. The obtained feature maps are passed to 2D Max Pooling layers. Equation 2 shows the spatial patch extraction performed on a batch of images x.

$$fm_s = MaxPool2D\left(ReLU\left(Conv2d(x)\right)\right) \quad (2)$$

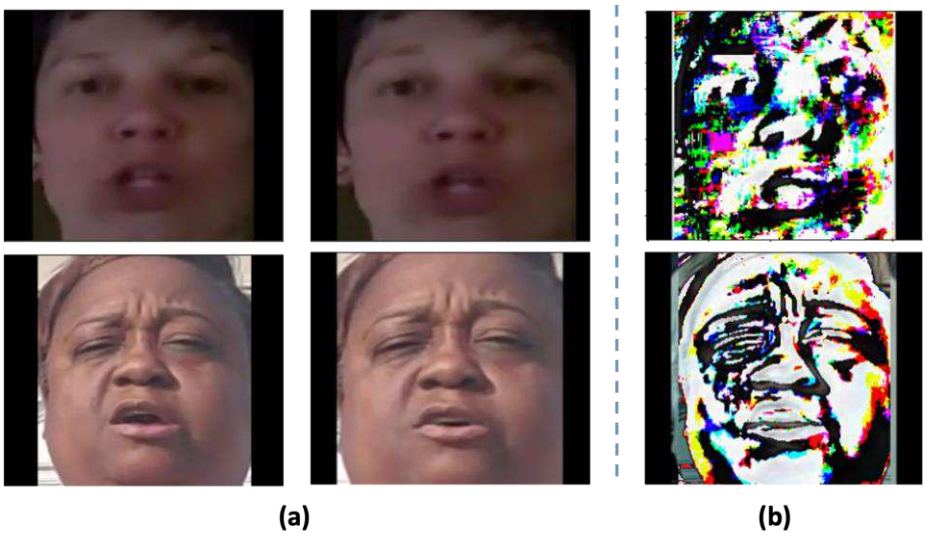

Figure 3: Differential Module used in the Temporal Transformer. (a) consecutive frames from DFDC, (b) the resultant differential image.

4.2.2. Temporal Feature Extraction

In order to detect irregularities in the transition between frames, temporal features are calculated. We adopted a differential module for detecting and classifying moving targets in real-time video samples from Lipton et al. (1998). Unlike raw feature extraction, the differential module makes the model insensitive to changes in illumination. A differential module, *d*, is introduced which computes the difference between consecutive frames, f_i and f_{i+1}, for all frames in sequence x. Figure 3 demonstrates the resultant differential image from sequential frames obtained using Equation 3.

$$d = [f_i - f_{i+1} \; \forall \; f_i \in S] \tag{3}$$

where S is the sequence of frames. The differential d is then fed to a 3D Convolution followed by 3D Max Pooling to extract the features which focus on temporal inconsistencies, as given in Equation 4.

$$fm_t = MaxPool3D\left(ReLU\left(Conv3d(d)\right)\right) \tag{4}$$

where fm_t are the resultant temporal feature maps. Conv3D, ReLU, and MaxPool3D are the same operations as in the spatial feature extraction process but in this case, they are applied to a sequence of images or tubelets instead of 2D images.

4.2.3. Spatiotemporal Feature Extraction

Spatiotemporal features are extracted to identify discrepancies in both images and transitions between consecutive frames. The sequence of frames, S, is directly passed to a 3D Convolution Layer and 3D Max Pooling, as given in Equation 5.

$$fm_{st} = MaxPool3D\left(ReLU\left(Conv3d(S)\right)\right) \quad (5)$$

where fm_s, fm_t, fm_{st} are the resultant spatiotemporal feature maps. The same functions (Conv3D, ReLU and MaxPool3D) are again performed, but in this case, they are applied directly to a sequence of images or tubelets without passing them to the differential module.

4.3. Attention based Tokens Extraction

Multi-head, self-attention-based Transformer encoders are used to provide a wider range of context for the features obtained in the feature extraction layer. The extracted f_{ms}, f_{mt}, and f_{mst} feature maps are fed into a separate, attention-based architecture comprised of a stack of transformer encoders. Each encoder layer contains two sub-layers: A Multi-head Self-Attention Layer (MSA) and a Multi-Layer Perceptron (MLP). Layer normalization (LN) is performed on the input features before feeding them to the MSA, as shown in Equation 6, along with a residual connection to the inputs, Equation 7, followed by a residual connection.

$$z_l' = MSA(LN(z_{l-1}) + z_{l-1}) \quad (6)$$

$$z_l = MLP(LN(z_l') + z_l') \quad (7)$$

$$y = LN(LN(z_l') + z_l) \quad (8)$$

The respective inconsistency dimension features, spatial y_S, temporal y_T, and spatiotemporal y_{ST}, are received from their respective networks as shown in Equation 8, where the MSA allows the model to jointly attend the information from different representation subspaces. Multi-Head Attention is defined as:

$$MultiHead(Q, K, V) = Conc.(head_1, \dots, head_h)W^O \quad (9)$$

$$head_i = Attention\left(QW_i^Q, KW_i^K, VW_i^V\right) \quad (10)$$

where Q, K, and V are the query, key, and value, respectively.

4.4. Feature Fusion and Sequence Pooling

The spatial, temporal and spatiotemporal tokens, taken from the transformer layers, are concatenated to form a single vector, v, as given in Equation 11.

$$v = Conc.(y_S, y_T, y_{ST}) \quad (11)$$

where y_S, y_T, y_{ST} are the tokens from spatial, temporal and spatiotemporal

encoders. For concatenation purposes, the output vectors from the transformer encoders all have the same shape. The concatenated tokens, *v*, from Equation 11 are then fed to a linear layer $g(v) \epsilon R^{d \times 1}$ followed by Softmax activation, as given in 12,

$$X_L = softmax(g(v)^T) \epsilon R^{b \times \mathbb{1} \times n} \tag{12}$$

The attention-weights, X_L, obtained from Equation 12 are multiplied with *v*, as shown in Equation 13.

$$z = X_L \times v \epsilon R^{b \times 1 \times d} \tag{13}$$

After pooling, the second dimension, $z \epsilon R^{b \times d}$, is squeezed. Function *g(v)* applies a linear transformation to the incoming data.

$$y = W v^T + B \tag{14}$$

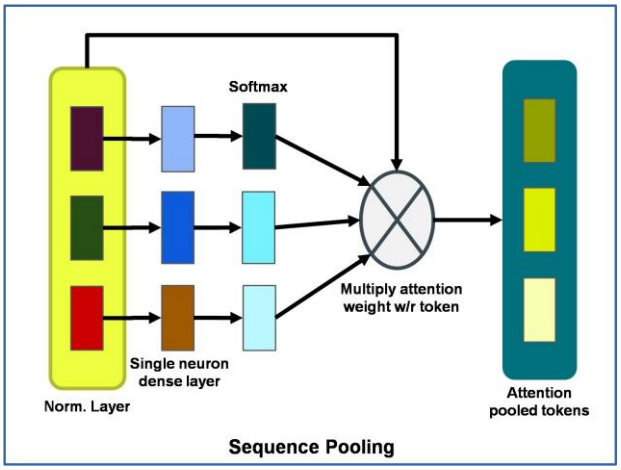

Figure 4: The architectural flow for sequence pooling, where normalized tokens are passed to a single neuron dense layer followed by softmax activation. The input tokens are passed through a single-neuron neural network which is then softmaxed and attention is multiplied with the input tokens.

where *W* is a weight matrix and *B* is the bias matrix, while the softmax function rescales an n-dimensional feature vector so that all elements of the n-dimensional tensor lie in the range of [0,1] and the sum is equal to 1 along a particular axis, as in 15

$$Softmax(x_i) = \frac{exp(x_j)}{\sum_j exp(x_j)} \tag{15}$$

As given in Figure 4, sequence pooling allows the network to weigh the sequentialtokens in the latent space produced by the transformer encoder and correlates data across the input tokens. The sequence pooling module attends to the sequential data in order to assign attention weights across the sequence of

tokens. By employing softmax activation to the attended tokens, the proposed model gives higher attention-weights to tokens that have more information related to downstream classification.

4.5. Downstream Classification

The output from the sequence pooling layers is fed to a fully-connected linear classifier, essentially a multi-layer perceptron (MLP), which encodes the features using hidden layers and gives a prediction in the last layer. Dropout and L2 regularization are used for better generalization.

$$Classifier(x_i) = \sum_i w_i x_i + bias_i \tag{16}$$

Sigmoid activation is used for binary classification, which has a characteristic S-shaped curve.

$$S(x) = \frac{1}{1+e^{-x}} \tag{17}$$

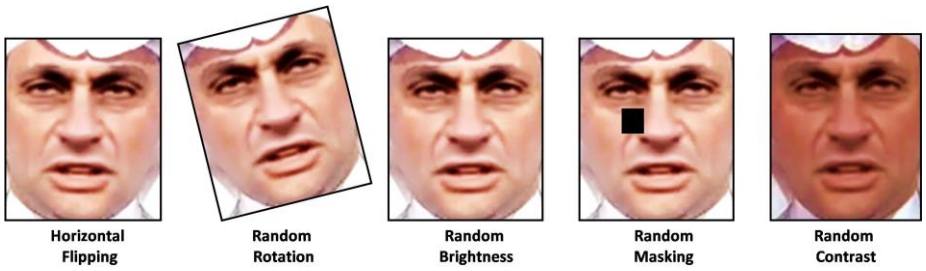

Figure 5: Visualization of different augmented samples after face detection and cropping.

5. Experiments and Results

This section provides a detailed discussion of the performance of the proposed method over various evaluation matrices, datasets used for experimentation, the preprocessing pipeline, and the training phases. All implementation and experiments were performed in a distributed manner on a High-Performance Computing Cluster with 4 Tesla v100 GPUs. TensorFlow was used for designing and training the proposed architecture, and for experimenting with larger batch sizes, automatic mixed precision was employed.

5.1. Data Preparation and Preprocessing

Released in 2019, Facebook's Deepfake Detection Challenge (DFDC) Dolhansky et al. (2020) is the largest publicly-available dataset of face swap videos, with more than 100,000 clips created via various Deepfake, non-learned, and GAN-based methods. Another dataset, FaceForensics++ Rossler et al. (2019), is composed of bonafide and forged video clips synthesized using a number of different generative methods. For evaluation based on FF++, we used the videos generated in the Face2Face, Deepfakes, FaceSwap, FaceShifter and Neural Textures subsets. We also used Celeb-DF Li et al. (2020c), itself comprised of 590

bonafide video samples sourced from online platforms with subjects of diverse age, gender, and ethnic group, and 5639 forged videos. Details of these datasets are shown in Table 1. To conduct a fair comparison with SoTA methods, we used the same split ratio released by each respective dataset.

For experimentation, face positions are extracted from deepfake video samples using MTCNN and cropped to a fixed dimension of 224×224 pixels without losing the aspect ratio, creating sequences of 30 consecutive frames. When cropping, empty areas of the image are filled with black pixels. For better generalization, extensive data augmentation techniques are used, including random brightness/contrast, masking patches of images with black pixels, and horizontal flipping. Figure 5 shows some of the augmented samples.

Table 1: Dataset details used for evaluation. DF: Deepfake, F2F: Face2Face, FS: FaceSwap, NT: Neural Textures

Datasets	Celeb-DF		FF++	c23		DFDC
Manipulation	DF	DF	F2F	FS	NT	FS
Training set	50k	24k	24k	24k	24k	100k
Testing set	32k	6k	6k	6k	6k	2.5k

5.2. Evaluation Metrics

Area Under the Curve (AUC): AUC measures the entire two-dimensional area under the receiver operating characteristic (ROC) curve, which is a plot that shows the performance of a classifier at all classification thresholds. It plots two parameters: i.e., True Positive Rate (TPR) and False Positive Rate (FPR). These are defined as:

$$TPR = \frac{TP}{TP+FN} \tag{18}$$

$$FPR = \frac{FP}{FP+TN} \tag{19}$$

where FP and TN refer to false positive and true negative, respectively.

Accuracy is the percentage of predictions classified correctly by a given model. For binary classification tasks like deepfake detection, accuracy may also be calculated in terms of positives and negatives as follows:

$$accuracy = \frac{n_{correct}}{n_{total}} \tag{20}$$

where $n_{correct}$ is the number of correct predictions and n_{total} is the total number of samples.

5.3. Training and Hyperparameter Setting

Training of the proposed system is conducted in two phases. The first phase is pretraining, where dimension-specific models are trained, and the second is training in the joint space. In the first phase, the model is focused on one inconsistency dimension of the data, and these sequences are passed through the temporal and spatiotemporal models. For the spatial model, sequences are rearranged to obtain images. The spatial, temporal and spatiotemporal models are pre-trained with sequence pooling and independent classification heads with a binary cross-entropy loss, along with the Adam optimizer with weight decay. Consequently, the spatial model learns to give a prediction based on images, while the temporal and spatiotemporal models learn from a sequence of frames. The embedding dimension for all three models is kept the same for ease of concatenation. The pre-trained models are frozen and the embeddings are concatenated, as described in Section 3.3, and passed to a sequence pooling layer, which in turn is passed to a fully connected classifier, as shown in Figure 1. For the spatial model, sequences are rearranged and the resultant images are used as input. Sequences of frames are then fed to the temporal and spatiotemporal models. These regularization techniques on individual models improved the overall AUC from 0.915 to 0.926 on the DFDC Dataset.

Table 2: A comparative analysis of the proposed system with SoTA methods on the DFDC dataset. S: Spatial Method, ST: spatiotemporal Method, STM: Spatial, Temporal and spatiotemporal

Method	Arch.	# P	AUC	F1-score
CViT Wodajo and Atnafu (2021)	S	89 M	0.8458	77.0%
TEI Liu et al. (2020)	ST	30. 4M	0.8697	-
ViT Distillation Heo et al. (2021)	S	373 M	0.978	91.9%
XceptionNet-avg Rossler et al. (2019)	S	22.8 M	0.843	-
I3D	ST	25 M	0.8082	-
EfficViT Coccomini et al. (2022)	S	109 M	0.919	83.8%
D-FWA Li and Lyu (2018)	S	-	0.8511	-
ADDNet-3D Zhao et al. (2021)	ST	-	0.7966	-
HolisticDFD (Our)	STM	11.5 M	0.926	92.64%

For finding optimal hyperparameters, we performed the experiments with the following embedding dimensions [64, 128, 256, 384] and U-net encoder-like convolution layers for feature extraction. Performance was optimal when the embedding size was set to 256. The learning rate was set to 10−3 at the start and later lowered to 10−5, along with weight decay of 10−4. We used 3 transformer layers inside for the spatiotemporal and temporal modules, and 6 transformer layers for the spatial module. We trained the model over 300 epochs, stopping early at the lowest validation loss, as required. We also experimented with different batch sizes from 8 to 16 on each device with distributed data-parallel.

5.4. Performance Analysis

Extant deepfake detection algorithms may be classified as either frame-based spatial (S in Table 2) or sequence-based spatiotemporal (ST in Table 2). For frame-based techniques, we chose Xception Rossler et al. (2019), Convolution- ViT Wodajo and Atnafu (2021), ViT Distillation Heo et al. (2021), EfficientVit Coccomini et al. (2022), and SelimEfficientNet. In contrast, frame-based methods are either based on Convolutional Neural Networks or Vision Transformers. To perform a fair comparison, we took the mean of the frame-level predictions in order to give a final prediction for the entire video. The sequence-based detectors used in the comparison are 3D convolution based C3D Liu et al. (2021), 2D-Convolution-with-RNN Graves (2012), and TEI Liu et al. (2020), for spatial-temporal modeling on a 2D CNN, D-FWA Li and Lyu (2018), ADDNet-3D Zhao et al. (2021), and S-IML-T Li et al. (2020a).

Table 3: Performance evaluation of the proposed framework on Celeb-DF and FaceForensics++ c23.

Method	Celeb-DF	FF++ c23
XceptionNet-avg Rossler et al. (2019)	0.9944	0.9940
Two-Branch Masi et al. (2020)	-	0.9643
D-FWA Li and Lyu (2018)	0.9858	-
I3D Spatiotemporal	0.9923	0.9826
ADDNet-3D Zhao et al. (2021)	0.9517	-
Meso-4 Afchar et al. (2018)	-	0.8310
LSTM based Network	0.9573	0.9482
Patch-DFD Yu et al. (2022)	-	0.9565
S-IML-T Li et al. (2020a)	0.9884	-
Bayar and Stamm (2016)	-	0.8297
Xia et al. (2022)	-	0.9100
HolisticDFD (our)	0.9624	0.9415

Table 4: Performance evaluation (accuracy) on different subsets of FaceForensics++ c23, DF: Deepfake, F2F: Face2Face, FS: FaceSwap, NT:Neural Texture.

Method	DF	F2F	FS	NT
C3D Liu et al. (2021)	0.9286	0.8857	0.9179	0.8964
XceptionNet-avg Rossler et al. (2019)	0.9893	0.9893	0.9964	0.9500
I3D Carreira and Zisserman (2017)	0.9286	0.9286	0.9643	0.9036
LSTM Tariq et al. (2020)	0.9964	0.9929	0.9821	0.9393
TEI Liu et al. (2020)	0.9786	0.9714	0.9750	0.9429
FaceNetLSTM Sohrawardi et al. (2019)	0.8900	0.8700	0.9000	-
DeepRhythm Qi et al. (2020)	0.9870	0.9890	0.9780	-
Comotion-35 Wang et al. (2020)	0.9595	0.8535	0.9360	0.8825
Comotion-70 Wang et al. (2020)	0.9910	0.9325	0.9830	0.9045
ADDNet-3d Zi et al. (2020)	0.9214	0.8393	0.9250	0.7821
HolisticDFD (Our)	0.98	0.95	0.944	0.965

As shown in Table 2, the proposed architecture achieves 92.64% AUC on the test set of the DFDC dataset. Our model is trained on the training set of the DFDC and we chose the model weights with the lowest loss on the validationset. When compared with the SoTA methods, our model's ROC-AUC curve has a comparable area (0.9264) to the SoTA (0.978) without using any distillation or transfer learning from other tasks. As shown in Table2, the Heo et al. (2021) models which demonstrate performance closest to the proposed method utilize distillation or transfer learning from large models such as EfficientNet.

Table 5: Cross-dataset generalization. The models are trained on the DFDC and Celeb-DF datasets and tested on other datasets

Train	Test	ROC-AUC
DFDC	FaceForensics++	0.761
DFDC	Celeb-DF	0.701
Celeb-DF	FaceForensics++	0.782

The proposed method was also evaluated on Celeb-DF and FaceForensics++ c23, and shows significant performance on the Neural Textures subset of Face-Forensics++, compared to the SoTA video and image level methods, as shown in Table 3. When tested on the subsets of FaceForensics++ c23, the proposed model gave SoTA accuracy of 0.965 on the Neural Textures subset of FaceForensics++, as demonstrated in Table 4. The proposed method gives an AUC of 0.9624 on Celeb-DF and 0.9415 on FaceForensics++ with c23 compression, as shown in Table 4. It is important to note that the proposed method performs better due to its ability to capture inconsistencies in small regions of the frames, such as lips, which is the case in Neural Texture.

For cross-dataset evaluation, as shown in Table 5, the proposed framework achieved an AUC of 0.761 on FaceForensics++ with the model trained on DFDC. The same model gave an AUC of 0.701 on Celeb-DF. If trained on Celeb-DF, the model showed an AUC of 0.782 on FaceForensics++.

We also performed time-complexity analysis of the HolisticDFD. The time was calculated by doing multiple forward passes through the framework with batch sizes ranging from 2 to 12 and a weighted average, as displayed in Figure 6. We also found that HolisticDFD has an approximately linear time complexity with respect to batch size. The proposed method took an average inference time of 14.63 seconds with an uncertainty of ± 3.06 seconds on a CPU and an average time of 250.46 milliseconds with an uncertainty of ± 106 milliseconds on GPU machines.

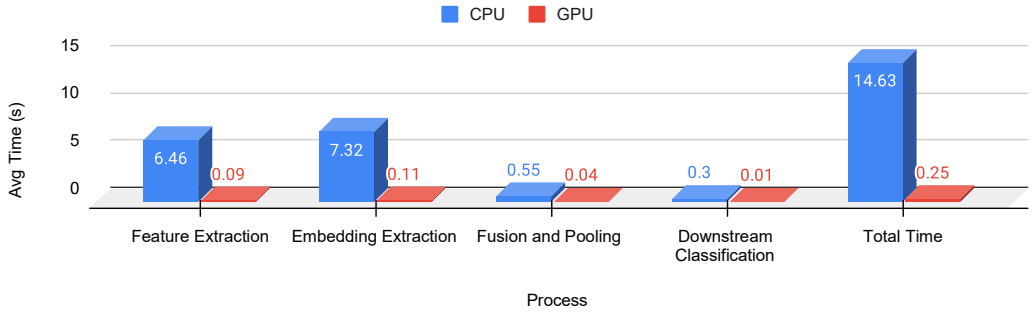

Figure 6: Time consumed by each module of the proposed framework on GPU and CPU.

5.5. Ablation Study

To test the effectiveness of the proposed method, we performed an ablation study by removing individual components and experimenting with binary combination of spatial, temporal and spatiotemporal feature-types. We also experiment with a majority voting ensemble of the three independent pipelines. The proposed frame- work obtained an AUC of 0.8968, 0.8801, and 0.901 using the spatial, temporal, and spatiotemporal transformers, individually. We also experimented with the majority ensemble of these three models and the obtained an AUC of 0.918. Although performance improved, the final configuration chosen for the proposed framework where we intelligently combined embeddings outperformed the majority voting ensemble. The performance of various combinations of spatial, temporal and spatiotemporal feature extractors in the proposed network is shown in Table 6. This performance gradually improved as embeddings from different models were infused, showing the enhanced performance of the joint model. The AUC score improved from 0.896 to 0.907 when the spatial and spatiotemporal modelsare concatenated, and further improved from 0.907 to 0.926 after concatenatingthe joint spatiotemporal model. This demonstrates that infusion enables better integrated pattern recognition by concatenating the embeddings from multiple dimensions.

Table 6: Ablation study of the proposed framework on DFDC. S: Spatial Transformer, T: Temporal Transformer, ST: spatiotemporal Transformer

Component	AUC	Params(#)
Only Spatial Transformer	0.8968	2.3 M
Only Temporal Transformer	0.8801	4.6 M
Only spatiotemporal Transformer	0.901	4.6 M
Spatial and Temporal Transformer	0.9021	6.9 M
Spatial and spatiotemporal	0.907	6.9 M
Ensemble (S, T, ST)	0.918	-
HolisticDFD (Our)	0.926	11.5 M

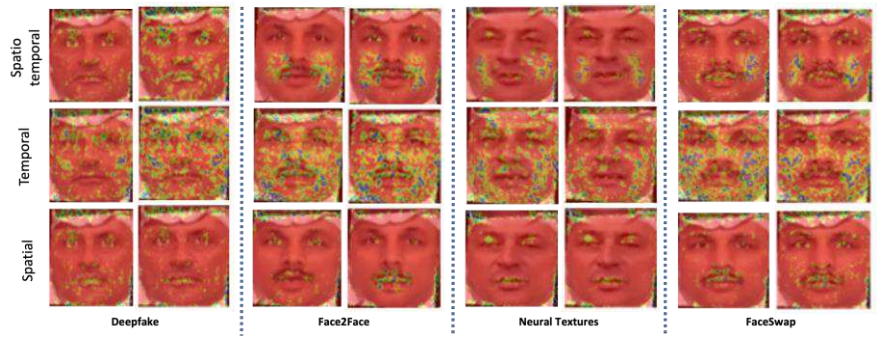

Figure 7: Integrated Gradient Analysis for different forgery techniques in the FaceForensics++ Dataset.

5.6. Integrated Gradient Analysis

To better explain the relationship between the prediction of each pipeline interms of its input images or sequences, we also performed integrated gradient analysis of the individual pipelines. Model gradients were computed with respect to inputs in order to determine attribution. Integrated gradient analysisof the spatial model was performed using a black image baseline and interpolating the target image in 10 steps. As given in Figure 7, the spatial model focused more on particular artifacts like eyes or lips. Next, integrated gradient analysis of the temporal model was performed on a sequence of frames with the same settings. From this analysis, it is evident that the temporal model focused onthe entire face. Similarly, an integrated gradient analysis of the spatiotemporal model focused on both particular position artifacts and small areas (e.g., surrounding of lips). The diversity of attention/attribution in the models demonstrate the strength of the proposed method.

Integrated gradient analysis also demonstrates the strength of the feature extractor with respect to different types of forgeries. For instance,Figure 7 shows the integrated gradient heatmaps of several forgery methods onthe DFDC dataset. The blue areas show the highest activation values on the heatmaps, which the proposed framework uses for predictions. As a face-based manipulation, FaceSwap transfers facial regions from the source to the target video, so the forged parts in the fabricated videos cover the entire facial regions. It can be seen in Figure 7 that the integrated gradient outputs of the temporal model are very high in FaceSwap when compared to other forgery methods. Face2Face is a reenactment system which modifies the expressions ina target video while maintaining identity information. It is clear from Figure7 that the spatiotemporal model focuses on the sides of the lips, which give a strong indication of expression-based modification. Similarly, the spatial model focuses on lips and eyes, markers for expression manipulation. The activation regions ofthe Neural Texture samples concentrated not only on lips but also onthe cheeks and forehead region, possibly

due to the texture rendering operation when synthesizing fake facial regions, resulting in irregular shadows and discordance in these areas. Examining the deepfakes in Figure 7, the forged facial regions are pasted back to the target face, providing obvious blending or visual artifacts near facial features like eyebrows. Activation areas in the corresponding heatmaps for spatial are thus concentrated around eyebrows and lips, while for temporal, on movements in the upper cheeks, and for spatiotemporal around eyebrows and the nose, which is in line with our hypotheses for this method.

In summary, Figure 7 demonstrates that the proposed framework has the ability to learn distinct features of different facial parts and forgery methods, which is a key step toward generalization of this framework to unseen deepfakes.

6. Conclusion

In this paper, we propose a novel, lightweight, and compact transformer-based, deepfake detection method which intelligently combines the embeddings from the spatial, temporal and spatiotemporal dimensions to differentiate a suspected deepfake from bonafide video. We show that by adding spatial, temporal, and spatiotemporal views of the data, the model learns a better data representation and performance gradually improves. Our model preforms competitively on the Celeb- DF and FaceForensics++ datasets and shows near SoTA performance on the DFDC dataset when compared on AUC. In addition, it outperformsall baselines on the basis of F1 Score. Cross-corpus evaluation of the proposed method is comparable to SoTA methods, which demonstrates the generalizability of the proposed method. More importantly, the proposed model is significantly lightweight, using just 3% of the parameters of SoTA deepfake detection methods. A performance evaluation of the model shows employing the spatial, temporal and spatiotemporal latent joint space and learned attention weights significantly improves the capability of deepfake detectors. A comparative analysis of the framework with existing techniques shows significant improvement when deployed on large datasets, such as the DFDC, and specific forgeries, such as neural texture.

Acknowledgement

This material is based upon work supported by the National Science Foundation (NSF) under award number 1815724, 2231619 and Michigan Transnational Research and Commercialization (MTRAC), Advanced Computing Technologies (ACT) award number 292883. Any opinions, findings, conclusions or recommendations expressed in this material are those of the author(s) and do not necessarilyreflect the views of the NSF and MTRAC ACT.

References

Afchar, D., Nozick, V., Yamagishi, J., Echizen, I., 2018. Mesonet: a compact facial

video forgery detection network, in: 2018 IEEE international workshop on information forensics and security (WIFS), IEEE. pp. 1–7.

Bayar, B., Stamm, M.C., 2016. A deep learning approach to universal image manipulation detection using a new convolutional layer, in: Proceedings of the 4th ACM workshop on information hiding and multimedia security, pp. 5–10.

Carreira, J., Zisserman, A., 2017. Quo vadis, action recognition? A new model and the kinetics dataset. CoRR abs/1705.07750. URL: http://arxiv.org/abs/1705.07750, arXiv:1705.07750.

Chai, L., Bau, D., Lim, S.N., Isola, P., 2020. What makes fake images detectable? understanding properties that generalize, in: European conference on computer vision, Springer. pp. 103–120.

Chen, B., Li, T., Ding, W., 2022. Detecting deepfake videos based on spatiotemporal attention and convolutional lstm. Information Sciences 601, 58–70.

Choi, Y., Choi, M., Kim, M., Ha, J.W., Kim, S., Choo, J., 2018. Stargan: Unified generative adversarial networks for multi-domain image-to-image translation, in: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 8789–8797.

Coccomini, D.A., Messina, N., Gennaro, C., Falchi, F., 2022. Combining efficientnet and vision transformers for video deepfake detection, in: International Conference on Image Analysis and Processing, Springer. pp. 219–229.

Dhariwal, P., Nichol, A., 2021. Diffusion models beat gans on image synthesis. CoRR abs/2105.05233. URL: https://arxiv.org/abs/2105.05233, arXiv:2105.05233.

Dolhansky, B., Bitton, J., Pflaum, B., Lu, J., Howes, R., Wang, M., Canton-Ferrer, C., 2020. The deepfake detection challenge dataset. CoRR abs/2006.07397. URL: https://arxiv.org/abs/2006.07397, arXiv:2006.07397.

Graves, A., 2012. Long short-term memory. Supervised sequence labelling with recurrent neural networks, 37–45.

Gu, Z., Chen, Y., Yao, T., Ding, S., Li, J., Huang, F., Ma, L., 2021. Spatiotemporal inconsistency learning for deepfake video detection, in: Proceedings of the 29th ACM International Conference on Multimedia, pp. 3473–3481.

Haiwei, W., Jiantao, Z., Shile, Z., Jinyu, T., 2022. Exploring spatial-temporal features for deepfake detection and localization. arXiv preprint arXiv:2210.15872.

Hassani, A., Walton, S., Shah, N., Abuduweili, A., Li, J., Shi, H., 2021. Escaping the

big data paradigm with compact transformers. CoRR abs/2104.05704. URL: https://arxiv.org/abs/2104.05704, arXiv:2104.05704.

Heo, Y.J., Choi, Y.J., Lee, Y.W., Kim, B.G., 2021. Deepfake detection scheme based on vision transformer and distillation. arXiv preprint arXiv:2104.01353.

Heo, Y.J., Yeo, W.H., Kim, B.G., 2023. Deepfake detection algorithm based on improved vision transformer. Applied Intelligence 53, 7512–7527.

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

Javed, A., Malik, K.M., Irtaza, A., Malik, H., 2021. Towards protecting cyber-physical and iot systems from single-and multi-order voice spoofing attacks. Applied Acoustics 183, 108283.

Jia, G., Zheng, M., Hu, C., Ma, X., Xu, Y., Liu, L., Deng, Y., He, R., 2021. Inconsistency-aware wavelet dual-branch network for face forgery detection. IEEE Transactions on Biometrics, Behavior, and Identity Science 3, 308–319.

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

Khalid, F., Javed, A., Irtaza, A., Malik, K.M., 2023. Deepfakes catcher: A novel fused truncated densenet model for deepfakes detection, in: Proceedings of International Conference on Information Technology and Applications: ICITA 2022, Springer. pp. 239–250.

Khalid, H., Tariq, S., Kim, M., Woo, S.S., 2021. Fakeavceleb: a novel audio- video multimodal deepfake dataset. arXiv preprint arXiv:2108.05080.

Khan, A., Javed, A., Malik, K.M., Raza, M.A., Ryan, J., Saudagar, A.K.J., Malik, H., 2022. Toward realigning automatic speaker verification in the era of covid-19. Sensors 22, 2638.

Li, X., Lang, Y., Chen, Y., Mao, X., He, Y., Wang, S., Xue, H., Lu, Q., 2020a. Sharp multiple instance learning for deepfake video detection, in: Proceedings of the 28th ACM international conference on multimedia, pp. 1864–1872.

Li, X., Yu, K., Ji, S., Wang, Y., Wu, C., Xue, H., 2020b. Fighting against deepfake: Patch&pair convolutional neural networks (ppcnn), in: Companion Proceedings of the Web Conference 2020, pp. 88–89.

Li, Y., Lyu, S., 2018. Exposing deepfake videos by detecting face warping artifacts. arXiv preprint arXiv:1811.00656.

Li, Y., Yang, X., Sun, P., Qi, H., Lyu, S., 2020c. Celeb-df: A large-scale challenging dataset for deepfake forensics, in: Proceedings of the IEEE/CVF conference on

computer vision and pattern recognition, pp. 3207–3216.

Liang, P., Liu, G., Xiong, Z., Fan, H., Zhu, H., Zhang, X., 2023. A facial geometry based detection model for face manipulation using cnn-lstm architecture. Information Sciences 633, 370–383.

Lipton, A.J., Fujiyoshi, H., Patil, R.S., 1998. Moving target classification and tracking from real-time video, in: Proceedings fourth IEEE workshop on applications of computer vision. WACV'98 (Cat. No. 98EX201), IEEE. pp. 8–14.

Liu, J., Zhu, K., Lu, W., Luo, X., Zhao, X., 2021. A lightweight 3D convolutional neural network for deepfake detection. Int. J. Intell. Syst. 36, 4990–5004.

Liu, Z., Luo, D., Wang, Y., Wang, L., Tai, Y., Wang, C., Li, J., Huang, F., Lu, T., 2020. Teinet: Towards an efficient architecture for video recognition, in: Proceedings of the AAAI Conference on Artificial Intelligence, pp. 11669– 11676.

Masi, I., Killekar, A., Mascarenhas, R.M., Gurudatt, S.P., AbdAlmageed, W., 2020. Two-branch recurrent network for isolating deepfakes in videos, in: European conference on computer vision, Springer. pp. 667–684.

Masood, M., Nawaz, M., Malik, K., Javed, A., Irtaza, A., Malik, H., 2022. Deepfakes generation and detection: state-of-the-art, open challenges, countermeasures, and way forward. Applied Intelligence, 1–53doi:10.1007/s10489- 022-03766-z.

Nguyen, X.H., Tran, T.S., Nguyen, K.D., Truong, D.T., et al., 2021. Learning spatiotemporal features to detect manipulated facial videos created by the deepfake techniques. Forensic Science International: Digital Investigation 36, 301108.

Qi, H., Guo, Q., Juefei-Xu, F., Xie, X., Ma, L., Feng, W., Liu, Y., Zhao, J., 2020. Deeprhythm: Exposing deepfakes with attentional visual heartbeat rhythms, in: Proceedings of the 28th ACM international conference on multimedia, pp. 4318–4327.

Rossler, A., Cozzolino, D., Verdoliva, L., Riess, C., Thies, J., Nießner, M., 2019. Faceforensics++: Learning to detect manipulated facial images, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 1–11.

Schwarcz, S., Chellappa, R., 2021. Finding facial forgery artifacts with parts- based detectors, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 933–942.

Sohrawardi, S.J., Chintha, A., Thai, B., Seng, S., Hickerson, A., Ptucha, R., Wright, M., 2019. Poster: Towards robust open-world detection of deepfakes, in: Proceedings of the 2019 ACM SIGSAC conference on computer and communications security, pp. 2613–2615.

Tariq, S., Lee, S., Woo, S.S., 2020. A convolutional LSTM based residual network

for deepfake video detection. CoRR abs/2009.07480. URL: https://arxiv.org/abs/2009.07480, arXiv:2009.07480.

Tian, C., Luo, Z., Shi, G., Li, S., 2023. Frequency-aware attentional feature fusion for deepfake detection, in: ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE. pp. 1–5.

Wang, G., Jiang, Q., Jin, X., Cui, X., 2022. FFR_FD: Effective and fast detection of deepfakes via feature point defects. Information Sciences 596, 472–488.

Wang, G., Zhou, J., Wu, Y., 2020. Exposing deep-faked videos by anomalous co-motion pattern detection. arXiv preprint arXiv:2008.04848.

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

Xia, Z., Qiao, T., Xu, M., Zheng, N., Xie, S., 2022. Towards deepfake video forensics based on facial textural disparities in multi-color channels. Information Sciences 607, 654–669.

Yang, Z., Liang, J., Xu, Y., Zhang, X.Y., He, R., 2023. Masked relation learning for deepfake detection. IEEE Transactions on Information Forensics and Security 18, 1696–1708.

Yu, M., Ju, S., Zhang, J., Li, S., Lei, J., Li, X., 2022. Patch-dfd: Patch-based end-to-end deepfake discriminator. Neurocomputing.

Zhao, C., Wang, C., Hu, G., Chen, H., Liu, C., Tang, J., 2023. Istvt: interpretable spatial-temporal video transformer for deepfake detection. IEEE Transactions on Information Forensics and Security 18, 1335–1348.

Zhao, H., Zhou, W., Chen, D., Wei, T., Zhang, W., Yu, N., 2021. Multi attentional deepfake detection, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 2185–2194.

Zhao, T., Xu, X., Xu, M., Ding, H., Xiong, Y., Xia, W., 2020. Learning to recognize patch-wise consistency for deepfake detection. arXiv preprint arXiv:2012.09311 6.

Zi, B., Chang, M., Chen, J., Ma, X., Jiang, Y.G., 2020. Wilddeepfake: A challenging real-world dataset for deepfake detection, in: Proceedings of the 28th ACM international conference on multimedia, pp. 2382–2390.