AVFakeNet: A unified end-to-end Dense Swin Transformer deep learning model for audio-visual deepfakes detection

Hafsa Ilyas[1], Ali Javed[1*], Khalid Mahmood Malik[2]

[1]Department of Software Engineering, University of Engineering and Technology-Taxila, 47050, Pakistan

[2]Department of Computer Science and Engineering, Oakland University, Rochester, MI, 48309, USA

[*]Corresponding author email: ali.javed@uettaxila.edu.pk

Abstract-Recent advances in the field of machine learning and social media platforms facilitate the creation and rapid dissemination of realistic fake content (i.e., images, videos, audios). Initially, the fake content generation involved the manipulation of either audio or video streams but currently, more realistic deepfakes content is being produced via modifying both audio-visual streams. Researchers in the field of deepfakes detection mostly focus on identifying the fake videos exploiting solely visual or audio modality. However, there exist a few approaches for audio-visual deepfakes detection but mostly are not evaluated on a multimodal dataset with deepfakes videos having the manipulations in both streams. The unified approaches evaluated on the audiovisual deepfakes dataset have reported low detection accuracies and failed when the faces are side-posed. Therefore, in this paper, we introduced a novel AVFakeNet framework that focuses on both the audio and visual modalities of a video for deepfakes detection. More specifically, our unified AVFakeNet model is a novel Dense Swin Transformer Net (DST-Net) which consists of an input block, feature extraction block, and output block. The input and output block comprises dense layers while the feature extraction block employs a customized swin transformer module. We have performed extensive experimentation on five different datasets (FakeAVCeleb, Celeb-DF, ASVSpoof-2019 LA, World Leaders dataset, Presidential Deepfakes dataset) comprising audio, visual, and audio-visual deepfakes along with a cross-corpora evaluation to signify the effectiveness and generalizability of our unified framework. Experimental results highlight the effectiveness of the proposed framework in terms of accurately detecting deepfakes videos via scrutinizing both the audio and visual streams.

Keywords: AVFakeNet, Audio-Visual Deepfake detection, Dense Swin Transformer, FakeAVCeleb, ASVSpoof-2019.

1. Introduction

In the last decade, we have seen tremendous growth in multimedia content on the Internet due to the economical prices of digital capturing devices and social media evolution. Nowadays, it has become very easy to manipulate content via different advanced multimedia editing tools [21]. Moreover, the availability of cutting-edge machine learning (ML) algorithms like GANs has made it possible to create highly realistic forged content (i.e., images, videos, and audios) to propagate disinformation through social networks (i.e., Facebook, Twitter, Instagram, etc.). As a result, disseminating fake content on social media platforms has become easier, making it more difficult to trust the media information. False information on social networks can affect the opinions and emotions of society and can also result in disruptive public acts based on misleading ideas. The generation of fake/synthesized content (including images, videos, and audios) using deep learning algorithms is well-known as deepfakes. Generative Adversarial Networks (GANs) [28] and Autoencoders (AEs) [29] based techniques are mainly used for the generation of synthesized videos and audios. Video deepfakes include the generation of fake/synthesized videos via replacing the person's face with another person (Face Swap), modifying the person's expression (Expression Swap), or synchronizing the person's lip movement with some sound (Puppet Mastery). While the audio deepfakes are the creation of cloned voices of a person depicting the individual speaking the things that are never spoken. Text-to-Speech Synthesis (TTS) and Voice Conversion (VC) are the two main techniques for audio deepfakes creation. In TTS synthesis, the person's natural voice is synthesized according to the given input text whereas VC is a technology in which the audio of the source person is modified to make it sound like the voice of the target person [21]. The deepfakes videos and audios generated using advanced AI algorithms have attained such realism that now it becomes difficult for humans to recognize the video or audio as a fake one. Thus, bring up the major privacy and security threats as fake voices can be used to fool the voice recognition system and spread fake news while fake videos can be used to defame a person or generate misinformation via impersonating a renowned personality. The example includes the fake video of Mark Zuckerberg posted on Instagram created with Canny AI's Video Dialogue Replacement (VDR) software [30].

In existing works, the researchers mainly focus on detecting the deepfakes through a single modality/stream (either video or audio). For instance, in [1], a deep learning model using the multi-layer perceptron (MLP) and convolution

neural network (CNN) was introduced for the detection of AI-generated deepfakes videos. Landmark features and frames were extracted from the input videos and fed to the MLP and CNN, respectively. At the classification stage, the output of MLP and CNN was combined to predict whether the video was fake or real. The model [1] was evaluated on a private dataset and achieved an accuracy and AUC score of 87% and 87.7%, respectively. Kohli et al. [2] presented a lightweight 3DCNN that extracted the spatial and temporal features using the optical flow method. A 4-depth matrix comprising two successive frames and their horizontal and vertical gradients was given to the model as an input. The model [2] was evaluated on the FaceForensics++ (FF++) dataset and showed good detection results. Likewise, for audio classification, a novel approach DeepSonar was introduced in [3], which monitored layer-wise neuron behavior to identify the AI-synthesized voices generated using text-to-speech and voice cloning systems. To evaluate the model [3], experiments were conducted on three datasets (FoR, Sprocket-VC, MC-TTS) covering English and Chinese languages. Hua et al. [4] demonstrated an end-to-end Time-domain Synthetic Speech Detection Net (TSSDNet) for the detection of audio deepfakes using deep learning features. TSSDNet was evaluated on a challenging ASVSpoof-2019 logical access (LA) dataset and attained an equal error rate (EER) of 1.64%. The model [4] shows good generalizability but is computationally complex.

Due to the lack of audio-visual deepfakes datasets, few unified models are presented in the literature for detecting deepfakes. Zhou et al. [31] introduced a joint detection framework for detecting deepfakes via audio and video modality. Similarly, [32] and [33] classified the videos either as fake or real by finding the dissimilarity between audio and visual streams. Due to the absence of a proper dataset, [31] utilized existing deepfakes datasets such as DFDC and applied the vocoders used in VC and TTS tasks to mimic the synthesized speech. [32] used the DFDC and DeepfakeTIMIT datasets. However, [33] evaluated the model on synchronous and asynchronous audio-visual pairs produced from VidTIMIT and DeepfakeTIMIT, respectively. The above-mentioned approaches are not evaluated on a dataset in which both audio and visual modalities are manipulated. To enhance the research on unified models for deepfakes detection, Khalid et al. [34] contributed a new audio-visual dataset FakeAVCeleb. In [35], the ensemble methods based on five classifiers i.e., Meso-4, MesoInception-4, Xception, VGG16, and EfficientNet-B0 were evaluated on the FakeAVCeleb dataset. The VGG16 model achieved the highest accuracy of 78.04%, while, XceptionNet showed the worst performance with an accuracy of 43.94%. It can be concluded that none of the methods provided satisfactory performance demonstrating that they are not suitable for audio-visual deepfakes detection. Davide et al. [25] presented a POI-Forensics for deepfakes detection based on audio-visual identity verification. The model was trained only on the augmented real videos of the VoxCeleb2 dataset and attained an accuracy of 86.6% on the FakeAVCeleb dataset. This model [25] has a limitation of the requirement of some real videos of the target subject as a reference during the testing. Moreover, POI-Forensics failed on the side-posed faces and performed well only on frontal-posed faces. Based on empirical findings that faces and voices are more mismatched in fake videos as compared to the real ones, Cheng et al. [36] introduced a deepfakes detection method called Voice-Face matching Detection (VFD) via finding the consistency between the voice and face of a person. Three datasets DFDC, Deepfake TIMIT, and FakeAVCeleb were used to evaluate this approach [36]. VFD achieved an accuracy of 81.52% and an AUC of 86.11% on the FakeAVCeleb dataset. VFD fails to detect deepfakes in the cases when a face is side-posed and there is insufficient illumination where faces are not clearly visible.

These days, deepfakes are not just created by forging only one modality/stream (video or audio) rather, more convincing fake videos are produced in which forgery is applied on both modalities (video and audio), thus enhancing the threats and concerns associated with deepfakes. Detecting such videos in which both visual and audio stream is modified is a challenging task. Moreover, there is also a lack of such datasets which contain fake videos along with fake audio. Thus, limiting the development of the unified model that can detect the audio and video deepfakes simultaneously. Most of the unified frameworks reported in the literature are not evaluated on the multimodal deepfakes dataset such as FakeAVCeleb. Also, the models [25,36] evaluated on the FakeAVCeleb datasets have detection accuracies lesser than 90% and fail to detect deepfakes videos that contain the faces at different angular positions. To address the aforementioned limitations, we present a unified AVFakeNet model that by using the visual and acoustics features exploits the spatio-temporal characteristics of the input video for deepfakes detection. For this purpose, we proposed a unified Dense Swin Transformer Net (DST-Net) for the detection of deepfakes videos via analyzing both audio and visual streams. Our unified DST-Net has three blocks named input block, feature extraction block, and output block. The input block consists of dense layers while the output block contains the combination of dense and dropout layers. Feature extraction block comprises the modified swin transformer. For the evaluation of our proposed unified framework, we utilized an audio-video multimodal deepfakes detection dataset named as FakeAVCeleb. According to the best of our knowledge, it is the only publicly available dataset that has the cloned deepfakes audios along with deepfakes videos. The major contributions of our work are:

- We propose a novel unified framework AVFakeNet that is able to accurately detect the manipulation in both the audial and visual streams of deepfakes video.
- We propose a Dense Swin Transformer Net that computes the dense hierarchical features maps for better representation of the input videos and improves the deepfakes detection performance.
- Our proposed unified model is robust against the high-quality deepfakes videos with angled or side-posed faces having variations in illumination conditions, people's ethnicity, gender, and age groups.
- We have performed extensive experimentation on five different datasets comprising audio, visual, and audio-visual deepfakes along with a cross-corpora evaluation to signify the effectiveness and generalizability of our proposed unified framework.

2. Literature Review

To counter the threats introduced because of deepfakes video and audio generation, researchers have introduced many different deepfakes detection models and algorithms. In this section, we have reviewed the state-of-the-art methods for the detection of audio and video deepfakes.

2.1. Video deepfakes detection

For video deepfakes detection, some approaches focus on hand-crafted features [5,19,20] or physiological features [6,7,8,10]. For example, Geura et al. [5] introduced a no pixel-based approach in which feature vectors were constructed from the stream descriptors information of the videos. These feature vectors were then used to train the ensemble of SVM and random forest classifiers. AUC score of 98.4% was achieved on Media Forensics Challenge (MFC) dataset. Despite the good performance, this approach [5] fails to handle the video re-encoding attacks. Ciftci et al. [6] presented a method that used biological features such as heart rate estimation to identify the deepfakes videos. SVM and CNN-based classifiers were trained on features extracted using the remote photoplethysmography (rPPG) technique. Likewise, in [7,8], rPPG-based physiological features were used to discriminate fake videos from real ones.

Keeping in view the possible abuses of deepfakes videos, researchers have also introduced deepfakes detection models based on deep neural networks (DNNs). Chintha et al. [9] introduced a framework based on XceptionNet and Bidirectional LSTM. XceptionNet was used to extract the facial features whereas temporal sequence analysis was performed using Bidirectional LSTM. To distinguish the real video's features from the fake ones, the model was trained on the combination of KL divergence and Cross Entropy loss functions. Likewise, in [11], facial features extracted using VGG-11 from the video frames were fed to the LSTM to obtain the temporal sequence descriptors. These descriptors were then used to train CNN frameworks named I3D, ResNet, and R3D for recognizing fake videos. This approach [11] achieved decent detection accuracy on the Celeb-DF dataset however it is computationally complex.

2.2. Audio deepfakes detection

Traditional voice spoofing detection methods focus on feature engineering where the hand-crafted features are used to train the classifier for the detection of the audio deepfakes. For instance, in [12], a global modulation 2D-DCT features extractor was presented that captured global spectro-temporal modulation patterns for audio deepfakes detection. The approach [12] attained an EER of 4.03% on ASVspoof-2019 LA, however, the performance decreases on noisy samples. To increase the diversity of ASVspoof-2019 LA training data, Das et al. [13] applied a signal companding-based data augmentation technique before computing the constant Q transform (CQT) features and then used these features to train the LCNN classifier. This method [13] improves the detection accuracy but at the expense of extensive training data. In our prior work [14], we developed a robust method for the detection of multiple spoofing attacks including single and multi-order playback, voice synthesis, and cloned replay attacks. A novel acoustic ternary patterns-Gammatone cepstral coefficients (ATP-GTCC) features were introduced to better capture the dynamic traits of the real human voice, robotic noise, and distortion in the playback samples for the accurate detection of spoofing attacks on voice-driven systems. ATP features descriptor uses a fixed threshold for patterns generation and provides a lower performance in real-time scenarios. To overcome this limitation of ATP features, we presented the extended local ternary patterns (ELTP) and fused them with Linear Frequency Cepstral Coefficient (LFCC) features in [15] for detecting the TTS and VC spoofing attacks. ELTP calculated the threshold dynamically by locally computing the standard deviation of each audio frame. We also developed a unified voice spoofing detector [43] by proposing novel acoustic-ternary co-occurrence patterns (ATCoP) and fused them with GTCC patterns to accurately detect all types of voice spoofing attacks. This anti-spoofing framework [43] was evaluated on four different datasets including voice spoofing detection corpus (VSDC), ASVspoof-2019, Google's LJ Speech, and YouTube deepfakes datasets to demonstrate the accurate detection performance for various kinds of audio deepfakes.

Deep neural networks have also shown great performance while detecting spoofed voices or audio deepfakes. Alanis et al. [16] introduced a Light Convolutional Gated Recurrent Neural Network (LC-GRNN) to expose the spoofing attacks (i.e., text to speech, voice conversion, and replay) via extracting discriminative frame level and contextual features. Log magnitude spectrograms with 256 bins were fed to the model to identify the speech as fake or real. ASVspoof-2015, 2017, and 2019 were used to evaluate the model. This anti-spoofing system [16] is computationally efficient but not robust against unseen spoofing attacks. In [17], a self-supervised approach known as SSAD consisting of an encoder, regression, and binary workers was presented to detect the original and fake voices. This approach [17] was evaluated on the ASVspoof-2019 LA dataset and achieved an EER of 5.31%. Although the model [17] is computationally efficient, the detection accuracy needs to be further improved. Zhang et al. [18] presented a one-class learning model (based on ResNet-18 and one-class softmax) that detected unknown synthetic voices generated using TTS and voice conversion techniques. The model was trained on 60-dimensional LFCCs features and attained 2.19% EER and a min t-DCF of 0.059 on the ASVspoof-2019 LA dataset.

The state-of-the-art methods for the detection of deepfakes mainly focus on the fake audio or video detection separately and attained reasonable results as discussed above. Less attention is given to the field of a unified model for detecting deepfakes utilizing both audio and visual streams of a video. The models exploiting both streams of the video are either not evaluated on multimodal datasets or failed to perform well in case of varying lighting conditions, videos having side posed faces, etc. Moreover, models are not evaluated for cross corpora settings. In this paper, a unified framework is presented that identifies the deepfakes videos via analyzing audio and video streams and overcomes the above-mentioned limitations of the existing deepfakes detection methods.

3. Analysis of the mel-spectrogram of real and fake videos of FakeAVCeleb dataset

Deepfakes attacks can be very harmful and used to defame a person, spread fake news on social media, hack voice-controlled systems, and exploit society's peace by conveying misleading and disrupting ideas. Mostly deepfakes attacks can either include audio or video manipulation. These attacks could be more dangerous when both audio and video stream of a video is modified to generate more realistic fake content.

Based on our analysis and observation, we argue that fake audio can be different from the real human voice from many perspectives. For instance, the real human voice can have many natural characteristics such as respiration, expression, vocalism, change in pitch, and tone of voice. Voice recording also entails many factors such as background noise, distortions, etc. Contrarily, synthetic voices lack the human voice prosodic attributes, variation in the pitch and tone of the voice. Also, the speaking process is linear and seamless with void of distortions or background noise in fake audio due to its linear generation process. Synthetic and real voices are also different in terms of the pauses between the speech. There is no distortion or noise in the spoofed speech if there is a pause. However, breathing or background noise is present in the real voice during the pause.

The advancement in the quality of deepfakes audio generation methods/tools has reduced the potential discriminative attributes of audio used to classify between fake and real speech, thus increasing the difficulty to detect fake utterances. We suppose that spectrograms generated using high resolution 2048fft bins can demonstrate the imperceptible differences and also depict the above-mentioned dissimilarity between real and fake audio. So, we investigated the spectrograms in this research work for audio modality with the expectation of providing a good performance compared to the conventional acoustic feature extraction. To support our assumptions, samples of real and fake audio Mel-Spectrograms are shown in Fig. 1. From the highlighted patterns in dotted regions in Fig. 1, it can be seen that the corresponding region becomes entirely blank in the respective Mel-Spectrogram where the pause occurs in fake audio. While in real speech, when a pause occurs, the corresponding Mel-Spectrogram area is not completely blank, but it contains some patterns due to the presence of noise. Moreover, the bright horizontal patterns represent the pitch and emphasize on the words. These lines are brighter when the speaker emphasizes on a word or there is high background noise and also patterns are not linear but irregular due to the variations in the pitch. Whereas for the fake audio, the horizontal patterns are more linear compared to the real audio because of the same pitch of the voice throughout the audio sample. It is also notable that, in the synthetic speech, there appears a vertical portion for each word in the Mel-Spectrogram while such portions are less prominent in the real speech Mel-spectrograms. Our proposed DST-Net model exploits globally aware dense hierarchical deep learning features, so we hypothesize that such features enable our model to accurately detect the above-mentioned distinctive artifacts in the Mel-Spectrogram of fake and real speeches.

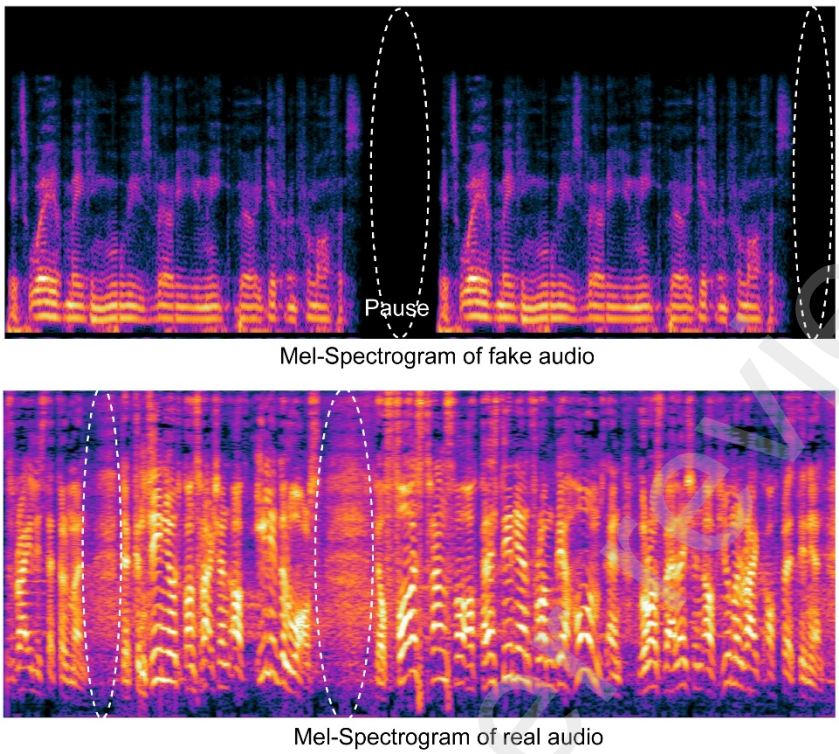

Fig. 1. Mel-Spectrograms of real and fake audios.

4. Proposed methodology

CNNs have been widely used in ML because of the effective feature extraction of convolution layers. The majority of deep learning-based deepfakes detection methods have employed CNNs. However, the scope of CNNs is limited owing to network depth and kernel size as the extremely deep neural networks induce gradient vanishing problems and large kernels increase computing costs. The transformers, on the other hand, have first achieved considerable success in the natural language processing arena via using self-attention setting, deeper mapping, and sequence-to-sequence model design. Thereafter, it has been employed in object detection and image recognition tasks.

Keeping in view the limitations of CNNs and the emerging use of transformers in the image recognition tasks, we utilize the swin transformer architecture with a modified MLP module. In our proposed network, we used the modified swin transformer as a feature extraction module. The purpose of this research work is to develop a unified model that can detect the manipulation in both audio and visual streams of deepfakes videos. The architectural details of the proposed model DST-Net and the workflow of the proposed framework are described in the subsequent sections.

4.1. Workflow of proposed unified framework

The detailed classification process of our unified framework for deepfakes detection is shown in Fig. 2. Our proposed framework is a two-stream network having an audio and video model that can classify both the audio and visual features extracted from a video. The video and audio models are trained on the frames and Mel-Spectrogram images, respectively and make the predictions individually. It is important to note that our proposed DST-Net is used to classify audio and visual features in both streams of the proposed framework. The model trained on audio Mel-Spectrograms is referred to as an audio model while the DST-Net trained on the extracted faces from the video frames is named as a video model. For testing, the input to our framework is a video along with its respective Mel-Spectrogram image. Frames of the video are extracted, and the face detection algorithm Multi-Task Cascaded Convolution Neural Network (MTCNN) [37] is used to detect the faces. However, Python librosa package is employed to generate Mel-Spectrogram with the following parameters: n_fft = 2048, hop_length = 512 and n_mels = 175. And then power_to_db function is used to convert the spectrum to decibel units. Mel-Spectrogram is an effective method to extract the hidden and useful features to visualize the audio as an image. In the next step, we resize and reshape the extracted faces from the video and a Mel-Spectrogram image. And then pass the frames and Mel-Spectrogram to the respective stream of the model containing the Dense Swin Transformer network. The video model provides the prediction for the faces extracted from the videos; therefore, we apply the majority voting rule to classify the overall visual stream either as fake or real.

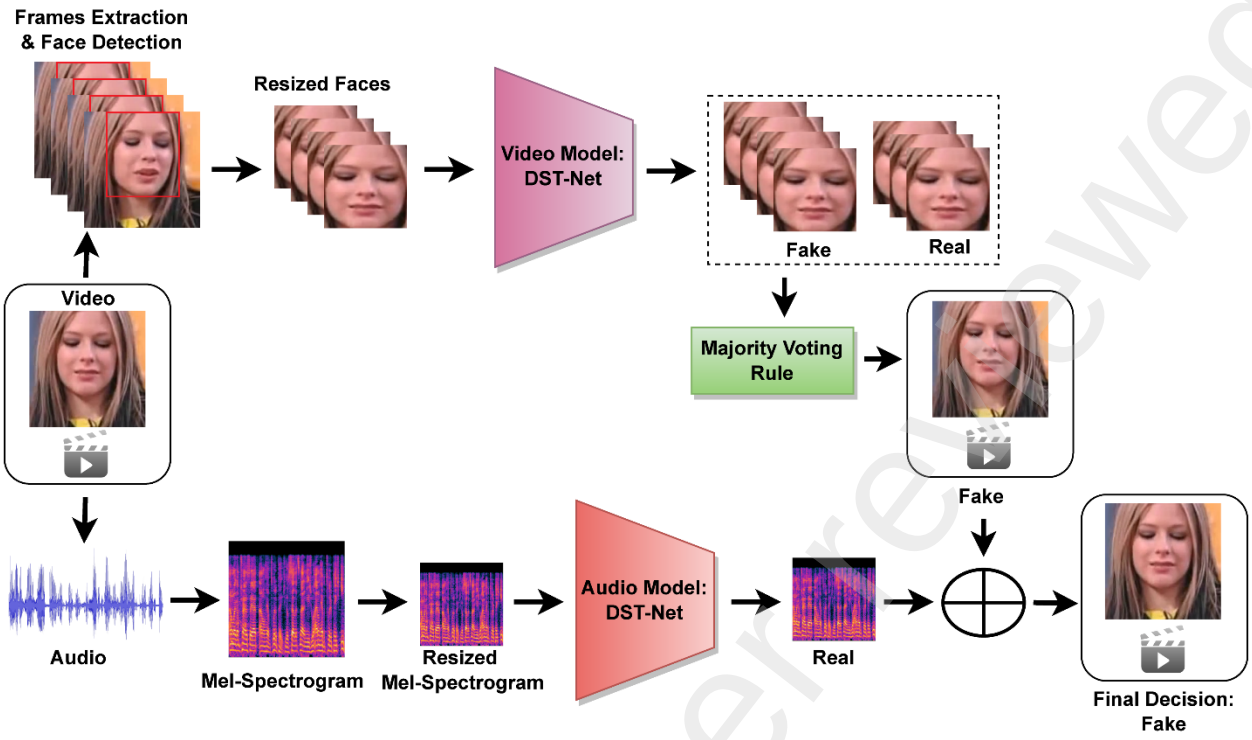

Fig. 2. Workflow of unified framework.

The majority voting rule is demonstrated in Eq. (1).

$$V_s = \max\{real, fake\} \tag{1}$$

where *Vs* denotes the label prediction from the video stream of the framework while *real* and *fake* indicates the real and fake frames count, respectively. Finally, we compare the predicted label from the audio and video stream of the framework, and based on the comparison shown in Eq. (2), we classify the video as fake or real.

$$L(v) = \begin{cases} Fake & \text{if} \quad A_s = fake \vee V_s = fake \\ Real & Otherwise \end{cases} \tag{2}$$

In Eq. (2), *L(v)* represents the overall predicted label for a video, and A_s indicates the label prediction from the audio stream of the network. The algorithm of the classification process is presented as Algorithm 1.

4.2. Dense Swin Transformer network

Our proposed unified model DST-Net consists of an input block (IB), feature extraction block (FEB), and output block (OB). IB has dense layers, FEB compose of a swin transformer module and OB comprises dense and dropout layers. The input and output blocks are placed at the start and end of the whole network while FEB is placed in between IB and OB. The whole network is shown in Fig. 3, and can be expressed as:

$$O_{IB} = Y_{IB}(x_i) \tag{3}$$

$$O_{FEB} = Y_{FEB}(O_{IB}) \tag{4}$$

$$O_{OB} = Y_{OB}(O_{FEB}) \tag{5}$$

where x_i indicates the input image. Y_{IB} (.), Y_{FEB}(.), and Y_{OB}(.) represent the input, feature extraction, and output block, respectively. O_{IB}, O_{FEB}, and O_{OB} indicate the output of the IB, FEB, and OB, respectively. The description of IB, FEB, and OB is provided in the following sections.

Algorithm 1: Classification process of proposed unified framework

Input: Video Repository, $V = \{v_1, v_2, v_3 \ldots, \ldots, v_n\}$
Mel-Spectrogram Image Repository, $M = \{m_1, m_2, m_3 \ldots, \ldots, m_n\}$
Output: Video prediction, V_p

Set $fake \leftarrow 0, real \leftarrow 0$
For each v in V do
  Extract the frames from v and detect the faces $F = \{f_1, f_2, f_3 \ldots, \ldots, f_n\}$ through *MTCNN()*.
  For each f in F **do**
    $f \leftarrow$ Resize (f) // Resized the detected facial frames
    $O_{IB} \leftarrow \mathrm{Y_{IB}}(f)$ // Input Block
    $O_{FEB} \leftarrow \mathrm{Y_{FEB}}(O_{IB})$ // Feature Extraction Block
    $O_{OB} \leftarrow \mathrm{Y_{OB}}(O_{FEB})$ // Output Block
    $\begin{cases} real \leftarrow \text{real} + + & \text{if} \quad O_{OB} == 0 \\ fake \leftarrow fake + + & Otherwise \end{cases}$
  End
  $V_s = \max\{real, fake\}$ // Majority Voting Rule
  Read Mel-Spectrogram Image m of the corresponding video v.
  $m \leftarrow$ Resize (m) // Resized the Mel-Spectrogram Image
  $O_{IB} \leftarrow \mathrm{Y_{IB}}(m)$ // Input Block
  $O_{FEB} \leftarrow \mathrm{Y_{FEB}}(O_{IB})$ // Feature Extraction Block
  $O_{OB} \leftarrow \mathrm{Y_{OB}}(O_{FEB})$ // Output Block
  $A_s \leftarrow O_{OB}$
  $V_p \leftarrow \begin{cases} Fake & \text{if} \quad A_s \leftarrow fake \vee V_s \leftarrow fake \\ Real & Otherwise \end{cases}$ // Final Decision
End

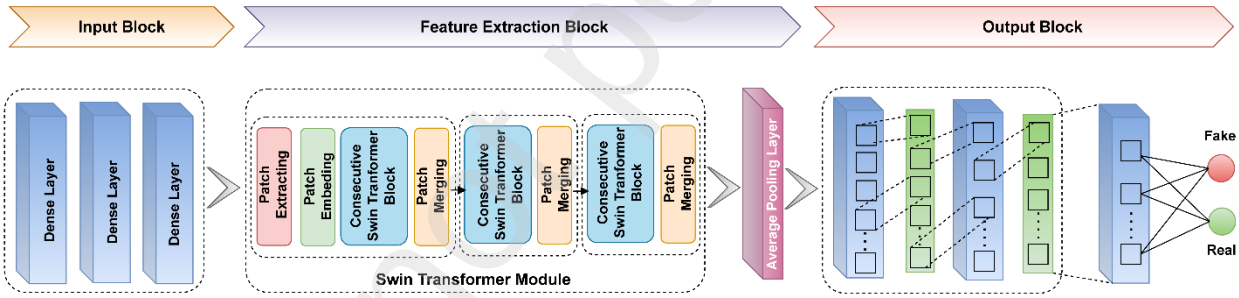

Fig. 3. Dense Swin Transformer network.

4.2.1. Input block

The input block comprises three dense layers, which are employed for primary visual processing and to extract the dense features from the input image. The dense layers transform the image space into the dense, high-dimensional feature space. These features encode the fine details of the input image which can be effective for improving the detection performance. Dense features are then passed to the feature extraction block based on the swin transformer (ST) for further processing.

4.2.2. Feature extraction block

FEB consists of a swin transformer module and a 1D-global average pooling layer at the end. Swin transformer module constructs the hierarchical feature maps starting with the small patches and gradually merging the patches as the network gets deeper in layers. The hierarchical features enable the model to learn effective global and local contextual representations and allows the model to perform dense prediction task. Moreover, the multi-head self-attention module captures the long-range dependencies and expands the receptive field with lesser parameters and lower computational complexity. This leads to better performance while detecting deepfakes videos.

The ST module comprises the patch extracting, patch embedding, two consecutive ST blocks, and patch merging. The patch extracting layer is used to split the incoming dense features into non-overlapping patches. Patch size is set to 3×3 and each patch is considered as a token. The tokens are mapped to vector data via the patch embedding layer,

which is subsequently utilized in transformer blocks. We used the embedded dimension of 64. After that, two consecutive ST blocks are applied to these tokens for the feature extraction. As the network grows, the patch merging layer is utilized to minimize the number of tokens.

Consecutive Swin Transformer blocks*:* The consecutive swin transformer blocks are presented in Fig. 4. Each ST block is composed of layer normalization (LN) layers, multi-head self-attention (MHSA) module, residual connection, and MLP module. Each module (MHSA and MLP) followed a residual connection however, an LN layer is applied before each of these modules. The two consecutive transformer blocks are different from each other in terms of the MHSA module. The first transformer block has a window-based MHSA (W-MHSA) module whereas in the second transformer block, shifted window-based MHSA (SW-MHSA) module is applied. Both modules conduct self-attention within non-overlapping windows, leading the computation complexity to become linear. However, the SW-MHSA module also allows cross-window interaction without any additional computational cost.

In both consecutive swin transformer blocks, the MLP is a four-layered module. The first two layers are identical and composed of Dense, ReLU activation function, and dropout. Similarly, the last two layers are also the same and consist of dense and dropout. Each preceding and succeeding dense layer is fully connected to each other thus enabling the dense feature extraction.

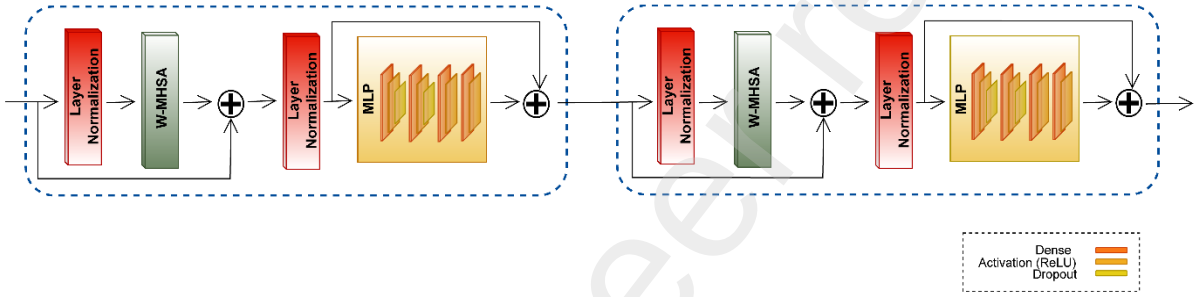

Fig. 4. Consecutive Swin Transformer blocks.

4.2.3. Output block

The feature vector obtained from the feature extraction block is passed to the output block, which transforms the high-dimensional feature space into the output image. OB consists of two dense layers (having ReLU activation function) followed by the dropout layer for regularization. Finally, to classify the input image as either real or fake, a fully connected layer with a softmax activation function is employed. The last fully connected layer has two output neurons for the classification. The softmax function in this layer transforms the neuron's value to 0 or 1 (0 for the real class whereas 1 for the fake class).

5. Experiments and results

In this section, details about the experimental setup and datasets used to evaluate the performance of the proposed DST-Net are provided. To justify the efficacy of our model, the discussion on the results and comparison with state-of-the-art methods are also given. Moreover, cross corpora evaluation of the unified model is also presented in the subsequent sections.

5.1. Dataset

For the detection of deepfakes, researchers have presented large and standard datasets such as FaceForensics++ [27] and Deepfakes Detection Challenge (DFDC) [26], but these datasets have some drawbacks. For instance, FF++ lacks the audios as it only contains manipulated videos with no audio. However, the DFDC dataset encompasses both fake audio and fake video, but the entire video is labeled as fake without specifying whether the audio or video is fake. Therefore, we utilized a recent FakeAVCeleb dataset [34] (comprises the videos having both visual and audio manipulation) for evaluating the performance of our proposed unified framework. Performance of our proposed model has also been evaluated on Celeb-DF [23] and ASVSpoof-2019 LA [38] datasets. Celeb-DF contains only the visual manipulation while ASVSpoof-2019 LA dataset includes only the manipulated speech samples. However, for cross-dataset evaluation, we used World Leader Dataset (WLD) [19] and Presidential Deepfakes Dataset (PDD) [22]. The description of these datasets is provided in the next subsections.

5.1.1. FakeAVCeleb dataset

FakeAVCeleb is an audio-video multimodal deepfake detection dataset having a lip-synced fake video along with synthesized audio. There are 500 real videos of celebrities in the dataset whereas the total number of fake videos is more than 20k. This dataset contains four subsets, RealAudioRealVideo (R_aR_v), FakeAudioFakeVideo (F_aF_v), RealAudioFakeVideo (R_aF_v), and FakeAudioRealVideo (F_aR_v). As the name suggests, R_aR_v includes real videos, the F_aF_v subset contains the fake videos having both audio and visual manipulation whereas F_aR_v and R_aF_v subsets contain the fake videos having only audio manipulation and visual manipulation, respectively. FakeAVCeleb dataset contains videos of individuals having diverse ages and ethnic backgrounds. Moreover, this dataset is unbiased in terms of gender and ethnicity as it contains the videos of both men and women belonging to four ethnic groups i.e., American, European, African, Asian (south), and Asian (east). The average duration of each video is 7 seconds and has only a single individual without any occlusion that might cover the person's face. A few samples of the FakeAVCeleb dataset are shown in Fig. 5.

Fig. 5. FakeAVCeleb dataset.

5.1.2. Celeb-DF (v2) dataset

Celeb-DF(v2) dataset contains visual manipulated deepfakes videos with no voice. The dataset consists of a total of 590 real videos of 59 celebrities gathered from youtube and 5639 deepfakes videos of corresponding real videos. Celeb-DF(v2) dataset includes individuals of various ethnicities, ages, and genders. Moreover, the dataset is challenging since it comprises high-resolution videos with different lighting conditions, orientations, and backgrounds. The frame rate of each video is 30fps and the average duration is appx. 13 seconds. Fig. 6 shows some frames of the Celeb-DF Dataset.

5.1.3. ASVSpoof-2019 LA dataset

ASVspoof-2019 LA dataset encompasses speech data that is captured from 107 individuals including 61 females and 46 males. The dataset is partitioned into three disjoint sets named training, development, and evaluation. The training and development sets include known attacks while the evaluation set contains 11 unknown and only 2 known spoofing attacks. The spoofed audio is generated using the 17 diverse VC, TTS, and hybrid systems.

5.1.4. Presidential Deepfakes dataset

PDD dataset consists of 32 videos of two US presidents Donald Trump and Joseph Biden. Half videos in this dataset are real while the other half are fake videos modified using impersonated audio, lip synchronization, and misleading content. So, in the fake videos, the speech of both presidents is fake as none of them actually spoke such statements as mentioned in the videos. The resolution of each video is 854×480 pixels, the frame rate is 30fps and the duration is between 15s to 30s. Some samples of the PDD dataset are shown in Fig. 7.

Fig. 6. Celeb-DF dataset.

Fig. 7. Presidential Deepfakes dataset.

5.1.5. World Leader dataset

WLD dataset contains the real and deepfakes videos of U.S politicians including Barack Obama, Joe Bidden, Donald Trump, Hillary Clinton, Bernie Sanders, and Elizabeth Warren. The corresponding comedic impersonator of each politician is used to create face-swapped and impersonated deepfake videos via GANs. For Obama, lip sync deepfakes videos are also included in the dataset. The duration of each video is 10 seconds while the frame rate is 30fps. Fig. 8 shows some frame samples of the WLD dataset.

Fig. 8. World Leader dataset.

5.2. Experimental setup and training parameters

The proposed DST-Net is trained from scratch with an image resolution of 128 × 128 (for extracted faces) and 175×175 (for Mel-Spectrograms). In order to find the optimized hyper-parameters for the proposed model, we performed extensive experimentation while tuning the hyper-parameters. After the detailed experiments, the optimized parameters values are: learning rate = 0.001, batch size = 16, label smoothing = 0.1, and weight decay = 0.0001. We trained the model using AdamW optimizer and Binary Cross Entropy loss along with label smoothing. The best model weights are stored using the early stopping on validation accuracy with the patience value of 5. All the experimentations are performed on high-performance computing clusters having the compute nodes with the following specifications: 40 CPU cores at 2.50 GHz and 192 GB RAM.

5.3. Performance evaluation measures

The performance of the proposed method is evaluated using standard metrics such as accuracy, Area under curve (AUC) score, precision, true positive rate (TPR), true negative rate (TNR), and F1-Score.

AUC measures the model's aptitude to distinguish between real and fake videos. The higher AUC indicates better performance of the model for discriminating between real and fake videos.

Accuracy is calculated by the sum of correctly predicted fake and real videos divided by the total number of videos in the test set. Accuracy is computed as follows:

$$Accuracy = \frac{TP+TN}{P+N} \tag{6}$$

TPR is the proportion of correctly predicted fake videos out of all fake videos. It indicates the model's ability to correctly predict the deepfake video as a fake one. TPR is calculated as follows:

$$TPR = \frac{TP}{TP+FN} \tag{7}$$

TNR is the fraction of correctly predicted real videos out of all real videos. It indicates the model's ability to correctly predict the real video as a real one. TNR is calculated as follows:

$$TNR = \frac{TN}{FP+TN} \tag{8}$$

Precision is the ratio of correctly predicted deepfake videos to the total number of fake predictions made by the model. It represents the quality of deepfakes videos prediction made by the model. We computed the precision as follows:

$$Precision = \frac{TP}{TP+FP} \tag{9}$$

F1-Score represents the harmonic mean of precision and TPR (recall) by combining both into a single metric. It is used to assess the performance of models where one model has high precision, and the other model has a higher recall. It can be calculated as follows:

$$F1\ Score = 2 \times \frac{Precision \times TPR\ (Recall)}{Precision + TPR\ (Recall)} \tag{10}$$

where *TP* represents the correctly predicted deepfakes videos and *TN* indicates the correctly detected real videos. *FP* denotes the false predicted deepfake videos and *FN* represents the false detected real videos. *P* and *N* represent the total fake and real videos.

5.4. Detection performance on different spectrograms

We conducted an experiment to analyze the performance of our proposed DST-Net model on different spectrograms of an audio stream. The spectrogram depicts the visualization of the frequency range that the signal contains over time. This experiment is conducted on the FakeAVCeleb dataset. The subsets used for this experiment are F_aF_v and R_aR_v each containing 500 videos. We split the subsets into training and testing sets with a split ratio of 80:20. Chroma-CQT, Gammatone Cepstral Coefficients (GTCC), Mel-Frequency Cepstral Coefficients (MFCC), and Mel-Spectrograms of the videos are computed using the python package librosa. After that, the model is trained and assessed on these spectrograms and the results are demonstrated in Table 1. From Table 1, it can be clearly seen that our proposed model, when evaluated on Mel-Spectrograms provides the highest accuracy of 97.51% and AUC of 97.52%. While, on all other spectrograms (i.e., GTCC, MFCC, and Chroma-CQT) the detection accuracy and AUC are below 90%. Mel-Spectrogram is a Spectrogram converted to a Mel Scale which mimics the working of a human ear. Mel-Spectrogram provides the sound information in a visual form to the model which is similar to the pitches that

humans can perceive. So, the Mel-Spectrogram depicts the audio signal information in a more descriptive way resulting the higher detection accuracy. As a result of these findings, we used the Mel-Spectrograms of the audio stream for all other deepfakes detection experimentations.

Table 1 Performance of DST-Net on different spectrograms.

Spectrograms	Accuracy (%)	AUC (%)
Mel-Spectrograms	97.51	97.52
GTCC	89.5	89.5
MFCC	88.5	88.5
Chroma-CQT	80.10	80.10

5.5. Performance evaluation

To evaluate the efficacy of the proposed model for audio-visual deepfakes detection, we conducted multiple experiments on standard datasets, and details are provided in the subsequent sections. The experimentation protocol in terms of dataset splitting information is provided in Table 2.

Table 2 Datasets details.

Training			Testing		
Split	**Subsets**	**No. of Samples**	**Split**	**Subset**	**No. of Samples**
Audio-Video Dataset FakeAVCeleb					
Train (80%)	R_aR_v	400	Test (20%)	R_aR_v	100
	F_aF_v	8753		F_aF_v	2081
	R_aF_v	7841		R_aF_v	1866
	F_aR_v	400		F_aR_v	100
Video Dataset Celeb-DF (v2)					
Train (80%)	Real	472	Test (20%)	Real	118
	Fake	4511		Fake	1128
Audio Dataset ASVSpoof-2019 LA Dataset					
Subsets		**No. of Bonafide Samples**		**No. of Spoofed Samples**	
Training		2,580		22,800	
Development		2,548		22,296	
Evaluation		7,355		63,882	

5.5.1. Performance evaluation on FakeAVCeleb dataset

To show that our proposed DST-Net is a unified model and capable of reliably detecting both the audio and visual deepfakes, we evaluated the performance of our proposed model on the FakeAVCeleb dataset. For this purpose, we conducted experiments in three different stages. In the first stage, we evaluated the performance of DST-Net using only visual stream, and the model trained on visual stream/modality is termed a video model. In the second stage, performance is evaluated on audio stream only and the trained model is named as audio model. While at the third stage, our proposed unified framework is evaluated on the FakeAVCeleb dataset via utilizing both the audio and video models. So, we evaluated the performance of our proposed model on the FakeAVCeleb dataset for video only, audio only, and audio-video modality.

Augmentation techniques

Because the real subset comprises only 500 videos, therefore, we applied different augmentation techniques to increase the number of real videos to match the number of videos in the fake subsets (F_aF_v, R_aF_v, F_aR_v) of the FakeAVCeleb dataset. The applied video augmentation techniques are: horizontal flip, vertical flip, translation, sharpening, elastic deformation, dropout, gamma correction, gaussian blurring, average blurring, bilateral blurring, median blur, gaussian noise, salt and pepper noise, raise blue channel, raise green channel, raise red channel, raise hue, raise intensity and raise saturation. Few of the frames of the augmented videos in the same above-mentioned order are shown in Fig. 9. Whereas we applied the following audio augmentation techniques: white noise, time stretch, pitch scale, random gain, invert polarity, Gaussian noise, high pass filter, low pass filter, pitch shift, shift, bandpass filter, band-stop filter, high shelf filter, low shelf filter, peaking filter, gain transition, Gaussian noise and pitch shift, pitch shift and high pass filter, Gaussian noise and high pass filter. Some samples of Mel-Spectrograms images of augmented audios in the same above-mentioned order are presented in Fig. 10.

Fig. 9. Frames of the augmented videos.

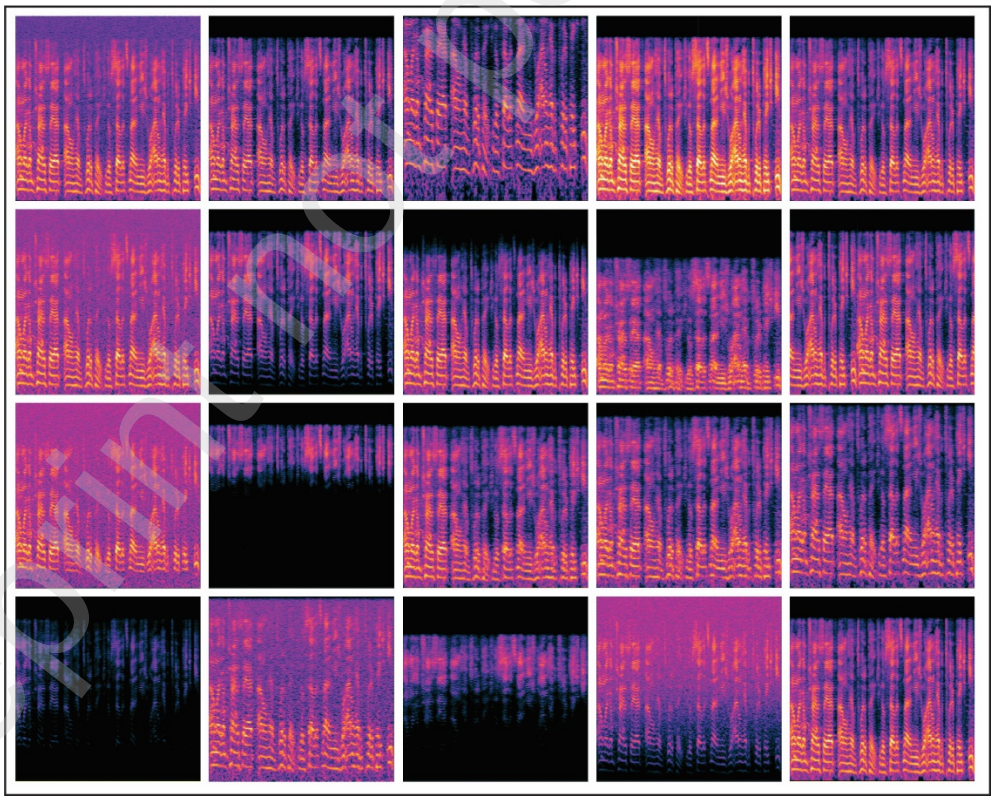

Fig. 10. Mel-spectrograms of augmented audios.

Evaluation on FakeAVCeleb for video-only modality

To evaluate the performance of our model for video-only modality on FakeAVCeleb dataset, we performed two experiments using three subsets (F_aF_v, R_aF_v, and R_aR_v) of the dataset. These subsets are further split into training and testing sets. In the first experiment, we used F_aF_v subset videos as fake and R_aR_v as real videos and trained the model on the extracted frames of training set videos of these subsets. For the second experiment, we used R_aF_v subset videos as fake and R_aR_v videos as the real ones to train the model. The trained models are then evaluated on the videos present in their respective testing sets. The results are shown in Table 3. From the results, it is seen that the video model has achieved an accuracy of 90.94% on F_aF_v and 85.29% on R_aF_v subsets illustrating that the model has the capability to detect the identity swapped and reenacted fake videos accurately. Both subsets contain the face-swapped visual content generated through different techniques i.e., DeepFaceLab [40], FaceSwap [41], and FSGAN [42]. Moreover, in F_aF_v subset, the mouth region is also modified to make it synced with the fake audio using the facial reenactment technique Wav2Lip. For both experiments, the TPR and TNR indicated that the video model predicts the fake videos more accurately as compared to the real ones.

Evaluation on FakeAVCeleb for audio-only modality

To check the effectiveness of DST-Net for audio-only modality, we also performed two experiments using subsets (F_aF_v, F_aR_v, and R_aR_v) of FakeAVCeleb dataset. Both experiments are different in terms of subsets used to train the model. For the first experiment, we used F_aF_v and R_aR_v whereas, for the second experiment, we used F_aR_v and R_aR_v subsets. For both experiments, we computed the Mel-Spectrograms of each video using python library librosa and stored them as 3-channel images. The proposed model is then trained and assessed on the extracted Mel-Spectrograms and the results are presented in Table 3. It can be observed from Table 3 that the audio model attained 98.75% and 94.5 % accuracy on F_aF_v and F_aR_v subsets, respectively indicating that our audio model can accurately detect the fake voices generated using TTS systems. TPR is 99.62% and TNR is 97.72% on F_aF_v subset. Similarly, for F_aR_v subset, TPR is 92% and TNR is 97%. These TPR and TNR values clearly indicate that our audio model can accurately detect both the fake and real Mel-Spectrograms. The precision score on F_aF_v and F_aR_v subsets is 98.06% and 96.84%, respectively indicating the outstanding fake video prediction quality of the model. In both subsets, the fake audio is generated using a real-time voice cloning method named Multispeaker Text-to-Speech Synthesis (SV2TTS) [39]. The fake speech generated via SV2TTS lacks the human voice naturalness and cannot isolate the reference audio prosody from the speaker's voice. The outperforming results indicate that our model captures these differences between real and synthetic voices with high detection accuracy. And also proves our hypothesis that the analysis of Mel-Spectrogram can be effective for detecting deepfakes audio.

Evaluation on FakeAVCeleb for audio-video modality

To analyze the robustness of our unified model for both the audio and video modalities, we performed two experiments using all subsets (F_aF_v, F_aR_v, R_aF_v, R_aR_v) of the FakeAVCeleb dataset. For these experiments, training sets are the same as for video-only and audio-only modality experiments. But the testing sets contain the videos along with their respective Mel-Spectrogram images. In the first experiment, we utilized the video and audio model trained on F_aF_v and R_aR_v subsets of the FakeAVCeleb dataset and evaluated our unified framework on the testing set containing the F_aF_v subset videos along with its Mel-Spectrograms. For the second experiment, the video model is trained on R_aF_v and R_aR_v subsets whereas, the audio model is trained on F_aR_v and R_aR_v subsets. The trained audio and video models are then evaluated on the testing set containing the fake videos and their respective Mel-Spectrograms from the testing set of both subsets (F_aR_v and R_aF_v). Table 3 shows the results of these experiments. Our unified framework achieved the detection accuracy of 92.59% on F_aF_v and 93.41% on R_aF_v + F_aR_v subset. The model achieved the TPR of 99.95% for subset F_aF_v and 94.35% for R_aF_v + F_aR_v subset. Overall, the results shown in Table 3 indicate that our unified framework is more robust in detecting fake videos as compared to the real videos, in the case of the FakeAVCeleb dataset.

As can be seen from Table 3, the model trained on different subsets when evaluated on their respective test sets, most of them show exceptional TPR while some exhibit outstanding precision, so the F1-Score is shown in Table 3 for a more thorough analysis of the proposed model on different subsets in terms of precision and recall. Except for one experiment in video-only modality, F1-Score is above 90% on all experiments performed for video only, audio only, and both modalities, indicating that our proposed model performs well on the FakeAVCeleb dataset. It can also be seen from Table 3 that for audio-video modality experiment, the detection accuracy and F1-Score for F_aF_v subset are slightly lower than the other subset. This may be due to the reason that the facial reenactment technique such as Wav2Lip is applied on video in F_aF_v subset to generate more realistic videos having the facial features modified and lip movement synchronized with the fake audios.

Table 3 Performance evaluation on FakeAVCeleb dataset.

Models	Testing Subsets		Accuracy (%)	AUC (%)	TPR (Recall) (%)	TNR (%)	Precision (%)	F1-Score (%)
	Real	Fake						
Video only	R_aR_v	F_aF_v	90.94	90.65	94.57	86.03	88.61	91.49
	R_aR_v	R_aF_v	85.28	85.09	94.80	75.45	80.04	86.79
Audio only	R_aR_v	F_aF_v	98.73	98.66	99.62	97.72	98.06	98.83
	R_aR_v	F_aR_v	94.5	94.5	92	97	96.84	94.36
Both (Audio and video)	R_aR_v	F_aF_v	92.59	92.01	99.95	84.08	87.91	93.55
	R_aR_v	$R_aF_v + F_aR_v$	93.41	84.67	94.35	75	98.67	96.46

Evaluation on FakeAVCeleb for angled or side-posed faces

To evaluate the performance of our proposed model specifically on the angled or side posed faces, we designed an experiment where we gathered the videos having the side pose of a person from the FakeAVCeleb dataset. After that, we evaluated our trained model on these videos to show their effectiveness on the angled faces. Our model classifies the videos accurately with 100% accuracy and AUC which indicates that DST-Net has the capability to accurately detect the angled fake faces if present in the videos while detecting the deepfakes. Few samples of the angled face from the videos are shown in Fig. 11.

In the extreme side-posed faces, only half region of the face is visible resulting in the loss of significant facial features information and thus making it more difficult to detect the synthetic face accurately. Our proposed DST-Net captures the global long-term dependency and dense hierarchical features which enable them to correctly classify the side-posed faces. Dense layer encodes fine details about the input image and swin transformer in the network architecture extracts the feature maps that have global aware attributes and also establishes the relationship between different image features. Due to these facts, our proposed model is able to detect the fake videos having the extreme side posed faces of the person. Furthermore, there are certain frames in the video when the person's face is at an angle or is looking at the camera rather than being severely side-posed at all times. Such frames can also aid in the reliable identification of real or fake videos with extreme side-facing poses.

Fig. 11. Angled or side-posed faces.

5.5.2. Performance evaluation on Celeb-DF dataset

To evaluate the performance of our proposed model on a diverse, challenging, and only visual manipulated dataset, we designed an experiment to analyze the performance of our model on the Celeb-DF dataset. For this purpose, we split the dataset into training and testing sets. In order to train the DST-Net model, we extracted the faces from the frames of the videos present in the training set. From the training set, 20% of the extracted faces are used for validation purposes during the training. After training the model, we evaluated it on the videos present in the testing set. The model was able to achieve an accuracy of 73.05% and an AUC score of 75.64% on the testing set videos. However, TPR and TNR are 72.07% and 79.21%, respectively. Moreover, the model attained the F1-Score of 82.20% and a precision of 95.65%. The low detection performance on Celeb-DF (v2) dataset is attributed to the fact that this dataset is highly unbalanced, and the dataset is also biased towards the male gender as only 30% of the dataset is comprised of females. Moreover, the dataset has less statistical difference between the real and fake videos. As there is no mismatch of skin color and illumination difference in the swapped fake faces, which may also affect the detection performance.

5.5.3. Performance evaluation on ASVSpoof-2019 LA dataset

In order to investigate the model behavior for a large-scale and standard audio-only dataset, we conducted an experiment where we evaluated the performance of our proposed DST-Net on the ASVSpoof-2019 dataset. More specifically, we used the Logical Access subset of the ASVSpoof-2019 dataset for the assessment of our model. We first generate the Mel-Spectrograms image of the audios present in the training, development, and evaluation set as our model demands images as input. Then, we trained the model on these Mel-Spectrograms using the training and development sets. The development set is used for validation purposes. After that, the trained model is evaluated on the Mel-Spectrograms present in the evaluation set. On ASVSpoof-2019 LA dataset, our model attained the EER of 0.13, accuracy, and AUC of 86.77% of 88.34%, respectively. The LA dataset contains the fake speech samples generated through voice cloning and synthetic speech generation methods. The fake audio generated using VC systems is more difficult to detect as compared to the audio generated via TTS systems. VC systems utilize the human voice as a source, conversely, TTS methods generate synthetic speech using digitized text. So, VC systems generated voice can sustain the prosodic characteristics of a person that the synthetic speech may lack, making the fake speech more realistic. In the presence of such challenging audio samples, the results indicate the effectiveness of our proposed model while detecting the spoofed audios generated through different VC and TTS techniques.

5.6. Comparison with state-of-the-art methods on FakeAVCeleb dataset

To justify and measure the effectiveness of our unified framework, we performed a comparative analysis of our DST-Net with the existing state-of-the-art methods on the FakeAVCeleb dataset. We compared the accuracy of our DST-Net with the methods reported in [25, 36, 35] for video only, audio only, and both (audio and video) modalities. The results of the proposed and existing models in terms of accuracy are provided in Table 4. The proposed method outperforms the existing contemporary models by attaining the highest accuracy of 90.94% for video-only modality, 98.73% for audio-only, and 92.59% for both (video and audio) modalities. In the case of video-only modality, Meso-4 is the worst performing model while the VGG16 performed the second best. For audio-only modality, our proposed DST-Net outperforms the second-best model with an average accuracy gain of 22%. XceptionNet is the worst performer for the detection of deepfakes via both modalities (audio and video). However, POI-Forensics is the second-best performing model for audio-visual deepfakes detection, but it has the limitation of requiring reference real video of the target subject at the testing time. We can conclude from this comparative analysis that the proposed framework outperforms the existing models and is capable of accurately identifying the deepfakes video via detecting manipulation in both streams (audio and video). It is important to mention that our proposed DST-Net also performed better over the baseline models for deepfakes detection using audio-only and video-only modalities of the FakeAVCeleb dataset.

Table 4 Comparison with existing models on FakeAVCeleb dataset.

Models	Accuracy (%)		
	Audio and Video Modality	**Video only Modality**	**Audio Only Modality**
XceptionNet [35]	43.94	73.06	76.26
Meso-4 [35]	45.93	43.15	50.36
EfficientNet-B0 [35]	63.18	59.64	50
MesoInception-4 [35]	72.87	77.88	53.96
VGG16 [35]	78.04	81.03	67.14
VFD [36]	81.52	--	--
POI-Forensics [25]	86.6	--	--
DST-Net (proposed)	92.59	90.94	98.73

5.7. Comparison with existing methods on ASVSpoof-2019 LA dataset

To investigate the performance of our model against the existing acoustic features extraction methods on the LA dataset, we evaluated our DST-Net with state-of-the-art (SOTA) methods [14,15,24,43]. The purpose of this analysis is to show that the Mel-Spectrograms can also be worthwhile for fake audio detection besides the acoustic features used for classifying synthetic speech. The performance comparison based on EER is shown in Table 5.

From Table 4, it is observable that our model achieves the EER of 0.13%, which is 0.61% less than the second-best performing model. However, our model performs almost equivalent to our prior method [14] and shows that it is remarkably good at the detection of logical access attacks. According to our expectations, DST-Net shows incredible classification performance and proves that Mel-Spectrograms provide good performance compared to conventional acoustic feature extraction for the detection of fake audios. Thus, it can be concluded that the image visualization of audios in terms of Mel-Spectrogram can also be effective for classifying fake audios.

Table 5 Performance comparison with existing SOTA methods.

Method	EER (%)
Hassan et al. [24]	3.05
Javed et al. [43]	0.75
Arif et al. [15]	0.74
Javed et al. [14]	0.1
Proposed Model	**0.13**

5.8. Cross corpora evaluation

We performed cross corpora evaluation to evaluate the generalization ability of our proposed unified framework, which is further subclassified as cross-set evaluation and cross-dataset evaluation. In cross-set evaluation, the models are trained on one subset and tested on another subset of the FakeAVCeleb dataset. Whereas in cross-dataset evaluation, models trained on subsets of the FakeAVCeleb dataset are used to test the videos of other datasets (i.e., PDD, WLD). The main goal of cross corpora evaluation is to analyze the potential and applicability of our proposed unified model in real-world scenarios for deepfakes detection.

5.8.1. Cross-set evaluation

The cross-set evaluation experiment is carried out to demonstrate the generalizability of the proposed model on different subsets of the FakeAVCeleb dataset. For this purpose, experimental protocols are kept the same as mentioned for audio-video modality experiment in Section 5.5.1. This experiment is conducted in two phases. In the first phase, audio and video models (trained on F_aF_v subset) are used to evaluate the testing test containing the videos and respective Mel-Spectrograms of R_aF_v and F_aR_v subsets. Similarly, in the second phase, the audio model (trained on F_aR_v subset) and video model (trained on R_aF_v subset) are used to assess the testing set containing the videos and Mel-Spectrograms of F_aF_v subset of the FakeAVCeleb dataset. The results of the cross-set evaluation are provided in Table 6.

From Table 6, it is seen that when evaluated on $R_aF_v + F_aR_v$ subset, the video model and audio model (trained on F_aF_v subset) have attained the precision of 99.29%. However, the video model (trained on R_aF_v) and audio model (trained on F_aR_v subset), when tested on F_aF_v subset, have achieved a TPR of 99.42%. F1-Score is reported in Table 6 for better understanding as one unified model achieves high recall and the other attains high precision. The F1-Score of 87.29% and 88.36% on $R_aF_v + F_aR_v$ and F_aF_v subsets, respectively demonstrates that the model is quite effective at detecting the deepfake videos. Table 6 shows that under cross-set evaluation settings, the proposed framework achieved an AUC score of 83.43% on $R_aF_v + F_aR_v$ subset and 84.87% on F_aF_v subset. The difference in the accuracy and AUC for $R_aF_v + F_aR_v$ testing subset is attributed to the fact that the class imbalance problem exists as the fake videos are greater in number as compared to the real ones. The results highlight that the proposed unified framework has great generalization aptitude for detecting deepfakes videos under a cross-set evaluation setting. It can also be inferred that the models trained on the videos having manipulation in both streams can reliably detect the videos having either fake audio or fake visual content.

Table 4 Cross-set evaluation on FakeAVCeleb dataset.

Training Subsets	Testing Subset	Accuracy (%)	AUC (%)	TPR (%)	TNR (%)	Precision (%)	F1-Score (%)
Video Model: F_aF_v Audio Model: F_aF_v	$R_aF_v + F_aR_v$	78.41	83.43	77.87	89	99.29	87.29
Video Model: R_aF_v Audio Model: F_aR_v	F_aF_v	85.94	84.87	99.42	70.32	79.52	88.36

5.8.2. Cross-dataset evaluation

The main purpose of the cross-dataset evaluation is to analyze the generalizability of the unified framework over completely different datasets. In this experiment, the audio and video models are trained on the F_aR_v and R_aFv subsets of the FakeAVCeleb dataset, respectively. Experimental protocols for training the audio and video models are the same as mentioned in Section 5.5.1. The trained models are then evaluated on unseen datasets in three phases. In the first phase, videos of different subsets of WLD are tested. In the second phase, models are evaluated on the PDD dataset. In the third phase, we applied different augmentation techniques to the PDD dataset and then tested the augmented videos. The cross-dataset evaluation results are shown in Table 7.

Table 5 Cross-dataset evaluation.

Training Subsets	Testing Subset	Accuracy (%)	AUC (%)
Video Model: R_aF_v Audio Model: F_aR_v	WLD - FaceSwap	73.98	74.52
	WLD - Imposter	61.74	60.15
	WLD- LipSync	69.32	53.69
	PDD - full	78.12	78.12
	PDD-aug-full	62.34	62.34

The real-world scenarios for the fake videos are included in the WLD and PDD datasets. In the FaceSwap subset of the WLD dataset, a more realistic fake video is created by swapping the face of the leader with their respective imposter. The accuracy of 73.98% on such realistic fake videos indicates that our proposed unified model is capable of accurately detecting the totally unseen real-world face-swapped videos. On Imposter and LipSync subsets, the accuracies are expected to be lower because the Imposter subset involves the real person impersonating himself as a leader making it harder for the model to identify the impersonated video. However, the LipSync subset consists of lip-synced videos of Obama in which only the mouth area is modified according to the speech. Therefore, it is more challenging for the model to detect manipulated videos due to very little semantic change in the lip-synced video. It is also important to note that, in all these subsets the audio stream is not manipulated. The accuracies of 61.74% and 69.32% on Imposter and LipSync subsets demonstrate that the proposed model performed fairly well on these subsets in a cross-dataset setting. The PDD dataset contains the fake videos of Donald Trump and Joseph Bidden which are lip synchronized according to the impersonated audio. The audio in these videos is not synthesized using any fake audio generation techniques, however, voice-over actors are used for producing impersonated audio making them more difficult to detect. Both of the leaders appear to be saying things that they have not really spoken about. Our unified model detects such misleading content with an accuracy of 78.12%. As the PDD dataset is very small, so we utilized augmentation techniques such as noise, blurring, etc., to extend it by making it more diverse and challenging which causes a decrease in the detection accuracy of our model on the augmented videos of the PDD dataset. Additionally, all the datasets used in cross-dataset evaluation are diverse and distinct from each other in terms of illumination conditions, video capturing devices, and manipulation techniques. It can be concluded from the detection accuracies reported in Table 7 that our model is generalizable and can be used to reliably detect real-world fake videos.

6. Conclusion

In this paper, we have presented a unified framework that is able to detect the deepfakes via identifying the manipulation in audio and visual streams of a video. We proposed a novel unified DST-Net model that accurately detects both audio and video deepfakes. DST-Net is evaluated on a challenging and diverse FakeAVCeleb dataset for audio only, video only, and both (audio and video) modalities. Our proposed model not only identifies the deepfakes videos accurately but also outperforms the contemporary models. To show the effectiveness of our model for visual-only and audio-only manipulation, we evaluated it on challenging Celeb-DF and ASVSpoof-2019 LA datasets. We have also conducted a cross corpora evaluation of our unified framework on FakeAVCeleb, PDD, and WLD datasets to demonstrate its efficacy and applicability in real-world scenarios. Extensive experimentations show that the proposed approach is effective and robust in detecting deepfakes videos having manipulation in both the audio and visual streams. In future research, we intend to further improve the performance of our model for cross-corpora evaluation.

Funding

This work was supported by the grant of the Punjab Higher Education Commission (PHEC) of Pakistan via Award No. (**PHEC/ARA/PIRCA/20527/21**) and NSF of USA via Award No. **1815724**.

Acknowledgment

This work was supported by the Multimedia Signal Processing (MSP) research lab at the University of Engineering and Technology (UET) Taxila. We would like to thank Prof. Hany Farid from the University of California Berkeley to provide us with their World Leaders Dataset for performance evaluation.

References

[1] Kolagati, S., Priyadharshini, T. and Rajam, V.M.A., 2022. Exposing deepfakes using a deep multilayer perceptron–convolutional neural network model. *International Journal of Information Management Data Insights*, *2*(1), p.100054.

[2] Kohli, A. and Gupta, A., 2022. Light-weight 3DCNN for DeepFakes, FaceSwap and Face2Face facial forgery detection. *Multimedia Tools and Applications*, pp.1-13.

[3] Wang, R., Juefei-Xu, F., Huang, Y., Guo, Q., Xie, X., Ma, L. and Liu, Y., 2020, October. Deepsonar: Towards effective and robust detection of ai-synthesized fake voices. In *Proceedings of the 28th ACM international conference on multimedia* (pp. 1207-1216).

[4] Hua, G., Teoh, A.B.J. and Zhang, H., 2021. Towards end-to-end synthetic speech detection. *IEEE Signal Processing Letters*, *28*, pp.1265-1269.

[5] Güera, D., Baireddy, S., Bestagini, P., Tubaro, S. and Delp, E.J., 2019. We need no pixels: Video manipulation detection using stream descriptors. *arXiv preprint arXiv:1906.08743*.

[6] Ciftci, U.A., Demir, I. and Yin, L., 2020. Fakecatcher: Detection of synthetic portrait videos using biological signals. *IEEE transactions on pattern analysis and machine intelligence*.

[7] Qi, H., Guo, Q., Juefei-Xu, F., Xie, X., Ma, L., Feng, W., Liu, Y. and Zhao, J., 2020, October. Deeprhythm: Exposing deepfakes with attentional visual heartbeat rhythms. In *Proceedings of the 28th ACM international conference on multimedia* (pp. 4318-4327).

[8] Hernandez-Ortega, J., Tolosana, R., Fierrez, J. and Morales, A., 2020. Deepfakeson-phys: Deepfakes detection based on heart rate estimation. *arXiv preprint arXiv:2010.00400*.

[9] Chintha, A., Thai, B., Sohrawardi, S.J., Bhatt, K., Hickerson, A., Wright, M. and Ptucha, R., 2020. Recurrent convolutional structures for audio spoof and video deepfake detection. *IEEE Journal of Selected Topics in Signal Processing*, *14*(5), pp.1024-1037.

[10] Fernandes, S., Raj, S., Ortiz, E., Vintila, I., Salter, M., Urosevic, G. and Jha, S., 2019. Predicting heart rate variations of deepfake videos using neural ode. In *Proceedings of the IEEE/CVF international conference on computer vision workshops* (pp. 0-0).

[11] de Lima, O., Franklin, S., Basu, S., Karwoski, B. and George, A., 2020. Deepfake detection using spatiotemporal convolutional networks. *arXiv preprint arXiv:2006.14749*.

[12] Gao, Y., Vuong, T., Elyasi, M., Bharaj, G. and Singh, R., 2021. Generalized spoofing detection inspired from audio generation artifacts. *arXiv preprint arXiv:2104.04111*.

[13] Das, R.K., Yang, J. and Li, H., 2021, June. Data augmentation with signal companding for detection of logical access attacks. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 6349-6353). IEEE.

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

[15] Arif, T., Javed, A., Alhameed, M., Jeribi, F. and Tahir, A., 2021. Voice spoofing countermeasure for logical access attacks detection. *IEEE Access*, *9*, pp.162857-162868.

[16] Gomez-Alanis, A., Peinado, A.M., Gonzalez, J.A. and Gomez, A.M., 2019, September. A light convolutional GRU-RNN deep feature extractor for ASV spoofing detection. In *Proc. Interspeech* (Vol. 2019, pp. 1068-1072).

[17] Jiang, Z., Zhu, H., Peng, L., Ding, W. and Ren, Y., 2020. Self-Supervised Spoofing Audio Detection Scheme. In *INTERSPEECH* (pp. 4223-4227).

[18] Zhang, Y., Jiang, F. and Duan, Z., 2021. One-class learning towards synthetic voice spoofing detection. *IEEE Signal Processing Letters*, *28*, pp.937-941.

[19] Agarwal, S., Farid, H., Gu, Y., He, M., Nagano, K. and Li, H., 2019, June. Protecting World Leaders Against Deep Fakes. In *CVPR workshops* (Vol. 1, p. 38).

[20] Jung, T., Kim, S. and Kim, K., 2020. Deepvision: Deepfakes detection using human eye blinking pattern. *IEEE Access*, *8*, pp.83144-83154.

[21] Masood, M., Nawaz, M., Malik, K.M., Javed, A., Irtaza, A. and Malik, H., 2022. Deepfakes Generation and Detection: State-of-the-art, open challenges, countermeasures, and way forward. *Applied Intelligence*, pp.1-53.

[22] Sankaranarayanan, A., Groh, M., Picard, R. and Lippman, A., 2021. The presidential deepfakes dataset. In *Proceedings of the AIofAI Workshop at the International Joint Conference on Artificial Intelligence*.

[23] Li, Y., Yang, X., Sun, P., Qi, H. and Lyu, S., 2020. Celeb-df: A large-scale challenging dataset for deepfake forensics. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 3207-3216).
[24] Hassan, F. and Javed, A., 2021, April. Voice spoofing countermeasure for synthetic speech detection. In *2021 International Conference on Artificial Intelligence (ICAI)* (pp. 209-212). IEEE.
[25] Cozzolino, D., Nießner, M. and Verdoliva, L., 2022. Audio-Visual Person-of-Interest DeepFake Detection. *arXiv preprint arXiv:2204.03083*.
[26] Dolhansky, B., Bitton, J., Pflaum, B., Lu, J., Howes, R., Wang, M. and Ferrer, C.C., 2020. The deepfake detection challenge (dfdc) dataset. *arXiv preprint arXiv:2006.07397*.
[27] Rossler, A., Cozzolino, D., Verdoliva, L., Riess, C., Thies, J. and Nießner, M., 2019. Faceforensics++: Learning to detect manipulated facial images. In *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 1-11).
[28] Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A. and Bengio, Y., 2014. Generative adversarial nets. *Advances in neural information processing systems*, *27*.
[29] Kingma, D.P. and Welling, M., 2013. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*.
[30] Westerlund, M., 2019. The emergence of deepfake technology: A review. *Technology Innovation Management Review*, *9*(11).
[31] Zhou, Y. and Lim, S.N., 2021. Joint audio-visual deepfake detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* (pp. 14800-14809).
[32] Chugh, K., Gupta, P., Dhall, A. and Subramanian, R., 2020, October. Not made for each other-audio-visual dissonance-based deepfake detection and localization. In *Proceedings of the 28th ACM international conference on multimedia* (pp. 439-447).
[33] Gu, Y., Zhao, X., Gong, C. and Yi, X., 2020, November. Deepfake Video Detection Using Audio-Visual Consistency. In *International Workshop on Digital Watermarking* (pp. 168-180). Springer, Cham.
[34] Khalid, H., Tariq, S., Kim, M. and Woo, S.S., 2021. FakeAVCeleb: a novel audio-video multimodal deepfake dataset. *arXiv preprint arXiv:2108.05080*.
[35] Khalid, H., Kim, M., Tariq, S. and Woo, S.S., 2021, October. Evaluation of an audio-video multimodal deepfake dataset using unimodal and multimodal detectors. In *Proceedings of the 1st workshop on synthetic multimedia-audiovisual deepfake generation and detection* (pp. 7-15).
[36] Cheng, H., Guo, Y., Wang, T., Li, Q., Ye, T. and Nie, L., 2022. Voice-Face Homogeneity Tells Deepfake. *arXiv preprint arXiv:2203.02195*.
[37] Zhang, K., Zhang, Z., Li, Z. and Qiao, Y., 2016. Joint face detection and alignment using multitask cascaded convolutional networks. *IEEE signal processing letters*, *23*(10), pp.1499-1503.
[38] Wang, X., Yamagishi, J., Todisco, M., Delgado, H., Nautsch, A., Evans, N., Sahidullah, M., Vestman, V., Kinnunen, T., Lee, K.A. and Juvela, L., 2020. ASVspoof 2019: A large-scale public database of synthesized, converted and replayed speech. *Computer Speech & Language*, *64*, p.101114.
[39] Jia, Y., Zhang, Y., Weiss, R., Wang, Q., Shen, J., Ren, F., Nguyen, P., Pang, R., Lopez Moreno, I. and Wu, Y., 2018. Transfer learning from speaker verification to multispeaker text-to-speech synthesis. *Advances in neural information processing systems*, *31*.
[40] Perov, I., Gao, D., Chervoniy, N., Liu, K., Marangonda, S., Umé, C., Dpfks, M., Facenheim, C.S., RP, L., Jiang, J. and Zhang, S., 2020. DeepFaceLab: Integrated, flexible and extensible face-swapping framework. *arXiv preprint arXiv:2005.05535*.
[41] Korshunova, I., Shi, W., Dambre, J. and Theis, L., 2017. Fast face-swap using convolutional neural networks. In *Proceedings of the IEEE international conference on computer vision* (pp. 3677-3685).
[42] Nirkin, Y., Keller, Y. and Hassner, T., 2019. Fsgan: Subject agnostic face swapping and reenactment. In *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 7184-7193).
[43] Javed, A., Malik, K.M., Malik, H. and Irtaza, A., 2022. Voice spoofing detector: A unified anti-spoofing framework. *Expert Systems with Applications*, *198*, p.116770.