Universiteit
Leiden
The Netherlands

Bachelor Computer Science

TimeSync:

Deepfake Detection with Temporal Lip Sync

Daan Griffioen

Supervisors:
Dr. Erwin M. Bakker
Prof. Dr. Michael S.K. Lew

BACHELOR THESIS

Leiden Institute of Advanced Computer Science (LIACS)
www.liacs.leidenuniv.nl

24/7/2024

Abstract

Deepfakes are Artificial images or videos produced by Deep Neural Networks (DNNs) in which subjects do or say things they never did. These deepfakes make it possible to alter or completely fabricate visuals and audio in videos and images. Deepfakes have become more sophisticated up to the point that many people cannot tell they are fake. Many deepfake methods are easily accessible online and are easy to use by almost anyone, leading to the online proliferation of highly realistic deepfaked materials. Therefore it is vital that methods exist to detect deepfakes to make sure that people are not being misled. This work, an improvement upon an existing deepfake detection method by adding a Temporal Transformer Network component to amplify the differences between genuine and deepfaked videos in order to boost performance. This novel method called TimeSync has a similar performance on parts of the benchmark datasets and sometimes outperforms the baseline deepfake detection method.

Contents

1 Introduction

There are various deepfake generation methodologies, ranging from audio to video to text generation. All of these generation methodologies have become more sophisticated in recent years. These methods are mostly built upon Generative Adversarial Networks (GANs)[GPAM+14]. With this underlying framework, deepfake generation methods can produce highly realistic deepfakes. Furthermore these generation methods are freely available and so easy to use that almost anyone is able to produce these realistic deepfakes. The result is that there are many deepfaked videos and images online, some of these deepfakes have become so realistic that people cannot tell the difference. This is a problem considering that malicious actors can use deepfakes to spread misinformation in an attempt to erode trust in governments or journalism or incite violence. Deepfakes are also being used in more direct approaches such as spreading misinformation about specific people, generating revenge pornography, or using cloned voices in sophisticated phishing attacks to mislead people over the phone [MAA+23].

While the improvement in deep learning techniques opened the door for ever more realistic deepfakes, it also facilitates more robust and precise detection methods. However, the field of deepfake generation is in a constant state of development, making deepfakes more realistic. It is therefore vital that deepfake detection keeps up. As a consequence the field of deepfake detection is in a constant state of development, where the availability of large datasets like Facebook's Deepfake Detection Challenge [DBP+20], FakeAVCeleb [KTW21] and DeeperForensics [JLW+20] play an essential role. Making it possible to train and test deepfake detection methods and gauge their effectiveness on unseen deepfake generation methods.

In this thesis, an addition to an existing deepfake detection method is proposed to attempt to increase its effectiveness at multimodal deepfake detection. To do this firstly the LipForensics deepfake detection method proposed by Haliassos et al. [HVPP21] will be modified following the proposed implementation by Shahzad et al. [SHK+22]. Furthermore, a Temporal Transformer Network will be added to the network, the latter is inspired by Lohit et al. [LWT19] showing that this module is capable of producing warped representations that increase variance between classes, these representations have shown improved results on Temporal Convolutional Networks (TCN). TCNs are used in the approaches of Haliassos et al. [HVPP21] and Shahzad et al. [SHK+22]. This novel method called TimeSync has a similar performance on parts of the benchmark datasets and sometimes outperforms the deepfake detection method by Shahzad et al. [SHK+22]. It is very capable of detecting visual deepfakes. However, it is less adept at detecting deepfakes that are solely audio-based. TimeSync is trained and tested on the FakeAVCeleb dataset [KTW21] with an addition of videos from VoxCeleb [CNZ18]. The main contributions of this research are

- Timesync, an audio-visual deepfake detection model pre-trained on lipreading that can distinguish between real and deepfaked videos using the discrepancies between high-level semantic features extracted from the lips of a target video and synthetically generated lips from a Wav2Lip model.
- Extensive experiments on each part of the FakeAVCeleb dataset demonstrate that TimeSync can outperform state-of-the-art visual deepfake detection models on several parts of the multimodal FakeAVCeleb dataset.

The rest of the paper is organized as follows; in Section 2 the related work and state-of-the-art are discussed. In Section 3 the fundamentals are introduced. The baseline implementation is discussed

in Section 4. The proposed implementation and architecture of TimeSync are discussed in Section 5. The details on the experimental setup, dataset, preprocessing, hyperparameter selection, and data splits are discussed in Section 6. In Section 7 the experimental results are given. Finally, conclusions and recommendations for further research are discussed in Section 8.

2 Related Work

In this section, several state-of-the-art deepfake detection methodologies will be discussed as well as several important datasets used in deepfake detection research.

There are many different approaches used to generate deepfakes, mostly these generation techniques utilize Generative Adversarial Networks (GAN) [GPAM+14], Variable Autoencoders (VAE) [KW13] and Autoencoders (AE) [KW13]. These methods coupled with an enormous amount of data make it possible to construct realistic and convincing deepfakes. There are numerous inventive deepfake detection methods that focus on different aspects and traces of these generation methods. A recent development in the field of deepfake detection is the work of Yang et al. [YZC+23] which besides proposing a method that surpasses the state-of-the-art also defines the state-of-the-art with comprehensive experiments on multiple deepfake detection models and datasets. There are two main groups of detection methodologies, unimodal and multimodal methods.

Unimodal methods
Haliassos et al. utilize a spatio-temporal network pre-trained on lipreading which outputs internal representations relating to mouth movement. These representations are then used by a Multi-Scale Temporal Convolutional Network (MS-TCN) to detect deepfaked videos. This approach is robust to various common corruptions as well as generalization to unseen deepfake forgery methods, this method is vulnerable to lip obstruction or deepfakes where the mouth has not been altered [HVPP21].
Zhao et al. propose Multi-attentional Deepfake Detection, a novel method that approaches deepfake detection not as a binary classifier but as a fine-grained classification problem. It is implemented by using multiple spatial attention heads to force the network to capture multiple local discriminative features from multiple face attentive regions, a textural feature enhancement block to extract and enhance textural information and bilinear attention pooling for aggregating textural and semantic features [ZWZ+21].
Rössler et al. released FaceForensics++, a dataset containing a large number of deepfaked videos. They also released an XceptionNet model trained on that dataset, XceptionNet is a convolutional Neural Network that is based on depthwise separable convolutional layers with residual connections. XceptionNet is pre-trained on ImageNet and finetuned on Rössler et al.'s dataset [RCV+19].

Multimodal methods
Shahzad et al. propose Lip Sync Matters, an approach built on LipForensics proposed by Haliassos et al.[HVPP21] by utilizing synthetic lips made using a Wav2Lip model alongside genuine lips. This approach adds the use of another modality to the approach proposed by Haliassos et al. With which multimodal deepfakes can be tackled. This multimodal approach showed promising results, improving upon the state-of-the-art on the FakeAVCeleb dataset[SHK+22].

Lomnitz et al. proposed a method that involves an ensemble of unimodal methodologies to make one multimodal method [LHASS20], while others focus on directly combining audio-visual features like Agarwal et al. who focus on the inconsistencies between sound and the lips of the deepfaked face [AFFA20].
Mittal et al. propose an audio-visual deepfake detection method that compares the similarity of the audio and visual modalities in a video and also compares the affective cues corresponding to perceived emotions from the two modalities within a video. A shortcoming of this model is that it fails to classify a fake video as fake because the deepfake does not contain a mismatch between the two modalities. This is also caused by the difference in which humans express emotion [MBC+20].
More recently Yang et al. proposed audio-visual joint learning for Detecting Deepfakes (AVoiD-DF) that focuses on audio-visual inconsistencies at a temporal and spatial scale. Experimental results show that AVoiD-DF outperforms both uni-and multi-modal state-of-the-art deepfake detection methods on multiple datasets[YZC+23].

Datasets
In deepfake detection research, there are many available datasets each containing different types and amounts of deepfake content. For example Facebook's Deepfake Detection Challenge (DFDC) dataset, which contains eight different facial modification algorithms and more than 100.000 videos. This dataset was released in 2019 to measure the progress of deepfake detection technology and is widely used in deepfake detection research [DBP+20].
More recently FakeAVCeleb was released, a multimodal deepfake dataset that contains over 20.000 videos divided across four different ethnic groups. Caucasian, Black, Asian (Southern) and Asian (Eastern). The male and female ratio of each ethnic group is 50%. The dataset is split into 4 categories, FakeVideo-FakeAudio, FakeVideo-RealAudio, RealVideo-FakeAudio and RealVideo-RealAudio. This dataset is widely used in multimodal deepfake detection research [KTW21].
FaceForensics++ is a dataset that contains 1000 original videos that have been manipulated with four different facial forgery methods. The dataset also provides differing levels of compression to simulate lower-quality videos that could occur in the real world [RCV+19].
A database often used in deepfake detection and generation research is the VoxCeleb2 database that contains more than 1.000.000 samples of unaltered speech from more than 7.000 speakers. With speakers from different ethnicities, ages and accents [CNZ18].

In this research the Lip Sync Matters method proposed by Shahzad et al. [SHK+22] is altered by adding a Temporal Transformer Network (TTN), this network aims to reduce variance inside classes and stimulate variance between classes. This method has been shown to provide improved results for TCN and LSTM classifiers [LWT19]. Which is the type of classifier that is used in Lip Sync Matters proposed by Shahzad et al. [SHK+22]. This is a novel approach to multimodal deepfake detection that incorporates the latest advancements in transformer architectures in a state-of-the-art technique utilizing temporal data results in an improvement in accuracy and robustness.

3 Fundamentals

The metrics used in this research to compare the performance of the different deepfake detection models are precision, recall, F1-score and Accuracy. These are defined by the following equations:

$$Precision = \frac{tp}{tp + fp} \tag{1}$$

$$Recall = \frac{tp}{tp + fn} \tag{2}$$

$$Accuracy = \frac{tp + tn}{tp + tn + fp + fn} \tag{3}$$

$$F1 - score = \frac{2 \times precision \times recall}{precision + recall} \tag{4}$$

Where 'tp' stands for the number of true positives, 'tn' for the number of true negatives, 'fp' for the number of false positives and 'fn' for the number of false negatives.

Another metric that is used is Area Under Curve (AUC), which is a measure calculated from the Receiver Operating Characteristics (ROC) curve. This ROC-curve depicts the model's ability to distinguish the positive and negative classes. A higher curve means better prediction ability of the model, and the higher the curve. From the area under this curve the AUC measure is calculated, with a larger area under the curve meaning better performance.
These metrics are widely used in the field of deepfake detection research.

4 Baseline Lip Sync Matters [SHK+22]

In this section, the baseline deepfake detection methodology by Shahzad et al. will be discussed.

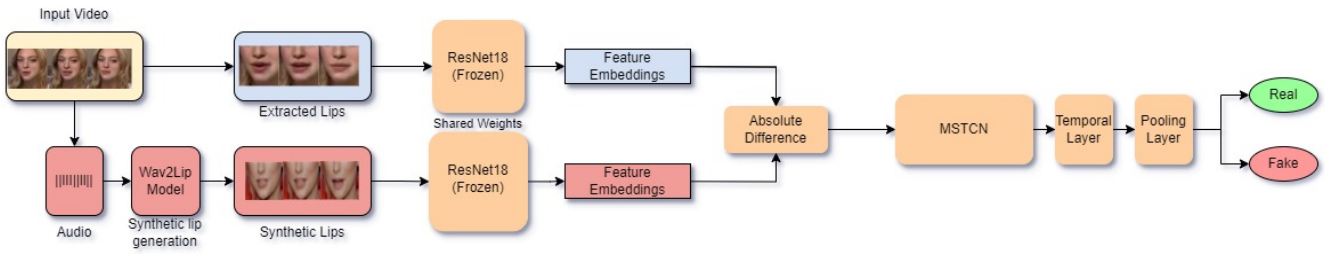

Figure 1: Lip Sync Matters deepfake detection architecture as proposed by Shahzad et al. [SHK+22]

The baseline detection method is a proposed deepfake detection method proposed by Shahzad et al. [SHK+22]. Shahzad et al. introduce a modified version of the face forgery detector proposed by Haliassos et al. [HVPP21]. This modified version utilizes two ResNet18 modules with a preceding 3D convolutional layer as feature extractors which output 512-D embeddings for each frame of the input. The input for the first ResNet18 module is a clip of 25 frames of cropped lips from a video. The input for the second ResNet18 module is a clip of 25 frames of cropped lips which have been generated from the audio using a Wav2Lip model adapted from [HPM+21]. Following the Resnet18 modules the absolute difference between the two resulting embeddings is calculated

and fed into a Multi-Scale Temporal Convolutional Network (MS-TCN) module, which serves to capture patterns over multiple time scales. The MS-TCN module is followed by a global average temporal pooling layer, and finally a linear classifier. The ResNet modules and MS-TCN are all pre-trained on lipreading tasks, during training the ResNet modules are frozen while the MS-TCN module is finetuned. This architecture is depicted in Figure 1

Because the code for the implementation of Shahzad et al. was not publicly available the baseline used in this research is a recreation adapted from LipForensics by Haliassos et al. [HVPP21]. The baseline follows the alterations described by Shahzad et al. [SHK+22] as closely as possible.

5 TimeSync

In this section our proposed novel deepfake detection method TimeSync will be described in detail.

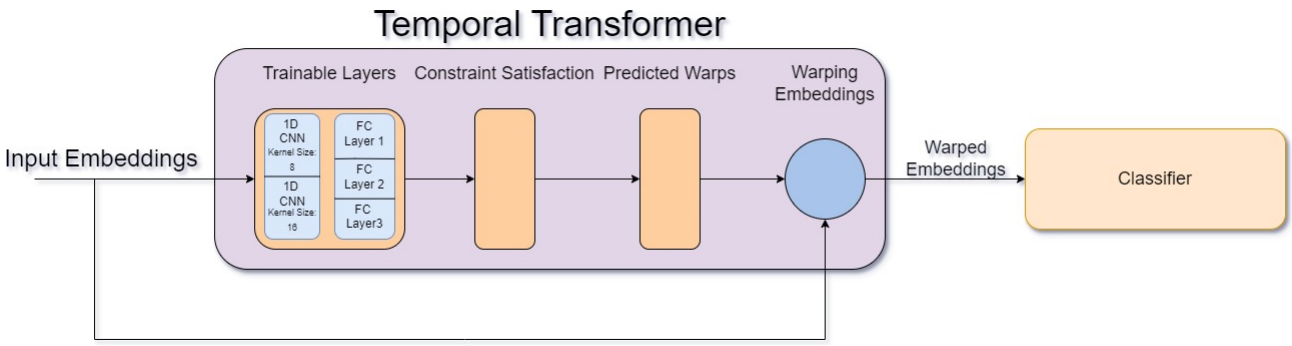

Figure 2: Temporal Transformer Network [LWT19]

In TimeSync a Temporal Transformer Network (TTN) is introduced, a TTN is a module proposed by Lohit et al. [LWT19] that learns warping functions to reduce in-class variability and increase between-class variability. A TTN is shown in Figure 2. The TTN used in this approach is made up of three parts, the first part, called the trainable layers consists of 2 1D convolutional neural networks with kernel sizes 8 and 16 respectively. This layer takes in the 512-D embeddings ouput of the preceding ResNet18 module. Following this layer there are three fully connected linear layers that are applied to the output of the convolutional neural networks. Each layer is followed by a Rectified Linear Unit (ReLU). The trainable layers output a vector with a length equal to the input, this vector needs to be converted into a valid warping function. This is done in the constraint satisfaction layers, in these layers the output from the trainable layers is normalized, squared and used to define the derivative of the warping function. The cumulative sum of this function is calculated and multiplied by the length of the input sequence to form the warping function. The next part is the differentiable temporal resampling, in this part the warping function is applied to the input sequence using linear interpolation to ensure the frames of the warped sequence are defined at regular intervals [LWT19].

TimeSync consists of the architecture introduced by Shahzad et al. where TTNs are added after each of the ResNet18 modules which output 512-D embeddings. The resulting outputs from the

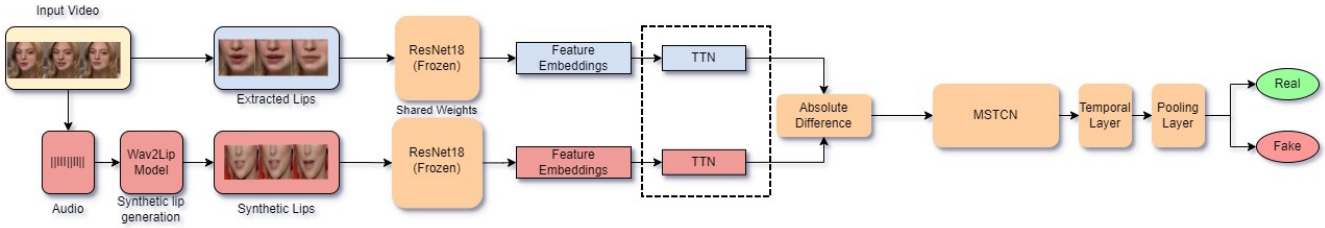

Figure 3: Proposed architecture of TimeSync built upon the implementation of the detection model as proposed by [SHK+22] with two added TTN modules before the MS-TCN.

TTN modules are then differentiated by calculating the absolute difference between them and fed into the MS-TCN followed by a pooling layer and a linear classifier. This architecture is shown in Figure 3. This is inspired by the work of N. Popov [K0l], whose work this implementation is partly built upon. His implementation can be seen in Figure 4 mainly his description of the trainable part of the TTN in combination with the TTN implementation by Lohit et al. [LWT19] was useful.

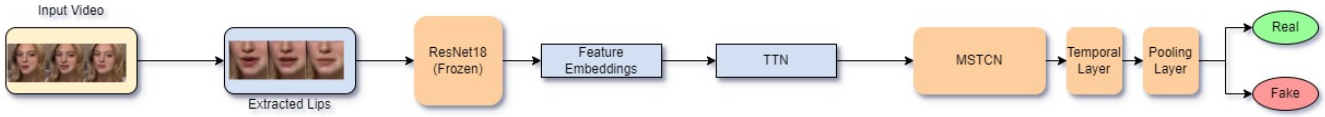

Figure 4: Architecture proposed by N. Popov [K0l].

The difference between TimeSync and the work of N. Popov is that Popov applied a single TTN to the unimodal face forgery detection approach by Haliassos et al. [HVPP21]. In TimeSync two TTNs are applied to a multimodal adaption of the face forgery detector made by Haliassos et al. [HVPP21], furthermore the outputs of these two TTNs are combined and fed into the MS-TCN module.

6 Experimental Setup

In this section we will discuss the datasets FakeAVCeleb and VoxCeleb, their contents and distributions, and what preprocessing has taken place. Furthermore we discuss the hyperparameters used for our methods, how the performance of the recreated baseline and TimeSync will be evaluated, how the models are trained, what metrics will be applied to compare them and what train-test split has been used.

6.1 Dataset and Preprocessing

Firstly we will discuss the FakeAVCeleb and VoxCeleb datasets used in our experiments. The kind of deepfake techniques that are contained in this dataset are described, what kind of preprocessing is done, the number of real and fake videos, and the oversampling strategy used.

6.1.1 Dataset

The datasets used in this research are the FakeAVCeleb and VoxCeleb datasets, both datasets are widely used in the field of deepfake detection research. The FakeAVCeleb dataset is used in research by Yang et al. [YZC+23] to measure the performance of the state-of-the-art in deepfake detection. The FakeAVCeleb database was chosen because it contains numerous deepfake generation methods, namely Fsgan, Fsgan-Wav2Lip, RTVC, Faceswap, Faceswap-Wav2Lip, and Wav2Lip. Furthermore, the videos are spread across different genders and ethnicities. This stimulates a more generalizable, less biased deepfake detection model. FakeAVCeleb contains 21.566 videos, of which 500 are genuine and 21.066 include some alteration. These fakes are divided into three categories, namely 'FakeVideo-RealAudio', 'FakeVideo-FakeAudio', and 'RealVideo-FakeAudio' as can be seen in Figure 5.
Because of the small number of unaltered videos in this dataset a number of videos were taken from the VoxCeleb dataset [CNZ18] which only contains real (i.e. genuine) videos. With the addition of these 3.600+ videos the class of real videos now contains 4.100+ videos. During training the real videos class was oversampled by allowing for more than one clip to be extracted from real videos per epoch, while only one clip per epoch could be extracted from videos in the fake class. This brings the ratio of real videos to fake videos from 1:4 to 1:1. This mirrors the approach taken by Shahzad et al. where genuine videos were taken from the VoxCeleb dataset to add to the real class [SHK+22].

6.1.2 Deepfake Generation Methods

In this section the different types of deepfake generation methods used in the FakeAVCeleb dataset will be discussed.

- Faceswap
  - Faceswap [KSDT17] is a deepfake generation method that involves swapping faces using Convolutional Neural Networks resulting in videos where a person's face is placed on another person's body while maintaining movement and expressions.
- FSGAN
  - Face Swapping GAN (FSGAN) [NKH19] is a face-swapping method using Recurrent Neural Networks that focuses on face reenactment adjusting for pose and expression variations aiming to create high-quality realistic deepfakes.
- Wav2Lip
  - Wav2Lip [PMNJ20] is a lip-synching deepfake generation method using Generative Adversarial Networks that generates realistic lip movements to match audio.
- RTVC
  - Real Time Voice Cloning (RTVC) [JZW+18] is a deepfake audio generation method that creates synthetic voices that closely match a target speaker's voice. It extracts

features from a short clip to capture the characteristics of a speaker's voice in order to synthetically recreate it.

- Faceswap-Wav2Lip
  - Wav2Lip applied on a video with deepfaked audio using RTVC and deepfaked video using Faceswap, creating a faked video with faked audio that is lip-synced [KTW21].
- FSGAN-Wav2Lip
  - Wav2Lip applied on a video with deepfaked audio using RTVC and deepfaked video using FSGAN [KTW21]..

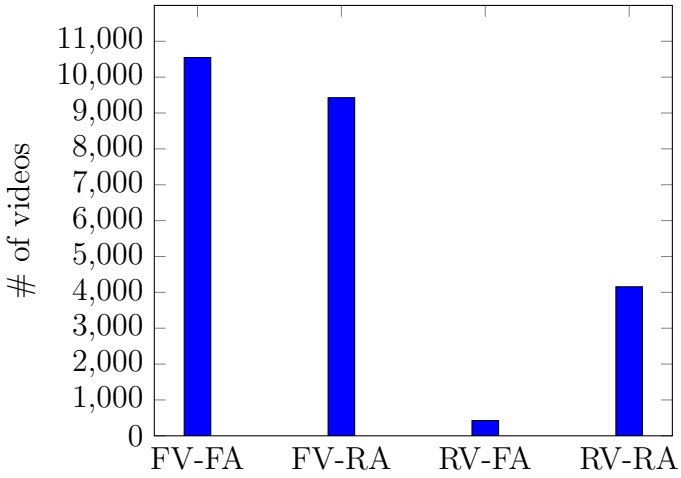

Figure 5: Distribution of videos in the training set, 'R' stands for real, 'F' for fake, 'V' for video and 'A' for audio, note that the RV-RA category is made up of FakeAVCeleb real videos and VoxCeleb real videos

The FakeAVCeleb dataset is very skewed as can be seen in Figure 5 and Figure 6. Even with 3.600+ additional videos from the VoxCeleb dataset the deepfaked videos still outnumber the genuine videos. To combat this oversampling of the minority class must be applied to prevent the model from overfitting to the fake category.

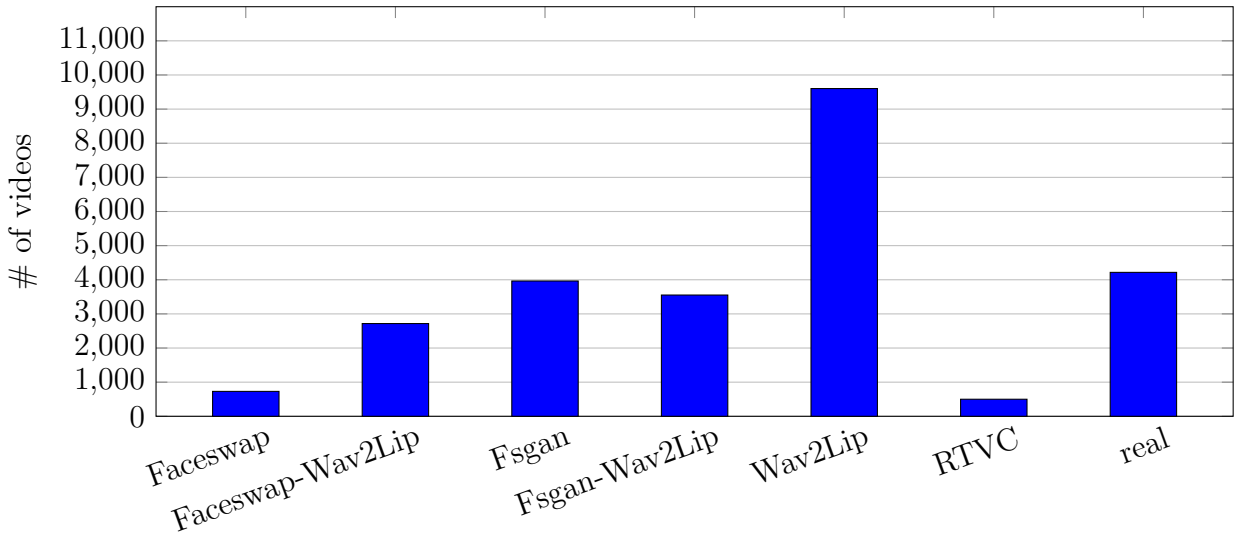

Figure 6: Distribution of videos in the training set, divided into individual methods, note that the 'real' category includes around 3.600 VoxCeleb videos

6.1.3 Preprocessing

In this section, the preprocessing used in both the Lip Sync Matters baseline as well as in TimeSync will be discussed.

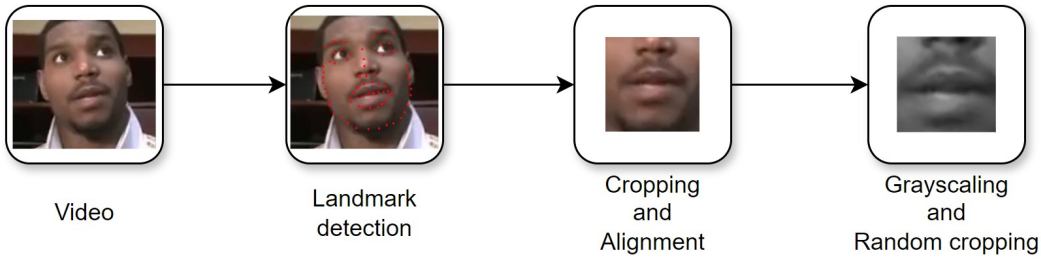

Figure 7: Overview of the preprocessing done on videos used in LipForensics, Lip Sync Matters and TimeSync.

The preprocessing on the FakeAVCeleb dataset with added videos from the VoxCeleb dataset is done by first taking the audio from the video and using a Wav2Lip model to generate synthetic lips for each video which will be further explained in detail in Section 6.1.4. The videos are cropped as individual 96×96 frames. This is done for both the unaltered (original) video and the video with synthetic lips. These crops are made by first calculating landmarks of the face using FAN [BT17], these landmarks are then smoothed over 12 frames to counteract any jitter and each frame is warped to the mean face and then are used to align the face and crop a 96×96 region around the mouth. These videos are grayscaled, randomly cropped to 88×88 and flipped horizontally with

a probability of 0.5. Subsequently they are fed into the model as tensors consisting of 25 frames with shape $1 \times 25 \times 88 \times 88$ This is visualized in Figure 7.

6.1.4 Lip Generation

The lip generation model used for generating synthetic lips was taken from [HPM+21]. This model was trained using an existing lip synthesis model as a teacher to train. The teacher that was used is the Wav2Lip [PMNJ20] speech-to-lip synthesis model. The result is a student model that can filter out noise and produce accurate synthetic lips. The model was trained using the LRS3 dataset that consists of 400+ hours of video taken from TED and TEDx talks. Unless stated otherwise the pre-trained model, which is publicly available here[1], was used.

6.2 Hyperparameters

Lip Sync Matters baseline
The hyperparameters for the baseline Follow the hyperparameters used in the Lip Sync Matters deepfake detection model proposed by Shahzad et al. [SHK+22]. The hyperparameters were taken from the Lip Sync Matters implementation to replicate the implementation them as closely as possible. The recreated baseline model was trained using the Adam optimizer, a learning rate of 2×10^{-4} and a batch size of 32. Because of the imbalance in the number of real and fake videos 3.600+ videos from the VoxCeleb dataset [CNZ18] were added to the real videos class. As the specific oversampling strategy was not mentioned by Shahzad et al. [SHK+22] an oversampling strategy was implemented which is described in Section 6.1.

TimeSync
TimeSync was trained using the Adam optimizer and a learning rate of 2×10^{-4} and a batch size of 32. Similarly to Lip Sync Matters the imbalance in the number of real and fake videos was rectified by adding 3.600+ from the VoxCeleb dataset [CNZ18] to the real videos class. The oversampling strategy that was implemented is described in Section 6.1.1.
Following the implementation by Shahzad et al. [SHK+22] the baseline Lip Sync matters model was trained using the Adam optimizer, a learning rate of 2×10^{-4}. The TTN uses a learning rate of 2×10^{-5} as is recommended by Lohit et al. [LWT19]. The batch size is 32. Because of the imbalance between the number of real and fake videos, the real videos were supplemented with 3.600+ real videos from the VoxCeleb dataset. Furthermore the real videos were oversampled by allowing the model to extract multiple clips from real videos per epoch and only one from fake videos per epoch. This ensures that the number of real and fake clips is balanced at 16 real and 16 fake videos per batch [SHK+22]. Except for the TTN, the hyperparameters used in TimeSync are the same as the hyperparameters used in Lip Sync Matters as to provide a reliable baseline comparison.

6.3 Training

Lip Sync Matters Baseline
The training setup for the recreated baseline follows the training setup of the Lip Sync Matters implementation by Shahzad et al. [SHK+22]. This is done to recreate Lip Sync Matters as closely

[1] https://github.com/Sindhu-Hegde/pseudo-visual-speech-denoising

as possible. The recreated baseline was pre-trained on the lipreading in the wild dataset [CZ16], which is a dataset containing 500.000+ samples uttered by hundreds of different subjects in various poses. The pretraining model used in the recreated baseline is publicly available here [2]. provided by Martinez et al. [MMPP20].
During training the ResNet18 modules remain frozen while the MS-TCN is finetuned, this is done as to not disturb the pretraining of the ResNet18 modules. The model is trained until it does not improve anymore.

TimeSync
In TimeSync the TTNs are trained first while the MS-TCN and ResNets remain frozen. The starting point of training is the recreated baseline model for Lip Sync Matters, the weights for the MS-TCN and ResNet modules are loaded and the weights for the TTNs are randomized. The TTNs have a learning rate of one-tenth of the MS-TCN. When the TTN When the TTN has stopped improving the MS-TCN is unfrozen and finetuned together with the TTN. This two part approach is taken to minimize the disruption to the pre-training of the MS-TCN, as the random initialization of the TTN can cause unstable outputs. Which may disrupt the pre-trained weights of the MS-TCN. By keeping the weights of the MS-TCN frozen while the TTN is first trained, the MS-TCN can maintain the effectiveness of its pre-training.

6.4 Train-test split

Following the train-test split used by Shahzad et al.[SHK+22] a number of test sets were constructed, each containing a different deepfake generation method. In total eight test sets were constructed each with 140 videos, separated into two even classes consisting of 70 fake and 70 real videos. 6 of these test sets contain fake videos sourced from one method, these are Fsgan, Fsgan-Wav2Lip, RTVC, Faceswap, Faceswap-Wav2Lip, and Wav2Lip. One test set, Even-Mix-1 which contains the same amount of faked videos per deepfake method present in the FakeAVCeleb dataset. The Even-Mix-2 test set contains the same amount of faked videos per deepfake category (FakeVideo-FakeAudio, RealVideo-FakeAudio, FakeVideo-RealAudio).

7 Results

In this section the experimental results of the recreated baseline, proposed TimeSync model and different state-of-the-art methods will be discussed.

[2]https://github.com/mpc001/Lipreading_using_Temporal_Convolutional_Networks

Test set	**Class**	**Baseline**				**Class**	**TimeSync**			
		Precision	Recall	F1-score	Accuracy		Precision	Recall	F1-score	Accuracy
Faceswap	Fake	0.76	0.69	0.72	0.74	Fake	0.97	0.87	0.92	**0.92**
	Real	0.71	0.79	0.75		Real	0.88	0.97	0.92	
Faceswap-Wav2Lip	Fake	0.96	1.0	0.98	0.98	Fake	0.99	1.0	0.99	**0.99**
	Real	1.0	0.95	0.97		Real	1.0	0.98	0.99	
Fsgan	Fake	0.95	1.0	0.97	0.97	Fake	0.99	0.99	0.99	**0.99**
	Real	1.0	0.94	0.97		Real	0.99	0.99	0.99	
Fsgan-Wav2Lip	Fake	0.9	1.0	0.95	0.94	Fake	0.96	1.0	0.98	**0.98**
	Real	1.0	0.86	0.92		Real	1.0	0.95	0.97	
RTVC	Fake	0.82	0.82	0.64	**0.71**	Fake	0.79	0.22	0.34	0.58
	Real	0.65	0.89	0.75		Real	0.55	0.94	0.69	
Wav2Lip	Fake	0.95	1.0	0.97	0.95	Fake	0.92	1.0	0.96	**0.96**
	Real	1.0	0.9	0.95		Real	1.0	0.9	0.95	
Even-Mix-1	Fake	0.96	0.86	0.91	0.89	Fake	0.99	0.89	0.94	**0.92**
	Real	0.8	0.94	0.86		Real	0.84	0.99	0.91	
Even-Mix-2	Fake	0.91	0.86	0.88	0.88	Fake	0.95	0.83	0.89	**0.89**
	Real	0.86	0.91	0.88		Real	0.84	0.95	0.89	

Table 1: Results for the recreated baseline implementation of Lip Sync Matters by Shahzad et al. [SHK+22] and the proposed TimeSync method

Test set	**Class**	**Lip Sync Matters**[SHK+22]				**TimeSync**			
		Precision	Recall	F1-score	Accuracy	Precision	Recall	F1-score	Accuracy
Faceswap	Fake	0.95	0.76	0.84	0.86	0.97	0.87	0.92	**0.92**
	Real	0.8	0.96	0.87		0.88	0.97	0.92	
Faceswap-Wav2Lip	Fake	0.96	1.00	0.98	0.98	0.99	1.0	0.99	**0.99**
	Real	1.00	0.96	0.98		1.0	0.98	0.99	
fsgan	Fake	0.94	0.69	0.79	0.82	0.99	0.99	0.99	**0.99**
	Real	0.75	0.96	0.84		0.99	0.99	0.99	
fsgan-Wav2Lip	Fake	0.96	0.99	0.97	0.97	0.96	1.0	0.98	**0.98**
	Real	0.99	0.96	0.97		1.0	0.95	0.97	
rtvc	Fake	0.95	0.83	0.89	**0.89**	0.79	0.22	0.34	0.58
	Real	0.85	0.96	0.90		0.55	0.94	0.69	
Wav2Lip	Fake	0.96	0.96	0.96	**0.96**	0.92	1.0	0.96	**0.96**
	Real	0.94	0.90	0.92		1.0	0.9	0.95	
Even-Mix-1	Fake	0.96	0.91	0.93	**0.94**	0.99	0.89	0.94	0.92
	Real	0.92	0.96	0.94		0.84	0.99	0.91	
Even-Mix-2	Fake	0.96	0.93	0.94	**0.94**	0.95	0.83	0.89	0.89
	Real	0.93	0.96	0.94		0.84	0.95	0.89	

Table 2: Results for the official Lip Sync Matters as reported by Shahzad et al. [SHK+22] and the proposed TimeSync method

Baseline

The results of the recreated baseline together with the results of TimeSync evaluated on the eight test sets are shown in Table 1. The baseline model achieved good predictions for most of the test sets, matching the performance of the Lip Sync Matters implementation by Shahzad et al. [SHK+22], which is shown in Table 2. The recreated baseline closely matched Lip Sync Matters' performance for the Faceswap-Wav2Lip, Fsgan-Wav2Lip and Wav2Lip test sets. The recreated baseline came close to the performance of Lip Sync Matters in the two Even-Mix test sets falling

.05 and .06 short in predicting Even-mix-1 and Even-mix-2 respectively. The recreated baseline outperforms the detection of Fsgan deepfakes in comparison to Lip Sync Matters. The recreated baseline underperforms in detecting Faceswap and RTVC when compared to Lip Sync Matters.
The reason for the poor performance of RTVC and Faceswap is that the Faceswap category contains only 730 videos and RTVC contains only 500 videos in the FakeAVCeleb dataset. The distribution of deepfake methods in the training set is shown in Figure 6. Which was also reported by Shahzad et al. [SHK+22] the difference in results between Lip Sync Matters and the recreated baseline can also be attributed to the implementation by Shahzad et al. not being publicly available. Making it necessary to write essential parts like the training script, sampler and oversampling strategy from scratch. Which could have had a significant impact on the performance and generalizability of the baseline method.

TimeSync
The results of the TimeSync model show an improvement in visual deepfake detection over the recreated Lip Sync Matters baseline as well as the original Lip Sync Matters by Shahzad et al. [SHK+22] shown in Table 2. The largest improvement over the recreated baseline is seen when looking at the Faceswap test set where a big leap of .18 in detection accuracy can be seen. Furthermore when looking at the other test sets it is clear that the addition of the TTNs provide a modest improvement over the baseline. However RTVC detection accuracy has dropped dramatically. This could be due to the TTN modules learning a warping function that enhances the differences between real and fake visual aspects, but decrease them between real and fake audio. Also because of the low number of RTVC deepfakes included in the FakeAVCeleb dataset as can be seen in Figure 6. The poor accuracy of detecting RTVC deepfakes has also led to the accuracy of the Even-Mix-2 dataset being lower.
However, apart from Wav2Lip where its performance is equal to Lip Sync Matters by Shahzad et al. TimeSync surpasses both the recreated baseline and Lip Sync Matters in visual deepfake detection.

7.1 Comparison to other models

In this section TimeSync will be compared to several state-of-the-art deepfake detection models.

From Table 3 and Table 4 it becomes evident that TimeSync surpasses every method tested on the Even-Mix-2 test set which encompasses the entire dataset, except Lip Sync Matters by Shahzad et al. However from Table 2 it is evident that TimeSync surpasses the state-of-the art Lip Sync Matters model in visual deepfake detection.

Test set	Modality	Class	Precision	Recall	F1-score	Accuracy
LipForensics [HVPP21]	V	Fake	0.90	0.56	0.69	0.76
		Real	0.69	0.94	0.80	
LipForensics [HVPP21]	A	Fake	0.49	0.99	0.66	0.49
		Real	0.0	0.0	0.0	
Lip Sync Matters [SHK+22]	A	**Fake**	**0.96**	**0.93**	**0.94**	**0.94**
		Real	**0.93**	**0.96**	**0.94**	
Lip Sync Matters our implementation	AV	Fake	1.0	0.75	0.86	0.87
		Real	0.8	1.0	0.89	
RealForensics [HMPP22]	AV	Fake	0.27	0.91	0.42	0.603
		Real	0.97	0.56	0.71	
FTCN [ZBC+21]	AV	Fake	0.73	0.76	0.74	0.73
		Real	0.74	0.71	0.72	
Exploiting visual artifacts [MRS19]	AV	Fake	0.5	0.09	0.15	0.49
		Real	0.49	0.89	0.63	
TimeSync	AV	Fake	0.95	0.83	0.89	0.89
		Real	0.84	0.95	0.89	

Table 3: Comparison of the proposed model with other models on the Even-Mix 2 test set. The modality column shows what modality the model focuses on, 'v' for visual, 'a' for audio and 'av' for audio-visual. Note that TimeSync is the second best performer.

Method	Modality	ACC(%)	AUC(%)
LipForensics [HVPP21]	V	80.1	82.4
Lip Sync Matters [SHK+22]	AV	**94.0**	-
Emotions Don't Lie [MBC+20]	AV	78.1	79.8
Xception [RCV+19]	V	67.9	70.5
AVoiD-DF[YZC+23]	AV	83.7	89.2
Lip Sync Matters our implementation	AV	88.2	**96.5**
TimeSync	**AV**	88.9	**96.5**

Table 4: Performance of several state-of-the-art deepfake detection models on the FakeAVCeleb dataset reported by Yang et al. [YZC+23] and TimeSync note that AUC was not reported for Lip Sync Matters by Shahzad et al.

The reported results from Yang et al. in Table 4 have been included to show the results of several state-of-the-art methods run in an extensive benchmark experiment, Yang et al. use a randomly selected test set that has balanced the positive and negative samples using oversampling. Furthermore the experiments were repeated for multiple rounds and the average value was recorded. TimeSync was run on the Even-Mix-2 test set in this table.
This comparison is fair because Yang et al. used a randomly selected test set consisting of 30% of the dataset which was oversampled for an equal representation of the positive and negative samples. Moreover the experiments repeated for multiple rounds [YZC+23]. The Even-Mix-2 test set contains an equal distribution of all categories in the FakeAVCeleb dataset, ensuring an even and representative comparison.

7.2 Discussion

From the above tables we can see that TimeSync offers results that outperform the recreated baseline, Lip Sync Matters by Shahzad et al. and several previous state-of-the-art methods in visual deepfake detection. Making TimeSync a more effective and robust visual deepfake detector. While the performance increase is not major it shows that the TTNs do offer an improvement and provide more robust deepfake detection when added to a detection pipeline involving MS-TCNs. There are however shortcomings inherent to the approach taken in this research. Namely the reliance on a clear view of the face of the speaker, this means that the model is vulnerable to lip obstruction and low quality video. And may perform poorly under these circumstances.

8 Conclusions and Further Research

In this research TimeSync is proposed, a novel, updated approach for multimodal deepfake detection. Two Temporal Transformer Networks (TTNs) were added to a recreated baseline of Lips Sync Matters by Shahzad et al. [SHK+22] to increase the separation between genuine and deepfaked videos from the FakeAVCeleb dataset. TimeSync offers improvement on the recreated baseline and the original Lip Sync Matters method by Shahzad et al. Improving upon visual detection capability of the model.
Further research could explore more advanced preprocessing techniques like refined lip extraction and different lip generation techniques. Furthermore adding TTNs to the Lip Sync Matters method by Shahzad et al. [SHK+22] could be explored if it is ever publicly released.
The main contributions in this research are

- Timesync, an audio-visual model pre-trained on lipreading that can distinguish between real and deepfaked videos using the discrepancies between high level semantic features extracted from the lips of a target video and synthetically generated lips from a Wav2Lip model.
- Extensive experiments on each part of the FakeAVCeleb dataset demonstrate that TimeSync can outperform state-of-the-art deepfake detection models on visual deepfakes such as Faceswap, FSGAN and Wav2Lip in the multimodal FakeAVCeleb dataset.

References

[AFFA20] Shruti Agarwal, Hany Farid, Ohad Fried, and Maneesh Agrawala. Detecting Deep-Fake Videos from Phoneme-Viseme Mismatches. *IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 6 2020.

[BT17] Adrian Bulat and Georgios Tzimiropoulos. How far are we from solving the 2d & 3d face alignment problem? (and a dataset of 230,000 3d facial landmarks). In *International Conference on Computer Vision*, 2017.

[CNZ18] J. S. Chung, A. Nagrani, and A. Zisserman. Voxceleb2: Deep speaker recognition. In *INTERSPEECH*, 2018.

[CZ16] J. S. Chung and A. Zisserman. Lip reading in the wild. In *Asian Conference on Computer Vision*, 2016.

[DBP+20] Brian Dolhansky, Joanna Bitton, Ben Pflaum, Jikuo Lu, Russ Howes, Menglin Wang, and Cristian Canton Ferrer. The deepfake detection challenge (dfdc) dataset, 2020.

[GPAM+14] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014.

[HMPP22] Alexandros Haliassos, Rodrigo Mira, Stavros Petridis, and Maja Pantic. Leveraging real talking faces via self-supervision for robust forgery detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14950–14962, 2022.

[HPM+21] Sindhu B. Hegde, K.R. Prajwal, Rudrabha Mukhopadhyay, Vinay P. Namboodiri, and C.V. Jawahar. Visual speech enhancement without a real visual stream. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 1926–1935, January 2021.

[HVPP21] Alexandros Haliassos, Konstantinos Vougioukas, Stavros Petridis, and Maja Pantic. Lips Don't Lie: a generalisable and robust approach to face forgery detection. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5039–5049, 6 2021.

[JLW+20] Liming Jiang, Ren Li, Wayne Wu, Chen Qian, and Chen Change Loy. DeeperForensics-1.0: A large-scale dataset for real-world face forgery detection. In *CVPR*, 2020.

[JZW+18] Ye Jia, Yu Zhang, Ron J. Weiss, Quan Wang, Jonathan Shen, Fei Ren, Zhifeng Chen, Patrick Nguyen, Ruoming Pang, Ignacio López-Moreno, and Yonghui Wu. Transfer learning from speaker verification to multispeaker text-to-speech synthesis. *CoRR*, abs/1806.04558, 2018.

[K0l] K0l1ka. Github k0l1ka deepfake video detection: Deepfake video detection.

[KSDT17] Iryna Korshunova, Wenzhe Shi, Joni Dambre, and Lucas Theis. Fast face-swap using convolutional neural networks. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 3697–3705, 2017.

[KTW21] Hasam Khalid, Shahroz Tariq, and Simon S. Woo. Fakeavceleb: A novel audio-video multimodal deepfake dataset, 2021.

[KW13] Diederik P Kingma and Max Welling. Auto-Encoding Variational Bayes. *arXiv (Cornell University)*, 1 2013.

[LHASS20] Michael Lomnitz, Z. Hampel-Arias, Vishal Sandesara, and Hans-Uwe Simon. Multimodal Approach for DeepFake Detection. *2020 IEEE Applied Imagery Pattern Recognition Workshop (AIPR)*, 10 2020.

[LWT19] Suhas Lohit, Qiao Wang, and Pavan Turaga. Temporal transformer networks: Joint learning of invariant and discriminative time warping. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12418–12427, 2019.

[MAA+23] Rami Mubarak, Tariq Alsboui, Omar Alshaikh, Isa Inuwa-Dutse, Saad Khan, and Simon Parkinson. A survey on the detection and impacts of deepfakes in visual, audio, and textual formats. *IEEE Access*, 11:144497–144529, 1 2023.

[MBC+20] Trisha Mittal, Uttaran Bhattacharya, Rohan Chandra, Aniket Bera, and Dinesh Manocha. Emotions don't lie: An audio-visual deepfake detection method using affective cues. In *Proceedings of the 28th ACM International Conference on Multimedia*, MM '20, page 2823–2832, 2020.

[MMPP20] Brais Martinez, Pingchuan Ma, Stavros Petridis, and Maja Pantic. Lipreading using temporal convolutional networks. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6319–6323, 2020.

[MRS19] Falko Matern, Christian Riess, and Marc Stamminger. Exploiting visual artifacts to expose deepfakes and face manipulations. In *2019 IEEE Winter Applications of Computer Vision Workshops (WACVW)*, pages 83–92, 2019.

[NKH19] Yuval Nirkin, Yosi Keller, and Tal Hassner. Fsgan: Subject agnostic face swapping and reenactment. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 7183–7192, 2019.

[PMNJ20] K R Prajwal, Rudrabha Mukhopadhyay, Vinay P. Namboodiri, and C.V. Jawahar. A lip sync expert is all you need for speech to lip generation in the wild. In *Proceedings of the 28th ACM International Conference on Multimedia*, MM '20, page 484–492, 2020.

[RCV+19] Andreas Rössler, Davide Cozzolino, Luisa Verdoliva, Christian Riess, Justus Thies, and Matthias Nießner. FaceForensics++: Learning to detect manipulated facial images. In *International Conference on Computer Vision (ICCV)*, 2019.

[SHK+22] Sahibzada Adil Shahzad, Ammarah Hashmi, Sarwar Khan, Yan-Tsung Peng, Yu Tsao, and Hsin-Min Wang. Lip Sync Matters: a novel multimodal forgery detector. *2022 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, 11 2022.

[YZC+23] Wenyuan Yang, Xiaoyu Zhou, Zhikai Chen, Bofei Guo, Zhongjie Ba, Zhihua Xia, Xiaochun Cao, and Kui Ren. AVOID-DF: Audio-Visual Joint Learning for Detecting Deepfake. *IEEE transactions on information forensics and security*, 18:2015–2029, 1 2023.

[ZBC+21] Yinglin Zheng, Jianmin Bao, Dong Chen, Ming Zeng, and Fang Wen. Exploring temporal coherence for more general video face forgery detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15044–15054, 2021.

[ZWZ+21] Hanqing Zhao, Tianyi Wei, Wenbo Zhou, Weiming Zhang, Dongdong Chen, and Nenghai Yu. Multi-attentional deepfake detection. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2185–2194, 2021.