Comparison of Deepfakes Detection Techniques

Sonia Salman
Dept.of Computer Science
National University of Computer and Emerging Sciences
Karachi, Pakistan
0000-0001-8113-1552

Jawwad Ahmed Shamsi
Dept.of Computer Science
National University of Computer and Emerging Sciences
Karachi, Pakistan
jawwad.shamsi@nu.edu.pk

2023 3rd International Conference on Artificial Intelligence (ICAI) | 979-8-3503-2212-5/23/$31.00 ©2023 IEEE | DOI: 10.1109/ICAI58407.2023.10136659

Abstract— **Detection of fake audio and video is a challenging problem. Deepfake is popularly used for creating fake audio and video content using deep learning. Deepfakes, artificially created audiovisual interpretations can be used to degrade the reputation of a renowned person, hate-speech, or affect public belief. The development of novel methods for identifying various deepfake video types has received a significant amount of research throughout the years. In this research, we present a thorough comparative analysis of current state-of-the-art deepfake detection methods. The primary goal of our research is to identify the factors that contribute to the performance degradation of deepfake detection models currently being used when tested against a comprehensive dataset.**

Keywords—deepfake detection, deep learning, comparative revicew

I. INTRODUCTION

Digital manipulation of multimedia content such as photos and videos containing facial information generated by modification methods particularly using deep learning technologies has recently become a major social concern. The popular term "Deep fakes" refers to a deep learning technique for creating fake faces, videos and audio of a public renowned person mostly. In late 2017, a Reddit user [1] titled "deepfakes" when he transferred the celebrity faces into adult movies by an algorithm. Deepfakes can be used legitimately for entertainment purposes. Fake news, hoaxes, financial fraud, and victim defamation are just a few of the more damaging uses of deep fakes. The dissemination of these edited movies poses a great threat to businesses, reputations, etc.

The majority of DeepFake videos uploaded online are face-swapping videos, which are the focus of most DeepFake detection algorithms now in use. Other than face swapping, voice cloning and/or speech synthesis techniques have been used to change the speech of the target person in the video in several scenarios. The rare type of deepfakes consists of such videos in which the target person's face and voice both are swapped with the source person.

It's critical to build sophisticated deep learning systems to combat synthetically generated fake data. Researchers are striving to detect the deepfakes in an effective and efficient manner. Existing detection systems are inadequately capable of detecting both fake audio and facial manipulation. Researchers have also developed multi-modal frameworks for audio and video based deep fake detection.

979-8-3503-2212-5/23/$31.00 ©2023 IEEE

Our contributions in this paper are as follows:

- We gathered real and deepfake videos dataset for an in-depth analysis of the state-of-the-art deepfake detection models.
- We consider voice-swapped, voice-cloned videos as well as face-swapped deepfake videos in our dataset to make it more comprehensive and challenging to detect.
- We perform qualitative analysis by comparing the architecture, features and detection techniques of our selected deepfake detection models.
- We also perform quantitative comparison of deepfake detection models by assessing their performance on a comprehensive dataset. The performance of those detection techniques is assessed using widely used evaluation metrics including accuracy, Precision, Recall and F1 score.

II. BACKGROUND OF DEEPFAKE VIDEOS

A. Deepfake Types and their Generation

DeepFake has become a threat to society and humanity. Even without prior knowledge of the domain, it is now possible to easily edit or create fake videos of the target person using a single computer. Several applications, such as DeepFaceLab[2], FaceSwap[3], and others, do the same function. There is a deep neural network that underlies all of these applications. We will now examine the underlying architecture of each type of deepfake generation.

1) Face-Swapped DeepFakes:

This is the most common type of deepfake that typically use auto-encoders. Auto encoder has two components: encoder and decoder. The encoder is used to determine the latent features of source and target faces. These features, also known as the latent space, are the basic components of the face that are characteristic of the majority of people, such as the position of the eyes, nose, and mouth. The encoder determines the similarities between the two faces before compressing face A into the fundamental latent features that give it its uniqueness; these latent features are referred to as feature set. The auto encoder then sends that information to the decoder, the decoder then does the real work of merging and overlays source Face on top of target Face in the video.

Many applications, such as DeepFaceLab and FaceSwap, swap faces while also applying proper lip-syncing, making it impossible to detect with naked eyes.

2) Voice-Swapped DeepFakes:

In this type of deepfake, the target person's face remains the same in both the real and fake video, but the voice is swapped with the other person’s voice. This type of deepfakes is created by GANs (Generative Adverserial Network) [4]. First the video frames are extracted from the target video and then fed into the Face encoder. The face encoder discovers the latent features of the face. The source input audio that is to be swapped with the target speaker is also fed to the Audio encoder after converted into spectogram. The audio encoder encodes the audio features. After the concatenation of the face and audio features, lip-synced video of the target speaker is generated. The discriminator then calculates the synced losses such as constrastive loss or binary cross entropy loss between the face and the audio for the synchronization. If the loss is closer to zero, then the video is termed as real by the discriminator. In this way, voice swapped lip-synced video is generated.

Applications such as wav2lip [5], first-order motion model [6] and PCAVS [7], etc are commonly used to create such videos.

3) Synthetic Speech/Voice-cloned Deepfakes:

This type is similar to the voice swap deepfakes, but with one key difference. In such deepfake videos, the words of the target speaker are manipulated/changed even with his own voice.

First the audio samples of the speaker are collected which are then fed into real time voice cloning tool/software to get the same voice with different spoken words. Then that cloned voice is used to be inputted in the audio encoder, rest of the process is similar to the voice swapped deepfakes. Voice cloned deepfakes are more dangerous as compared to the voice swapped deepfakes. Synthetic deepfake audio is generated using applications such as Google Tacotron2 [8] and DeepVoice [9].

4) Face-Swapped with Synthetic Speech Deepfakes:

In this deepfake type, not only is the face of the target person swapped but also his uttered words are manipulated by speech synthesize methods. The fake video is generated first by implementing the same procedure of face swapping then that fake video is fed into the face encoder after converting into video frames. The synthetic audio is also inputted into the audio encoder after converting it into audio spectrogram. The audio encoder encodes the audio features. After the concatenation of the fake face and audio features, lip-synced video of the swapped face is generated. The discriminator then calculates the synced losses such as constrastive loss or binary cross entropy loss between the face and the audio for the synchronization. If the loss is closer to zero, then the video is termed as real by the discriminator. In this way, voice swapped lip-synced video is generated. Such deepfake videos are very rare as more time and training is required in its generation.

B. Deepfake Detection

Deepfakes pose a growing threat to democracy, society's security, and individual privacy. Methods for identifying deepfakes were suggested as soon as the threat posed by them was discovered. Early methods depended on hand-crafted features that were made from artefacts and errors in the modified video synthesis process. On the other hand, recent techniques used deep learning to automatically extract salient and discriminative features in order to detect deepfakes. Deepfake detection is often viewed as a binary classification problem where classifiers are used to differentiate between real and fake videos. To train classification models, this type of technique involves a huge database of original and manipulated videos. Following are the prominent deepfake detection methods:

1) Visible artifact-based detection methods

There can be visible colour differences and resolution inconsistencies between the internal face and background sections in the deepfake video. The boundaries are basically detectable because of these discrepancies. These artefacts are identified using CNN-based methods.

2) Biological signal-based detection methods

Latent biological signals in faces such as eye blinking are difficult to interpret using GAN, making it nearly impossible to generate human looks with appropriate behaviour. Biological signals are extracted based on this analysis to detect deepfake videos.

3) Facial expression-based detection methods

Deepfake videos generated in the wild have face emotions that may not align well with fake faces. Detection methods based on emotions extract the facial and audio features from the video and then classify them as real or fake.

4) Irregularities in video frames-based detection techniques

There can be irregularities among the consecutive video frames caused by the deepfake generation process. Temporal consistency-based detection methods recognise the consistency between consecutive frames that improves the detection performance.

5) GAN fingerprint-based detection techniques

Generated faces by GANs (Generative Adversarial Networks) frequently have GAN-generated fingerprints that are visible in the faked image.

6) Out-of-lips sync-based detection techniques

Speech and mouth are not well synced in voice-swapped deepfake videos. This discrepancy can be easily detected by correlating the speech to landmarks around the mouth.

III. LITERATURE REVIEW

There are various deepfake detection models, and the efforts are still continued on developing robust architecture with better accuracy. The following are a few of the models that are proven to be a good deepfake detection model models/technique. Such models are divided into fake face detection, fake audio detection and audio-visual deepfake detection categories.

A. Fake Face Detection Models

Dang et al. [10] developed their own huge dataset DFFD based on the combination of previous real faces datasets such as CelebA, FFHQ and FaceForensics++. They suggested a detection model that uses CNN models like Xception-Net and VGG16 as well as an attention technique to improve the classifier model's feature maps. The proposed attention map can be implemented using a single convolution layer, associated loss functions, and masking of the following high-dimensional characteristics.

Neves et al. [11] proposed GANPrintR detection model that caters to the above-mentioned limitation about the visibility of GAN fingerprints. Their model is based on an autoencoder to remove the GAN generated fingerprint to spoof the detection system. The system will consider such synthetic facial images as real. They also created the iFakeFaceDB database, which does not contain GAN fingerprint data.

Afchar et al. [12] proposed a successful model known as MesoNet that considers the mesoscopic features of faces and it consists of the two forms of CNN model employed with less layers known as Meso_4 and MesoInception_4.

B. Audio Spoof Detection Model

The technique proposed by Huang et al [13] for detecting audio spoofing. Firstly, short term zero crossing rate along with energy were used to detect the silent segments from every signal of speech. Next, the linear filter bank key points were calculated by the allocated parts in the comparatively higher frequency domain. Finally, an attention improved Dense Net-BiLSTM framework was developed to find the audio exploitation. This technique works well and at the cost of high computation.

C. Multi Modal Detection Systems

Following are the few multi modal detection techniques that can detect not only the fake faces in the video frames but also detect the manipulated audios in it if exist.

Chugh et al [14] proposed a dissonance-based architecture, if there is any dissimilarity in either the audio/visual part in video then its score is calculated as MDS modality dissonance score which is then aggregated on DFDC and DFTIMIT datasets. Contrastive loss is used for lip syncing issue which is closer for audio and video features and farther for fakes. The architecture also includes cross entropy loss. 3D ResNet, 3D

CNNs are used for recognizing action and ResNet for image classification. Extraction of audio features is done by Mel-Frequency Cepstral Coefficients MFCC. Both the audio and visual features are passed to visual and audio streams respectively that gives the cross-entropy loss for each, also the contrastive loss is calculated for each. Final loss is then calculated from both the losses.

Lomnitz et al [15] proposed a multi modal approach in which for single frame, transfer learning is done along with the two models trained in ImageNet [16]. ResNet-152 [17] is used as an Xception network. For multi frames, XceptionNet and BiLSTM with attention layer. SincNet is used for audio detection. OpenCv, along with the face detection algorithm BlazeFace is used for visual features extraction. MoviePy for audio features extraction.

IV. Comparison of the Deepfake Detection Techniques.

We analyze the following state-of-the-art deepfake detection techniques for the comparison with a comprehensive deepfake videos dataset.

A. MesoNet

Afchar et al. proposed a successful model known as MesoNet which is a CNN-based network that was influenced by InceptionNet [18] to detect face tampering in videos. It comprises of two distinct networks with just a few layers to concentrate on the mesoscopic characteristics of the images.:

- Meso-4 as shown in Fig. 1 is a CNN network with four convolutional layers and a fully connected layer, and
- MesoInception-4 was a modified version of the Inception module that was added to Meso-4.

Both the models are trained on deepfake datasets consisting of frames of faces from deepfake videos. MesoNet computes the mean output of a layer for sets of real and fake pictures and then analyzes the variations of activation and interprets the parts of the input images that are crucial for classification.

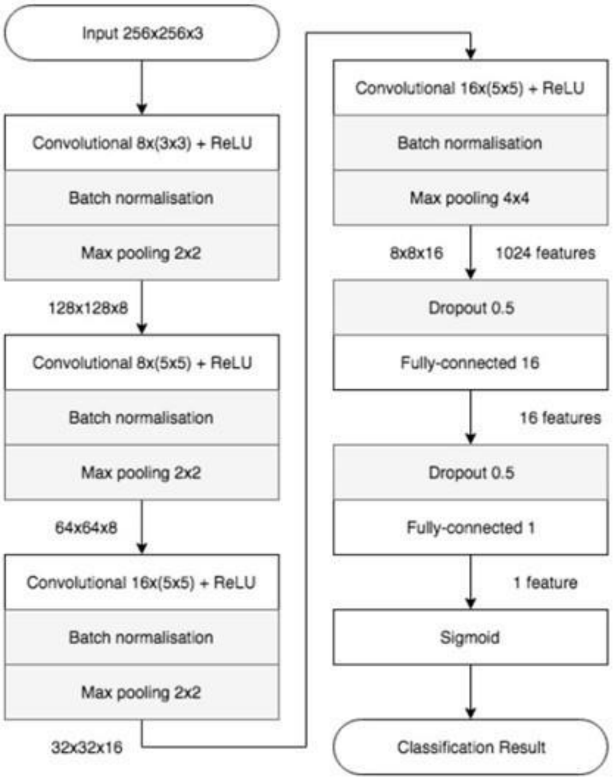

Fig. 1. The network architecture of Meso-4

B. SyncNet

Syncnet [19] is a language and speaker independent solution to the lip-sync problem without labeled processed both audio and video clips as inputs. A combined embedding between the audio and the mouth images can be trained using a two-stream ConvNet architecture as shown in Fig. 2. The audio input is made up of MFCC values, which are widely used in automatic speech and speaker recognition. A series of mouth regions represented as grayscale images with dimensions of 111×111×5 (W×H×T) dimensions for 5 frames serves as the visual network's input. The core principle is that for real videos, the output of the audio and video networks is equal, while it differs for fake videos. The Euclidean distance between the network outputs can then be determined. If there is more distance between the audio and video inputs, then the video is considered as deepfake.

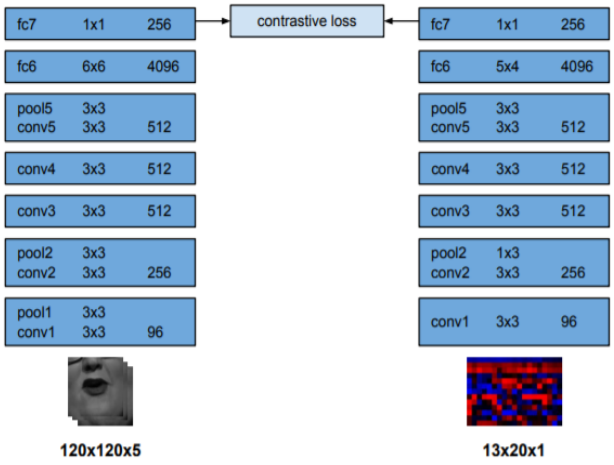

Fig. 2. The network architecture of SyncNet

C. Video Face Manipulation Detection through Ensemble of CNNs

This detection technique [20] addresses the issue of detecting facial manipulation in video sequences that target contemporary facial manipulation techniques. The method (Fig. 3) enhances a recently proposed solution and takes influence from the class of EfficientNet [21] models. It discusses a group of models that were developed utilizing two key ideas:

- an attention mechanism that produces an interpretation of the model that is understandable to humans while also enhancing the network's capacity for learning
- a triplet siamese training method that extracts detailed features from the data to improve classification results.

The results are evaluated on two known large datasets that show the efficiency by ensembling of different trained CNNs models.

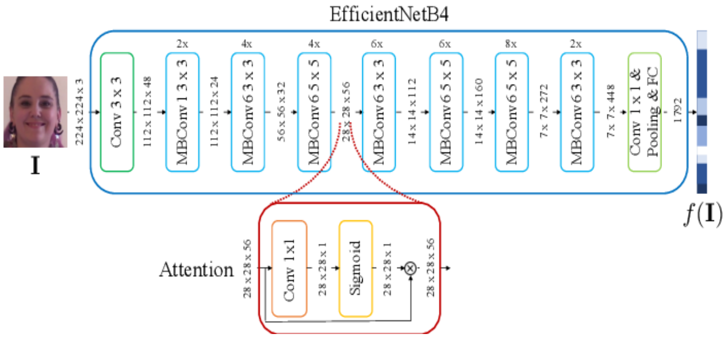

Fig. 3. The network architecture of EfficientNetB4Att model that comprised of blue block (EfficientNetB4) and an attention mechanism.

D. Evaluation Metrics

We compared the performance of all deepfake detection techniques by using the following widely used evaluation metrices:

1) Accuracy

Accuracy refers to the number of times the model was generally accurate.

$$Accuracy\ =\frac{TP+TN}{TP+TN+FN+FP}*100$$

2) Precision

Precision indicates how effectively the model foresees a particular category.

$$\text{Precision}=\left(\frac{TP}{TP+FP}\right)*100$$

3) Recall

Recall demonstrates how frequently the model can identify a specific class.

$$\text{Recall}==\left(\frac{TP}{TP+FN}\right)*100$$

4) F1-score

A model's performance is assessed using the F1-score, which takes precision and recall into consideration.

$$F1-Score==\left(\frac{2}{\frac{1}{\text{Precision}}+\frac{1}{\text{Recall}}}\right)*100$$

TP stands for True Positive when both the video and the model prediction are accurate. TN stands for True Negative when both the video and the predictions are fake. FN stands for False Negative, when the prediction is false, yet the video is real. False Positive (FP) occurs when a video is fake yet predicted to be real.

V. Results and Discussion

A. Qualitative Analysis

We compare three state-of-art deepfake detection methods with respect to their architecture, detection performance, features, dataset used, and limitations as summarized in table 1.

TABLE I. Quantitative Comparison of deepfake detection Techniques

Detection Models	**Layer Architecture**	**Detection Performed**	**Features**	**Dataset used**	**Limitations**
MesoNet	MesoInception-4	Face-Swapped images only	Deep (Mesoscopic) features	FaceForensics++	Performance decreases on low-quality videos
SyncNet	2-stream CNN	Out-of-lips sync detction	Audio/Video synchronization	VIDTIMIT	Performance degrades in real videos detection.
Video Face Manipulation Detection through Ensemble of CNNs	EfficientNetB4	Face-Swapped images only	Facial features	DFDC	Performance decreases in voice-swapped videos detection.

MesoNet and Ensemble of CNNs based methods are developed for the face-swapped detection. Whereas SyncNet is developed for the detection of out-of-lips sync videos. MesoNet comprises of MesoInception architecture derived from InceptionNet and detects facial manipulation by extracting the mesoscopic features of face. In contrast, the other method leveraged high-level prominent facial information from a group of CNNs to identify face-swapped deepfakes. SyncNet differs from the other two techniques in that it uses a 2-stream CNN model for synchronizing audio and video by separating the audio and video features and comparing them side by side using Euclidean distance.

MesoNet performs well in both the face-swapped and voice-swapped deepfakes. It minimizes the computational cost by the technique of down sampling the video frames. However, at the same time it also decreases the accuracy of detecting deepfakes, especially the low-quality frames. Ensemble of CNNs method works fairly well in face-swapped deepfakes as it extracts the high-level semantic features of the manipulated area of the face but didn't perform well in voice-swapped deepfakes. Whereas SyncNet also performs good in both the swapped types of deepfakes. On the other hand, it works by contrasting the mouth movement and audio of that specific frame. The original videos can exhibit out-of-sync mouth movements, making it impossible to tell them apart from fake ones.

B. Quantitative Analysis

For the comparative analysis, we collected around 1000 real and deepfake videos from different internet sources. The dataset comprises of face-swapped, voice-swapped, voice-cloned and real videos.

Table 2 below summarizes the evaluation metrics of the deepfake detection models.

TABLE II. QUANTITATIVE COMPARISON OF DEEPFAKE DETECTION TECHNIQUES

Detection Models	Accuracy	Precision	Recall	F1-Score
MesoNet	62.5	69	45	54.5
SyncNet	67	62.7	84	71.9
Video Face Manipulation Detection through Ensemble of CNNs	54.5	61.5	24	34.6

We analyzed that the SyncNet performs better in detecting all types of deepfakes as compared to the other detection models. Though it has been developed to detect the out-of-lips sync deepfakes, it also works well in face-swapped deepfakes. MesoNet performs well in detecting face-swapped deepfakes as it minimizes the computational cost by the technique of down sampling the video frames however at the same time it also decreases the accuracy of detecting deepfakes especially the low-quality frames. Whereas video face manipulation through ensemble of CNNs have least accuracy in predicting the audio-swapped deepfakes. The analysis shows that the 2-stream CNN model (used in SyncNet) for both audio and video modules perform better in face and voice-swap detection. The architecture works around the mouth regions to calculate the distance between the audio and the corresponding video. Since obvious artefacts around the lips can also be present in a face-swapped deepfake, this approach is more effective.

VI. CONCLUSION AND FUTURE WORK

The potential for exploitation of deepfake technology could lead to harm for many individuals. Therefore, a deepfake detection model that has a high accuracy in detecting both face and voice swapped deepfakes is needed. Deepfake detection methods are highly dependent on the applicability of deepfake detection methods. In order to detect more diverse patterns of deepfake content in the future, research must be done to develop deepfake detection techniques that are good at both detection and classification. We will continue our research to overcome the limitations of the discussed deepfake detection models by incorporating techniques in order to detect all types of deepfakes with high accuracy.

References

[1] T. T. Nguyen, C. M. Nguyen, D. T. Nguyen, D. T. Nguyen, and S. Nahavandi, "Deep Learning for Deepfakes Creation and Detection," arXiv preprint arXiv:1909.11573, 2019

[2] https://github.com/iperov/DeepFaceLab

[3] Faceswap: https://github.com/deepfakes/faceswap

[4] Goodfellow et al., "Generative adversarial nets," in Advances in neural information processing systems, 2014, pp. 2672-2680

[5] K. Prajwal, R. Mukhopadhyay, V. P. Namboodiri, and C. 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, 2020, pp. 484-492

[6] S. Aliaksandr, S. Lathuilière, S. Tulyakov, E. Ricci, and N. Sebe. "First order motion model for image animation." Advances in Neural Information Processing Systems 32 (2019).

[7] Zhou, H., Sun, Y., Wu, W., Loy, C. C., Wang, X., & Liu, Z. (2021). Pose-controllable talking face generation by implicitly modularized audio-visual representation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 4176-4186).

[8] W. Yuxuan, R. J. Ryan, D. Stanton, Y. Wu, R. Weiss, N. Jaitly, Z. Yang et al. "Tacotron: Towards end-to-end speech synthesis." arXiv preprint arXiv:1703.10135 (2017).

[9] DeepVoice3: https://arxiv.org/abs/1710.07654

[10] Dang, H., Liu, F., Stehouwer, J., Liu, X., & Jain, A. K. (2020). On the detection of digital face manipulation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 5781-5790)

[11] Neves, J. C., Tolosana, R., Vera-Rodriguez, R., Lopes, V., Proença, H., & Fierrez, J. (2020). Ganprintr: Improved fakes and evaluation of the state of the art in face manipulation detection. IEEE Journal of Selected Topics in Signal Processing, 14(5), 1038-1048.

[12] D. Afchar, V. Nozick, J. Yamagishi, and I. Echizen, "Mesonet: a compact facial video forgery detection network," in 2018 IEEE International Workshop on Information Forensics and Security (WIFS), 2018, pp. 1-7: IEEE.

[13] L. Huang and C.-M. Pun, "Audio Replay Spoof Attack Detection by Joint Segment-Based Linear Filter Bank Feature Extraction and Attention-Enhanced DenseNet-BiLSTM Network," IEEE/ACM Transactions on Audio, Speech, Language Processing, vol. 28, pp. 1813-1825, 2020

[14] Chugh, K., Gupta, P., Dhall, A., & 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).

[15] Lomnitz, M., Hampel-Arias, Z., Sandesara, V., & Hu, S. (2020, October). Multimodal Approach for DeepFake Detection. In 2020 IEEE Applied Imagery Pattern Recognition Workshop (AIPR) (pp. 1-9). IEEE.

[16] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei, "ImageNet Large Scale Visual Recognition Challenge," International Journal of Computer Vision (IJCV), vol. 115, no. 3, pp. 211–252, 2015.

[17] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," CoRR, vol. abs/1512.03385, 2015. [Online]. Available: http://arxiv.org/abs/1512.03385

[18] Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., & Wojna, Z. (2016). Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 2818-2826).

[19] Chung JS, Zisserman A. "Out of time: automated lip sync in the wild". In Asian conference on computer vision 2016 Nov 20 (pp. 251-263). Springer, Cham.

[20] Bonettini, N., Cannas, E. D., Mandelli, S., Bondi, L., Bestagini, P., & Tubaro, S. (2021, January). Video face manipulation detection through ensemble of cnns. In *2020 25th international conference on pattern recognition (ICPR)* (pp. 5012-5019). IEEE.

[21] Tan, M., & Le, Q. (2019, May). Efficientnet: Rethinking model scaling for convolutional neural networks. In *International conference on machine learning* (pp. 6105-6114). PMLR.