International Journal of Advances in Science Engineering and Technology, ISSN(p): 2321 –8991, ISSN(e): 2321 –9009
Volume-11, Issue-1, Jan.-2023, http://iraj.in

A PHONEME-BASED DEEP FAKE DETECTION SYSTEM

[1]JONAS KRAUSE, [2]HEITOR SILVERIO LOPES

Laboratory of Bioinformatics, The Federal University of Technology – Paraná
E-mail: hslopes@utfpr.edu.br

Abstract – What happens if we can no longer trust our senses? There is a general concern that Artificial Intelligence (AI) and Machine Learning (ML) algorithms could soon create extraordinarily realistic fake videos that trick our eyes and ears. Thinking on a counter-response, we propose a new deepfake detection system based on phonemes, their transcribed text, the associated mouth movements, and video-extracted features. As a proof-of-concept, we create a Brazilian Portuguese deepfake detection system using three presidential candidates of the 2022 elections and one of the authors. We also present a novel dataset of authentic and fake videos from these four individuals mixing their identities, which we used to extract features and feed our classification methods. Our classification methods achieved satisfactory results when authenticating (or not) testing videos that contain at least one of the trained phonemes from the Brazilian Portuguese language. In conclusion, we support the hypothesis that deepfake detection is possible due to the lack of expression in the target's mouth, especially in non-English language fake videos. And developing a deepfake detection system with individual-guided classification models may help authenticate videos of celebrities or politicians in future and online events.

Keywords - Deepfake, Phoneme, Portuguese, Brazil.

I. INTRODUCTION

Deepfakes are manipulated videos, images, and audio that makes someone realistically appear to be doing or saying something that they did not do. The popularization of deepfakes started in 2017 after VICE News's Samantha Cole published an article[1] showing a manipulated porn video that appeared to feature “Wonder Woman” actress Gal Gadot. From that point forward, deepfake videos focused on putting celebrity faces into pornographic movies. However, in 2019, deepfakes of North American Speaker of the House Nancy Pelosi, Facebook CEO Mark Zuckerberg, and “Game of Thrones” character Jon Snow went viral, requiring the attention of authorities. As a result, computer scientists have dedicated a lot of effort in the last few years to detect deepfakes. More specifically, most existing systems seek to evaluate any manipulated image, audio, or video.

One of these systems is the work of Güera and Delp [1], in which authors implement a Recurrent Neural Network (RNN) for deepfake detection. They propose a temporal-aware pipeline and use one Convolutional Neural Network (CNN) to extract frame-level features. These features train an RNN that learns to classify if a video has been manipulated or not.In searching for forensics clues, the work of Nguyen et al. [2] implements capsule networks (CNN-based models) to detect forged images and videos. Their method uses specific CNN models to detect various spoofs, from replay attacks using printed images to computer-generated videos. Nhu et al. [3] also presented a forensics face detection and used Generative Adversarial Networks (GANs) to create fake training data for their CNN-based analysis. In this way, authors generate faces with multiple resolutions and sizes to help data augmentation creating a deepfake recognition system with transferable weights for robust face feature extraction.

In 2019, Ciftci et al. [4] proposed the *FakeCatcher*, a system to detect synthetic portrait videos.In their system, authors extracted biological signals from facial regions on authentic and fake portrait video pairs.Applying transformations to compute the spatial coherence and temporal consistency, capture the signal characteristics in feature sets, and train a Support Vector Machine (SVM) and a CNN. In the end, they aggregate authenticity probabilities to decide whether the video is fake or not.Li and Lyu [5] also focused on signals from facial regions by detecting face-warping artifacts. According to their research, deepfake algorithms can only generate images of limited resolutions, further warped to match the original faces in the source video.Such transforms leave distinctive artifacts in the resulting deepfake videos, and state-of-the-art CNNs may detect these artifacts.On the other hand, manipulations such as video compression may help hide these distinctive clues.For instance, in the compression process, one cancompress areas left behind from the Deep Learning (DL) architecture (such as CNNs) to the size of a pixel.In this way, important distinctive artifacts may be lost in the resize process, making the deepfake detection approach an even more challenging task.

In Brazil, deepfake videos have been used mainly for humorous content. They are created basically with two different approaches: i) By using stunt people who pretend to be the target individual (reenactment), or ii) By inserting public faces in pre-existing music videos and memes (facial transfer). However, soon enough, Brazilian deepfakes will contain political messages and attacks on public individuals. Some of this “humorous content” already

[1] https://www.vice.com/en/article/gydydm/gal-gadot-fake-ai-porn

produces fake videos of local celebrities and government individuals with the objective of demoralizing them. And some of these deepfakes achieve an excellent quality being impossible to determine their veracity only by watching them. Soon enough, deepfake political videos will influence the Brazilian elections, and we need to prepare ourselves for these virtual attacks.To minimize the impact deepfakes may have on the Brazilian community, we started a project reviewing academic articles, creating deepfake videos based on the Brazilian language (Portuguese), and developing counter-fit measures to authenticate (or not) a target individual's video or online streaming.

II. METHODOLOGY

Our method discovers mouth reenactment videos (also known as "dubbing") by analyzing specific phonemes of the target language and the individual's mouth's movements during that phoneme pronunciation. We also exploit that most deepfake videos came from DL architectures previously trained with English videos and audio data. Furthermore, new technologies explore the unique traits of each language in the world. For example, the movie industry already uses this deepfake technology for realistic voice dubbing into another language and editing. The video game industry has found more use by creating virtual characters giving a natural lingo to their products. And this technology has already been used in educational media, reenacting historical figures and memorable speeches.

Nonetheless, unethical individuals are ready to use it as a weapon for misinformation and public opinion manipulation. For example, an attacker can impersonate a target individual to gain the trust of a family member, gaining access to undue money or some other asset. Someone can also generate embarrassing content about famous individuals for blackmailing purposes or generate content to affect the public opinion of a specific political leader. This technology may also tamper with surveillance footage or other archival imagery to plant false evidence in a trial. Ultimately, the attack can place online threats such as impersonating someone in a real-time conversation or fake media. As a result, mouth reenactment has been a widely explored topic in literature and imposes one of the biggest threats in this misinformation war.

In particular, Computer Science researchers have been using numerous DL methods with different approaches on this matter, such as one-to-one (identity to identity), many-to-one (multiple identities to single identity), or many-to-many (multiple identities to multiple identities) deepfake creation techniques. Table 1 summarizes these mouth manipulation techniques found in the literature. Grouping them by their approaches, we notice similarities when comparing their implemented DL techniques. For example, all papers that implemented a one-to-one approach used Generative Adversarial Networks (GANs). And more than half of the papers reviewed that use many-to-one approaches have also implemented GANs. One can notice the implementation of Recurrent Neural Networks (RNNs) when papers use many-to-one and many-to-many approaches. These DL architectures add a temporal awareness to the deepfake models adjusting pose and expression variations. The listed many-to-many approaches show the combination of multiple techniques and DL architectures. In these approaches, we can understand using Encoders/Decoders (EDs) to create deepfake videos from many sources to many targets. And the implementation of Gated Recurrent Units (GRUs) and RNNs to the audio and identity encoding.

Approach	Paper	Year	DL Architecture
One-To-One	[6]	2017	CycleGAN
	[7]	2018	RecycleGAN
	[8]	2019	RealisticFaceGAN
	[9]	2020	DeepFaceLabGAN
Many-To-One	[10]	2017	RNN+MFCC
	[11]	2018	RNN+Char2Wav
	[12]	2018	ReenactGAN
	[13]	2018	MoCoGAN
	[14]	2018	Vid2VidGAN
	[15]	2018	RNN+MFCC
	[16]	2019	RNN+CGAN
	[17]	2019	Pix2PixGAN+AdaIN
	[18]	2019	Vid2VidGAN+AdaIN
[19]	2015	RNN+MFCC	
Many-To-Many	[20]	2018	EDs+CNN
	[21]	2018	EDs+GRU
	[22]	2019	EDs+GRU+RNN
	[23]	2019	GAN+RNN+MFCC
	[24]	2019	CGAN
	[25]	2019	CNN+RNN+CGAN
	[26]	2021	EDs+GAN

Table 1
Mouth manipulation papers and their DL architectures.

In our method, we cover all the previously listed deepfake creation strategies and approaches by analyzing specific phonemes of any language and creating a correlation between the transcription of this sound and the mouth movement during its pronunciation. So we propose the first Phoneme-based Deepfake Detection System (PDDS) that creates classification models specialized in distinctive language phonemes. These mathematical models are ML algorithms trained with the dataset extracted over selected video clips of the spoken language associated with the specific phoneme. Implemented in Python 3.10, the following steps resume the testing process of our PDDS:

1) Extract the video's length (in seconds);

2) Transcribe audio to text with time stamps every 5 seconds;
3) Select phonemes to be analyzed;
4) Clip the corresponding 5 seconds sections;
5) Extract face images from clips;
6) Extract mouth landmarks and video features;
7) Classify data grouped by every 7 frames (one phoneme) inside clips as true or false.

For the first version of the PDDS, we limited the phoneme analysis to real and fake videos with no more than 60 seconds. We implemented the audio transcription in Python, similar to transcription websites[2]. Both techniques presented satisfactory results by creating perfect transcriptions of the Portuguese language.

After selecting any real or fake labeled video database, we localized the selected phoneme on each video (real or fake) inside a clip of 5 seconds as per its transcription (steps 4, 5, and 6).For testing our system, we will swap the 5 seconds clip looking for trained patterns of the mouth movement and video features (step 7).For training our ML algorithms, we manually extracted mouth landmarks and video data features from small clips with a fraction of a second (00:00:00.20) representing the phonemes that our PDDS has to classify.The small clips output seven frames when extracting face images, enough to create the pattern of mouth landmarks movement.By manually collecting this data, we ensure that the training database for the ML algorithms contains representative data of the selected phoneme. For the final classification, we associate the mouth movement and other video-related features to phonemes and search for this pattern in the five second-clip(s) that the transcription of a new video suggests.If the classification models find these patterns in the indicated clip of the target video, the classification models authenticate the video as legit. Otherwise, the PDDS could not verify its authenticity and classify it as a deepfake. In this way, our system may also authenticate parts of the target video in case some part of the analyzed speech has been altered.

In summary, for every selected phoneme, the PDDS will find its transcription within every range of five seconds and use the previously trained ML algorithms to identify mouth and video patterns that correspond to the pronunciation of the phoneme. And after having clips with the selected phonemes defined, we extract twenty Cartesian points representing the mouth from each of the frames of the clip. Each of the twenty (X, Y) points represents the mouth position in each frame of the video. We performed this action using the OpenCV[3] library for face landmark detection in images (or frames) but using only the mouth-related points.We also use Kinetics CNN architectures to extract video movement features. They are Two-Stream Inflated 3D ConvNet (I3D) based on 2D ConvNet inflation based on the work of Carreira and Zisserman [27]. Thus, our classification models can learn seamless spatiotemporal features from videos while leveraging successful CNN architecture designs and their parameters. And by combining these features (mouth landmarks and Kinetics I3D) extracted from phonemes videos, we produce recognizable patterns between the movement of the mouth and extracted numeric attributes. In this way, ML algorithms can learn these patterns and recognize (or, in this case, authenticate) if the same pattern exists in a video target swapping. If our classification models detect these patterns in the five-second clip indicated by the transcribed text, the PDDS authenticates the video as the original. If the phoneme exists in transcribed text and the classification models can not verify its presence, the PDDS indicates that the target video is fake. In other words, it does not have the correct mouth movement during the pronunciation of the selected phoneme and points to the target video as manipulated (or fake). Otherwise, if the target video does not have any selected phonemes, the current version of the PDDS cannot authenticate this video, which will require an update of the dataset and classification models according to the phoneme and language.

III. DATASET

By selecting the Brazilian Portuguese language, we direct the first version of the PDDS to the analysis of non-English deepfake videos. And the main difficulty in this research was finding realistic deepfake videos in Portuguese.As mentioned previously, in Brazil, deepfakes generally contain humorous advertisements. They mostly create celebrities and political individuals dancing and celebrating victory (or loss of their opponents).Hence, we produced a novel dataset of real and fake videos exclusively with Brazilian Portuguese audio focusing on creating the first PDDS.This dataset started with videos of four individuals, one of the authors (Jonas Krause), and three 2022 presidential candidates (Jair Bolsonaro, LuísInácio Lula da Silva, and Simone Tebet).Using two original videos (with less than 60 seconds) of each individual (a total of 8 original videos), we used two different deepfake existing techniques (*MyVoiceYourFace*[4] and *FaceSwap*[5]) to create 96 fake videos combining each individual and their original videos.These two deepfake techniques used to create the fake videos in Portuguese use similar inputs (one image or several images) and output the projection of the input(s) over the target video, also provided by the user.With only one image and one video, we combined eight original images

[2] https://sonix.ai/

[3] https://docs.opencv.org/3.4/

[4] https://www.myvoiceyourface.com/

[5] https://faceswap.dev/

with eight original videos to make sixty-four deepfakes using *MyVocieYourFace*. And using 7,000 internet images of each of the four individuals with eight original videos, we created another 32 deepfakes using the *FaceSwap*. We made all the originals and deepfakes(a total of 104 videos) available online in a GitHub repository[6].

IV. CLASSIFICATION MODELS

Machine Learning (ML) classification models are as good as the database used to train them. And the current database has a limited universe of 104 videos. However, we consider it plausible for the first version of the PDDS. We reinforce that we may retrain the implemented ML models for larger datasets and adapt them to different deepfake databases. With that in mind, a much more populated database would be necessary to fine-tune and precisely train the presented ML models (or new ones) for accuracy improvement.

Following our previous experience and the work of Ciftci et al. [4], we implement the Support Vector Machine (SVM) as a classification model for the Brazilian Portuguese version of the first PDDS.SVM is an ML algorithm trained to encounter the optimum hyperplane in N-dimensional space (with N being the number of features).It creates support vectors based on the points closest to the decision surface (or hyperplane).These bordering data points are the most difficult to classify, and they directly impact the optimum location of the decision surface.As an example, Figure 1 shows the SVM algorithm defining the optimal hyperplane (in this case, a line) of a two-dimensional space, the support vectors (dashed lines) of each class (squares and blue), and the maximum margin.As the number of dimensions N increases, these margins allow the hyperplane to adapt to each feature or component.And consequently, it creates a more accurate classification model.

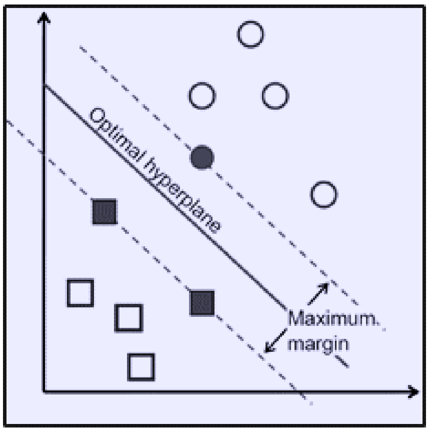

Figure 1: SVM with Linear Classification Kernel

We also explore the multiple classification kernels other than the Linear one presented in Figure 1.So we retrain the classification models with the Polynomial and Radial Basis Function (RBF) classification kernels using the same SVM algorithm. Figure 2 visually presents these other two kernels and their hyperplane and decision surface.

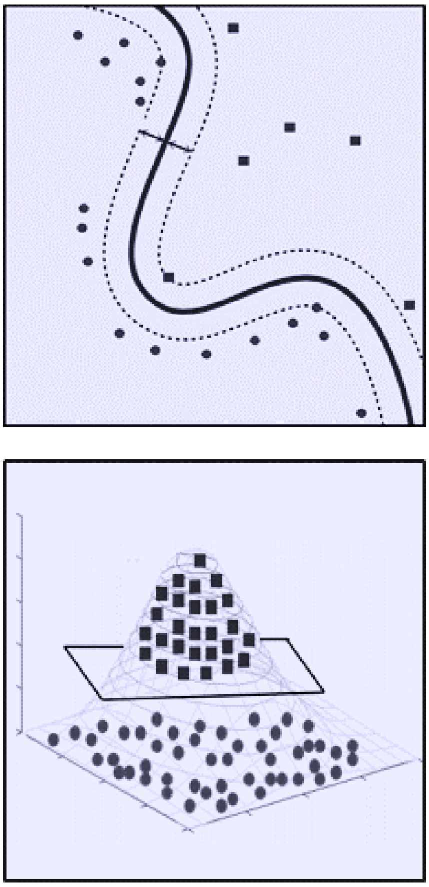

Figure 2: SVM with Polynomial and RBF Classification Kernels

V. RESULTS

From the Brazilian Portuguese dataset, we extract the length of each video, transcribe the audio, clip the phonemes, extract faces from the clip's frames (512 x 512 pixels), extract mouth landmarks (280 integer values), and I3D video features (1,024 float values). And from the 104 videos of the dataset, the PDDS could not fully extract the necessary mouth points from nine videos. By analyzing which are these videos, we noticed that all of the undetected mouth landmarks came from deepfakes created by the *MyVoiceYourFace* method. When watching some of the dataset videos, one can note that the created deepfakes of this method are not much realistic videos like the ones produced by the other method (*FaceSwap*). With that in mind, we may consider the extraction of mouth landmarks as the first classification step of the PDDS. And, if we cannot detect the necessary landmarks to represent the mouth

[6] https://github.com/jonaskrause/DeepFake-PortugueseDataset

in each frame of the speakable audio, we have strong indicators that the video is a deepfake.

In this paper, we focus on two specific phonemes of the Portuguese language: the *"ÃO"* and *"É"*. By the transcribed texts, we identified these phonemes 288 times throughout the 104 original and fake videos. From the 96 deepfake ones, the PDDS could not extract the necessary mouth landmarks of selected phonemes from 17 scenes, which indicates that these videos were possibly fabricated and do not project the correct mouth landmarks to simulate the corresponding mouth movement. Therefore, the first step of the PDDS correctly classified nine of the 104 dataset videos as a deepfake. But, as expected, the real challenge is to identify the more realistic deepfakes presented in the dataset. When we consider only these 95 videos with lifelike mouth movements, the PDDS collected the necessary data for the selected phonemes 271 times (136 *"ÃOs"* and 135 *"És"*). We use this data labeling them as real or fake in the training process of our classification models using the SVM algorithm. For this, we separated 80% of the data for the learning process and 20% for the validation process of the algorithms. To further explore the extracted data, we also implemented two preprocessing techniques to manipulate the raw data to ignore most of the magnitude difference and put in evidence the relationship among the data. With this intent, we applied the *Savitzky-Golay*filter with the first and second derivatives (*SG1d* and *SG2d*) to smooth out noisy signals with a large frequency span. Table 2 presents the results of the SVM classification models with different kernels, two preprocessing techniques, and the accuracy percentages over the validation data when using only mouth landmarks, only I3D, and both combined (Mouth+I3D).

SVM	Preprocess	Mouth	I3D	Mouth+I3D
Raw		75.9%	70.3%	77.7%
Linear	*SG1d*	74.0%	77.7%	75.9%
	SG2d	79.6%	74.0%	81.4%
	Raw	83.3%	83.3%	85.1%
Poly	*SG1d*	85.1%	83.3%	87.0%
	SG2d	81.4%	85.1%	85.1%
	Raw	83.3%	85.1%	85.1%
RBF	*SG1d*	85.1%	87.0%	**88.8%**
	SG2d	85.1%	85.1%	87.0%

Table 2
Accuracy results of SVM classification models during their validation process.

Analyzing Table 2, we note that the preprocessing technique improved most of the SVM models' accuracy, especially for non-linear kernels (Poly and RBF) with a first derivative filter (*SG1d*). It is also noticeable that grouping the mouth landmarks and the I3D features showed improvement in almost all the cases, which leads us to the best model for the first version of the PDDS: The SVM with RBF kernel. In the following steps, we may test other preprocessing techniques and new feature extraction methods, which we can aggregate to create a more robust deepfake detection system.

To better understand these results, we looked into which clips of the validation set (20% of 271 = 54 phoneme clips) belong to each video to summarize how many original and fake videos the best-trained model (SVM with RBF kernel and *SG1d*) classified. In this process, we implement the last stage of the PDDS and match the learned pattern of mouth movement using two feature extraction methods. By swapping the 5-second clip where we located the phoneme, the SVM model matches for a sequence of seven frames with a stride of one throughout all the extracted frames. In this way, every time one of the individuals speaks one of the selected phonemes, the PDDS can authenticate if the person in the video is making a mouth movement accordingly.The first analysis shows that the PDDS correctly authenticated all the original videos, which indicates that the SVM model learned the necessary components to recognize the extracted features in a video swap process. Nevertheless, as seen in Table 2, the SVM models authenticated four videos (or part of them) as real ones, even being deepfakes. We created them using the *FaceSwap* method, and three of them are the deepfake in which the individual is the input and target (but different files). It means that we produce deepfakes of one person over the video of that same person, making it easy for the process to create a much more realistic deepfake. And it worked, deepfakes of Jonas Krause being Jonas Krause, Jair Bolsonaro being Jair Bolsonaro, and LuísInácio Lula da Silva being LuísInácio Lula da Silva tricked the first version system. The last video misclassified by the SVM model also tricked the PDDS, which partially authenticated it. That means it correctly classified one (or more) 5-second clip(s) of the video but not the entire length, pointing to the wrong output or a partial authentication. In this case, one deepfake that projects the face of Jair Bolsonaro in one of the authors (Jonas Krause) tricked the system. And despite the satisfactory results of the SVM training process, we understand that PDDS needs further improvement by testing multiple classification models, implementing new preprocessing, and exploring other feature extraction techniques. Furthermore, the Brazilian Portuguese dataset needs to be populated with more individuals and realistic deepfakes.

VI. CONCLUSION

Experiments that we conducted in thispaper indicate that it is possible to identify deepfakes using phoneme-base patterns of the mouth and features extracted from the frames that comprise them. Here we introduced the first Brazilian Portuguese deepfake dataset with 104 videos (with 8 originals and 96 deepfakes) and the first Phoneme-based DeepfakeDetection System (PDDS) with a SVM algorithm with multiple kernels trained over this dataset.

As presented in the methodology section, the PDDS is an upgradable pipeline system that is flexible to new data preprocessing techniques and multiple classification models. Most important, as we designed the PDDS, the system can operate in any phoneme-based language. And it can take benefits when authenticating non-English videos with unique phonemes mainly because commonly listed creating deepfake approaches are based on AI approaches trained over English speakers' videos that do not have the same mouth expressions. In addition, the analysis of the presented results indicates that developing a deepfake detection system with individually guided classification models can be a robust and fast approach to identifying deepfakes in real-time transmissions and live speeches.

ACKNOWLEDGMENT

We would like to thank the National Council for Scientific and Technological Development (CNPq) for the scholarship (164765/2020-4) and support of our research.

REFERENCE

[1] D. Güera and E. J. Delp, “Deepfake video detection using recurrent neural networks,” in 2018 15th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS), 2018.

[2] H. H. Nguyen, J. Yamagishi, and I. Echizen, “Capsule-forensics: Using capsule networks to detect forged images and videos,” CoRR, vol. abs/1810.11215, 2018.

[3] T. Do Nhu, I. Na, and S. Kim, “Forensics face detection from gans using convolutional neural network,” in International Symposium on Information Technology Convergence (ISITC), 2018.

[4] U. A. Ciftci and I. Demir, “Fakecatcher: Detection of synthetic portrait videos using biological signals,” CoRR, vol. abs/1901.02212, 2019.

[5] Y. Li and S. Lyu, “Exposing deepfake videos by detecting face warping artifacts,” CoRR, vol. abs/1811.00656, 2018.

[6] R. Xu, Z. Zhou, W. Zhang, and Y. Yu, “Face transfer with generative adversarial network,” CoRR, vol. abs/1710.06090, 2017.

[7] Bansal, S. Ma, D. Ramanan, and Y. Sheikh, “Recycle-gan: Unsupervised video retargeting,” inProceedings of the European Conference on Computer Vision (ECCV), 2018.

[8] Z. Liu, H. Hu, Z. Wang, K. Wang, J. Bai, and S. Lian, “Video synthesis of human upper body with realistic face,” CoRR, vol. abs/1908.06607, 2019.

[9] Perov, D. Gao, N. Chervoniy, K. Liu, S. Marangonda, C. Um'e, M. Dpfks, C. S. Facenheim, L. RP, J. Jiang, S. Zhang, P. Wu, B. Zhou, and W. Zhang, “Deepfacelab: A simple, flexible and extensible face swapping framework,” CoRR, vol. abs/2005.05535, 2020.

[10] S. Suwajanakorn, S. M. Seitz, and I. Kemelmacher-Shlizerman, “Synthesizing obama: Learning lip sync from audio,” ACM Trans. Graph., vol. 36, no. 4, 2017.

[11] R. Kumar, J. Sotelo, K. Kumar, A. de Brebisson, and Y. Bengio, “Obamanet: Photo-realistic lip-sync from text,” 2017.

[12] W. Wu, Y. Zhang, C. Li, C. Qian, and C. C. Loy, “Reenactgan: Learning to reenact faces via boundary transfer,” CoRR, vol. abs/1807.11079, 2018.

[13] S. Tulyakov, M.-Y. Liu, X. Yang, and J. Kautz, “Mocogan: Decomposing motion and content for video generation,” in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 2018.

[14] T. C. Wang, M.Y. Liu, J.Y. Zhu, G. Liu, A. Tao, J. Kautz, and B. Catanzaro, “Video-to-video synthesis,” in Advances in Neural Information Processing Systems,Eds., vol. 31. Curran Associates, Inc., 2018.

[15] S. A. Jalalifar, H. Hasani, and H. Aghajan, “Speechdriven facial reenactment using conditional generative adversarial networks,” CoRR, vol. abs/1803.07461, 2018.

[16] O. Fried, A. Tewari, M. Zollhöfer, A. Finkelstein, E. Shechtman, D. B. Goldman, K. Genova, Z. Jin, C. Theobalt, and M. Agrawala, “Text-based editing of talking-head video,” CoRR, vol. abs/1906.01524,2019.

[17] E. Zakharov, A. Shysheya, E. Burkov, and V. S. Lempitsky, “Few-shot adversarial learning of realistic neural talking head models,” CoRR, vol. abs/1905.08233, 2019.

[18] T.-C. Wang, M.-Y. Liu, A. Tao, G. Liu, B. Catanzaro, and J. Kautz, “Few-shot video-to-video synthesis,” in Advances in Neural Information Processing Systems, H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, Eds., vol. 32. Curran Associates, Inc., 2019.

[19] T. Shimba, R. Sakurai, H. Yamazoe, and J.-H. Lee, “Talking heads synthesis from audio with deep neural networks,” in 2015 IEEE/SICE International Symposium on System Integration (SII), 2015.

[20] H. Zhou, Y. Liu, Z. Liu, P. Luo, and X. Wang, “Talking face generation by adversarially disentangled audio-visual representation,” CoRR, vol. abs/1807.07860, 2018.

[21] Vougioukas, S. Petridis, and M. Pantic, “End-to-end speech-drivenfacial animation with temporal gans,” 2018.

[22] Vougioukas, S. Petridis, and M. Pantic, “Realistic speech-driven facial animation with gans,” in Int. J. Comput. Vision, vol. 128, no. 5, p. 1398–1413, 2020.

[23] Y. Song, J. Zhu, D. Li, X. Wang, and H. Qi, “Talking face generation by conditional recurrent adversarial network,” 2019.

[24] Pumarola, A. Agudo, A. Martinez, A. Sanfeliu, and F. Moreno-Noguer, “Ganimation: One-shot anatomically consistent facial animation,” 2019.

[25] Chen, R. K. Maddox, Z. Duan, and C. Xu, “Hierarchical cross-modal talking face generationwith dynamic pixel-wise loss,” CoRR, vol. abs/1905.03820, 2019.

[26] Lahiri, V. Kwatra, C. Fr¨uh, J. Lewis, and C. Bregler, “Lipsync3d: Data-efficient learning of personalized 3d talking faces from video using pose and lighting normalization,” CoRR, vol. abs/2106.04185,2021.

[27] J. Carreira and A. Zisserman, “Quo vadis, action recognition? a new model and the kinetics dataset,” 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 4724–4733, 2017.

★★★