RESEARCH ARTICLE OPEN ACCESS

Deep Fake Image Detection Using Supervised Machine Learning Technique: A Study

Sharad Shrivastava*, Prof. Sandeep Piplotia**
*(M. Tech. Scholar, Department of CSE-AI/ML, OIST, Bhopal)
**(Assistant Professor, Department of CSE-AI/ML, OIST, Bhopal)

ABSTRACT

The rapid advancement of deep learning technologies has led to the emergence of deepfakes—synthetically generated images or videos that closely mimic real human appearances. These manipulated images pose a significant threat to digital security, privacy, and the credibility of online content. As a countermeasure, this study investigates the application of supervised machine learning techniques for the detection of deepfake images. A labeled dataset containing both authentic and deepfake images is used for training and evaluation. The proposed methodology involves key pre-processing steps followed by feature extraction techniques, including facial landmarks and texture descriptors. These features serve as input to various supervised learning algorithms such as Support Vector Machine (SVM), Random Forest, K-Nearest Neighbors (KNN), and Logistic Regression. The models are evaluated based on performance metrics such as accuracy, precision, recall, F1-score, and ROC-AUC. Experimental results indicate that SVM and Random Forest outperform other models in terms of accuracy and generalization, demonstrating strong potential for real-world deepfake detection tasks. The study highlights the effectiveness of classical machine learning methods when combined with appropriate feature extraction strategies. It concludes by suggesting future work directions, such as incorporating deep learning approaches, video-based detection, and real-time implementation for broader and more robust applications in digital forensics.

Keywords : — Deepfake, Supervised Learning, SVM, Image Forensics, Face Detection, Feature Extraction, Machine Learning

I. INTRODUCTION

Visual aids are commonly utilized across various industries such as law, medicine, and entertainment [1, 2]. However, the extensive usage of visual media also presents a risk of misuse. Media forgery has been prevalent in digital culture for a while, where software tools like Photoshop are used for manual manipulation of media content. With the recent advancements in Computer Vision (CV) and Machine Learning (ML) technologies, media forgery has become more accessible and widespread.

In 2012, the field of CV experienced a significant breakthrough when AlexNet, an AI model developed by Alex, outperformed other models in the image recognition challenge by a large margin. Since then, AlexNet, which is a classic convolutional neural network architecture, has been instrumental in many CV applications. Another leap forward in CV research was made in 2014 when Goodfellow introduced the Generative Adversarial Network (GAN). GAN enables the creation of realistic-looking images from scratch without human intervention or manual editing.

The rapid evolution of hardware that supports artificial neural network models' training has catalyzed the growth of deep learning. In 2017, a novel deep learning-based media forgery algorithm called 'Deepfake' emerged and wreaked havoc, threatening society's security and privacy. Deepfake is a synthetic technique that replaces the person in an existing image or video with someone else's likeness or characteristics. It is a portmanteau of 'deep learning' and 'fake'. It originated from an anonymous individual under the pseudonym 'deepfake' who uploaded numerous pornographic videos to the Reddit website. The actresses' faces in the videos were swapped with those of other celebrities using deep learning [3, 4].

Figure 1 outlines the examples of deepfake based on different generation methods. Based on the figure,

a. Puppet Master refers to the transfer of motion movement by synthesizing the motion of the source and regenerating onto the target output [5];
b. Face Swapping involves swapping the facial regions between two people from one to another [6];
c. Involve facial reenactment where Neural Textures focuses on deferred neural rendering to integrate neural textures in the parametric vectors for facial synthesis [7] while Face2Face uses GAN, such as CycleGAN and Star- GAN to achieve the synthesis output [8];
d. Present entire face synthesis to produce non-exist human outputs by the training on the different source data to capture their significant facial characteristics [9, 10];
e. Leverages GAN's capability to modify certain facial attributes on target.

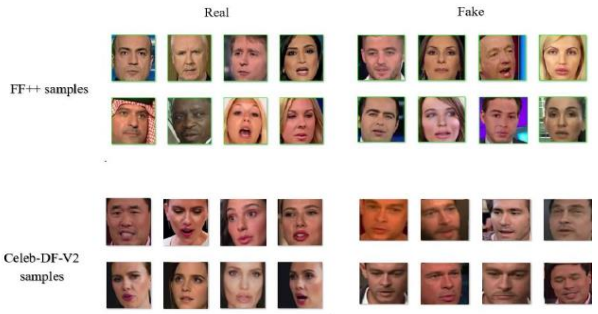

Figure 1: Examples of Deepfake

II. LITERATURE REVIEW

Masood et al. [1], easy access to audiovisual content on social media and the availability of modern tools such as Tensorflow and Keras, open source training models, economic computing infrastructure, and rapid development of profound (DL) methods have created new trends. Particularly, they are ready to use slightly available, generically controversial networks (geese), making it possible to deceive disinformation and revenge porn to partially or completely deceive deepfake media, financial fraud and other homes to interfere with government functions. Existing research focuses primarily on DeepFak image and video detection. This paper provides a comprehensive review and detailed analysis of existing tools and machine learning (ML)-based approaches, as well as methods used to identify such operations in both audio and video. For each category of DeepFak, we discuss information about operational approaches, current public data records, and key criteria and results for assessing the performance of DeepFak detection techniques. Additionally, it will explain open challenges, list future directions, and control researchers on topics that need to be considered to improve both the areas of Deepake generation and detection. This work is hoped to help readers understand how current limitations and where future research leads them to.

Waseem et al. [2], advancements in facial manipulation generation have led to relatively practical and indistinguishable face and expression change videos. However, this has additionally raised issues concerning the safety dangers related to deepfakes. In the sphere of multimedia forensics, the detection and particular localization of photo forgery has end up vital tasks. Current deepfake detectors carry out nicely with superb faces inside unique datasets, however regularly warfare to preserve their overall performance while evaluated throughout one of a kind datasets. To this end, we suggest an attention-primarily based totally multi-challenge method to enhance function maps for type and localization tasks. The encoder and the attention-primarily based totally decoder of our community generate localized maps that spotlight areas with statistics approximately the sort of manipulation. These localized capabilities are shared with the type community, enhancing its overall performance. Instead of the use of encoded spatial capabilities, attention-primarily based totally localized capabilities from the decoder`s first layer are mixed with frequency area capabilities to create a discriminative illustration for deepfake detection. Through tremendous experiments on face and expression change datasets, we show that our technique achieves aggressive overall performance in contrast to modern deepfake detection tactics in each in-dataset and cross-dataset scenarios.

Khan et al. [3], currently, Deepfake Media is extensive due to the slightly available tools and mobile apps. This allows you to generate realistic deep pope videos/images without the need for technical knowledge. As we are further advances in this technological field in the near future, we are also expected to see the amount and quality of deep-pake media, but deep-fuck media could create new practical means of spreading false/disinformation. Based on these concerns, you will need tools to detect Deepfake media. In this study, we propose a new hybrid transformer network using an initial fusion strategy for Deepfake video recognition. The model uses two different CNN networks, namely H.(1) XceptionNet and (2) efficient B, as feature extractors. Train both extracts together with the trance in FaceForensics++, DFDC benchmarks. Although our model has a relatively simple architecture, comparable results provide other, more advanced, cutting-edge approaches when evaluated in FaceForensics++ and DFDC benchmarks. We also propose new excerpts from face cutouts and random excerpts. The proposed augmentation has been shown to improve model identification performance and overadaptation. Furthermore, it shows that the model can be trained from a fairly small amount of data.

Ali Raza et al. [4], deepfake is used in engineered media to create counterfeit visual and sound substance in view of an individual's current media. The deepfake is replaces an individual's face and voice with counterfeit media to make it sensible looking. Counterfeit media content age is deceptive and a danger to the local area. These days, deepfakes are exceptionally abused in cybercrimes for wholesale fraud, digital coercion, counterfeit news, monetary misrepresentation, VIP counterfeit vulgarity recordings for extorting, and some more. As indicated by a new Sensity report, more than 96% of the deepfakes are of indecent substance, with most casualties being from the Unified Realm, US, Canada, India, and South Korea. In 2019, cybercriminals produced counterfeit sound substance of a CEO to call his association and request that they move $243,000 to their ledger. Deepfake wrongdoings are raising everyday. Deepfake media recognition is a major test and has popularity in computerized criminology. A high level exploration approach should be worked to shield the casualties from coercing by distinguishing deepfake content. The essential point of our exploration study is to identify deepfake media utilizing an effective structure. A novel deepfake indicator (DFP) move toward in view of a half and half of VGG16 and convolutional brain network design is proposed in this review. The deepfake dataset in view of genuine and counterfeit appearances is used for building brain network methods. The Xception, NAS-Net, Portable Net, and

VGG16 are the exchange learning methods utilized in examination. The proposed DFP approach accomplished 95% accuracy and 94% precision for deepfake recognition. Our novel proposed DFP approach beat move learning procedures and other cutting edge investigations. Our original exploration approach assists online protection experts with defeating deepfake-related cybercrimes by precisely distinguishing the deepfake content and saving the deepfake casualties from extorting.

Zobaed et al. [5], the fast progression in profound learning makes the separation of true and controlled facial pictures and video cuts phenomenally more enthusiastically. The hidden innovation of controlling facial appearances through profound generative methodologies, articulated as DeepFake that have arisen as of late by advancing countless malevolent face control applications. Consequently, the need of other kind of procedures that can evaluate the trustworthiness of computerized visual substance is undeniable to lessen the effect of the manifestations of DeepFake. An enormous group of exploration that are performed on DeepFake creation and discovery make an extent of pushing each other past the ongoing status. This study presents difficulties, research patterns, and headings connected with DeepFake creation and recognition procedures by assessing the remarkable examination in the DeepFake area to work with the improvement of additional powerful methodologies that could manage the more development DeepFake later on.

Thambawita et al. [6], late worldwide improvements highlight the conspicuous job large information have in present day clinical science. However, protection issues comprise a predominant issue for gathering and dividing information among scientists. Be that as it may, manufactured information created to address genuine information conveying comparable data and circulation might ease the security issue. In this review, we present generative ill-disposed networks (GANs) fit for creating reasonable engineered DeepFake 10-s 12-lead electrocardiograms (ECGs). We prepared the GANs with 7,233 genuine typical ECGs to deliver 121,977 DeepFake ordinary ECGs. By confirming the ECGs utilizing a business ECG translation program (Dream 12SL, GE Medical services), we show that the Pulse2Pulse GAN was better than the WaveGAN* to deliver practical ECGs. ECG spans and amplitudes were comparative between the DeepFake and genuine ECGs. Albeit these engineered ECGs imitate the dataset utilized for creation, the ECGs are not connected to any people and may subsequently be utilized unreservedly. The manufactured dataset will be accessible as open access for specialists at OSF.io and the DeepFake generator accessible at the Python Bundle List (PyPI) for producing engineered ECGs. All in all, we had the option to create reasonable engineered ECGs utilizing generative ill-disposed brain networks on ordinary ECGs from two populace studies, in this manner tending to the applicable security issues in clinical datasets.

Ahmed et al. [7], the objective of this study is to find whether openness to Deepfake recordings improves individuals at identifying Deepfake recordings and whether it is a superior methodology against battling Deepfake. For this study a gathering from Bangladesh has chipped in. This gathering were presented to various Deepfake recordings and posed resulting inquiries to confirm enhancement for their degree of mindfulness and identification in setting of Deepfake recordings. This study has been acted in two stages, where second stage was performed to approve any speculation. The phony recordings are customized for the particular crowd and where fit, are made without any preparation. At last, the outcomes are examined, and the review's objectives are surmised from the acquired information.

Das et al. [8], improvements in the way deep pope produces have made the creation of altered and manipulated faces more common. At the same time, we saw the development of identification models to distinguish the original faces of manipulated images or video content. This article focuses on identifying existing deep foot restrictions and defects. Through quantitative and qualitative analysis of existing methods and data records, we identified several important issues related to Deepfark's perception. I found the Deepfake data records to be very exaggerated. This means that models can easily match. Data records are created with a small set of real faces to generate some false examples. In training these data records, models tend to memorize the actor's faces and labels instead of learning fake traits. To alleviate this problem, we propose a simple data augmentation method known as facial disorders. Our method uses facial landmark information to dynamically block areas of the image. This helps the model selectively handle related regions of the input. Our evaluation experiments show that face allocation can successfully improve data variation and reduce overly adaptation problems. Our method achieves a protocol reduction of 15.2% to 35.3% on a variety of data records compared to other occlusion-based methods. Furthermore, we propose general data for preprocessing to train and evaluate existing architectures that can improve the generalizability of these models for deeper detection.

Brian Dolhansky et al. [9], present a see of the Deepfakes Recognition Challenge (DFDC) dataset comprising of 5K recordings including two facial change calculations. An information assortment crusade has been completed where partaking entertainers have gone into a consent to the utilization and control of their similarities in our production of the dataset. Variety in a few tomahawks (orientation, complexion, age, and so on.) has been thought of and entertainers recorded recordings with erratic foundations consequently bringing visual changeability. At last, a bunch of explicit measurements to assess the exhibition have been characterized and two existing models for distinguishing deepfakes have been tried to give a reference execution pattern.

Brian Dolhansky et al. [10], Deepfakes are a new off-the-rack control procedure that permits anybody to trade two personalities in a solitary video. Notwithstanding Deepfakes, an assortment of GAN-based face trading strategies have likewise been distributed with going with code. To counter this arising danger, we have developed a very huge face trade video dataset to empower the preparation of recognition models, and coordinated the going with DeepFake Discovery Challenge (DFDC) Kaggle rivalry. Critically, all recorded subjects consented to take part in and have their similarities adjusted during the development of the face-traded dataset. The DFDC dataset is by a wide margin the biggest presently and freely accessible face trade video dataset, with more than 100,000 all out cuts obtained from 3,426 paid entertainers, created with a few Deepfake, GAN-based, and non-learned techniques. As well as depicting the strategies used to develop the dataset, we give a definite examination of the top entries from the Kaggle challenge. We show in spite of the fact that Deepfake recognition is very troublesome despite everything a strange issue, a Deepfake location model prepared exclusively on the DFDC can sum up to genuine "in nature" Deepfake recordings, and such a model can be a significant examination device while breaking down possibly Deepfaked recordings.

III. CHALLENGES

The deepfake generation and detection can be compared to a cat-and-mouse game where improving the generator leads to the advancement of the detector. Conventional methods show that designing a detector based on a particular generator's weaknesses, such as traces or anomalies, is not a sustainable, reliable, and flexible solution. As deepfake generators aim to produce artifact-less results, the trend in detector research has shifted towards discrimination based on learned features instead of handcrafted ones. However, pre-trained CNN models may not perform well with different deepfake scenarios and can be vulnerable to malicious attacks. Addressing these drawbacks may bring the deepfake detector's performance to a higher level.

The creation of state-of-the-art deepfakes heavily relies on GAN technology. Researchers have improved deepfake network training by integrating tertiary concepts such as style transfer, motion transfer, biometric artifacts, and semantic segmentation to achieve more hyperrealistic and natural results with high confidence [11, 12]. However, current deepfakes are still imperfect and leave room for improvement. GAN training is time-consuming, resource-intensive, and susceptible to overfitting, and the output is not flawless enough to evade detection.

IV. IMPORTANCE OF MACHINE LEARNING IN DEEP FAKE DETECTION

Machine Learning plays a crucial role in detecting deepfakes due to its ability to automatically learn patterns and anomalies in visual data that are often imperceptible to the human eye. Deepfake images and videos are generated using complex algorithms like Generative Adversarial Networks (GANs), which can mimic facial expressions, lighting, and textures with high realism. Manual detection becomes increasingly difficult as these techniques evolve.

Machine Learning models, particularly supervised learning algorithms, can be trained on large datasets containing both real and fake images to distinguish between them based on extracted features such as facial landmarks, texture inconsistencies, or pixel-level artifacts. These models learn from examples and improve their accuracy over time, enabling faster, more scalable, and more accurate detection.

Additionally, ML-based detection systems can adapt to new types of deepfake content, making them more robust than traditional rule-based or manual methods. With the growing risk of deepfakes in media, security, politics, and online identity, ML provides an essential line of defense for maintaining the authenticity and trustworthiness of digital content.

V. MACHINE LEARNING

Machine Learning is a subset of Artificial Intelligence concerned with "teaching" computers how to act without being explicitly programmed for every possible scenario. The central concept in Machine Learning is developing algorithms that can self-learn by training on a massive number of inputs. Machine learning algorithms are used in various applications, such as email filtering and computer vision, where it is difficult or infeasible to develop conventional algorithms to perform the needed tasks [4]. Machine learning enables the analysis of vast amounts of information. While it usually delivers faster, more precise results to identify profitable prospects or dangerous risks, it may also require additional time and assets to train it appropriately. Merging machine learning with AI and perceptive technologies can make it even more effective in processing vast volumes of information. Machine learning is closely associated with computational statistics, which focuses on making predictions using computers. Machine learning approaches are conventionally divided into three broad categories, namely Supervised Learning, Unsupervised Learning & Semi-supervised Learning, depending on the nature of the "signal" or "feedback" available to the learning system.

Supervised Learning

A model is trained through a process of learning in which predictions must be made and corrected if those predictions are wrong. The training process continues until a desired degree of accuracy is reached on the training data. Input data is called training data and has a known spam / not-spam label or result at one time.

Unsupervised Learning

By deducting the structures present in the input data, a model is prepared. This may be for general rules to be extracted. It may be through a mathematical process that redundancy can be systematically reduced, or similar data can be organized. There is no labeling of input data, and there is no known result.

Semi-Supervised Learning

Semi-supervised learning fell between unsupervised learning (without any labeled training data) and supervised learning (with completely labeled training data). There is a desired problem of prediction, but the model needs to learn the structures and make predictions to organize the data. Input data is a combination of instances that are marked and unlabeled.

Technique

Nearest Neighbors Algorithm

The Nearest Neighbor (NN) rule differentiates the classification of unknown data point because of closest neighbor whose class is known. The nearest neighbor is calculated based on estimation of k that represents how many nearest neighbors are taken to characterize the data point class. It utilizes more than one closest neighbor to find out the class where the given data point belong termed as KNN. The data samples are required in memory at run time called as memory-based technique. The training points are allocated weights based on their distances from the sample data point. However, the computational complexity and memory requirements remained key issue. For addressing the memory utilization problem, size of data gets minimized. The repeated patterns without additional data are removed from the training data set [15].

Naive Bayes Classifier

Naive Bayes Classifier technique is functioned based on Bayesian theorem. The designed technique is used when dimensionality of input is high. Bayesian Classifier is used for computing the possible output depending on the input. It is feasible to add new raw data at runtime. A Naive Bayes classifier represents presence (or absence) of a feature (attribute) of class that is unrelated to presence (or absence) of any other feature when class variable is known. Naïve Bayesian Classification Algorithm was introduced by Shinde S.B and Amrit Priyadarshi (2015) that denotes statistical method and supervised learning method for classification. Naive Bayesian Algorithm is used to predict the heart disease. Raw hospital dataset is employed. After that, the data gets preprocessed and transformed. Finally by using the designed data mining algorithm, heart disease was predicted and accuracy was computed.

VI. CONCLUSIONS

Deepfake technology presents a growing challenge in the digital era, with its potential to manipulate and spread misinformation through highly realistic synthetic images. This study demonstrates the effectiveness of supervised machine learning techniques in detecting such manipulated content. By using appropriate feature extraction methods and training classifiers such as Support Vector Machine, Random Forest, and K-Nearest Neighbors, the models were able to distinguish between real and fake images with notable accuracy.

Among the tested algorithms, SVM and Random Forest showed superior performance in terms of precision and generalization, highlighting their suitability for deepfake detection tasks. The study confirms that even traditional machine learning methods, when combined with proper preprocessing and feature selection, can serve as a reliable solution for identifying deepfakes.

However, as deepfake generation techniques continue to advance, there is a need for continuous improvement in detection models. Future work may involve integrating deep learning approaches, expanding datasets to include more variations, and developing real-time detection systems. Ultimately, machine learning will remain a key component in safeguarding digital content integrity and combating the misuse of synthetic media.

REFERENCES

[1] Masood, M.; Nawaz, M.; Malik, K.M.; Javed, A.; Irtaza, A.; Malik, H. Deepfakes Generation and Detection: State-of-the-Art, Open Challenges, Countermeasures, and Way Forward. Appl. Intell. 2023, 53, 3974–4026.

[2] Waseem, S.; Abu-Bakar, S.A.R.S.; Omar, Z.; Ahmed, B.A.; Baloch, S.; Hafeezallah, A. Multi-Attention-Based Approach for Deepfake Face and Expression Swap Detection and Localization. EURASIP J. Image Video Process. 2023, 1, 14.

[3] Khan, S.A.; Dang-Nguyen, D.T. Hybrid Transformer Network for Deepfake Detection. In Proceedings of the 19th International Conference on Content-based Multimedia Indexing, Graz, Austria, 14–16 September 2022; pp. 8–14.

[4] Ali Raza, Kashif Munir and Mubarak Almutairi, “A Novel Deep Learning Approach for Deepfake Image Detection”, Apply Science, pp. 01-15, 2022.

[5] Zobaed, S.; Rabby, F.; Hossain, I.; Hossain, E.; Hasan, S.; Karim, A.; Hasib, K.M. Deepfakes: Detecting forged and synthetic media content using machine learning. In Artificial Intelligence in Cyber Security: Impact and Implications; Springer: Berlin/Heidelberg, Germany, 2021; pp. 177–201.

[6] Thambawita, V.; Isaksen, J.L.; Hicks, S.A.; Ghouse, J.; Ahlberg, G.; Linneberg, A.; Grarup, N.; Ellervik, C.; Olesen, M.S.; Hansen, T.; et al. DeepFake electrocardiograms using generative adversarial networks are the beginning of the end for privacy issues in medicine. Sci. Rep. 2021, 11, 21869.

[7] Ahmed, M.F.B.; Miah, M.S.U.; Bhowmik, A.; Sulaiman, J.B. Awareness to Deepfake: A resistance mechanism to Deepfake. In Proceedings of the 2021 International Congress of Advanced Technology and Engineering (ICOTEN), Taiz, Yemen, 4–5 July 2021; pp. 1–5.

[8] Das, S.; Seferbekov, S.; Datta, A.; Islam, M.S.; Amin, M.R. Towards Solving the Deepfake Problem: An Analysis on Improving Deepfake Detection Using

Dynamic Face Augmentation. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, BC, Canada, 11–17 October 2021; pp. 3776–3785.

[9] Brian Dolhansky, Russ Howes, Ben Pflaum, Nicole Baram, and Cristian Canton Ferrer. The Deepfake Detection Challenge (DFDC) Preview Dataset. arXiv eprints, page arXiv:1910.08854, 2019.

[10] Brian Dolhansky, Joanna Bitton, Ben Pflaum, Jikuo Lu, Russ Howes, Menglin Wang, and Cristian Canton Ferrer. The DeepFake Detection Challenge Dataset. arXiv e-prints, page arXiv:2006.07397, 2020.

[11] Md Shohel Rana, Mohammad Nur Nobi, Beddhu Murali, and Andrew H Sung. Deepfake detection: A systematic literature review. IEEE access, 10:25494–25513, 2022.

[12] Jin, B.; Cruz, L.; Gonçalves, N. Deep facial diagnosis: Deep transfer learning from face recognition to facial diagnosis. IEEE Access 2020, 8, 123649–123661.

[13] Lewis, J.K.; Toubal, I.E.; Chen, H.; Sandesera, V.; Lomnitz, M.; Hampel-Arias, Z.; Prasad, C.; Palaniappan, K. Deepfake Video Detection Based on Spatial, Spectral, and Temporal Inconsistencies Using Multimodal Deep Learning. In Proceedings of the 2020 IEEE Applied Imagery Pattern RecognitionWorkshop (AIPR),Washington DC, DC, USA, 13–15 October 2020; pp. 1–9.

[14] Lee, H.; Park, S.H.; Yoo, J.H.; Jung, S.H.; Huh, J.H. Face recognition at a distance for a stand-alone access control system. Sensors 2020, 20, 785.

[15] Z. He, W. Zuo, M. Kan, S. Shan, and X. Chen. Attgan: Facial attribute editing by only changing what you want. IEEE Transactions on Image Processing, 28(11): 5464–5478, 2019.

[16] Sergey Tulyakov, Ming-Yu Liu, Xiaodong Yang, and Jan Kautz. Mocogan: Decomposing motion and content for video generation. pages 1526–1535, 2018.

[17] Guha Balakrishnan, Amy Zhao, Adrian V. Dalca, Fr´edo Durand, and John Guttag. Synthesizing images of humans in unseen poses. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018.

[18] Soumya Tripathy, Juho Kannala, and Esa Rahtu. Icface: Interpretable and controllable face reenactment using gans. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2020.

[19] Wayne Wu, Yunxuan Zhang, Cheng Li, Chen Qian, and Chen Change Loy. Reenactgan: Learning to reenact faces via boundary transfer. In Proceedings of the European Conference on Computer Vision (ECCV), 2018.