International Journal of Recent Development in Engineering and Technology
Website: www.ijrdet.com (ISSN 2347 - 6435 (Online))Volume 4, Issue 6, June 2015)

Deepfake Image Detection using Deep Learning Technique: A Review

Yashkeerti Baderiya[1], Prof. Adarsh Raushan[2], Dr. Sadhna K Mishra[3]
[1]*M. Tech. Scholar, Department of Computer Science and Engineering, LNCT, Bhopal*
[2]*Assistant Professor, Department of Computer Science and Engineering, LNCT, Bhopal*
[31]*Head of Dept., Department of Computer Science and Engineering, LNCT, Bhopal*

Abstract: - Deepfake is an advanced synthetic media technology that can generate deceptively authentic yet forged images and videos by modifying a person's likeliness. The term "Deepfake" is a portmanteau of "Deep learning" and "Fake," which reflects the utilization of artificial intelligence and deep learning algorithms in creating deepfake. The deepfake generation involved training to learn the nuances of facial attributes, facial expressions, motion movement, and speech patterns to produce fabricated media that are indistinguishable from the actual footage. Deepfake is often used to manipulate human content, especially the invariant facial regions. The spatial relationship between the facial attributes is vital for generating a convincing hyper-realistic deepfake output. The subtle inconsistency between face features, such as eye spacing, skin color, and mouth shape, could be used as a telltale sign of deepfake disrimination. Although, a lot of techniques has been invented to detect deepfake but not all of them works perfectly and accurately for all cases, also, as more up to date deepfake creation strategies are grown, ineffectively generalizing methodologies should be continually refreshed to cover these new techniques. In this paper the study of different deepfake image detection paper and deep learning technique.

Keywords: - Deep Learning, Deepfake, Image Detection

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.

International Journal of Recent Development in Engineering and Technology
Website: www.ijrdet.com (ISSN 2347 - 6435 (Online))Volume 4, Issue 6, June 2015)

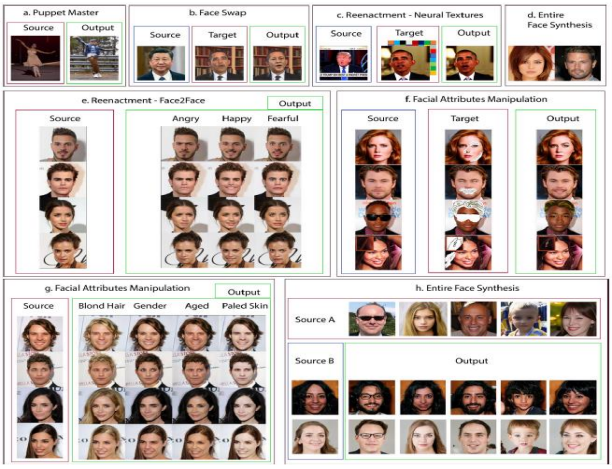

Figure 1: Examples of Deepfake

II. LITERATURE REVIEW

Ali Raza et al. [1], 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. [2], 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. [3], 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. [4], 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.

Brian Dolhansky et al. [5], 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. [6], 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.

Md Shohel Rana et al. [7], throughout recent many years, quick advancement in man-made intelligence, AI, and profound learning has brought about new methods and different apparatuses for controlling mixed media. However the innovation has been for the most part utilized in real applications, for example, for diversion and schooling, and so on., noxious clients have likewise taken advantage of them for unlawful or detestable purposes. For instance, great and practical phony recordings, pictures, or sounds have been made to spread falsehood and misleading publicity, instigate political strife and disdain, or even bother and extortion individuals. The controlled, great and practical recordings have become referred to as of late as Deepfake. Different methodologies have since been depicted in the writing to manage the issues raised by Deepfake. To give a refreshed outline of the exploration works in Deepfake recognition, we lead a methodical writing survey (SLR) in this paper, summing up 112 important articles from 2018 to 2020 that introduced different systems. We break down them by gathering them into four distinct classifications: profound learning-based procedures, old style AI based strategies, factual methods, and blockchain-based methods. We additionally assess the presentation of the location ability of the different techniques regarding different datasets and reason that the profound learning-based strategies outflank different strategies in Deepfake identification.

Jin et al. [8], the connection among face and illness has been talked about from a long time back, which prompts the event of facial conclusion. The goal here is to investigate the chance of distinguishing sicknesses from uncontrolled 2D face pictures by profound learning strategies. In this paper, we propose utilizing profound exchange gaining from face acknowledgment to play out the PC supported facial conclusion on different illnesses. In the trials, we play out the PC helped facial conclusion on single (beta-thalassemia) and different illnesses (beta-thalassemia, hyperthyroidism, Down disorder, and disease) with a somewhat little dataset. The general top-1 exactness by profound exchange gaining from face acknowledgment can arrive at more than 90% which beats the presentation of both conventional AI techniques and clinicians in the examinations. In functional, gathering sickness explicit face pictures is mind boggling, costly and tedious, and forces moral restrictions because of individual information therapy. Subsequently, the datasets of facial finding related investigates are private and for the most part little contrasting and the ones of

other AI application regions. The outcome of profound exchange learning applications in the facial conclusion with a little dataset could give a minimal expense and harmless way for sickness screening and location.

Lewis et al. [9], verification of advanced media has turned into a steadily squeezing need for present day culture. Starting from the presentation of Generative Antagonistic Organizations (GANs), engineered media has become progressively challenging to distinguish. Engineered recordings that contain modified faces or potentially voices of an individual are known as deepfakes and undermine trust and security in computerized media. Profound fakes can be weaponized for political benefit, defame, and to sabotage the standing of well known individuals. In spite of flaws of deepfakes, individuals battle to recognize legitimate and controlled pictures and recordings. Thusly, it is vital to have computerized frameworks that precisely and productively arrange the legitimacy of advanced content. Numerous new deepfake discovery strategies utilize single casings of video and spotlight on the spatial data in the picture to gather the validness of the video. A few promising methodologies exploit the transient irregularities of controlled recordings; be that as it may, research essentially centers around spatial highlights. We propose a crossover profound learning approach that utilizes spatial, phantom, and worldly substance that is coupled in a predictable manner to separate genuine and counterfeit recordings. We show that the Discrete Cosine change can improve deepfake identification by catching otherworldly elements of individual casings. In this work, we fabricate a multimodal network that investigates new highlights to recognize deepfake recordings, accomplishing 61.95% exactness on the Facebook Deepfake Recognition Challenge (DFDC) dataset.

Lee et al. [10], despite the fact that entrance control in light of human face acknowledgment has become famous in shopper applications, it actually has a few execution issues before it can understand an independent access control framework. Inferable from an absence of computational assets, lightweight and computationally productive face acknowledgment calculations are required. The regular access control frameworks require critical dynamic collaboration from the clients notwithstanding its non-forceful nature. The lighting/enlightenment change is one of the most troublesome and testing issues for human-face-acknowledgment based admittance control applications. This paper presents the plan and execution of an easy to use, independent access control framework in view of human face acknowledgment a good ways off. The nearby double example (LBP)- AdaBoost structure was utilized for face and eyes discovery, which is quick and invariant to enlightenment changes. It can recognize faces and eyes of shifted sizes a ways off. For quick face acknowledgment with a high precision, the Gabor-LBP histogram system was changed by subbing the Gabor wavelet with Gaussian subordinate channels, which decreased the facial component size by 40% of the Gabor-LBP-based facial elements, and was strong to critical enlightenment changes and confounded foundations. The trials on benchmark datasets created face acknowledgment correctnesses of 97.27% on an E-face dataset and 99.06% on a XM2VTS dataset, separately. The framework accomplished a 91.5% genuine acknowledgment rate with a 0.28% bogus acknowledgment rate and found the middle value of a 5.26 casings/sec handling speed on a recently gathered face picture and video dataset in an indoor office climate.

Problem Formulation

Watching viral videos of Texas Senator Ted Cruz with his face swapped for that of actor Paul Rudd, or actress Jennifer Lawrence answering questions at the Golden Globes — but with the face of actor Steve Buscemi — it might seem like politics and Hollywood should be the focus areas for combatting misleading videos, but as Deeptrace's report showed, targets for manipulation are no longer limited to government leaders or famous actresses. It doesn't have to be a politician to be a deepfake. It even might be your friend. It could be you that's targeted. "It doesn't have to be a politician to be a deepfake," Panetta said in agreement. "It even might be your friend. It could be you that's targeted. It doesn't have to be someone who's famous."

For example, with scheduled, public quarterly earnings calls that are recorded, it could be possible to take a CFO's voice recording and turn it into what sounds like an urgent directive to employees to share their bank information. Or imagine a similar recording but this time a CEO announces companywide layoffs; the market responds and stocks crash, all because of a deepfake.

"I'm not trying to sow paranoia here but we're trying to sort of be realistic about what could happen," Burgund said. "No doubt there are people working on ways to figure out how to obfuscate in certain ways ... it's an arms race."

Ajder said a big risk right now is defamation. Deepfake videos don't even have to be that good, as long as the person is recognizable and the graphics are good enough for a viewer to identify the person and see they're doing or saying something. That leaves an imprint, Ajder said, and can hurt someone's reputation especially if their name and face is part of negative video or audio real or a deepfake.

International Journal of Recent Development in Engineering and Technology
Website: www.ijrdet.com (ISSN 2347 - 6435 (Online))Volume 4, Issue 6, June 2015)

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. DEEP LEARNING

Deep learning is a subset of machine learning techniques focused on classification tasks and evolutionary algorithms [14]. There are three types of learning: supervised learning, semi-supervised and unsupervised. Deep-learning architectures incorporating deep learning models, fully connected networks, recurrent neural networks, and artificial neural networks were used in fields involving machine learning, artificial intelligence, computer vision, data analysis, realized, social media site filtering, computational linguistics, computational biology, drug design, information retrieval, and clear overview, among others [15]. Knowledge acquisition and decentralized organizational infrastructure in biological systems influenced artificial neural networks (ANNs). ANNs vary from the human brain in several ways. In particular, neural networks are constant and symbolic, whereas most functioning entities' biological brains are dynamic and analog.

Deep learning gets its name from the fact that it employs many layers in the network. Early research demonstrated that a linear perceptron cannot be used as a universal classifier but that a network with a non-polynomial input layer and one unrestrained width hidden layer may. Deep learning is a more recent variant involving many layers of bounded size, allowing for functional application and optimization while maintaining theoretical subjectivity under mild conditions. For execution, teachability, and comprehensibility, profound learning structures are additionally permitted to be assorted and drift away generally from experimentally educated connectionist models, consequently the "coordinated" segment [16].

Most of new profound learning procedures center around AI, particularly convolutional brain organizations (CNNs). They may likewise incorporate propositional equations or dormant factors organized layer-wise in profound generative models like profound conviction organizations and profound Boltzmann machines. Each degree of profound learning figures out how to transform the information it gets into a somewhat more unique and composite portrayal. The crude contribution to a picture acknowledgment program could be a grid of pixels; the principal delegate layer could extract the pixels and encode edges; the subsequent layer could make and encode edge plans; the third layer could encode a nose and eyes, and the fourth layer could perceive that the picture contains a face. Importantly, a deep learning algorithm may figure out which features belong to which level on its own.

The term "deep learning" refers to the number of layers that the data is transformed through. Deep learning systems, in particular, have a significant credit assignment path (CAP) depth [5]. The CAP is the input-to-output transition chain. CAPs are used to define possible causal relationships between input and output. The depth of the CAPs in a feedforward neural network is equal to the network's depth plus the number of hidden layers plus one. The CAP depth in recurrent neural networks, where a signal can propagate through a layer multiple times, is theoretically unlimited. Although no generally agreed-upon depth level separates shallow and deep learning, most researchers agree that deep learning needs a CAP depth greater than 2. In the sense that it can imitate any function, CAP of depth two is a universal approximate [7]. More layers, on the other hand, do not improve the network's ability to approximate functions. Extra layers aid in learning the features effectively because deep models can extract better features than shallow models. Deep convolutional layers can construct deep learning architectures in CNN. The DL can aid in the deconstruction of these abstractions and the identification of which features improve results. Deep learning methods eliminate feature engineering for supervised learning tasks by converting data into compact feature vectors analogous to factor loading and generating layered structures that reduce redundancy. Unsupervised learning tasks may benefit from deep learning algorithms. This is a significant advantage since unlabeled data is more plentiful than labeled data. ANN and deep belief networks are the two basic neural network that works like unsupervised learning approach. There are a few different types of deep learning algorithms, which are mentioned below [8].

V. CONCLUSION

Deepfake detection techniques will never be perfect. Accordingly, in the deepfakes weapons contest, even the best discovery techniques will frequently linger behind the most exceptional creation strategies. They use progressed man-made intelligence calculations to dissect and recognize deepfakes with great precision. Another test is that mechanical arrangements will have no effect when they aren't utilized. Given the disseminated idea of the contemporary biological system for sharing substance on the web, some deepfakes will definitely contact their target group without going through location programming.

REFERENCES

[1] Ali Raza, Kashif Munir and Mubarak Almutairi, "A Novel Deep Learning Approach for Deepfake Image Detection", Apply Science, pp. 01-15, 2022.
[2] 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.
[3] 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.
[4] 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.
[5] 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.
[6] 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.
[7] Md Shohel Rana, Mohammad Nur Nobi, Beddhu Murali, and Andrew H Sung. Deepfake detection: A systematic literature review. IEEE access, 10:25494–25513, 2022.
[8] Jin, B.; Cruz, L.; Gonçalves, N. Deep facial diagnosis: Deep transfer learning from face recognition to facial diagnosis. IEEE Access 2020, 8, 123649–123661.
[9] 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.
[10] 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.
[11] 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.
[12] Sergey Tulyakov, Ming-Yu Liu, Xiaodong Yang, and Jan Kautz. Mocogan: Decomposing motion and content for video generation. pages 1526–1535, 2018.
[13] 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.
[14] 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.
[15] 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.