A Deep Learning-Based Approach for Effective DeepFake Video Detection

Nishant Narjinary IIT2022060
Uttkarsh Malviya IIT2022061
Priyanshi Khataniya IIT2022064
Vaibhav Simha IIT2022067
Boinapally Rishika IIT2022113

***Abstract*—The rising density of deepfakes in online media is extremely alarming, unleashing a pandemic of misinformation and identity theft. Fuelling this plague is the rapid advancement of deepfake generative models based on GANs and Transformers that produce hyper-realistic deepfakes, indistinguishable by most people. Videos, being the most preferred medium for information, have become a prime target for these technologies. This underscores the need for effective deepfake detection techniques.**

However, most existing techniques lack generalizability, leading to a steep decline in accuracy when used on unseen datasets. Therefore, it is crucial to develop an effective yet generalizable deepfake detection technique. In our work, we present a robust deep-learning CNN-based architecture that utilizes the power of $ResNext50\_32x4d$ as a feature extractor for an LSTM unit over 40 epochs, trained on the massive DFDC dataset, achieving a respectable accuracy of 91.25

I. Introduction

With the emergence of potent deep learning-based generative approaches, there has been a conspicuous influx of manipulated images and videos which are widely circulated, undetected by most observers. It is especially alarming to note that often these generative models have been used for malicious purposes, ranging from defamation for instance, through the production of indecent representations of the targets to dissemination of misinformation for various causes that include leveraging of the slanderous mayhem for political advantage. These highly realistic and convincing manipulated media are referred to as DeepFakes and are usually in the form of images, audios and videos. In re- cent years, there were also malicious attacks involving this technology to tam per with medical images, particularly CT-scans. Considering the burgeoning number of people accessing the internet since the COVID-19 pandemic and more so of those reliant on it for their general awareness, development of ac- curate and efficient detection techniques is pivotal and the need of the hour. DeepFake imagery involving facial manipulations are primarily produced us- ing face-swapping wherein the face of the source identity is replaced with that of another identity. The popularity of videos as a medium of relaying information makes them the primary target of malicious manipulation. Effective DeepFake video detection is therefore indispensable to tackle misinformation and thwart the pernicious interests of the perpetrators. Given the widespread availability and sophistication of such tools, it is clear that DeepFake detection mechanisms must evolve at a comparable pace. Ad- vanced detection techniques leveraging deep learning, computer vision, and forensic analysis are critical to identifying these manipulations accurately. Additionally, efforts must focus on building robust, real-time detection systems capable of handling high volumes of data, particularly in videos, which remain the most popular medium for information dissemination. Collaboration among researchers, policymakers, and technology platforms is imperative to establish standardized detection frameworks, raise public awareness, and implement countermeasures to combat the growing threat of DeepFakes.

II. Literature Review

Deep Learning methods have advanced and improved the technology that is used for generating and processing multimedia content [1], especially that of videos and pictures. DeepFakes, a relatively new face of fakery, have come up to allow us to generate videos of highly realistic disposition in which the faces of the people are swapped and modified to alter the identity of individuals so as to depict them as doing activities that they had never actually done. In some cases, only the lips and the movement of the eyes are modified to achieve the effect. This effect, in most cases, is of malicious and compromising nature,

underscoring the importance of much-needed improvements in DeepFake de- tection methods to counter this new vessel of evil. Such content is especially

devastating for public figures such as influencers, politicians, leaders and busi- nessmen [1].

Further exacerbating the situation is the proactive role of social media in peddling this kind of misinformation, allowing it global reach and impact. Novel and powerful synthesized video generation technologies have come up, namely Face2Face [2], Deep video Portraits [3], and StarGAN [4]. On the other front, significant research has been done on distinguishing real media

from synthetic by analysing various parameters such as possible inconsisten- cies within the RGB frames in the video [5, 6, 7]. In many cases, popular pre- trained CNN models are directly used to perceive characteristic artefacts from

each frame of a video. In [8], media subjected to the manipulation of the face is detected through a recurrent convolutional approach in which a set of frames

is dealt with as an ensemble. Different considerations were taken such as phys- ical traits as in the case of [9] in which researchers accentuated the role of the

blinking of the eye as an artefact that can used to discern fakes from real videos. Similarly, in [10], facial expressions have been used as a basis to perform the same classification of fake and real videos. A relatively new approach of its time was proposed in [1] with its reliance on the sequence and not merely the

frames, bringing in a temporal dimension to the discernment process, espe- cially when it comes to dissimilarities observed in a video with respect to time.

Therein, optical flow fields have been gleaned to make use of the correlation between frames as input of the CNN classifiers. Currently, there are two major categories of DeepFake detection models: image-level (frame-level) and video-level detectors. Almost all the detectors

of the former kind make use of spatial artifacts such as inconsistencies in tex- ture (Zhao et. al, 2021), colour distortion (Kumar et al., 2020). Dong et al. relied

on spatial identity inconsistencies between the inner face region and the outer

face region, but this proved to be not so effective for DeepFake video detec- tion (Liu et al., 2024). DeepFake detection is, in general, pictured as a binary

arrangement problem wherein movies are to be told apart based on whether they are reliable or interfering (Singh et al., 2021). For this, there is a need for an enormous volume of real and fake videos to train these models. Previous research on detection was primarily based on two methods: CNN and RCNN (Region-based CNN).In CNN-based architectures, pictures of faces cropped from the frames of videos are fed into a CNN for training and classification on the level of an

image. These techniques, however, take into consideration spatial informa- tion from a particular frame. The models using RCNN, meanwhile, need to be

trained on a series of video frames so as to churn out a video-level result. RCNN combines both the architectures of CNN and RNN (Tariq et al., 2021). This allows RCNN-based models to exploit temporal and spatial information

to detect deepfakes (Chung et al., 2015). Other deepfake detection architec- tures rely on Machine Learning techniques such as the Support Vector Machine

for classification and for extracting characteristics (Yazdinejad et al., 2020). In Y. Mirsky et al, 2019, a GAN-based model called CT-GAN was introduced to tamper with CT-scans of lungs, inserting and removing cancers. MedNet (S. Albahli et al., 2023) is a modified Efficient-Net architecture with an added attention module to detect tampered CT-scans. The applicability of general deepfake video detection models on 3D-images is still not well-explored. Of late, Transformers have also been employed in detection architectures. Junke Wang et al. (2022) present Multi-modal Multi-scale Transformers for deepfake image detection that works on patches of various sizes, looking out for local inconsistencies in images at different spatial levels. Coming to models that use temporal identity features, LRNet (Zekun Sun et al., 2021) utilises these features from videos for detection. It has a calibration module for refining the extracted facial landmarks which are then run through a two-branch RNN for analysing temporal patterns of identity landmarks. This, however, has poor generalization ability. ID-Reveal (Jiankang Deng et al., 2019) makes use of 3DMM to capture temporal identity features whose pattern is then analysed by a temporal ID network through adversarial training on reference videos. Identity Inconsistency Transformer (ICT) (Xiaoyi Dong et al., 2022) uses spatial facial identities inconsistency of the inner and outer facial areas in face-swapped images.

III. Methodologies

A. Dataset

For the purposes of our project, we have chosen the DeepFake Detection Challenge (DFDC) dataset, a extensive dataset of labelled real and deepfake videos that have been generated using a variety of techniques including GAN-based and non-learned methods. Firstly, the videos are all of actors, numbering 3,426, who agreed to be a part of the data collection process. Deepfake Autoencoder (DFAE) [19], MM/NN Face Swap [20], NTH (Neutral Talking Head) model [21], FSGAN [22] and StyleGAN [23] were used to create deepfakes of the real source data. In all, there are 128,154 10-second video clips featured 960 unique subjects. Of these, 104,500 are synthetically generated videos which have been created using 8 different deepfake methods. For the implementation of our approach, we considered a smaller segment of 400 videos in which 323 are deepfakes and 77 are real.

Table 1: Quantitative comparison of various Deepfake datasets

Dataset	Unique fake videos	Total videos	Unclear rights	Agreeing subjects[a]	Total subjects	Methods	No. perturb.	No. benchmarks[b]
DF-TIMIT [17]	640	960	×	0	43	2	-	4
UADFV [30]	49	98	×	0	49	1	-	6
FF++ DF [23]	4,000	5,000	×	0	?	4	2	19
Google DFD [6]	3,000	3,000	✓	28	28	5	-	-
Celeb-DF [18]	5,639	6,229	×	0	59	1	-	-
DeeperForensics-1.0 [14]	1,000	60,000	×	100	100	1	7[c]	5
DFDC Preview[5]	5,244	5,244	✓	66	66	2	3	3
DFDC	**104,500**	**128,154**	✓	**960**	**960**	**8**[d]	**19**	**2,116**

B. Data Pre-Processing

Each video is split into frames and the first 150 frames are considered. Then these frames are run through a process of face detection, followed by each of them being cropped to the detected facial landmarks. For this, the frames are processed in batches of size 2 to prevent overloading of the memory. The cropped images are then resized to standard dimensions of 112 x 112 pixels. The face recognition is performed using the face recognition library whereas cv2 is used for frame extraction and resizing.

C. Model Architecture

The pre-processed data, specifically the cropped frames, are fed into a pretrained ResNext50 model of which the last two layers are removed so as to only retain the feature extractor. The cropped frames, each of 112 x 112 pixels, undergo feature extraction with the output feature map having the latent dimension of 2048. This is then passed through an LSTM unit whose function is to discern the temporality of the data and produce temporal embeddings which are to be averaged

out. These averaged-out embeddings are then run through a fully connected layer for the final out of classification as to whether the video is fake or real. The loss criterion considered is Cross Entropy Loss. The formula for general cross entropy is as follows:

$$L = -\frac{1}{N}\sum_{i=1}^{N}\sum_{c=1}^{C} y_{i,c}\log(\hat{y}_{i,c})$$

For Binary Cross Entropy, the calculation is done as follows:

$$L = -\frac{1}{N}\sum_{i=1}^{N}\left[y_i \log(\hat{y}_i) + (1-y_i)\log(1-\hat{y}_i)\right]$$

Where:

- N: Number of samples.
- C: Number of classes.
- $y_{i,c}$: True label for sample i and class c (one-hot encoded).
- $\hat{y}_{i,c}$: Predicted probability for class c of sample i (softmax output).
- $\log$: Natural logarithm.

The Long Short-Term Memory (LSTM) model is a specialized recurrent neural network designed to capture long-term dependencies in sequential data, overcoming the vanishing gradient problem of traditional RNNs. Its key innovation lies in the use of gatesforget, input, and outputthat regulate the flow of information through a memory cell. These gates selectively decide what to retain, update, or output, enabling the model to effectively manage temporal relationships in

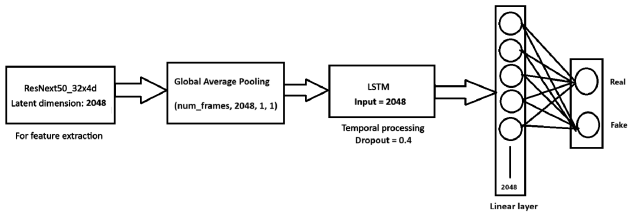

Fig. 1. Architecture of the ResNext50_32x4d-LSTM model

complex sequences. The cell state acts as a persistent memory, while the hidden state encodes information for the next time step or layer, making LSTMs particularly suited for tasks like time-series forecasting, text generation, and speech recognition. In our research, we leverage the LSTM's ability to process sequential data efficiently, enabling the model to capture dependencies that span across long time frames. This is critical for our dataset, which involves analyzing patterns where historical context significantly influences predictions. By incorporating techniques like stacked LSTMs for learning hierarchical patterns and bidirectional LSTMs for capturing dependencies in both forward and backward directions, we aim to enhance the model's ability to generalize and deliver accurate results.

stage	output	ResNet-50	**ResNeXt-50 (32×4d)**
conv1	112×112	7×7, 64, stride 2	7×7, 64, stride 2
conv2	56×56	3×3 max pool, stride 2 [1×1, 64 3×3, 64 1×1, 256] ×3	3×3 max pool, stride 2 [1×1, 128 3×3, 128, C=32 1×1, 256] ×3
conv3	28×28	[1×1, 128 3×3, 128 1×1, 512] ×4	[1×1, 256 3×3, 256, C=32 1×1, 512] ×4
conv4	14×14	[1×1, 256 3×3, 256 1×1, 1024] ×6	[1×1, 512 3×3, 512, C=32 1×1, 1024] ×6
conv5	7×7	[1×1, 512 3×3, 512 1×1, 2048] ×3	[1×1, 1024 3×3, 1024, C=32 1×1, 2048] ×3
	1×1	global average pool 1000-d fc, softmax	global average pool 1000-d fc, softmax
# params.		**25.5**×10^6	**25.0**×10^6
FLOPs		**4.1**×10^9	**4.2**×10^9

Fig. 2. ResNext-50 architectures

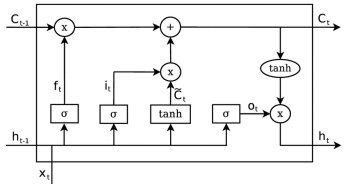

Fig. 3. LSTM Architecture

IV. Result Disscussion

A. Results

The model achieved an accuracy of 91.25is crucial to contextualize it within the practical requirements of deepfake detection, where high accuracy is essential for real-world deployment due to the potential consequences of misclassifications. The precision of 95.38misclassifies real content as fake. This attribute is particularly valuable in scenarios where erroneously labeling real content as fake could lead to distrust or misinformation. The recall of 93.94Together, these metrics paint a favorable picture of the model's balance between identifying fakes and avoiding unnecessary false alarms. The confusion matrix exhibits a reasonable diagonal relationship, indicating that the majority of predictions fall into the correct categories. This confirms that the model maintains a good trade-off between precision and recall, without heavily favoring one at the expense of the other. The training and validation losses exhibit convergence over the 30 epochs, with the training loss stabilizing at 0.22 and the validation loss at 0.31 by the final epoch. This steady decline and eventual narrowing of the loss gap suggest that the model is well-trained and not significantly overfitting, as the validation loss does not deviate considerably from the training loss. The slight difference in these values is expected and indicates that the model generalizes effectively to unseen data. The composition of the dataset has played a great role in making the model generalisable, owing to the great variety of deepfake generation techniques featured, even though this has rendered the dataset quite imbalanced.

B. Performance Metrics

For evaluating the performance of the model, we used the metrics of Accuracy, Precision, Recall and F1-score which we shall define in this section. Accuracy measures the ratio of correctly predicted samples to the total samples. Precision,

or also known as the positive predictive value, measures the proportion of correctly predicted positive samples out of all samples predicted as positive. It focuses on the quality of positive predictions. Recall measures the proportion of correctly predicted positive samples out of all actual positive samples. It focuses on capturing as many positives as possible. Finally, F1-Score is the harmonic mean of precision and recall. It balances the trade-off between the two metrics. Refer to the formulae below.

$$\text{Accuracy} = \frac{\text{Number of Correct Predictions}}{\text{Total Number of Predictions}} \tag{1}$$

$$\text{Precision} = \frac{\text{True Positives (TP)}}{\text{True Positives (TP)} + \text{False Positives (FP)}} \tag{2}$$

$$\text{Recall} = \frac{\text{True Positives (TP)}}{\text{True Positives (TP)} + \text{False Negatives (FN)}} \tag{3}$$

$$\text{F1-Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{4}$$

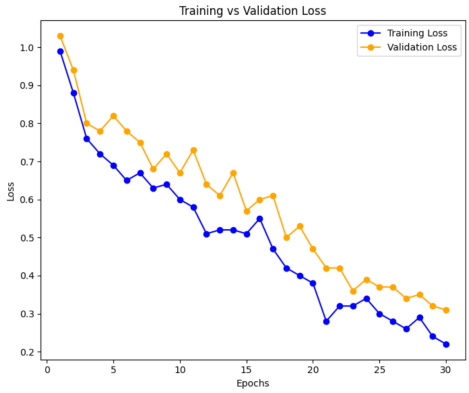

Fig. 4. Plot of Training Loss vs Validation Loss

C. Challenges

Data imbalance, with its being greatly askew towards fake data, made it difficult for models to minimise bias and improve accuracy. However, the inherent nature of the dataset being rich with deepfakes produced by 8 different kinds of deepfake techniques allowed it to generalise well, as per observational analysis.

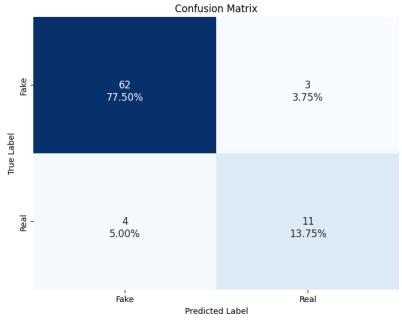

Fig. 5. Confusion Matrix

V. Future Scope and Conclusion

A. Future Scope and Areas of Improvement

While the proposed approach has demonstrated high accuracy and generalizability in detecting human-centric DeepFake videos, there remain several avenues for enhancement and exploration. One significant area of improvement involves evaluating the model's performance on DeepFake videos of nonhuman entities, such as animals. Manipulations involving non-human subjects pose unique challenges, as artifacts and temporal inconsistencies may manifest differently. A thorough analysis of these videos will provide insights into the model's adaptability and identify areas requiring refinement to broaden its applicability. Another promising direction lies in expanding the scope of detection to include modalities beyond traditional video data. The comparative analysis of detection methods for both videos and 3D imagery, such as CT scans, remains underexplored. Medical imagery, often subjected to tampering, represents a critical application area for DeepFake detection. Modifying the architecture to handle 3D datasets and comparing its performance against existing methods will offer valuable insights into its robustness and flexibility. To push the boundaries of detection capabilities, incorporating advanced techniques such as Generative Adversarial Networks (GANs), Transformers, and Diffusion-based models presents an exciting opportunity. These models, known for their ability to capture intricate patterns and relationships, can enhance the detection of subtle inconsistencies across spatial and temporal dimensions. Additionally, exploring alternative sequential frameworks beyond LSTMs, such as GRUs or attention-based mechanisms, could improve temporal modeling while potentially reducing computational demands. The ultimate objective remains the creation of a well-rounded, generalizable DeepFake detection system that can accurately distinguish not only manipulated videos but also tampered audio. Integrating audio analysis within the framework will address the multimodal nature of many DeepFakes, further improving detection reliability. Simultaneously, efforts to develop computationally efficient models will be critical for real-world deployment, particularly in resource-constrained environments.

B. Conclusion

In this study, a robust framework for DeepFake detection has been developed, achieving a commendable accuracy of 91.25 The simplicity of the architecture is another notable achievement. By combining the feature extraction prowess of a pretrained ResNext50 model with the sequential learning capabilities of an LSTM, the framework effectively balances computational efficiency and detection accuracy. ResNext50, known for its lightweight and efficient design, extracts rich spatial features without excessive computational overhead, while the LSTM module adeptly captures temporal dynamics across frames. This fusion allows the model to exploit

both spatial and temporal inconsistencies inherent in DeepFake videos, offering a more comprehensive analysis than frame-by-frame methods. Temporal analysis, in particular, has proven indispensable in this work. Unlike approaches focused solely on spatial artifacts, which may vary across different DeepFake generation methods, temporal inconsistencies are more universal and intuitive indicators of manipulation. By examining the sequential relationships across frames, the model effectively identifies anomalies that persist over time, significantly enhancing its detection capabilities. Moreover, the approach prioritizes practicality. The preprocessing pipeline, which includes frame extraction, face detection, cropping, and resizing, is streamlined for efficiency without compromising on the quality of input data. The choice of Cross-Entropy Loss as the criterion further supports a balanced performance between precision and recall, ensuring the model minimizes both false positives and negatives. In conclusion, this work not only demonstrates high accuracy in DeepFake detection but also highlights the importance of dataset diversity, simplicity in architecture, and the critical role of temporal analysis. The results suggest that such an approach can be a valuable tool for real-world DeepFake detection, paving the way for further enhancements in scalability and deployment

VI. References

1) Amirini et al., 2019; Deepfake Video Detection through Optical Flow based CNN
2) J. Thies, M. Zollhofer, M. Stamminger, C. Theobalt, and M. Niessner. Demo of face2face: Real-time face capture and reenactment of RGB videos. In *ACM SIGGRAPH 2016 Emerging Technologies*, SIGGRAPH 16, pages 5:15:2, New York, NY, USA, 2016. ACM.
3) H. Kim, P. Garrido, A. Tewari, W. Xu, J. Thies, M. Niessner, P. Pterez, C. Richardt, M. Zollhlofer, and C. Theobalt. Deep video portraits. *ACM Trans. Graph.*, 37(4):163:1–163:14, July 2018.
4) Y. Choi, M. Choi, M. Kim, J. Ha, S. Kim, and J. Choo. StarGAN: Unified generative adversarial networks for multi domain image-to-image translation. *CoRR*, abs/1711.09020, 2017.
5) A. Rlossler, D. Cozzolino, L. Verdoliva, C. Riess, J. Thies, and M. Niessner. Faceforensics: A large-scale video dataset for forgery detection in human faces. *CoRR*, abs/1803.09179, 2018.
6) A. Rlossler, D. Cozzolino, L. Verdoliva, C. Riess, J. Thies, and M. Niessner. Faceforensics++: Learning to detect manipulated facial images. *CoRR*, abs/1901.08971, 2019.
7) D. Afchar, V. Nozick, J. Yamagishi, and I. Echizen. Mesonet: a compact facial video forgery detection network. pages 17, 12 2018.
8) E. Sabir, J. Cheng, A. Jaiswal, W. AbdAlmageed, I. Masi, and P. Natarajan. Recurrent convolutional strategies for face manipulation detection in videos, 05 2019.
9) Y. Li, M. Chang, and S. Lyu. In ictu oculi: Exposing AI generated fake face videos by detecting eye blinking. *CoRR*, abs/1806.02877, 2018.
10) S. Agarwal, H. Farid, Y. Gu, M. He, K. Nagano, and H. Li. Protecting world leaders against deep fakes. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2019.
11) Tianchen Zhao, Xiang Xu, Mingze Xu, Hui Ding, Yuanjun Xiong, and Wei Xia. Learning self-consistency for deepfake detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15023–15033, 2021.
12) Prabhat Kumar, Mayank Vatsa, and Richa Singh. Detecting face2face facial reenactment in videos. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 2589–2597, 2020.
13) Baoping Liu, Bo Liu, Ming Ding, Tianqing Zhu, Xin Yu. TI2Net: Temporal Identity Inconsistency Network for Deepfake Detection. *Computer Vision and Pattern Recognition*, Computer Vision Foundation, 2023.
14) Yoseph Hailemariam, Abbas Yazdinejad, Reza M Parizi, Gautam Srivastava, and Ali Dehghantanha. An empirical evaluation of AI deep explainable tools. In *IEEE Globecom Workshops (GC Wkshps)*, 2020.
15) Saleh Albahli and Marriam Nawaz. MedNet: Medical deepfakes detection using an improved deep learning approach. *Multimedia Tools and Applications*, 2023.
16) Zekun Sun, Yujie Han, Zeyu Hua, Na Ruan, Weijia Jia. Improving the Efficiency and Robustness of Deepfakes Detection through Precise Geometric Features, 2021.
17) Siying Cui, Jiankang Deng, Jia Guo, Xiang An, Yongle Zhao, Xinyu Wei, Ziyong Feng. IDAdapter: Learning Mixed Features for Tuning-Free Personalization of Text-to-Image Models, 2024.
18) Xiaoyi Dong, Jianmin Bao, Dongdong Chen, Ting Zhang, Weiming Zhang, Nenghai Yu, Dong Chen, Fang Wen, Baining Guo. Protecting Celebrities from DeepFake with Identity Consistency Transformer, 2022.
19) Brian Dolhansky, Joanna Bitton, Ben Pflaum, Jikuo Lu, Russ Howes, Menglin Wang, Cristian Canton Ferrer. The DeepFake Detection Challenge (DFDC) Dataset. *arXiv:2006.07397v4*, 2020.
20) Dong Huang and Fernando de la Torre. Facial action transfer with personalized bilinear regression. In *Proc. of the European Conference on Computer Vision (ECCV)*, SpringerVerlag, 2012.
21) Egor Zakharov, Aliaksandra Shysheya, Egor Burkov, and Victor Lempitsky. Few-shot adversarial learning of realistic neural talking head models. In *Proc. of the IEEE International Conference on Computer Vision (ICCV)*, 2019.
22) Yuval Nirkin, Yosi Keller, and Tal Hassner. FSGAN: Subject agnostic face swapping and reenactment. In *Proc. of the IEEE International Conference on Computer Vision (ICCV)*, 2019.
23) Tero Karras, Samuli Laine, and Timo Aila. A style-

based generator architecture for generative adversarial networks. *Proc. of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.