DLDFD: Recurrence Free 2D Convolution Approach for Deep Fake Detection

Jag Mohan Singh and Raghavendra Ramachandra
Norwegian Biometrics Laboratory, Norwegian University of Science and Technology, Gjøvik, Norway

Keywords: DeepFakes Detection, Resnet, Deep Learning Architecture.

Abstract: Deep Fake images, which are digitally generated either through computer graphics or deep learning techniques, pose an increasing risk to existing face recognition systems. This paper presents a Deep-Learning-based Deep Fake Detection (DLDFD) architecture consisting of augmented convolutional layers followed by Resnet-50 architecture. We train DLDFD end-to-end with low-resolution images from the FaceForensics++ dataset. The number of images used during different phases includes approximately 1.68 million during training, 315k during validation, and 340k during testing. We train DLDFD in three different scenarios, combined image manipulation where we achieve an accuracy of 96.07% compared to 85.14% of state-of-the-art (SOTA), single image manipulation techniques where we get 100% accuracy for neural textures, and finally, cross-image manipulation techniques where we achieve an accuracy of 94.28% on the unseen category of face swap much higher than SOTA. Our approach requires only 2D convolutions without recurrence as compared to SOTA.

1 INTRODUCTION

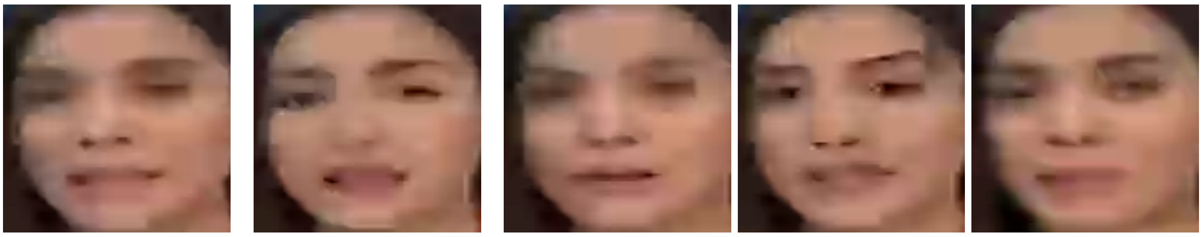
Researchers have achieved ever-increasing photorealism from deep-learning techniques with the advent of computer-generated imagery (CGI). CGI techniques especially face synthesis, a.k.a deepfakes, pose an increasing threat to manual and automatic face recognition systems. To overcome the risks posed by deepfakes, researchers are developing deep-fake detection (DFD) methods. Recently, the digital AVATAR from FaceMe (Fac, 2018) was deployed at Auckland Airport to answer bio-security questions (Auc, 2018). The main effort in producing photo-real digital humans currently lies in face synthesis. The amount of realism in face-synthesis is increasing over time. The initial techniques for a deep-fake generation were based on computer graphics posing it as a forward face synthesis problem such as the one done by authors in (NVi, 2011) where they produced a life-like rendering of a human head requiring an expensive setup. Hu et al. (Hu et al., 2017), and Yamaguchi et al. (Yamaguchi et al., 2018) have approached face-synthesis as a forward-technique using an expensive setup called Light-Stage (Deb, 2000).

With the advent of deep learning, the requirement for an expensive setup for face-synthesis has reduced significantly. Karras et al. (Karras et al., 2018) did one of the initial works in this direction using Generative Adversarial Networks (GANs). The quality of synthesized images was further improved by Juefei-Xu et al. (Juefei-Xu et al., 2018). Karras et al. (Karras et al., 2019) also enhanced the quality of face-synthesis produced even more photo-real digital face. When we say that face quality has improved during synthesis, it needs to be pointed out that it includes hair and freckles of the skin. The results from Karras et al. (Karras et al., 2019) are available on a public website (per,). However, recently the limitation of using an expensive setup for acquiring Light Field for face-synthesis was overcome in work by Sengupta et al. (Sengupta et al., 2021) where it is obtained from a person watching videos on a regular computer.

DeepFakes (dee, 2019), one of the popular techniques for replacing one person's face with another, leverages computer graphics & visualization techniques and has been used for defaming persons. We propose the use of DLDFD architecture based on deep learning, and the contributions of our proposed approach are as follows:

- Our proposed approach is based on a recurrence-free 2D convolution architecture DLDFD, which involves augmented layers followed by Resnet-50 architecture, unlike the previous SOTA based on 3D convolution or recurrence-based 2D convolution.

Singh, J. and Ramachandra, R.
DLDFD: Recurrence Free 2D Convolution Approach for Deep Fake Detection.
DOI: 10.5220/0010880500003124
In *Proceedings of the 17th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications (VISIGRAPP 2022) - Volume 4: VISAPP*, pages 568-574
ISBN: 978-989-758-555-5; ISSN: 2184-4321



(a) Real Face (b) Deep Fakes (c) Face2Face (d) Face Swap (e) Neural Textures

Figure 1: Low-Resolution Image Manipulation from Faceforensics++ dataset (Rössler et al., 2018). Note the low-resolution of real face makes the classification challenging as distinction real v/s fakes becomes difficult.

- We perform an extensive evaluation of FaceForensics++ low-resolution dataset, including combined image-manipulation techniques, single image-manipulation techniques, and cross-evaluation of image-manipulation techniques. It needs to be pointed out that there are few related works for low-resolution evaluation of FaceForensics++ compared to extensive literature for DeepFakes in general.
- In terms of results, we achieve an accuracy of 96.07% compared to 85.14% of SOTA for combined image manipulation. We achieve a 100% accuracy for neural textures in single image manipulation techniques. Finally, we achieved an accuracy of 94.28% for the unseen category of face swap, much higher than SOTA for cross-image manipulation techniques.

In the rest of the paper, we provide the literature review of the critical papers in Deep Fakes Detection in Section 2, followed by the proposed method in Section 3. We describe experimental setup & results in Section 4, and conclude the paper by providing conclusions & future work for Deep Fakes Detection in Section 5.

2 RELATED WORK

In this section, we present the related work in manipulated face images. Digital Face Manipulation can be grouped into expression swap (face reenactment) and identity swap, as mentioned by Rossler et al. (Rössler et al., 2019). Identity Swap techniques replace a genuine user's face with another person, including methods like FaceSwap (Kowalski, 2016) and DeepFakes (dee, 2019). Expression Swap techniques don't change the person's identity but change a real user's expressions by expressions obtained by another person, including Face2Face (Thies et al., 2016) and NeuralTextures (Thies et al., 2019). The manipulations of Face2Face, & FaceSwap are based on computer graphics techniques, whereas Neural Textures, & DeepFakes are based on machine-learning.

We now review the essential works in Deep Fake Detection (DFD). Li et al. (Li et al., 2018) did one of the initial works in DFD where they proposed the use of temporal pattern of eye blinking using a combination of Convolutional Neural Network (CNN) and a long-term recurrent neural network (CNN) on a custom dataset achieving an area under the curve (AUC) of 0.99. Li et al. (Li and Lyu, 2019) overcame the limitation of training pairs of real and fake samples for DFD by using face warping artifacts for generating deepfakes from real images. They used CNN for deepfake detection of public datasets of DeepfakeTIMIT (Korshunov and Marcel, 2018), and UADFV (Li et al., 2018), and achieved an AUC of 97.4% on UADFV, and for DeepfakeTIMIT an AUC of 99.9% for Low Quality, & 93.2% for High Quality. Li et al. (Li et al., 2020) achieved SOTA results for the FaceForensics++ dataset based on the observation that there is a blending boundary in the forged image, which is absent in the real image. They achieved an accuracy of 99% on single image manipulation techniques and 97-98% on cross image manipulation techniques, but their approach requires a mask in addition to the image during training. Jung et al. (Jung et al., 2020) proposed the use of Generative Adversarial Networks (GAN) for detection of significant eye-blinking patterns in a video and have an accuracy of 87.5% around a different type of deepfake videos. Sun et al. (Sun et al., 2021) proposed the use of geometric features for Deep Fake Detection where they first calibrate the geometric features to achieve a more precise location. This is followed by using a two-stream Recurrent Neural Network (RNN) to extract temporal features to achieve an AUC of 99.9% on the Faceforensics++ dataset. However, this method provides a single evaluation and does not cross-evaluate manipulation techniques. A summary of challenges for deepfake detection was mentioned by Lyu et al. (Lyu, 2020) where it is noted that deepfakes datasets have visual artifacts present in them, and performance evaluation of deepfakes algorithms is binary classification & uses fixed techniques, unlike real-world deepfakes. The videos on social media platforms like Facebook and Instagram are usually stripped of metadata and compressed for bandwidth optimization, making the deepfakes classification difficult. A comprehensive survey about DeepFakes is done by Tolosana et al. (Tolosana et al., 2020) where they provide an overview of current SOTA methods for deepfake detection, datasets, and open challenges for the area. Mirsky et al. (Mirsky and Lee, 2021) provided an overview of both creation and detection algorithms for deepfakes.

We now specifically focus on classifying low-resolution manipulated face images, which is a challenging area. In a more recent work, Sabir et al. (Sabir et al., 2019) used recurrent neural network (RCN) for DFD on FaceForensics++ dataset (Rössler et al., 2019), and achieve an accuracy of 96.9% on Deep Fakes, 94.35% on Face2Face, & 96.3% on FaceSwap on low-resolution videos from the dataset. It needs to be pointed out that low-resolution image classification is challenging in general, and for manipulated face images in particular (Wang and Dantcheva, 2020) due to the high compression factor. The high compression factor of low-resolution image-manipulation techniques is shown in Figure 1. We now describe the related work for low-resolution image-manipulation techniques, where the current SOTA is by Wang et al. (Wang and Dantcheva, 2020) where they performed manipulated video classification on FaceForensics++. They used 3d convolution-based CNNs for their proposed approach which included 3D Resnet, & 3D Resnext (Hara et al., 2018), and I3D (Carreira and Zisserman, 2017). Wang et al. evaluated base networks of 3D Resnet, 3D Resnext, and I3D without modification which is a limitation of their technique.

Furthermore, 3D convolution is more memory intensive to train compared to 2d convolution-based deep-learning networks. Liu et al. (Liu et al., 2021) reduced the computation budget of 3D convolution by the use of Spatial rich model (SRM) features. However, they generally evaluated the Faceforensics++ dataset and were not specific to low quality.

3 PROPOSED METHOD

In this section, we present our proposed method where the proposed architecture is chosen to give high accuracy with minimal increase in computational cost over the Resnet-50 architecture. This is achieved by the use of augmented layers followed by Resnet-50 architecture. The design of augmented layers, which includes three 3×3 convolutions, followed by one 1×1 convolution, is inspired from the inception module of the Inception Network (Szegedy et al., 2015). The augmented layers in the proposed network architecture are followed by Resnet-50 (He et al., 2016), and it is chosen due to its high generalization capability as pointed out by He et al. (He et al., 2020). The augmented layers are selected so that the computational cost does not increase significantly, and we don't use the complete inception module. Our proposed approach consists of the different stages as indicated in the block diagram in Figure 2.

3.1 Face Detection & Image Normalization

We process the video dataset (Rössler et al., 2019) to extract frames, followed by face detection using MTCNN (Zhang et al., 2016). The input face during training is passed through the following transformations, which include random horizontal flipping and normalizing the image to mean, & variance of the Imagenet dataset (Krizhevsky et al., 2012). The main idea for this step is to remove background as this can increase the chances for the proposed deep network to misclassify. The image normalization is performed so that real and fakes have normalized data, resulting in better classification.

3.2 Proposed CNN Architecture (DLDFD)

Our architecture consists of taking the input image at a resolution of 224×224, which is followed by three convolution layers ($3 \times 3 \times 64$), which are concatenated and passed through 1×1 Convolution ($1 \times 1 \times 192$). This is followed by the use of Resnet-50 (He et al., 2016) network architecture. The output from Resnet-50 is passed through a fully connected layer to generate labels. The main idea behind the use of augmented layers at the top of the Resnet-50 architecture is to allow the use of sparsity (Szegedy et al., 2015) in the proposed convolution neural network, which results in improved performance with minor gain in the computational cost. Three additional convolution layers are used to minimize the computational cost and achieve improved performance. Further, the choice of 1×1 Convolution (Lin et al., 2013) results in dimensionality reduction and allows for patch-wise discrimination where the latter is important for deepfakes classification.

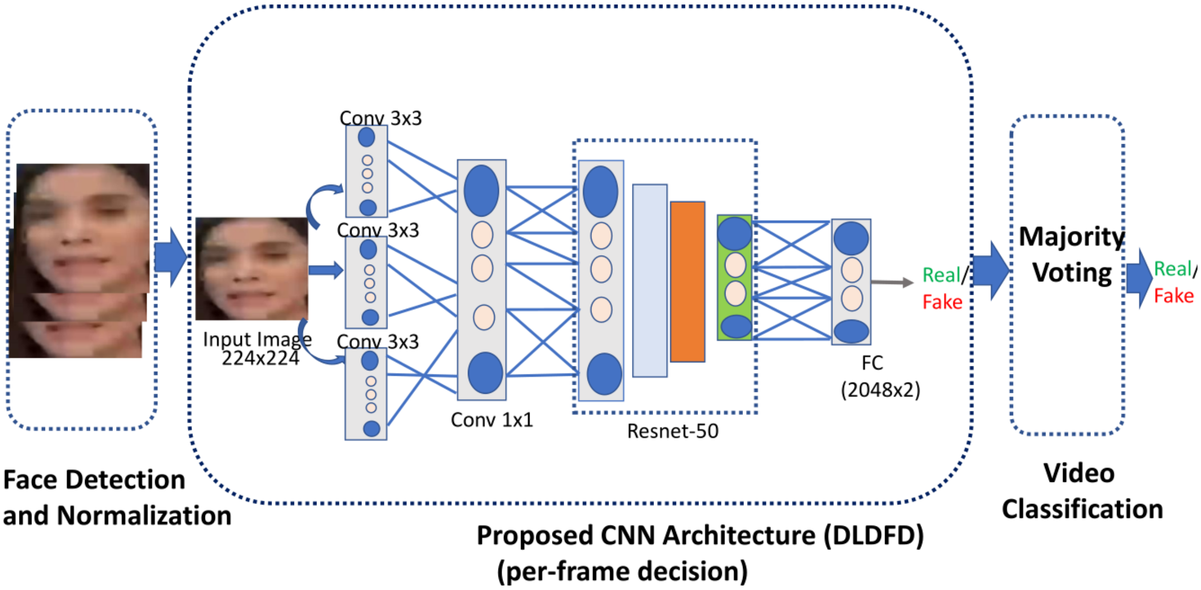


Figure 2: Proposed Architecture for Manipulated Image Classification.

3.3 Video Classification

Once the network is trained, we extract the prediction label from the fully connected layer for each video frame. This is then followed by majority voting to compute the video-level predictions. The loss function used during training is weighted binary cross-entropy, and we use Adam Optimizer with a base learning rate of 0.001, the momentum of 0.9, and 25 epochs. We choose weighted binary cross-entropy as the loss function as it helps in the unbalanced classification as mentioned by Wang et al. (Wang and Dantcheva, 2020). The Resnet-50 architecture weights are initialized by the model trained on Imagenet from He et al. (He et al., 2016).

4 EXPERIMENTAL SETUP & RESULTS

We first describe the public dataset FaceForensics++ (Rössler et al., 2018) used for the comparison of our proposed method with SOTA. FaceForensics++ dataset consists of the face manipulations in video format and has 1000 real videos and four image manipulations, each with 1000 videos. It consists of image manipulations of DeepFakes, FaceSwap, Face2Face, and Neural Textures in high-resolution and low-resolution formats. We perform the experiments on Faceforensics++ Dataset after extracting frames from the videos, and we divide the dataset into training, validation, and test sets. The training set consists of 367228 pristine images, and in manipulated images, we have 291789 from neural textures, 367009 deep fake, 292320 face swap, and 366631 face2face manipulations. The validation set consists of 68857 pristine images, and in manipulated images, we have 54617 from neural textures, 68664 deep fake, 54624 face swap, and 68854 face2face manipulations. The testing set consists of 73768 pristine images, and in manipulated images, we have 59670 from neural textures, 73766 deep fake, 59672 face swap, and 73770 face2face manipulations. The dataset is summarized in Table 1, both in the form of images and videos. We perform different experiments on the Faceforensics++ dataset on the lines of those devised by Wang et al. (Wang and Dantcheva, 2020) where we report True Classifications Rates (TCR) in them.

4.1 Combined Manipulation Techniques

We present the results in Table 2 with combined manipulation techniques. The training and the testing are performed as real v/s combined fakes which include DeepFakes (DF), Neural Textures (NT), Face2Face (FF), and FaceSwap (FS). We pose the classification as a two-class problem of real v/s combined fakes. Our proposed method outperforms the current state-of-the-art (SOTA) in this evaluation, as shown in Table 2.

Table 1: Details of Faceforensics++ Dataset used.

	Images				
Phase	Pristine	NT	DF	FS	F2F
Training	367228	291789	367009	292320	366631
Validation	68857	54617	68664	54624	68854
Testing	73768	59670	73766	59672	73770
	Videos				
Training	720	720	720	720	719
Validation	140	140	140	140	140
Testing	140	140	140	140	140

Table 2: Results on Low-Resolution Images from Faceforensics++ Dataset, with image-manipulations of (DeepFakes (DF), Neural Textures (NT), Face2Face (FF), and FaceSwap (FS)).

<table>
<tr><th colspan="6">Combined Evaluation of Low-Resolution Faceforensics++</th></tr>
<tr><td>Algorithm</td><td colspan="4">Train and Test</td><td>TCR%</td></tr>
<tr><td>3D Resnet (Wang and Dantcheva, 2020)</td><td colspan="4">FS,DF,F2F,NT</td><td>83.86</td></tr>
<tr><td>3D ResneXt (Wang and Dantcheva, 2020)</td><td colspan="4">FS,DF,F2F,NT</td><td>85.14</td></tr>
<tr><td>Proposed Method</td><td colspan="4">FS,DF,F2F,NT</td><td><b>96.07</b></td></tr>
<tr><th colspan="6">Single Evaluation of Low-Resolution Faceforensics++</th></tr>
<tr><td colspan="2">Algorithm</td><td>DF%</td><td>F2F%</td><td>FS%</td><td>NT%</td></tr>
<tr><td colspan="2">3D Resnet (Wang and Dantcheva, 2020)</td><td>91.81</td><td>89.6</td><td>88.75</td><td>73.5</td></tr>
<tr><td colspan="2">3D Resnext (Wang and Dantcheva, 2020)</td><td>93.36</td><td>86.06</td><td>92.5</td><td>80.5</td></tr>
<tr><td colspan="2">I3D (Wang and Dantcheva, 2020)</td><td>95.13</td><td>90.27</td><td>92.25</td><td>80.5</td></tr>
<tr><td colspan="2">RCNN w Densenet (Sabir et al., 2019)</td><td><b>96.9</b></td><td><b>94.35</b></td><td><b>96.3</b></td><td></td></tr>
<tr><td colspan="2">Proposed Method</td><td>94.64</td><td>81.4</td><td>94.28</td><td><b>100</b></td></tr>
<tr><th colspan="6">Cross Evaluation of Low-Resolution Faceforensics++</th></tr>
<tr><td>Train</td><td>Test</td><td>3D Resnet (Wang and Dantcheva, 2020)</td><td>3D Resnext (Wang and Dantcheva, 2020)</td><td>I3D (Wang and Dantcheva, 2020)</td><td>Proposed Method</td></tr>
<tr><td>FF, DF, F2F</td><td>NT</td><td>64.29</td><td><b>68.57</b></td><td>66.79</td><td>49.2</td></tr>
<tr><td>FS, DF, NT</td><td>F2F</td><td><b>74.29</b></td><td>70.71</td><td>68.93</td><td>40.0</td></tr>
<tr><td>FS, F2F, NT</td><td>DF</td><td><b>75.36</b></td><td>75.00</td><td>72.50</td><td>69.28</td></tr>
<tr><td>DF, F2F, NT</td><td>FS</td><td>59.64</td><td>57.14</td><td>55.71</td><td><b>94.28</b></td></tr>
</table>

4.2 Single Manipulation Techniques

We present the results in Table 2 with single manipulation techniques. The training and the testing are performed as real v/s individual fakes, which includes DeepFakes (DF), Neural Textures (NT), Face2Face (FF), and FaceSwap (FS). Our proposed method outperforms the current state-of-the-art (SOTA) in this evaluation for FS and NT. We get 100% classification accuracy on Neural Textures, attributed to our proposed network identifying the neural renderings. Further, our proposed network performed on par with SOTA for deep fakes and face swap. Thus, our proposed network can recognize changes in the image when either the identity changes or a neural rendering is performed. However, only on face2face, our proposed method performs slightly lower than SOTA, with only expression change in this. This could be attributed to the fact that there are local changes at the image level during expression transfer.

4.3 Cross Manipulation Techniques

We present the results in Table 2 with cross manipulation techniques. The training is performed as real v/s cross fakes, which includes DeepFakes (DF), Neural Textures (NT), Face2Face (FF), and FaceSwap (FS), for, e.g., training includes real images, and fake images from FF, DF, & F2F, and testing is performed on NT. Our proposed method outperforms the current state-of-the-art (SOTA) in this evaluation when face swap is used as identity change happens during this testing. It performs on par with SOTA when deep fakes are used during test time. However, for the evaluation, when neural textures or face2face are used during test time, our proposed network cannot generalize that well. This can be attributed to our proposed network's performance when an identity change happens during the test time, which is the case for deep fakes, and face swap test tasks.

4.4 Analysis of Results

We now analyze the results presented in previous subsections. Our proposed method works well on the combined image-manipulation and single-image manipulation techniques (Table 2). The case of combined and single image manipulation is the seen class scenario, and DLDFD performs well mainly due to augmented layers. DLDFD (Table 2) achieves comparable accuracy with SOTA for the cross image-manipulation technique result and is better than SOTA for face-swap image-manipulation. This is the unseen class case, and DLDFD generalizes well only when identity changes are used during test time.

5 CONCLUSIONS & FUTURE-WORK

In this paper, we proposed the use of DLDFD architecture based on 2d convolution with augmented layers to achieve better results for the evaluation of multiple image-manipulation techniques, two categories for the assessment of single image-manipulation technique, and one category for the evaluation of cross image-manipulation technique compared with SOTA based on 3d convolution (Wang and Dantcheva, 2020). 2d convolution in DLDFD achieves better performance at a lower computational cost.

We would extend the current technique to improve cross image manipulation results specifically to generalize for expression change during test time in future work. We would perform a more extensive comparison with the method by Sabir et al. (Sabir et al., 2019) as they have provided only a few results with single image manipulation techniques. The modification to improve cross image manipulation results would involve redesigning the proposed network architecture and using appropriate loss functions specifically for improving generalization.

ACKNOWLEDGMENT

This work is carried out under the partial funding of the Research Council of Norway (Grant No. IKT-PLUSS 248030/O70).

REFERENCES

This person does not exist. https://thispersondoesnotexist.com/. (Accessed on 08/26/2021).

(2000). http://gl.ict.usc.edu/LightStages/.

(2011). http://ict.usc.edu/prototypes/digital-ira/.

(2018). Auckland Airport deploys avatar to answer biosecurity questions. https://bit.ly/3ISXJ4w.

(2018). Face Me. https://www.faceme.com/.

(2019). Deepfakes code. https://github.com/deepfakes/faceswap. (Accessed on 08/26/2021).

Carreira, J. and Zisserman, A. (2017). Quo vadis, action recognition? a new model and the kinetics dataset. In *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6299–6308.

Hara, K., Kataoka, H., and Satoh, Y. (2018). Can spatiotemporal 3d cnns retrace the history of 2d cnns and imagenet? In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 6546–6555.

He, F., Liu, T., and Tao, D. (2020). Why resnet works? residuals generalize. *IEEE Transactions on Neural Networks and Learning Systems*, 31(12):5349–5362.

He, K., Zhang, X., Ren, S., and Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778.

Hu, L., Saito, S., Wei, L., Nagano, K., Seo, J., Fursund, J., Sadeghi, I., Sun, C., Chen, Y.-C., and Li, H. (2017). Avatar digitization from a single image for real-time rendering. *ACM Trans. Graph.*, 36(6):195:1–195:14.

Juefei-Xu, F., Dey, R., Bodetti, V. N., and Savvides, M. (2018). RankGAN: A Maximum Margin Ranking GAN for Generating Faces. In *Proceedings of the Asian Conference on Computer Vision (ACCV)*.

Jung, T., Kim, S., and Kim, K. (2020). Deepvision: Deepfakes detection using human eye blinking pattern. *IEEE Access*, 8:83144–83154.

Karras, T., Aila, T., Laine, S., and Lehtinen, J. (2018). Progressive growing of GANs for improved quality, stability, and variation. In *International Conference on Learning Representations*.

Karras, T., Laine, S., and Aila, T. (2019). A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4401–4410.

Korshunov, P. and Marcel, S. (2018). Deepfakes: a new threat to face recognition? assessment and detection. *CoRR*, abs/1812.08685.

Kowalski, M. (2016). Faceswap code. (Accessed on 08/26/2021).

Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25:1097–1105.

Li, L., Bao, J., Zhang, T., Yang, H., Chen, D., Wen, F., and Guo, B. (2020). Face x-ray for more general face forgery detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5001–5010.

Li, Y., Chang, M.-C., and Lyu, S. (2018). In ictu oculi: Exposing ai created fake videos by detecting eye blinking. In *2018 IEEE International Workshop on Information Forensics and Security (WIFS)*, pages 1–7.

Li, Y. and Lyu, S. (2019). Exposing deepfake videos by detecting face warping artifacts. In *IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*.

Lin, M., Chen, Q., and Yan, S. (2013). Network in network. *arXiv preprint arXiv:1312.4400*.

Liu, J., Zhu, K., Lu, W., Luo, X., and Zhao, X. (2021). A lightweight 3d convolutional neural network for deepfake detection. *International Journal of Intelligent Systems*, 36(9):4990–5004.

Lyu, S. (2020). Deepfake detection: Current challenges and next steps. pages 1–6.

Mirsky, Y. and Lee, W. (2021). The creation and detection of deepfakes: A survey. *ACM Comput. Surv.*, 54(1).

Rössler, A., Cozzolino, D., Verdoliva, L., Riess, C., Thies, J., and Nießner, M. (2018). Faceforensics: A large-scale video dataset for forgery detection in human faces. *CoRR*, abs/1803.09179.

Rössler, A., Cozzolino, D., Verdoliva, L., Riess, C., Thies, J., and Nießner, M. (2019). Faceforensics++: Learning to detect manipulated facial images. *arXiv preprint arXiv:1901.08971*.

Sabir, E., Cheng, J., Jaiswal, A., AbdAlmageed, W., Masi, I., and Natarajan, P. (2019). Recurrent convolutional strategies for face manipulation detection in videos. *CoRR*, abs/1905.00582.

Sengupta, S., Curless, B., Kemelmacher-Shlizerman, I., and Seitz, S. M. (2021). A light stage on every desk. *CoRR*, abs/2105.08051.

Sun, Z., Han, Y., Hua, Z., Ruan, N., and Jia, W. (2021). Improving the efficiency and robustness of deepfakes detection through precise geometric features. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3609–3618.

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., and Rabinovich, A. (2015). Going deeper with convolutions. In *Computer Vision and Pattern Recognition (CVPR)*.

Thies, J., Zollhöfer, M., and Nießner, M. (2019). Deferred neural rendering: Image synthesis using neural textures. *ACM Transactions on Graphics (TOG)*, 38(4):1–12.

Thies, J., Zollhofer, M., Stamminger, M., Theobalt, C., and Nießner, M. (2016). Face2face: Real-time face capture and reenactment of rgb videos. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2387–2395.

Tolosana, R., Vera-Rodriguez, R., Fierrez, J., Morales, A., and Ortega-Garcia, J. (2020). Deepfakes and beyond: A survey of face manipulation and fake detection. *Information Fusion*, 64:131–148.

Wang, Y. and Dantcheva, A. (2020). A video is worth more than 1000 lies. comparing 3dcnn approaches for detecting deepfakes. In *2020 15th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2020)*, pages 515–519. IEEE.

Yamaguchi, S., Saito, S., Nagano, K., Zhao, Y., Chen, W., Olszewski, K., Morishima, S., and Li, H. (2018). High-fidelity facial reflectance and geometry inference from an unconstrained image. *ACM Trans. Graph.*, 37(4):162:1–162:14.

Zhang, K., Zhang, Z., Li, Z., and Qiao, Y. (2016). Joint face detection and alignment using multitask cascaded convolutional networks. *IEEE Signal Processing Letters*, 23(10):1499–1503.