

Impact of Benign Modifications on Discriminative Performance of Deepfake Detectors

Yuhang Lu, Evgeniy Upenik, and Touradj Ebrahimi
Multimedia Signal Processing Group (MMSPG)
École Polytechnique Fédérale de Lausanne (EPFL)
CH-1015 Lausanne, Switzerland
Email: firstname.lastname@epfl.ch

Abstract—Deepfakes are becoming increasingly popular in both good faith applications such as in entertainment and maliciously intended manipulations such as in image and video forgery. Primarily motivated by the latter, a large number of deepfake detectors have been proposed recently in order to identify such content. While the performance of such detectors still need further improvements, they are often assessed in simple if not trivial scenarios. In particular, the impact of benign processing operations such as transcoding, denoising, resizing and enhancement are not sufficiently studied. This paper proposes a more rigorous and systematic framework to assess the performance of deepfake detectors in more realistic situations. It quantitatively measures how and to which extent each benign processing approach impacts a state-of-the-art deepfake detection method. By illustrating it in a popular deepfake detector, our benchmark proposes a framework to assess robustness of detectors and provides valuable insights to design more efficient deepfake detectors.

I. INTRODUCTION

In recent years, face swapping has become one of the most popular face manipulation techniques due to the great concerns with the deepfake generation. The deep learning-based tools and open source software have simplified the creation of such manipulated contents, posing serious public concerns. It is crucial to develop deepfake detection systems that can automatically and effectively identify manipulated videos and images. However, while the current state-of-the-art detectors achieve promising results on target databases, they are trained and tested in rather simple scenarios and often show poor performance in post-processed contents. In realistic situations, videos and images on the Internet and social networks constantly undergo benign processing to ease transmission, including but not limited to compression, denoising, enhancement and resizing. A deepfake detector should be robust to such modifications. Moreover, hackers can also make use of such drawbacks and easily fool the forensic detection by post-processing.

Some recent works [1], [2], [3] have considered this problem when creating the database. For example, FaceForensics++ database [1] compresses raw video footage based on H.264 at two different quality factors. Jiang et al. [2] extensively applied

real-world perturbations to enlarge the dataset and meanwhile obtain a more challenging and diverse benchmark. The most recent large scale deepfake database DFDC [3] introduces various overarching types of augmentations. Despite their advancement, neither these database providers nor researchers working in this area, systematically compare and summarize the impact of perturbations by means of post-processing. A more comprehensive and systematic approach with in-depth analysis of how the benign processing impacts the detection performance is desired.

In this work, an extensive evaluation of the performance of a popular deepfake detection approach towards different modifications is presented as a first step towards explainable artificial intelligence behind deepfake generation and detection. We conduct our study on FaceForensics++ dataset. To generate a broad range of variations on the test data, popular post-processing operations such as video compression, image transcoding, smoothing, and enhancement re applied to deepfakes. In addition, we also modify the image with additive Gaussian noise. As learning-based image processing algorithms have become very popular, we also include learning-based compression and super-resolution methods in our framework. While a performance degradation is expected after applying one of these processing methods, our framework quantifies the impact of each type of processing by showing to which extent the detector under analysis is sensitive to such modifications.

The paper is organized as follows: Section 2 provides a short overview of deepfake creation, detection and popular databases. Section 3 details the proposed assessment framework. Results are discussed in Section 4 and Section 5 draws conclusions.

II. RELATED WORK

The related work can be organized in three categories, namely, deepfake generation, deepfake detection, and datasets. For a broader review readers are refer to [4].

A. Deepfake Generation

Even though studies on face manipulation appeared as early as in the first decade of this century [5], the visual quality of the produced deepfakes only reached acceptable levels

The authors acknowledge support from CHIST-ERA project XAIface (CHIST-ERA-19-XAI-011) with funding from the Swiss National Science Foundation (SNSF) under grant number 20CH21_195532

with introduction of deep convolutional (CNN) and generative adversarial (GAN) neural networks [6].

Currently, among the successful deepfake generation methods one can note, particularly: *face swapping*, *lip syncing*, and *motion models*. Face swapping methods typically substitute the face of a subject with that of a target [7]. Lip syncing only changes the lip movements of a face in a way that the generated movements correspond to a particular speech [8], [9]. Motion models take a still image of a face as an input and animate it according to a reference from another face [10].

B. Deepfake Detection

One of the first methods for deepfake detection was proposed by Zhou et al. in [11]. They use two parallel networks: one detects image tampering and another performs steganalysis to account for camera processing and local noise characteristics. Afchar et al. [12] propose an approach that is claimed to be robust to compression artifacts, thanks to excluding explicit noise from consideration.

In [13], Nguyen et al. proposed a convolutional neural network with a Y-shaped autoencoder that can quickly adapt to deal with unseen attacks by using only a few samples for fine-tuning. In [14], the same authors introduce a capsule network that can detect various types of attacks, ranging from presentation attacks using printed images to attacks using fakes created by deep learning.

C. Available Datasets

Various datasets have been proposed and used by the research community for training deepfake detectors. Korshunov and Marcel [15] were among the first to release a public dataset of deepfake videos in 2018 that was created using a GAN-based face-swapping algorithm. A bigger and more popular dataset *FaceForensics++* was released a year later by Rossler et al. [1]. This dataset consists of more than 1.8 million images that were face-manipulated as well as their ground truth. In early 2020, major companies partnered with academia to organize the *Deepfake Detection Challenge*¹ that led to the creation of a dataset [3]. Finally, Li et al. [16] created the *Celeb-DF* dataset by face-swapping videos using synthesized images of celebrities.

III. ASSESSMENT FRAMEWORK

A. Assessment Database

FaceForensics++ [1] was used as the main database for our assessment framework. FaceForensics++ is one of the most popular large-scale databases in the media forensic detection field. It is an extension of the original FaceForensics database [17], which was initially focused on facial reenactment through Face2Face [18] manipulation. FaceForensics++ contains 1000 real videos collected from YouTube and 4000 fake videos generated using both classical computer graphics

approaches, i.e. FaceSwap² and Face2Face [18], and learning-based approaches, i.e. Deepfake³, NeuralTextures [19]. Every two pristine videos were paired up and swapped faces using the four manipulation methods respectively.

For a realistic setting, all videos are compressed using the H.264 codec, which is common in real-world social media and websites, in particular. In addition to raw (no compression), high quality, and low quality compressed videos were generated using a quantization parameter equal to 23 and 40. To our knowledge, FaceForensics++ was one of the few databases that considers different levels of video quality.

The dataset was split into 720 videos for training, 140 for validation, and 140 for testing. To ease the assessment and to establish a uniform evaluation framework, the default split up was performed and focused on the 140 testing videos. In addition, the first 10 frames of each testing video was selected and cropped around facial areas. The cropped images were then resized to the same dimensions and saved in PNG format.

B. Detection Method

The publicly available detector called Capsule-Forensics and proposed by Nguyen et al. [14], was adopted.⁴

The Capsule-Forensics approach was selected in our study for several reasons. It combines the traditional CNN and the Capsule Networks, which require fewer parameters with reasonable computational resources when compared to approaches that are based on CNN only, while achieving similar detection performance. The lightweight architecture is more time efficient during evaluation, making it a suitable candidate for our exhaustive assessment framework. Moreover, this approach demonstrates high accuracy on FaceForensics++ dataset and provides a reasonable baseline for the proposed framework.

Furthermore, the original authors used VGG19 [20] pre-trained on ImageNet [21] as feature extractor, followed by several primary capsules and two output capsules ('real' and 'fake'). The input images were pre-processed by cropping the face area using a face detection algorithm, then resized to 300x300, which they claim to be large enough to provide reliable results. In addition, two regularizations were introduced, to reduce overfitting, by adding random noise and a dropout operation during training. In this paper, the pretrained model released by the authors was directly applied in our experiments.

C. Processing Operations

To create a comprehensive and systematic assessment framework, consider multiple benign operations based on both conventional and deep learning-based approaches were considered. As most of the current forgery detection models are evaluated in trivial settings, the main objective here is to see how commonly used operations in real situations affect the state-of-the-art detectors. The details of all operations used

²<https://github.com/MarekKowalski/FaceSwap/>

³<https://github.com/deepfakes/faceswap>

⁴<https://github.com/nii-yamagishilab/Capsule-Forensics-v2>

¹<https://www.kaggle.com/c/deepfake-detection-challenge/>

in evaluations are described below with the illustration of a typical example provided in Figure 1.

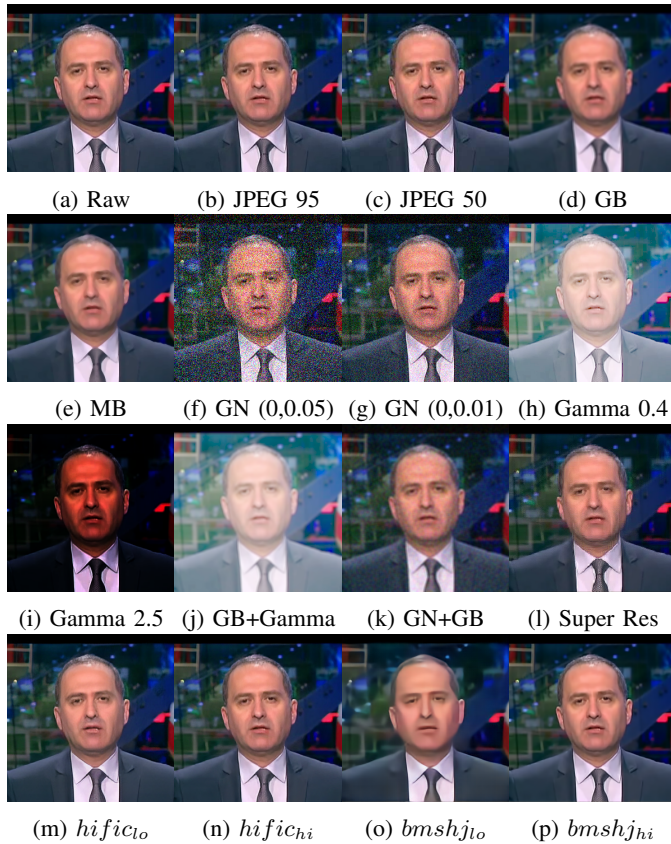


Fig. 1: Example of a typical frame in the dataset, after applying various operations. GB: Gaussian blur with kernel size 5; MB: Median blur with kernel size 5; GN: Gaussian noise; Super Res: Learning-based super resolution; *hifio* and *bshj*: Two learning-based image compression algorithms.

Lossy compression refers to the class of data encoding methods that remove unnecessary or less important information and only uses partial data to represent the content. These techniques are used to reduce data size for efficient storage and transmission content and are widely applied to images and videos on social networks and websites. It is interesting to observe how the distortion or artifacts by lossy data compression would affect the performance of a deepfake detector. The FaceForensics++ database provides two compressed versions based on H.264. An additional version was created using libx265 with the parameter 28, that typically corresponds, visually speaking, to libx264 at CRF 23. For images, JPEG compression with three quality factors was applied to reveal the impact of conventional image compression. In image processing, blurring, also known as smoothing, is a widely used operation, typically to reduce noise which at the same time results in reduction of details. Raw frames were processed with different filters such as Gaussian, and other popular blur operations with varying kernel sizes. A contrasting operation is the addition a Gaussian noise to frames in the dataset, which

often occurs during data acquisition by image sensors. We also considered gamma correction as an image enhancement technique, which corrects the brightness of a frame by using a non-linear transformation between the input values and the mapped output. Moreover, a mixture of two or three operations was also considered, such as combining Gaussian noise and Gaussian blur techniques, making the database to better reflect more complex real-world scenarios.

Recently, thanks to breakthroughs in both hardware and machine learning, a large number of image processing techniques have been proposed based on deep learning, including image upsampling, image compression, image enhancement and so on. For example, Fabian Mentzer et al. [22] directly optimize the neural network for a better rate-distortion trade-off and leverage GANs to prevent compression artifacts, which yields reconstructions with higher fidelity. With promising outcomes from these methods, it is also important to analyze the impact of learning-based processing on deepfake detection. Therefore, in addition to the conventional approaches, a data-driven super resolution technique [23] and two generative lossy compression algorithms [22], [24] were evaluated, in comparison to linear interpolation for resizing and the classical JPEG, H.264 and H.265 compression.

D. Implementation Details

The entire assessment framework is implemented in PyTorch with Python. The checkpoint used were trained using the Adam optimizer [25] with $\beta_1 = 0.9$, $\beta_2 = 0.999$, and a learning rate of 5×10^{-4} . A total of 720 pristine videos and 2880 deepfakes were used for training. The first 100 frames of each training video was chosen to train the model for 25 epochs. Notably, the detector was trained on purely unprocessed genuine and fake contents.

The preprocessing step is comprised of video compression, image processing, face detection and resizing. The proposed framework uses the entire test set of FaceForensics++ database, specifically 140 real videos and the same amount of fake videos for each of the four manipulations. We then extract the first 10 frames of each video and use a face detection to crop and align the frames. Afterwards, multiple benign processing operations are applied independently and constitute the dataset for different assessments.

For a fair comparison with the results obtained by a baseline configuration, the binary detection accuracy is used to rate the performance. Considering the imbalanced test data and to better estimate the system robustness, Area under the Curve (AUC) of Receiver Operating Characteristic Curve (ROC) and F1-score as the metrics are also added to experimental results.

IV. EXPERIMENTAL RESULTS AND DISCUSSION

The performance of the state-of-the-art forensic detection methods towards multiple benign modifications has been assessed. The findings show that preprocessing with even benign operations can have a negative impact on the robustness of a deepfake detector and generally results in noticeable decline

of detection performance. In Table I, the accuracy, AUC and F1-score are reported for the model Capsule-Forensics [14].

TABLE I: Detection performance measured by Accuracy, AUC and F1-score, after applying different types of operations to the dataset. GN: Gaussian Noise, GB: Gaussian Blur, GC: Gamma Correction. For bmsj compression, 1 represents lowest compression quality and 8 the highest.

	Operations	Acc	AUC	F1-Score
Raw	-	80.25	71.35	87.40
Video Compression	H265_c28	76.51	61.49	85.49
	H264_c23	76.38	67.96	84.87
	H264_c40	68.23	60.04	78.77
Image Transcoding	JPEG 95	78.88	70.40	86.61
	JPEG 75	76.57	70.33	84.65
	JPEG 50	72.14	69.17	80.97
Image Smoothing	Gaussian Blur ks=3	75.50	66.91	84.14
	Gaussian Blur ks=5	72.60	62.60	82.23
	Mean Blur ks=5	72.63	62.51	82.26
	Median Blur ks=5	78.25	60.83	86.86
Additive Noise	Gaussian Noise (0,0.01)	56.61	55.47	67.89
	Gaussian Noise (0,0.05)	27.66	49.58	22.35
Gamma Correction	$\gamma=0.4$	78.33	58.74	87.09
	$\gamma=2.5$	79.75	69.27	87.26
Combination	GN (0,0.01) + GB ks=5	54.33	60.05	63.89
	GB ks=5 + GC $\gamma=0.4$	74.44	57.06	84.34
	GB ks=5 + JPEG 95	70.48	62.54	80.42
	GC $\gamma=0.4$ + JPEG 95	78.07	60.30	86.78
Resizing	Linear Interp (scale=1.3)	83.18	75.30	89.37
	Super Res (scale=1.3)	78.65	68.02	86.53
AI-based Compression	hifc_low	76.93	63.12	85.66
	hifc_med	78.19	65.96	86.36
	hifc_high	77.76	66.82	85.95
	bmsj_1	73.28	56.63	83.48
	bmsj_4	76.46	59.39	85.65
	bmsj_8	78.69	65.42	86.79

With the exception of resizing, the best performance is achieved on raw videos with an accuracy of 80.25%. The performance is somewhat lower than the reported score in the original paper because of differences in face detection and training, among others. However, this does not impact the conclusions as the goal is to assess impact on performance after applying various benign operations.

Performance w.r.t data compression and transcoding It is expected that the performance of the detector degrades as the raw videos are compressed with lower quality (larger coding parameters). The model achieves 76.38% accuracy when the test videos are compressed based on libx264 encoder at CRF 23 and it drops to 68.23% when at CRF 40. It is interesting to observe that the detector’s performance on libx265 encoded video at CRF 28 is similar as on H.264 codec at CRF 23, while it results in about half the bitrate. Similar reduction trends can be observed in JPEG as well as in learning-based compression. The performance only decreases by 1.37% on JPEG compressed images with high quality (95%), while the performance drops quickly as we lower the image quality. The trend is shown in Figure 2. For both learning-based lossy compression algorithms, we evaluate the lowest, medium, and highest quality levels. It is remarkable that the high quality learning-based image compression results in comparable performance as high image quality JPEG compression.

Performance w.r.t image smoothing Three blurring techniques are evaluated in our framework and all of them lead to performance degradation. The median filter brings less negative impact and reduces the accuracy by only 2%, while the Gaussian and mean filters reduce by 7.65% and 7.62%.

Performance w.r.t noise Gaussian noise has a significantly negative impact on performance when compared to all other operations. The detector performs only with 56.61% accuracy when test images are corrupted by additive Gaussian noise with mean 0 and variance 0.01, nearly close to random guesses. The results are even worse after increasing the variance, potentially posing great threat to current deepfake detectors. We evaluate a mixed operation that adds Gaussian noise first, followed by a denoising Gaussian filter. Interestingly, the additional denoising operation makes the result even worse. Overall, the combination of two or three processing operations degrade the detection accuracy further when compared to only considering one operation independently.

Performance w.r.t resizing The test image were resized with a scale of 1.3 which unexpectedly brings improvements detection accuracy, which is consistent with the observations of the original author of Capsule-Forensics. On the contrary, the performance on re-scaled images through learning-based super resolution technique results in slightly decreased performance, possibly due to the accompanying artifacts. However, more experiments are expected to investigate the reasons behind.

Performance w.r.t image enhancement It was also considered to change the contrast by underexposing and overexposing frames using gamma operations with $\gamma = 0.4$ and $\gamma = 2.5$. As shown from Table I, both operations decrease the performance by a rather small margin and the detector achieves accuracy of 78.33% and 79.75% respectively.

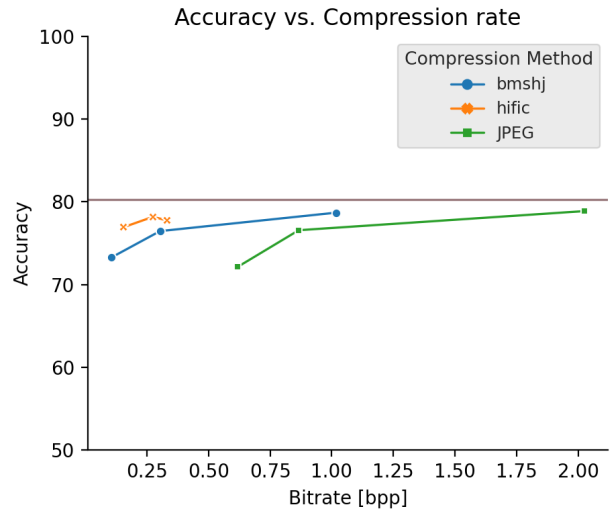


Fig. 2: Detection accuracy with respect to three compression algorithms. The horizontal line is the performance on raw data.

V. CONCLUSIONS

In this work, a rigorous and systematic assessment framework for deepfake detectors in realistic situations was presented. The impact of both conventional and learning-based benign modifications on deepfake detectors were taken into account. The usefulness of the proposed framework has been illustrated by conducting an extensive evaluation on Capsule-Forensics approach with processed FF++ database. The benchmark identifies and quantifies the different influences that each benign modification can bring to a deepfake detector. The proposed framework can be adapted to a wider variety of detectors and recognition tasks, and applied to different datasets, in order to obtain more comprehensive and insightful conclusions to other types of detectors and recognition algorithms. Such analysis can become useful for design of more efficient and robust detectors and recognition tasks.

REFERENCES

- [1] A. Rössler, D. Cozzolino, L. Verdoliva, C. Riess, J. Thies, and M. Niessner, "FaceForensics++: Learning to Detect Manipulated Facial Images," in *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, Oct. 2019, pp. 1–11, iSSN: 2380-7504.
- [2] L. Jiang, R. Li, W. Wu, C. Qian, and C. C. Loy, "DeeperForensics-1.0: A Large-Scale Dataset for Real-World Face Forgery Detection," in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2020, pp. 2886–2895, iSSN: 2575-7075.
- [3] B. Dolhansky, J. Bitton, B. Pfau, J. Lu, R. Howes, M. Wang, and C. C. Ferrer, "The DeepFake Detection Challenge (DFDC) Dataset," *arXiv:2006.07397 [cs]*, Oct. 2020, arXiv: 2006.07397. [Online]. Available: <http://arxiv.org/abs/2006.07397>
- [4] R. Tolosana, R. Vera-Rodriguez, J. Fierrez, A. Morales, and J. Ortega-Garcia, "Deepfakes and beyond: A Survey of face manipulation and fake detection," *Information Fusion*, vol. 64, pp. 131–148, 2020.
- [5] D. Bitouk, N. Kumar, S. Dhillon, P. Belhumeur, and S. K. Nayar, "Face swapping: automatically replacing faces in photographs," in *ACM SIGGRAPH 2008 papers*, ser. SIGGRAPH '08. New York, NY, USA: Association for Computing Machinery, Aug. 2008.
- [6] I. Korshunova, W. Shi, J. Dambre, and L. Theis, "Fast Face-Swap Using Convolutional Neural Networks," in *2017 IEEE International Conference on Computer Vision (ICCV)*, Oct. 2017, pp. 3697–3705, iSSN: 2380-7504.
- [7] I. Perov, D. Gao, N. Chervoniy, K. Liu, S. Marangonda, C. Umé, M. Dpfks, C. S. Facenheim, L. RP, J. Jiang, S. Zhang, P. Wu, B. Zhou, and W. Zhang, "DeepFaceLab: A simple, flexible and extensible face swapping framework," *arXiv:2005.05535 [cs, eess]*, May 2020, arXiv: 2005.05535. [Online]. Available: <http://arxiv.org/abs/2005.05535>
- [8] K. R. Prajwal, R. Mukhopadhyay, J. Philip, A. Jha, V. Nambodiri, and C. V. Jawahar, "Towards Automatic Face-to-Face Translation," in *Proceedings of the 27th ACM International Conference on Multimedia*, ser. MM '19. New York, NY, USA: Association for Computing Machinery, Oct. 2019, pp. 1428–1436.
- [9] K. R. Prajwal, R. Mukhopadhyay, V. P. Nambodiri, and C. V. Jawahar, "A Lip Sync Expert Is All You Need for Speech to Lip Generation in the Wild," in *Proceedings of the 28th ACM International Conference on Multimedia*. New York, NY, USA: Association for Computing Machinery, Oct. 2020, pp. 484–492.
- [10] A. Siarohin, S. Lathuilière, S. Tulyakov, E. Ricci, and N. Sebe, "First Order Motion Model for Image Animation," *Advances in Neural Information Processing Systems*, vol. 32, 2019.
- [11] P. Zhou, X. Han, V. I. Morariu, and L. S. Davis, "Two-Stream Neural Networks for Tampered Face Detection," in *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, Jul. 2017, pp. 1831–1839, iSSN: 2160-7516.
- [12] D. Afchar, V. Nozick, J. Yamagishi, and I. Echizen, "MesoNet: a Compact Facial Video Forgery Detection Network," in *2018 IEEE International Workshop on Information Forensics and Security (WIFS)*, Dec. 2018, pp. 1–7, iSSN: 2157-4774.
- [13] H. H. Nguyen, F. Fang, J. Yamagishi, and I. Echizen, "Multi-task Learning for Detecting and Segmenting Manipulated Facial Images and Videos," in *2019 IEEE 10th International Conference on Biometrics Theory, Applications and Systems (BTAS)*, Sep. 2019, pp. 1–8, iSSN: 2474-9699.
- [14] H. H. Nguyen, J. Yamagishi, and I. Echizen, "Use of a Capsule Network to Detect Fake Images and Videos," *arXiv:1910.12467 [cs]*, Oct. 2019, arXiv: 1910.12467. [Online]. Available: <http://arxiv.org/abs/1910.12467>
- [15] P. Korshunov and S. Marcel, "DeepFakes: a New Threat to Face Recognition? Assessment and Detection," *arXiv:1812.08685 [cs]*, Dec. 2018, arXiv: 1812.08685. [Online]. Available: <http://arxiv.org/abs/1812.08685>
- [16] Y. Li, X. Yang, P. Sun, H. Qi, and S. Lyu, "Celeb-DF: A Large-Scale Challenging Dataset for DeepFake Forensics," in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2020, pp. 3204–3213, iSSN: 2575-7075.
- [17] A. Rössler, D. Cozzolino, L. Verdoliva, C. Riess, J. Thies, and M. Nießner, "Faceforensics: A large-scale video dataset for forgery detection in human faces," *arXiv preprint arXiv:1803.09179*, 2018.
- [18] J. Thies, M. Zollhofer, M. Stamminger, C. Theobalt, and M. Nießner, "Face2face: Real-time face capture and reenactment of rgb videos," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 2387–2395.
- [19] J. Thies, M. Zollhöfer, and M. Nießner, "Deferred neural rendering: Image synthesis using neural textures," *ACM Transactions on Graphics (TOG)*, vol. 38, no. 4, pp. 1–12, 2019.
- [20] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.
- [21] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei, "ImageNet Large Scale Visual Recognition Challenge," *International Journal of Computer Vision (IJCV)*, vol. 115, no. 3, pp. 211–252, 2015.
- [22] F. Mentzer, G. D. Toderici, M. Tschannen, and E. Agustsson, "High-fidelity generative image compression," *Advances in Neural Information Processing Systems*, vol. 33, 2020.
- [23] X. Wang, K. Yu, S. Wu, J. Gu, Y. Liu, C. Dong, Y. Qiao, and C. C. Loy, "EsrGAN: Enhanced super-resolution generative adversarial networks," in *The European Conference on Computer Vision Workshops (ECCVW)*, September 2018.
- [24] J. Ballé, D. Minnen, S. Singh, S. J. Hwang, and N. Johnston, "Variational image compression with a scale hyperprior," *arXiv preprint arXiv:1802.01436*, 2018.
- [25] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.