Practical Manipulation Model for Robust Deepfake Detection

Supplementary Material

Overview

This document provides additional information omitted from the main paper for brevity. Appendix A explains the inference process on the images in Fig. 1. Appendix B provides additional example images of our Practical Manipulation Model, in the same way as in Fig. 2. Additional details for our method are described in Appendix C, whereas Appendix D provides more details on the ablation study. Appendix E provides additional robustness examples, similar to Sec. 4.4 in the main paper and Appendix F qualitatively evaluates robustness using GradCAM [35]. Finally, Appendix G provides more qualitative examples of our model and LAA [29], while Appendix H presents a metric for the hardness of our degradations and Appendix I describes some details of our implementation.

A. Inference on the assassination attempt image

Figure 1 shows a real image [28] and its corresponding real-world deepfake [13], which can avoid detection due to being slightly blurry. In the image, the faces of the two Secret Service agents have been modified to make them smile.

We tested the detection on the agent on the right because he is more obviously visible, properly exposed, and looking at the camera. Since both detectors are meant to be applied to crops of faces, we manually created a crop around the agent's head and performed detection on that.

B. More example images

Figure A shows some additional training images. The table follows the same concept as Fig. 2 in the main paper, providing more examples that could not be included for space reasons.

C. Method details

This section provides explanations of two details of our method, which were omitted from the paper due to space constraints.

C.1. Noise

Channel-wise correlated Gaussian noise is given by

$$\Sigma' = \left|(l_2' \cdot s)^2 \cdot U^T D U\right| \tag{1}$$

for some orthogonal matrix U and diagonal matrix D. The diagonal elements of D are randomly sampled from $Uniform(0, 1)$. U is entirely randomly sampled from $Uniform(0, 1)$ and then orthogonalized. $l_2' = \frac{l_2}{255}$ is the noise level in range $[0, 1]$. When calculating in range $\{0, 1, ..., 255\}$, the result has to be scaled back up: $\Sigma = 255 \cdot \Sigma'$

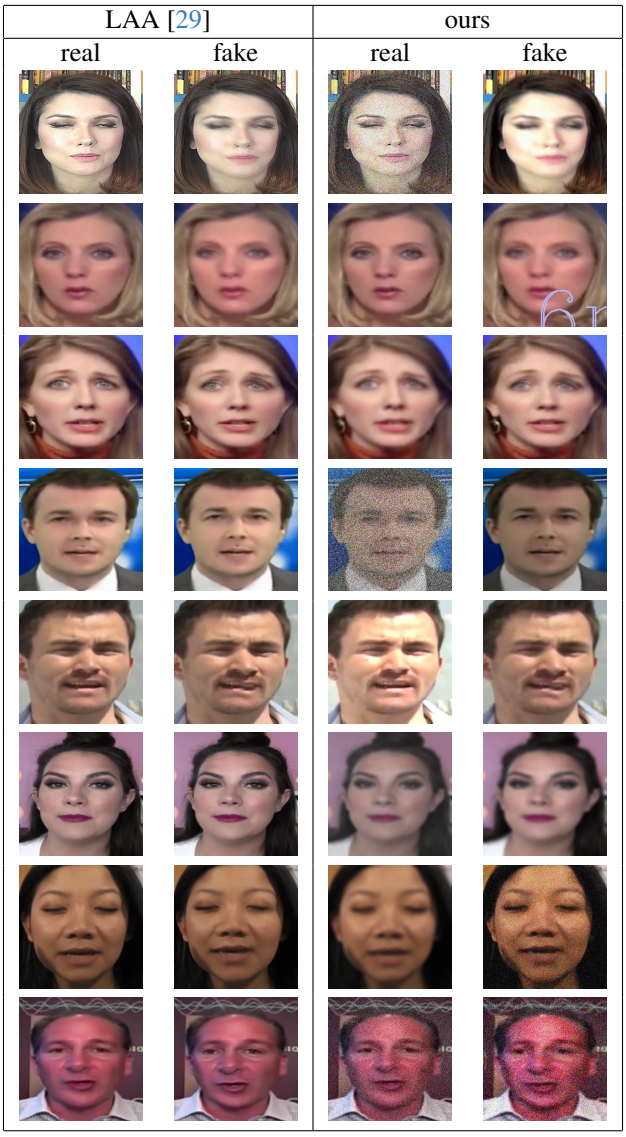


Figure A. Additional example images from our method and LAA [29]. This is an extension of Fig. 2. The original images are taken from the FaceForensics++ dataset [33].

C.2. Distractors

Multiple distractors per image. As described in the paper, multiple distractors may be added to a single image.

Variant	example training images
Original (LAA)	
+ Degradations	
+ Additional Masks	
+ Poisson blending	
+ Distractors	
+ Generator Artifacts (PMM)	

Table A. **Sample images during our ablation study.** The images are examples of the kind of fake data the model sees during the ablation. Note that not all changes are visible to the human eye. Images are selected to emphasize the effect of the visible steps, avoiding cases where a change is randomly not used. Original images are taken from FaceForensics++ [33].

The number of distractors is geometrically distributed. This means they are added until the random decision with probability p_d fails. Alternatively, we limit the number of distractors to 10 to avoid cluttering the image too much for large values of p_d. However, since we use $p_d = 0.2$, this case is only expected to happen with a probability of $0.2^{11} \approx 2 \cdot 10^{-8}$ and should, therefore, not have much effect in our experiments.

Text settings. The following settings are chosen for the overlaid text:

1. The text itself is a concatenation of $n \sim Uniform(0, 10)$ randomly selected characters, chosen from the list of all printable characters.
2. The text is placed at position (lower left corner) $x \sim Uniform(-100, WIDTH)$ and $y \sim Uniform(0, HEIGHT + 100)$. Positions outside of the image region result in the text being partially visible.
3. Font face is uniformly chosen from $\{0, ..., 7\}$.
4. Font scale is chosen from $Uniform(0, 8)$.
5. Color is uniformly chosen from $\{0, ..., 255\}^3$.
6. Line thickness is uniformly chosen from $\{1, 2, ..., 8\}$.
7. Line type is uniformly chosen from $\{0, 1, 2\}$.

D. Details on the ablation study

Table A shows examples of the types of images our model sees while training each of the ablation study variants. The images are selected to show the change of each variant (where possible), so the distribution is skewed towards our changes and away from the baseline. However, all images shown can appear during training.

Parameters. During training, we use degradations with $p = 50\%$ and strength $s = 0.5$. Poisson blending is used $p_p = 50\%$ of the time, and distractors are placed in $p_d = 20\%$ of the images, as described in Appendix C.2. Finally, we use both types of Generator Artifacts in $p_g = 25\%$ of the images, each.

From our testing, the model is relatively robust to hyperparameter choice, except for $p \approx 1$, where clean images appear too rarely. Especially small values of p and s have never been observed to hurt performance (*e.g.* $+3.42\%$ AUC on DFDCP for $p = s = 0.3$).

E. Additional robustness evaluation

We present an extended version of Fig. 4 in Fig. B, where we additionally tested the robustness of LAA [29], SBI [36]

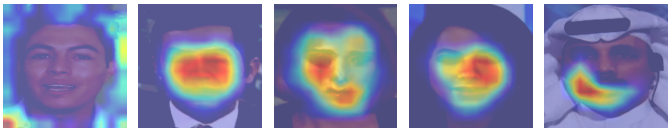
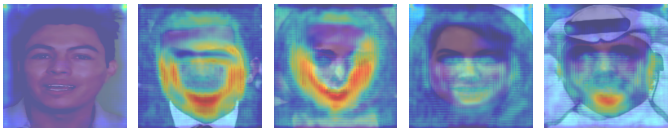
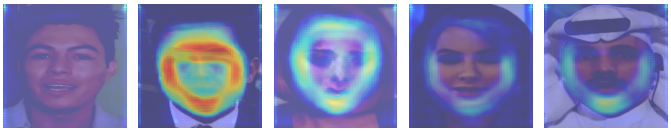
Deepfake-detector	Gaussian Noise $\sigma = 20$	real Original	fake DF [9]	FS [19]	F2F [39]	NT [40]
SBI [36]						
SBI [36]	✓					
LAA [29]						
LAA [29]	✓					
LAA+PMM (ours)						
LAA+PMM (ours)	✓					

Table B. **GradCAM visualization of our PMM compared to SBI [36] and LAA [29].** Note that for noisy images, both SBI and LAA fail to capture useful information and therefore classify all images as real. Images are taken from the FaceForensics++ dataset [33].

(EfficientNet [38] and Xception [6] backbones), and ours (+PMM) to speckle noise and low resolution. For a description of the Gaussian noise, Gaussian blur, JPEG compression, and motion blur, see Sec. 4.4. The diagram shown here is just an enlarged version for better readability.

Speckle noise. Similar to the results for Gaussian noise, both SBI and LAA lose performance, even at small noise levels. At $\sigma = 20$, LAA performs at chance level (0.54 AUC), whereas SBI (both) is better at > 0.6 AUC. However, our methods can beat both baselines by a large margin: SBI by $\geq 26\%$ and LAA by $\geq 35\%$. Even at the strongest settings $\sigma = 60$ (much larger than the $25 \cdot s = 12.5$ used during training), our methods still beat all baselines by $\geq 25\%$.

Resolution. For large scale-factors of $\frac{1}{4}$ and above, all tested models are relatively robust, showing more than 0.89 AUC, except for Xception, which only scores 0.85 AUC. For small scale factors of $\frac{1}{16}$ and $\frac{1}{32}$, our model clearly outperforms the baselines, *e.g.* by 7.7%, 10.8% and 11.8% for LAA+PMM vs. LAA, EfficientNet and Xception respectively at $\frac{1}{32}$.

Leave-one-out training. As described in the main paper, we also test models that use the entire PMM model, except for the specific test degradation, during training (PMM\deg). They perform worse than the full PMM model, but better than LAA, indicating generalization to unseen degradations. Gaussian noise is similar but about three to 5 percent worse than PMM until $\sigma = 20$. For stronger noise, the no-Gaussian-noise model drops off significantly, reaching about chance level at $\sigma = 60$. This is still much better than LAA, which drops to chance performance by $\sigma = 10$. For Gaussian blur, performance is better than LAA (up to 12%), but worse than full PMM (up to 6.7%). JPEG compression also leads to a smaller performance drop than LAA, but larger than PMM (e.g. 68% vs. 63% vs. 58% for PMM, PMM\deg and LAA, respectively, at quality 0). For speckle noise, performance is worse than PMM, by at most 5.6%, but much better than vanilla LAA, which is at chance performance, already at $\sigma \leq 10$. For small resize factors ($\leq \frac{1}{8}$), the model trained without resize is about 2.5% worse than PMM. However, this is still about 6% better than LAA.

These results suggest that due to the variety of degrada-

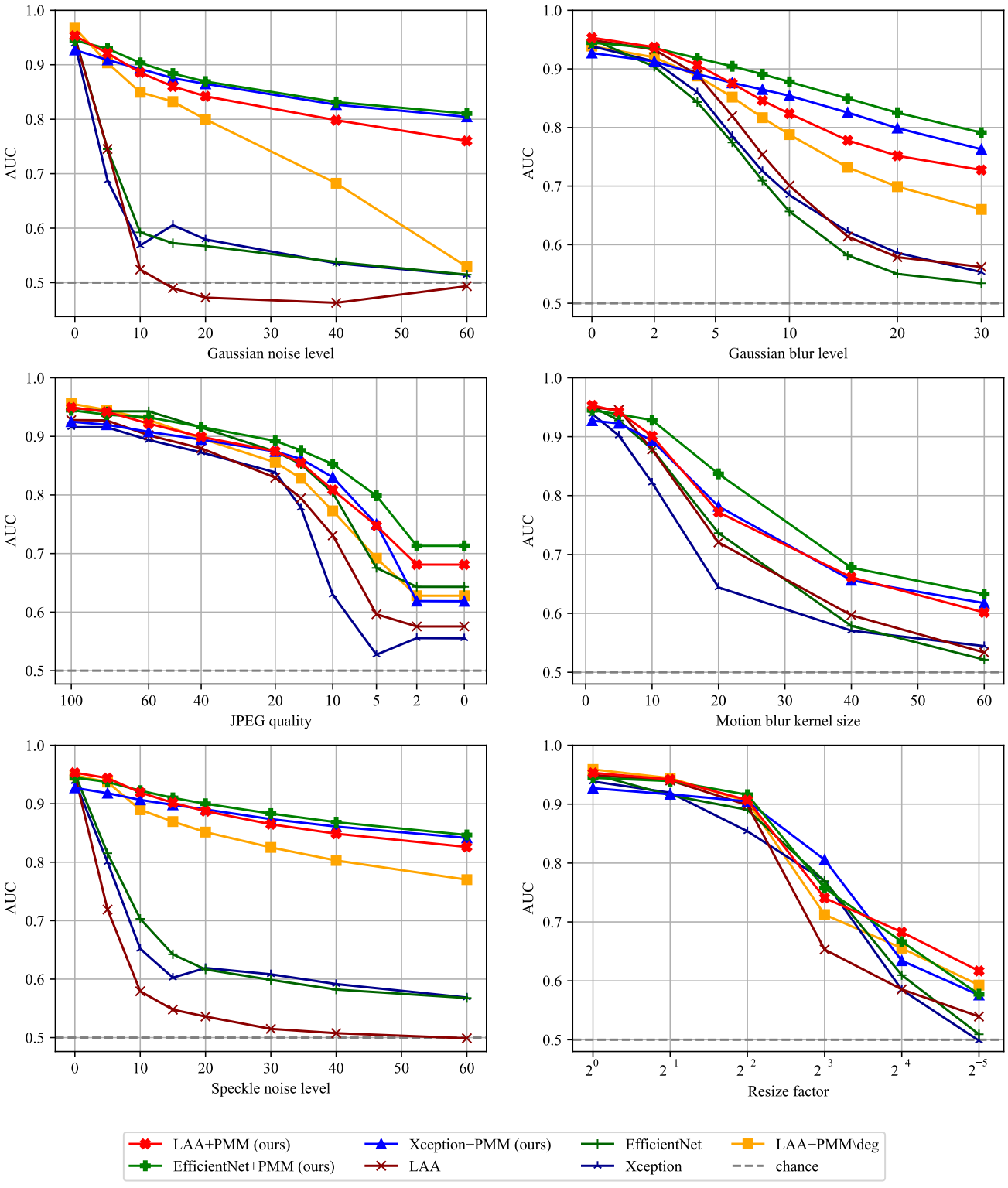


Figure B. **Additional robustness evaluation** of the s-o-t-a deepfake detectors LAA [29] and SBI [36] (EfficientNet [38] and Xception [6] backbones) vs. ours (+PMM), tested on the test-split of FF++ [33]. The first four plots show the same data as Fig. 4 in the main paper. Similar to these, our models outperform the baselines for the low-quality settings of Speckle noise and Resizing.

tions in our PMM, it is capable of generalizing to unseen degradations that were not part of the training data. This also aligns with our findings for motion blur. We therefore expect our model to generalize to novel degradations that we did not model during training.

New dataset. As describes in the main paper, our method can outperform LAA on the CDF based part of the face-swap and face-reenactment subsets of DF40 [45]. We also tested on the FF-based data. Here, LAA performs better than LAA+PMM (0.942 vs. 0.925 AUC). However, using data from the same source as the training data is not a true cross-dataset test and therefore not a good approximation to a real-world scenario, as it cannot be assumed that the real world follows the training distribution. The result is therefore only given for completeness.

F. GradCAM evaluation

Following related work [5, 23, 29, 36, 42], we visualize the inner workings of our model using GradCAM [35], comparing to SBI [36] and LAA [29]. The results are shown in Tab. B. All models work well on clean images, but ours focuses more cleanly on the blending boundary, with very few other points in the image activated. This indicates that our method focuses almost exclusively on the artifact regions and neither on the face (like SBI) nor the background (like LAA).

We also test on images, with added Gaussian noise ($\sigma = 20$), to explore the failure cases of the other models. In this case, both models focus on the image corners instead of the face. The focus of our model barely changes and only loses some strength. This is consistent with the observation that we can still correctly classify the fake images, whereas LAA and SBI cannot.

G. Qualitative demonstration

Figure C shows additional qualitative evaluations of LAA [29] and our model. All images are taken from the Celeb-DF-v2 dataset [22].

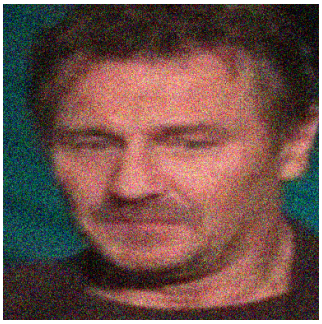
Figure Ca shows a failure case for LAA. Despite the image being sourced from the high-quality Celeb-DF-v2 dataset, the image is slightly blurry. LAA fails to recognize the image as fake, whereas ours can recognize it.

Figure Cb shows a failure case for both models. Again, the image is slightly blurry, but it also seems to be a very good fake, as neither LAA nor our model can recognize it as fake. From a human perspective, we also cannot tell any visible signs of the image being fake. This emphasizes the need to also continue developing models towards high-quality fakes. In this paper, we mainly concern ourselves with robustness to low-quality data.

Finally, Fig. Cc shows an example of the increased robustness of our model. We add Gaussian noise ($\sigma = 40$),

(a) Fake image: LAA [29] predicts real, ours predicts fake.

(b) Fake image: LAA [29] predicts real, ours predicts real.



(c) Fake image: LAA [29] predicts real, ours predicts fake.

Figure C. **Qualitative demonstration of LAA [29] and our model.** All images are taken from the Celeb-DF-v2 dataset [22]. For image (c), we added strong Gaussian noise to showcase the robustness of our model.

and yet, our model can still recognize the image as fake, while LAA considers it real. Again, from a human perspective, we cannot tell a sign of a fake, given that the image is extremely noisy.

H. Recognizability

To measure the difficulty of our degradations, we provide a metric that shows how recognizable faces remain. For this purpose, we train an EfficientNet-b4 [38] to recognize the identities of the faces in the FaceForensics++ [33] train split (720 identities). Then, we test the model under several kinds of degradations. Figure D shows the results under our degradation model. At the PMM settings, the face detector still achieves an accuracy of 67%, so our settings are of

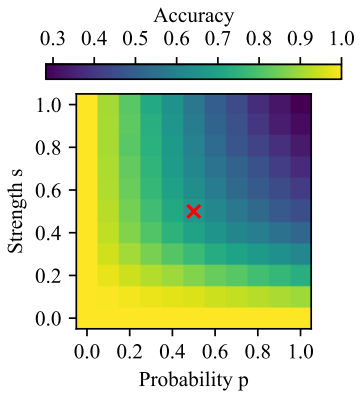


Figure D. Recognizability of faces under our degradation model. ✖ marks our settings during PMM training.

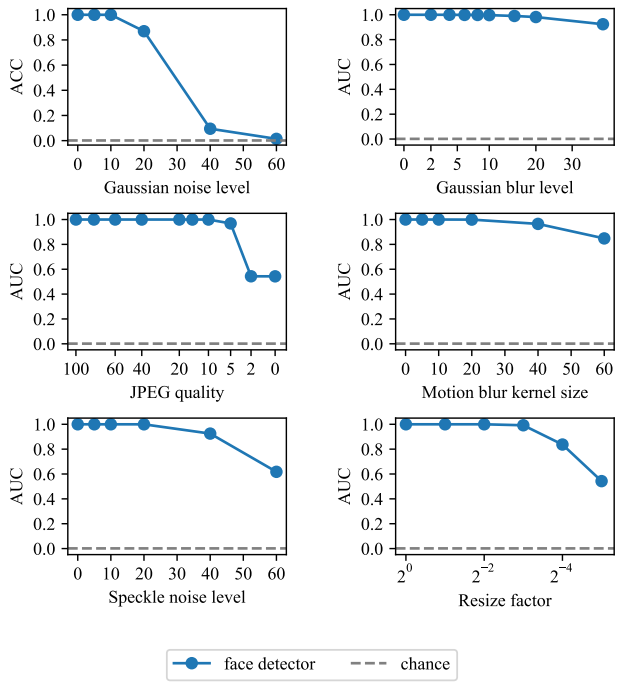


Figure E. Robustness of the face detector to the degradations from Fig. 4 and Fig. B

reasonable hardness. Furthermore, we can see (Fig. E) that most individual degradations shown in Figures 4 and B still allow for an accuracy of $> 50\%$ (at 720 possible identities). The only exception is noise; however, here, the face detector is still robust (accuracy $> 85\%$) until a value of $\sigma = 20$, where previous deepfake detectors already fail.

We can, therefore, show that our degradations are of reasonable difficulty, as faces still remain recognizable under these conditions.

I. Implementation details

We mainly use the settings from LAA [29] with some exceptions. We use a batch size of 7 to fit into the 24GB of VRAM of the NVIDIA RTX 3090 that we use for training. Furthermore, we use a learning rate scheduler, which reduces the learning rate by a factor of 0.2 every time the validation loss does not decrease for 10 epochs. We make this change to avoid having to tune the schedule for our experiments manually. Since we do not make any changes to the model itself, we can start from the pretrained checkpoint provided by [29], to save compute and ensure we start from the s-o-t-a point.

References

[1] Darius Afchar, Vincent Nozick, Junichi Yamagishi, and Isao Echizen. Mesonet: a compact facial video forgery detection network. *2018 IEEE International Workshop on Information Forensics and Security (WIFS)*, pages 1–7, 2018. 1

[2] Saifuddin Ahmed. Who inadvertently shares deepfakes? analyzing the role of political interest, cognitive ability, and social network size. *Telematics and Informatics*, 57:101508, 2021. 1

[3] Denis Bobkov, Vadim Titov, Aibek Alanov, and Dmitry Vetrov. The devil is in the details: Stylefeatureeditor for detail-rich stylegan inversion and high quality image editing, 2024. 5

[4] G. Bradski. The OpenCV Library. *Dr. Dobb's Journal of Software Tools*, 2000. 4

[5] Junyi Cao, Chao Ma, Taiping Yao, Shen Chen, Shouhong Ding, and Xiaokang Yang. End-to-end reconstruction-classification learning for face forgery detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4113–4122, 2022. 2, 5, 6

[6] François Chollet. Xception: Deep learning with depthwise separable convolutions. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1800–1807, 2016. 3, 6, 7, 4

[7] Alex Clark. Pillow (pil fork) documentation, 2015. 4

[8] Davide Cozzolino, Justus Thies, Andreas Rössler, Christian Riess, Matthias Nießner, and Luisa Verdoliva. Forensictransfer: Weakly-supervised domain adaptation for forgery detection. *ArXiv*, abs/1812.02510, 2018. 1

[9] Deepfakes. deepfakes_faceswap. https://github.com/deepfakes/faceswap, 2019. 5, 3

[10] Brian Dolhansky, Russ Howes, Ben Pflaum, Nicole Baram, and Cristian Canton-Ferrer. The deepfake detection challenge (dfdc) preview dataset. *ArXiv*, abs/1910.08854, 2019. 2, 3, 5, 6, 7

[11] Brian Dolhansky, Joanna Bitton, Ben Pflaum, Jikuo Lu, Russ Howes, Menglin Wang, and Cristian Canton-Ferrer. The deepfake detection challenge dataset. *ArXiv*, abs/2006.07397, 2020. 2, 3, 4, 5, 6

[12] faceman. Donald trump with boris johnson's haircut. https://faceswaponline.com/donald-trump-with-boris-johnsons-haircut, 2020. date accessed: 2024-11-07. 8

[13] Beatriz Farrugia. How ai tools fueled online conspiracy theories after trump assassination attempt. https://dfrlab.org/2024/08/23/how-ai-tools-fueled-online-conspiracy-theories-after-trump-assassination-attempt/, 2024. date accessed: 2024-11-07. 1

[14] Yingjie Guo, Cheng Zhen, and Pengfei Yan. Controllable guide-space for generalizable face forgery detection. *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 20761–20770, 2023. 2, 5, 6

[15] Jeffrey T Hancock and Jeremy N Bailenson. The social impact of deepfakes, 2021. 1

[16] Ahmed Abul Hasanaath, Hamzah Luqman, Raed Katib, and Saeed Anwar. Fsbi: Deepfakes detection with frequency enhanced self-blended images. *ArXiv*, abs/2406.08625, 2024. 2

[17] Yang He, Ning Yu, Margret Keuper, and Mario Fritz. Beyond the spectrum: Detecting deepfakes via re-synthesis. *ArXiv*, abs/2105.14376, 2021. 2

[18] Tack hyun Jung, Sangwon Kim, and Keecheon Kim. Deepvision: Deepfakes detection using human eye blinking pattern. *IEEE Access*, 8:83144–83154, 2020. 1

[19] Marek Kowalski. Faceswap. https://github.com/MarekKowalski/FaceSwap, 2018. 5, 3

[20] Nicolas Larue, Ngoc-Son Vu, Vitomir Struc, Peter Peer, and Vassilis Christophides. Seeable: Soft discrepancies and bounded contrastive learning for exposing deepfakes. *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 20954–20964, 2022. 2, 5, 6

[21] Lingzhi Li, Jianmin Bao, Ting Zhang, Hao Yang, Dong Chen, Fang Wen, and Baining Guo. Face x-ray for more general face forgery detection. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5000–5009, 2019. 1, 2, 5, 6

[22] Yuezun Li, Xin Yang, Pu Sun, Honggang Qi, and Siwei Lyu. Celeb-df: A large-scale challenging dataset for deepfake forensics. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3204–3213, 2019. 6, 5

[23] Honggu Liu, Xiaodan Li, Wenbo Zhou, Yuefeng Chen, Yuan He, Hui Xue, Weiming Zhang, and Nenghai Yu. Spatial-phase shallow learning: Rethinking face forgery detection in frequency domain. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 772–781, 2021. 2, 5, 6

[24] Yuhang Lu and Touradj Ebrahimi. A new approach to improve learning-based deepfake detection in realistic conditions, 2022. 2, 3, 4

[25] Yuhang Lu and Touradj Ebrahimi. Assessment framework for deepfake detection in real-world situations. *ArXiv*, abs/2304.06125, 2023. 2, 3, 4

[26] Yucheng Luo, Yong Zhang, Junchi Yan, and Wei Liu. Generalizing face forgery detection with high-frequency features. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16312–16321, 2021. 2

[27] Iacopo Masi, Aditya Killekar, Royston Marian Mascarenhas, Shenoy Pratik Gurudatt, and Wael AbdAlmageed. Two-branch recurrent network for isolating deepfakes in videos. *ArXiv*, abs/2008.03412, 2020. 2

[28] moreechampion.com.au. Fbi names shooter after trump assassination attempt. https://www.moreechampion.com.au/story/8694792/fbi-names-shooter-after-trump-assassination-attempt/, 2024. date accessed: 2024-11-07. 1

[29] Dat Nguyen, Nesryne Mejri, Inder Pal Singh, Polina Kuleshova, Marcella Astrid, Anis Kacem, Enjie Ghorbel, and Djamila Aouada. Laa-net: Localized artifact attention network for quality-agnostic and generalizable deepfake detection. In *Proceedings of the IEEE/CVF Conference on*

Computer Vision and Pattern Recognition, pages 17395–17405, 2024. 1, 2, 3, 5, 6, 7, 4

[30] Patrick Pérez, Michel Gangnet, and Andrew Blake. Poisson image editing. *ACM SIGGRAPH 2003 Papers*, 2003. 5

[31] Yuyang Qian, Guojun Yin, Lu Sheng, Zixuan Chen, and Jing Shao. Thinking in frequency: Face forgery detection by mining frequency-aware clues. *ArXiv*, abs/2007.09355, 2020. 2

[32] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models, 2021. 5

[33] Andreas Rössler, Davide Cozzolino, Luisa Verdoliva, Christian Riess, Justus Thies, and Matthias Nießner. FaceForensics++: Learning to detect manipulated facial images. In *International Conference on Computer Vision (ICCV)*, 2019. 1, 2, 3, 5, 6, 7, 4

[34] Ekraam Sabir, Jiaxin Cheng, Ayush Jaiswal, Wael AbdAlmageed, Iacopo Masi, and P. Natarajan. Recurrent convolutional strategies for face manipulation detection in videos. In *CVPR Workshops*, 2019. 1

[35] Ramprasaath R. Selvaraju, Abhishek Das, Ramakrishna Vedantam, Michael Cogswell, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. *International Journal of Computer Vision*, 128:336 – 359, 2016. 5, 1

[36] Kaede Shiohara and T. Yamasaki. Detecting deepfakes with self-blended images. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 18699–18708, 2022. 1, 2, 3, 4, 5, 6, 7

[37] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, D. Erhan, Ian J. Goodfellow, and Rob Fergus. Intriguing properties of neural networks. *CoRR*, abs/1312.6199, 2013. 3, 8

[38] Mingxing Tan and Quoc V. Le. Efficientnet: Rethinking model scaling for convolutional neural networks. *ArXiv*, abs/1905.11946, 2019. 3, 5, 6, 7, 4

[39] Justus Thies, Michael Zollhöfer, Marc Stamminger, Christian Theobalt, and Matthias Nießner. Face2face: Real-time face capture and reenactment of rgb videos. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2387–2395, 2016. 5, 3

[40] Justus Thies, Michael Zollhöfer, and Matthias Nießner. Deferred neural rendering. *ACM Transactions on Graphics (TOG)*, 38:1 – 12, 2019. 5, 3

[41] Sheng-Yu Wang, Oliver Wang, Richard Zhang, Andrew Owens, and Alexei A. Efros. Cnn-generated images are surprisingly easy to spot... for now. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8692–8701, 2019. 2

[42] Zhiyuan Yan, Yuhao Luo, Siwei Lyu, Qingshan Liu, and Baoyuan Wu. Transcending forgery specificity with latent space augmentation for generalizable deepfake detection. *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8984–8994, 2023. 2, 5, 6

[43] Zhiyuan Yan, Yong Zhang, Yanbo Fan, and Baoyuan Wu. Ucf: Uncovering common features for generalizable deepfake detection. *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 22355–22366, 2023. 2, 5, 6

[44] Zhi Yan, Yong Zhang, Xinhang Yuan, Siwei Lyu, and Baoyuan Wu. Deepfakebench: A comprehensive benchmark of deepfake detection. *ArXiv*, abs/2307.01426, 2023. 2, 5, 6

[45] Zhiyuan Yan, Taiping Yao, Shen Chen, Yandan Zhao, Xinghe Fu, Junwei Zhu, Donghao Luo, Li Yuan, Chengjie Wang, Shouhong Ding, et al. Df40: Toward next-generation deepfake detection. *arXiv preprint arXiv:2406.13495*, 2024. 6, 8, 5

[46] K. Zhang, Jingyun Liang, Luc Van Gool, and Radu Timofte. Designing a practical degradation model for deep blind image super-resolution. *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 4771–4780, 2021. 2, 3, 4, 6, 7

[47] Tianchen Zhao, Xiang Xu, Mingze Xu, Hui Ding, Yuanjun Xiong, and Wei Xia. Learning self-consistency for deepfake detection. *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 15003–15013, 2020. 2

[48] Wanyi Zhuang, Qi Chu, Zhentao Tan, Qiankun Liu, Haojie Yuan, Changtao Miao, Zixiang Luo, and Nenghai Yu. Uia-vit: Unsupervised inconsistency-aware method based on vision transformer for face forgery detection. *ArXiv*, abs/2210.12752, 2022. 2, 5, 6