F. Boutros, N. Damer, M. Fang, M. Gomez-Barrero, K. Raja,
C. Rathgeb, A. Sequeira, and M. Todisco (Eds.): BIOSIG 2024,
Lecture Notes in Informatics (LNI), Gesellschaft für Informatik, Bonn 2024

2D-Malafide: Adversarial Attacks Against Face Deepfake Detection Systems

Chiara Galdi[1], Michele Panariello[1], Massimiliano Todisco[1] and Nicholas Evans[1]

Abstract: We introduce 2D-Malafide, a novel and lightweight adversarial attack designed to deceive face deepfake detection systems. Building upon the concept of 1D convolutional perturbations explored in the speech domain, our method leverages 2D convolutional filters to craft perturbations which significantly degrade the performance of state-of-the-art face deepfake detectors. Unlike traditional additive noise approaches, 2D-Malafide optimises a small number of filter coefficients to generate robust adversarial perturbations which are transferable across different face images. Experiments, conducted using the FaceForensics++ dataset, demonstrate that 2D-Malafide substantially degrades detection performance in both *white-box* and *black-box* settings, with larger filter sizes having the greatest impact. Additionally, we report an explainability analysis using GradCAM which illustrates how 2D-Malafide misleads detection systems by altering the image areas used most for classification. Our findings highlight the vulnerability of current deepfake detection systems to convolutional adversarial attacks as well as the need for future work to enhance detection robustness through improved image fidelity constraints.

Keywords: deepfake detection, adversarial attacks, lightweight adversarial attacks, convolutional filters, image perturbations.

1 Introduction

In recent years, deep learning-based image recognition systems have achieved remarkable success across various applications, from face recognition to autonomous driving [Op24, Ja20]. However, these systems are vulnerable to adversarial attacks, namely deliberate manipulations designed to deceive the model [GJ20, VNR21]. Adversarial noise can typically be applied with subtle or seemingly insignificant perturbations to pixel values [GSS15], involving even only small portions of the image. The perturbations are specially crafted to exploit model vulnerabilities and provoke erroneous outputs. Even if the perturbed image is indistinguishable to the eye from the original image, there can be drastic influences upon the model output.

Most adversarial attacks involve additive noise, where image-specific perturbations are learned and directly added [Am23]. Fortunately, these approaches are unsuitable for real-time implementation and exhibit high sensitivity to the specific input image. Typically, these methods are trained and tested using the same set of deepfake data, with no assurances of effectiveness against *unseen* deepfakes — a property often referred to as *generalisability*. Some adversarial attacks, whether additive [SSP23] or involving spatial transformations [Zh20], have partially solved the problem of generalisation but come at the cost of high complexity.

[1] EURECOM, Digital Security Department, Sophia Antipolis, France, [firstname.lastname]@eurecom.fr

In this work, we propose the first adversarial attack which attempts to fulfil the *generalisability* property through convolutive noise while still being computationally lightweight. The former goal is met by optimising the adversarial perturbation over multiple samples. The latter is achieved by reducing the number of learnable parameters thanks to simple, yet effective modelling choices.

Building on a previous work, named Malafide [Pa23] which explored adversarial perturbation attacks against voice anti-spoofing solutions, we have tailored and implemented a novel adversarial attack named 2D-Malafide against image deepfake detection systems. This technique allows the attack to be mounted independently to the specific input image, and requires the optimisation of only a small number of filter coefficients. While the attack is agnostic to the type of classifier and image, e.g. be they face, fingerprint, or iris images, etc, in this paper we report its application specifically to face images and face deepfake detection.

Our experiments demonstrate that 2D-Malafide significantly degrades the performance of recent face deepfake detectors. The attack remains effective in both *white-box* settings, where the filter is specifically trained to manipulate a particular detector model, as well as *black-box* settings, and hence poses a substantial threat to the reliability of such detection systems.

2 Related Work

The concept of *adversarial attacks* against neural networks was originally introduced in [Sz14, GSS15] in the context of image classification tasks. The term usually refers to the introduction of perturbations to the input image of a neural network so as to manipulate the output or decision. Such perturbations can be crafted by optimising the pixel values of the input image via a gradient descent-based technique to maximise the output probability of an arbitrary, incorrect class.

Adversarial attacks have since been explored in a wide variety of different domains, including deepfake detection. Early investigations showed that deepfakes can be rendered undetectable by deepfake detection algorithms using specially crafted adversarial perturbations [CF20, GJ20, Hu21, Ji22]. However, these studies focused on crafting individual adversarial perturbations for each deepfake sample, a computationally intensive process.

More recent adversarial attack techniques have since been proposed to overcome this issue. The authors of [FHD24] proposed the use of generative adversarial networks (GANs) [Go20] to produce adversarial attacks for arbitrary deepfake samples. In [Ho23], adversarial perturbations are modelled as a linear combination of image transformations whose weights are optimised across multiple deepfake images in order to minimise the chances of detection. Using a similar objective function, the work in [Ne21] demonstrates how a video deepfake detection system can be manipulated by using a single layer of additive noise with bounded amplitude applied to each image frame.

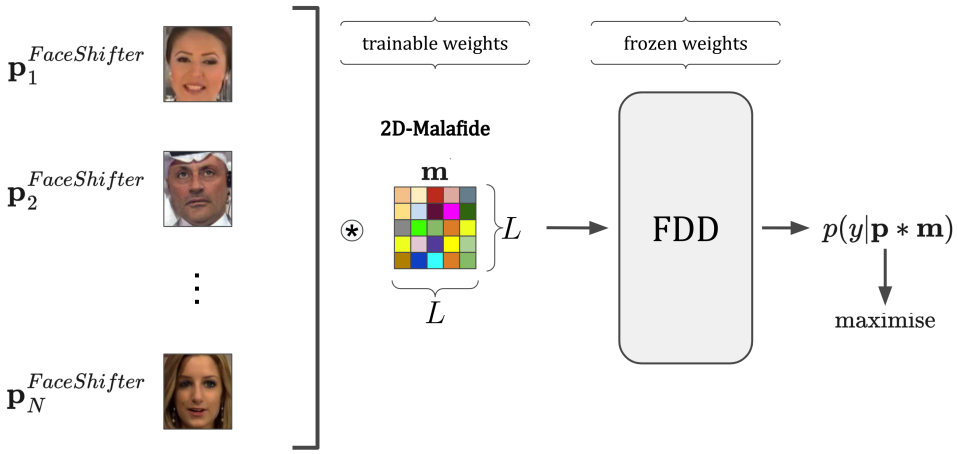

Fig. 1: The training procedure of the 2D-Malafide filter $\mathbf{m}$ for face images generated with the attack $a =$ FaceShifter against the face deepfake detector (FDD).

To the best of our knowledge, the only work that explores the generation of *generalisable* adversarial perturbations against deepfake detectors is [Li24]. The authors propose a GAN-based technique to produce shadows which are introduced to an image deepfake to conceal generated artefacts. Nonetheless, this technique involves the training of two generative neural networks and requires considerable computing capabilities.

3 2D-Malafide

In this section we describe the adaptation and implementation of 2D-Malafide for adversarial attacks against face deepfake detection (FDD) systems.

Let $\mathbf{P}^{(a)} = \{\mathbf{p}_1^{(a)}, \mathbf{p}_2^{(a)} \dots \mathbf{p}_N^{(a)}\}$ be a set of deepfake/spoofed images generated by algorithm a. Each image is designed to deceive a deepfake detection system to increase the likelihood of false accept decisions. Let $\text{FDD}(\mathbf{I}) = s(y \mid \mathbf{I})$ be a deepfake detector model which assigns a score y to image $\mathbf{I}$, where higher scores reflect greater support for the bona fide class and lower scores for the deepfake class. For spoofed images $\mathbf{p}_i^{(a)}$, $\text{FDD}(\mathbf{p}_i^{(a)})$ should hence produce low scores. 2D-Malafide attacks involve the optimisation of a 2D linear time-invariant (LTI), non-causal filter. The coefficients are optimised to provoke the misclassification of deepfake images as bona fide. The 2D LTI, $L \times L$ filter $\mathbf{m}^{(a)}$ is designed to maximise $\text{FDD}(\mathbf{p}_i^{(a)} * \mathbf{m}^{(a)})$, where $*$ denotes the 2D convolution operator. In the case of several different deepfake algorithms $a_1 \dots a_K$, an attacker can optimise an equivalent number of filters $\mathbf{m}^{(a_1)} \dots \mathbf{m}^{(a_K)}$. The filter should then be tuned to counter the reliance of the FDD system upon attack-specific artefacts. Filter coefficients $\mathbf{m}^{(a)}$ can be optimised with conventional gradient descent using the set of spoofed images $\mathbf{P}^{(a)}$. The objective function is given by

$$\max_{\mathbf{m}^{(a)}} \sum_i \text{FDD}(\mathbf{p}_i^{(a)} * \mathbf{m}^{(a)}) \tag{1}$$

A graphical depiction of the training procedure is shown in Fig. 1 for an attack $a =$ FaceShifter. The filter is optimised independently for each attack so as to manipulate the behaviour of a common FDD.

Without constraints, 2D-Malafide filtering can cause excessive image degradation. For detection settings in the absence of a human observer, this may have little consequence. However, where the FDD system is deployed alongside other systems, the distortion introduced to compromise the FDD system might also interfere with the behaviour of any other auxiliary system, e.g. an automatic face recognition system. In this case, for instance, it might even *improve* its resistance to attack, e.g. if image quality is significantly degraded.

Accordingly, $\mathbf{m}^{(a)}$ should be constrained to balance the maximisation of (1) and the preservation of image fidelity, e.g., clarity, detail, or key features. This can be achieved by tuning the filter size $L \times L$. Larger filters allow for greater manipulation and stronger attacks but can also introduce greater distortion. Conversely, smaller filters can be configured so that they introduce less distortion at the expense of a weaker attack. We apply image normalisation after filtering in order to ensure that pixel values do not surpass the maximum quantisation level.

4 Experimental Setup

All experiments were conducted using the FaceForensics++ (FF++) dataset [Ro19]. It contains 1000 bona fide videos in addition to 5000 corresponding fakes generated with 5 different algorithms. The first two are computer graphics-based approaches. **Face2Face** [Th18] is a facial reenactment system which transfers expressions from a source video to a target video while retaining the target face identity. **FaceSwap** [Ko24] transfers the face region from a source to a target video using facial landmarks to fit a 3D model which is then back-projected, blended, and colour corrected. There are three deep learning-based approaches. The first, Deepfakes, was implemented using the open-source implementation *deepfakes faceswap*[3] and requires training with a pair of videos of source and target subjects. The second, **NeuralTextures** [TZN19], learns a neural texture of the target person using a photometric reconstruction loss combined with an adversarial loss for training. The last is the two-stage face-swapping method **FaceShifter** [Li20] which uses a pair of input images (a source for identity and a target for attributes like pose and expression) and a two-stage framework (AEINet and HEARNet) for high-fidelity face swaps.

Although the FF++ dataset contains videos, the selected FDD systems operate on individual frames hence, in the remainder of this paper, mentions of the dataset refer to the collection of *frames* extracted from FF++ videos. The attacker is assumed to have access only to the test partition of the dataset. Thus, the FF++ test partition was used for training and testing 2D-Malafide attacks. Attack-specific filters were trained according to (1), using subsets of FF++ for each deepfake method. The FF++ test partition was split into 70% for training (Part 1) and 30% for testing (Part 2), with 1399 images in Part 1 and 599 images in Part 2. 2D-Malafide filters were trained using Part 1 and tested using Part 2. This setup

[3] `https://github.com/deepfakes/faceswap`

simulates offline filter training and online attacks. 2D-Malafide filters were trained using only deepfake images.

Each attack-specific 2D-Malafide filter is trained using the Adam algorithm [24]. The learning rate and weight decay are tuned separately for each FDD system. The maximum number of epochs is set to 100 since, for all but a single experiment, training reaches the stop condition before 100 epochs, where the stop condition is defined by an equal error rate (EER) in excess of 50%. The resulting filter is then applied to Part 2 for evaluation. A batch size of 32 was chosen because it was suitable for the GPUs used for our experiments. During optimisation of 2D-Malafide, the weights of the FDD pre-trained models are frozen. We explored different filter sizes $L = (3, 9, 27, 81)$ in order to analyse the impact on performance. Our implementation is available as open-source and can be used to reproduce our results.[4]

To determine the effectiveness of the adversarial filter attack we used the following two FDD systems.

CADDM [Do23][5] is a deepfake detection system developed to address the problem of *Implicit Identity Leakage*. The authors observed that deepfake detection models supervised using only binary labels are sensitive to identity. Thus, they propose a method, termed an *ID-unaware Deepfake Detection Model*, to reduce the influence of the identity representation. This is achieved by guiding the model to focus on local rather than global (whole image) features. Intuitively, by forcing the model to focus only on local areas of the image, less attention will be paid to global identity information.

Self-Blended Images (SBIs) [SY22][6] is a deepfake detection system which leverages training data augmentation to improve generalisability. The key idea behind SBIs is that the use of more general and barely recognisable fake samples encourage classifiers to learn generic and robust representations without overfitting to manipulation-specific artefacts. Fake samples are generated by blending pairs of *pseudo source* and *target* images, obtained using different image augmentation transformations, thereby increasing the difficulty of the face forgery detection task and encouraging the learning of more generalisable models.

The implementations of both CADDM and SBIs used in this work support the use of different backbone architectures. For our experiments, both methods use EfficientNet convolutional neural networks, the only difference being that we use efficientnet-B3 for CADDM, but efficientnet-B4 for SBIs. Models pre-trained using the FF++ training dataset are used for both methods and are available on the respective GitHub repositories.

5 Experimental Results

Results presented in Table 1 show EER values for CADDM and SBI FDD systems with and without the application of 2D-Malafide filters under *white-box* (tested using the same

[4] `https://github.com/eurecom-fscv/2D-Malafide`
[5] `https://github.com/megvii-research/CADDM`
[6] `https://github.com/mapooon/SelfBlendedImages`

Tab. 1: Comparison in terms of EER [%] of the baseline performance without filtering and the performance of different sizes of 2D-Malafide filters under *white-box* (trained and tested on the same FDD) and *black-box* (tested on different FDDs) settings. The results are shown for five attack types.

Baseline Deepfake Detection System - CADDM (C) / SBI (S)										
Attack type	Deepfakes		Face2Face		FaceShifter		FaceSwap		NeuralTextures	
FDD	C	S	C	S	C	S	C	S	C	S
No filter	0.00	0.71	1.34	1.43	1.34	7.14	0.67	1.43	2.50	5.00

2D-Malafide trained on CADDM and tested on CADDM / SBI - (W)hite box / (B)lack box										
Filter size	W	B	W	B	W	B	W	B	W	B
3x3	3.17	6.51	2.83	5.33	2.83	6.34	4.34	9.68	4.84	6.34
9x9	3.17	7.34	7.50	8.84	6.49	4.66	8.68	9.02	6.68	7.68
27x27	46.41	8.01	49.83	7.16	50.17	7.16	46.41	7.68	51.92	6.01
81x81	47.08	7.34	55.50	10.33	64.00	2.16	48.08	7.68	62.10	4.51

2D-Malafide trained on SBI and tested on SBI / CADDM - (W)hite box / (B)lack box										
Filter size	W	B	W	B	W	B	W	B	W	B
3x3	6.18	3.17	13.17	2.83	11.00	2.83	8.01	3.67	13.86	4.51
9x9	13.86	1.34	28.83	1.50	34.17	0.67	31.39	2.00	33.06	2.84
27x27	6.85	2.01	40.17	2.83	43.34	0.67	39.40	3.34	45.24	3.50
81x81	29.05	3.17	26.67	2.83	45.99	2.83	30.05	5.01	29.22	3.84

countermeasure used for 2D-Malafide training) and *black-box* (tested using an unseen countermeasure) settings. Results are shown separately for the baseline FDD system (top block), then 2D-Malafide attacks trained using CADDM (middle) and SBI (bottom). Baseline FDD results show detection error rates for the five different attack types.

For the CADDM *white-box* setting (denoted W in Table 1), the application of 2D-Malafide filters leads to a significant increase in EER, especially with larger filter sizes (27×27 and 81×81). This indicates a substantial degradation in FDD performance, demonstrating the effectiveness of the adversarial filters in deceiving the detection system. For the corresponding *black-box* setting, for which the model is trained using CADDM but tested using SBI, results show that most filters provoke an increase in the baseline EER. However, in some cases (highlighted in red), filtering instead reduces the EER, indicating that they made it easier for the FDD system to detect the underlying attack.

For the SBI *white-box* setting the 2D-Malafide filters again lead to notable increases in the EER, particularly for the 27×27 filter size. We note that, for the 81×81 filter, the EERs decrease slightly, showing that the largest filter size is less effective for SBI than for CADDM. For the corresponding *black-box* setting, filtering generally increases the baseline EER. However, the impact is less pronounced compared to CADDM, indicating that adversarial training performed using SBI does not generalise well.

Overall, results indicate that FDD systems are vulnerable to 2D-Malafide attacks, with the greatest impact observed under *white-box* settings. The impact varies with filter size. Larger filters (27×27 and 81×81) tend to cause the most significant degradation in de-

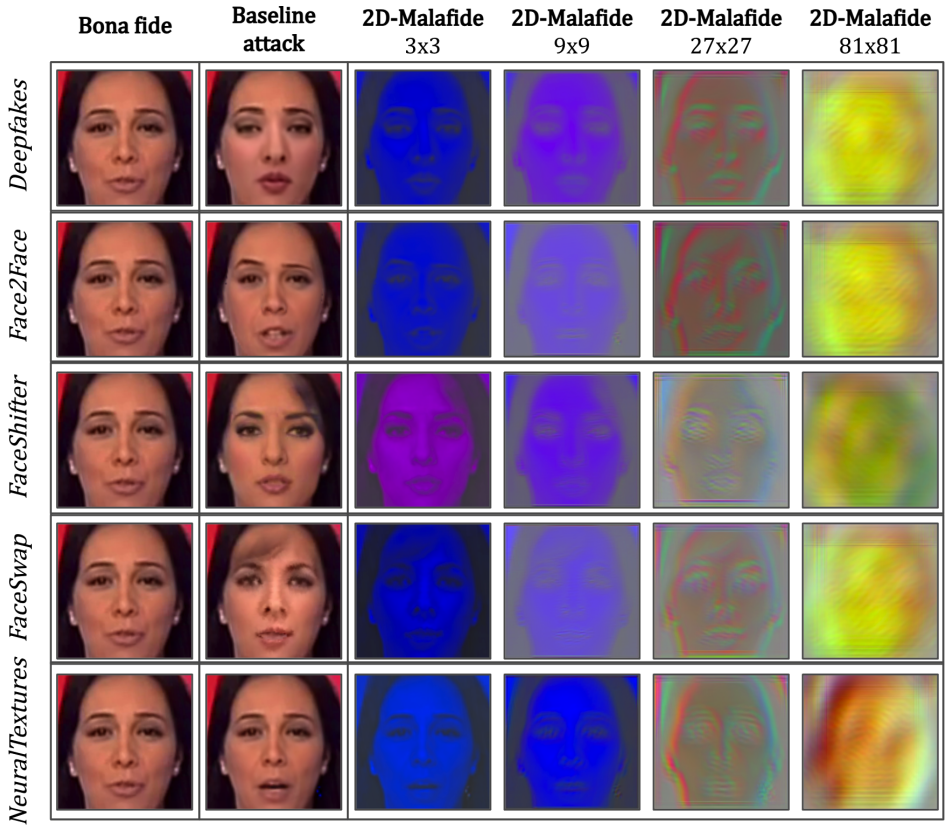

Fig. 2: Examples of bona fide, baseline attack and four configurations of 2D-Malafide filter for the five deepfake attacks. Results are taken from training based on CADDM system.

tection performance, particularly for CADDM. Under *black-box* settings, while filtering generally provokes an increase in error rates, there are instances where detection performance improves, suggesting that adversarial filtering does not always generalise well to unseen detectors.

Last, Fig. 2 shows a comparison of bona fide images, the corresponding attacks and then after application of four different 2D-Malafide filters, for the CADDM FDD system. For smaller filter sizes, the face is still recognisable even if the colours are unnatural. For larger filters, the face is significantly distorted or even unrecognisable. This finding in itself highlights a critical limitation in face deepfake detection in that they can be compromised so easily with images which do not even resemble natural faces. This raises concerns about the robustness of such systems when dealing with altered or degraded images, with obvious implications for both the security and reliability of the technology.

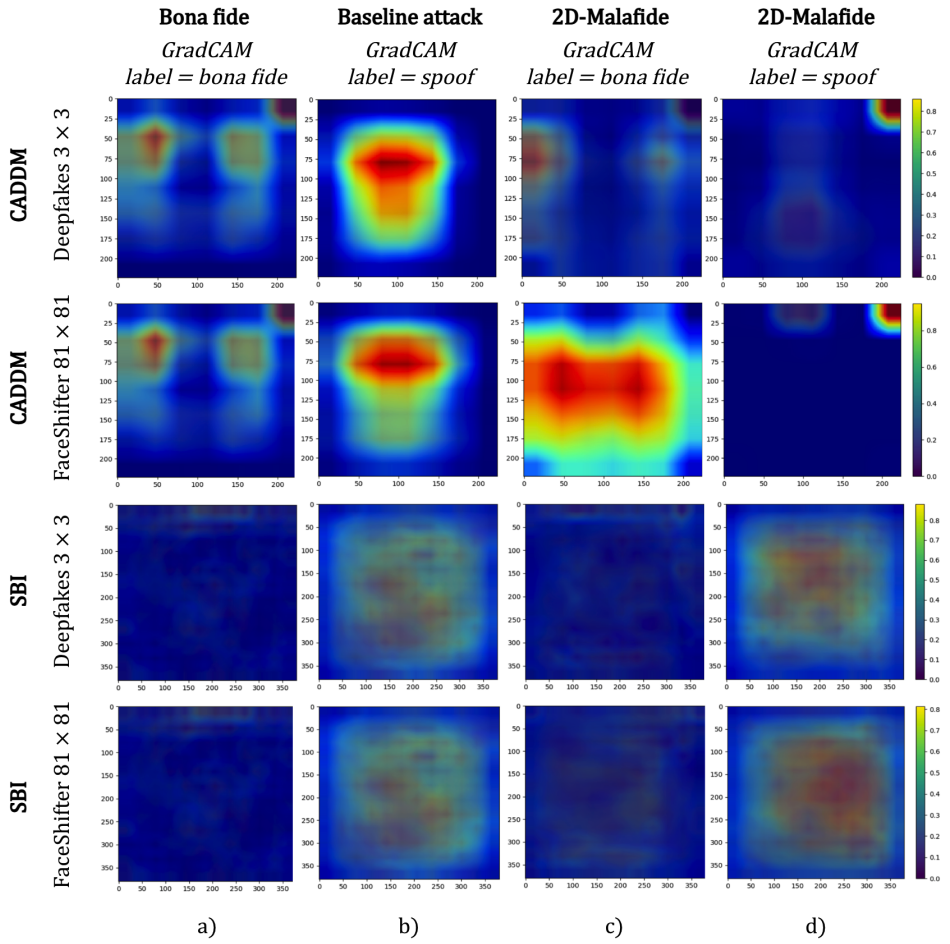

Fig. 3: GradCAM explainability results for Deepfakes 3×3 and FaceShifter 81×81 image samples classified with CADDM and SBI FDD systems applied on bona fide (a), baseline attack (b), and 2D-Malafide attacks processed with GradCAM label bona fide (c) and spoof (d).

6 Explainability Analysis

In order to gain deeper insights into the impact of 2D-Malafide filtering upon the deepfake detectors, we also report an explainability analysis. We report the GradCAM [Gc21] heatmaps for a pair of different attacks and filter sizes when using the CADDM and SBI detectors, specifically Deepfakes 3×3 and FaceShifter 81×81. GradCAM is applied to each image in the test set and for each category: *bona fide*, *spoof*, *spoof + malafide*. The resulting heatmaps are averaged to show predominant activation patterns.

The heatmaps in Fig. 3 indicate the areas of the face images where the model focuses its attention according to the input label, hence revealing features relevant to either *bona fide*

or *spoof* classes. The first row of Fig. 3 shows results for CADDM and the Deepfakes 3×3 attack. The left-most heatmap in column (a), shows that significant facial landmarks which correspond to the contours of the face are most informative for the classification of images as bona fide. In constrast, in the case of fake images, shown in column (b), facial landmarks corresponding to areas of the eyes and eyebrows are most informative. Visual inspections reveal that these areas often correspond to visible artefacts, e.g. double eyebrows, resulting from the application of Deepfakes.

Heatmaps in columns (c) and (d) display results after application of 2D-Malafide and for fake face images when using *bona fide* and *spoof* labels respectively. Whereas heatmaps (a) and (c) exhibit similar patterns, heatmaps (b) and (d) are notably different. 2D-Malafide hides fake image artefacts upon which the detector relies, namely those in the central part of the face. There are no obvious activations in this area in heatmap (d), hence why the model is misled into classifying the fake as bona fide.

The second row of Fig. 3 shows results for FaceShifter 81×81 attacks, again for CADDM. The heatmap in column (c) shows that the CADDM model focuses on the sides of the face image, but with greater intensity than for bona fide images. Heatmap (d) remains similar to that for the Deepfakes 3×3 attack. Not only does 2D-Malafide hide fake artefacts, it also provokes a greater rate in the misclassification of fake images by causing the detector to focus more on sides of the face. This finding accounts for results reported in Table 1, in particular cases for which 2D-Malafide is more efficient the largest filter size. The dominant spot to the upper right might be due to the Multi-scale Detection Module (MSDM) of the CADDM architecture. The MSDM uses predefined anchor boxes which are tiled across the image. The level of activations in this area might correspond to the location of the last analysed anchor box.

Heatmaps in rows 3 and 4 of Fig. 3 show results for the SBI detector. For the Deepfakes 3×3 attack, the detector focuses on small parts of bona fide images at different positions, hence the seemingly flat heatmap. In contrast and in the case of fakes, the model focuses predominantly on central areas of the face, albeit in a less localised manner compared to CADDM. After application of 2D-Malafide filtering, there are few differences between results for bona fide images (a) and filtered bona fide images (c), and also between those for fakes (b) and filtered fakes (d). However, a closer look revels how attention for attacks without filtering, shown in column (b), is more concentrated to the bottom left of the central part of the face. Instead, for fake images processed by 2D-Malafide, attention is concentrated more to the top right, and more so for FaceShifter 81×81 attacks.

7 Conclusions

In this article we introduce 2D-Malafide, an adversarial attack which uses 2D convolutional filtering to deceive face deepfake detection systems. The attack significantly increases the EER of state-of-the-art deepfake detectors in both *white-box* and *black-box* settings and highlights the vulnerability of current FDD systems to such attacks. Larger filters (27×27 and 81×81) cause substantial performance degradation. Moreover, the

generalisability of 2D-Malafide ensures robustness across various image inputs, making for a versatile threat. Colour information is the first to be impacted by the application of 2D-Malafide showing that the FFDs considered in this work fail to recognise simple, even unnatural changes in colour.

GradCAM explainability analysis reveals that 2D-Malafide misleads FDD systems by altering the areas of an image they use for classification, thereby increasing false acceptance rates. Attack success varies across different FDD systems, indicating some level of generalisability but also a dependency on the specific architecture.

The results emphasise the need for comprehensive and diverse training datasets to improve FDD robustness. Future research should focus on enhanced image fidelity constraints, including colour consistency, to counter such adversarial attacks. Overall, 2D-Malafide demonstrates the critical need for ongoing advancements in FDD technology to ensure the security and reliability of deepfake detection systems.

References

[Am23] Ambati, Rahul; Akhtar, Naveed; Mian, Ajmal; Rawat, Yogesh S: PRAT: PRofiling Adversarial a Ttacks. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. S. 3667–3676, 2023.

[CF20] Carlini, Nicholas; Farid, Hany: Evading Deepfake-Image Detectors with White- and Black-Box Attacks. In: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). S. 2804–2813, 2020.

[Do23] Dong, Shichao; Wang, Jin; Ji, Renhe; Liang, Jiajun; Fan, Haoqiang; Ge, Zheng: Implicit Identity Leakage: The Stumbling Block to Improving Deepfake Detection Generalization. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). S. 3994–4004, June 2023.

[FHD24] Fan, Bing; Hu, Shu; Ding, Feng: Synthesizing Black-Box Anti-Forensics Deepfakes With High Visual Quality. In: ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). S. 4545–4549, 2024.

[Gc21] Gildenblat, Jacob; contributors: PyTorch library for CAM methods. `https://github.com/jacobgil/pytorch-grad-cam`, 2021.

[GJ20] Gandhi, Apurva; Jain, Shomik: Adversarial Perturbations Fool Deepfake Detectors. In: 2020 International Joint Conference on Neural Networks (IJCNN). S. 1–8, 2020.

[Go20] Goodfellow, Ian; Pouget-Abadie, Jean; Mirza, Mehdi; Xu, Bing; Warde-Farley, David; Ozair, Sherjil; Courville, Aaron; Bengio, Yoshua: Generative adversarial networks. Commun. ACM, 63(11):139–144, oct 2020.

[GSS15] Goodfellow, Ian J.; Shlens, Jonathon; Szegedy, Christian: Explaining and Harnessing Adversarial Examples. In (Bengio, Yoshua; LeCun, Yann, Hrsg.): 3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings. 2015.

[Ho23] Hou, Yang; Guo, Qing; Huang, Yihao; Xie, Xiaofei; Ma, Lei; Zhao, Jianjun: Evading DeepFake Detectors via Adversarial Statistical Consistency. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). S. 12271–12280, June 2023.

[Hu21] Hussain, S.; Neekhara, P.; Jere, M.; Koushanfar, F.; McAuley, J.: Adversarial Deepfakes: Evaluating Vulnerability of Deepfake Detectors to Adversarial Examples. In: 2021 IEEE Winter Conference on Applications of Computer Vision (WACV). IEEE Computer Society, Los Alamitos, CA, USA, S. 3347–3356, jan 2021.

[Ja20] Janai, Julius; Guney, Fatma; Ranjan, Anurag; Black, Michael J.; Geiger, Andreas: Computer Vision for Autonomous Vehicles: Problems, Datasets and State-of-the-Art. Foundations and Trends in Computer Graphics and Vision, 12(1–3):1–308, 2020.

[Ji22] Jia, Shuai; Ma, Chao; Yao, Taiping; Yin, Bangjie; Ding, Shouhong; Yang, Xiaokang: Exploring Frequency Adversarial Attacks for Face Forgery Detection. In: 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). S. 4093–4102, 2022.

[Ko24] Kowalski, Marek: GitHub repository - MarekKowalski/FaceSwap, August 2024. original-date: 2016-06-19T00:09:07Z.

[Li20] Li, Lingzhi; Bao, Jianmin; Yang, Hao; Chen, Dong; Wen, Fang: Advancing High Fidelity Identity Swapping for Forgery Detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). June 2020.

[Li24] Liu, Jiatong; Zhang, Mingcheng; Ke, Jianpeng; Wang, Lina: AdvShadow: Evading DeepFake Detection via Adversarial Shadow Attack. In: ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). S. 4640–4644, 2024.

[Ne21] Neekhara, Paarth; Dolhansky, Brian; Bitton, Joanna; Ferrer, Cristian Canton: Adversarial Threats to DeepFake Detection: A Practical Perspective. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops. S. 923–932, June 2021.

[Op24] Opanasenko, Volodymyr Mykolaevich; Fazilov, Shavkat Khayrullaevich; Mirzaev, Olimjon Nomazovich; Kakharov, Shukrullo Sa'dullo ugli: An Ensemble Approach To Face Recognition In Access Control Systems. Journal of Mobile Multimedia, 20(03):749–768, May 2024.

[Pa23] Panariello, Michele; Ge, Wanying; Tak, Hemlata; Todisco, Massimiliano; Evans, Nicholas: Malafide: a novel adversarial convolutive noise attack against deepfake and spoofing detection systems. In: Proc. INTERSPEECH 2023. S. 2868–2872, 2023.

[Ro19] Rossler, Andreas; Cozzolino, Davide; Verdoliva, Luisa; Riess, Christian; Thies, Justus; Nießner, Matthias: Faceforensics++: Learning to detect manipulated facial images. In: Proceedings of the IEEE/CVF international conference on computer vision. S. 1–11, 2019.

[SSP23] Stanly, Hamil; S., Mercy Shalinie; Paul, Riji: A review of generative and non-generative adversarial attack on context-rich images. Engineering Applications of Artificial Intelligence, 124:106595, 2023.

[SY22] Shiohara, Kaede; Yamasaki, Toshihiko: Detecting Deepfakes with Self-Blended Images. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. S. 18720–18729, 2022.

[Sz14] Szegedy, Christian; Zaremba, Wojciech; Sutskever, Ilya; Bruna, Joan; Erhan, Dumitru; Goodfellow, Ian J.; Fergus, Rob: Intriguing properties of neural networks. In (Bengio, Yoshua; LeCun, Yann, Hrsg.): 2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Conference Track Proceedings. 2014.

[Th18] Thies, Justus; Zollhöfer, Michael; Stamminger, Marc; Theobalt, Christian; Nießner, Matthias: Face2Face: real-time face capture and reenactment of RGB videos. Commun. ACM, 62(1):96–104, dec 2018.

[TZN19] Thies, Justus; Zollhöfer, Michael; Nießner, Matthias: Deferred neural rendering: image synthesis using neural textures. ACM Trans. Graph., 38(4), jul 2019.

[VNR21] Vakhshiteh, Fatemeh; Nickabadi, Ahmad; Ramachandra, Raghavendra: Adversarial Attacks Against Face Recognition: A Comprehensive Study. IEEE Access, 9:92735–92756, 2021.

[Zh20] Zhang, Yanghao; Ruan, Wenjie; Wang, Fu; Huang, Xiaowei: Generalizing Universal Adversarial Attacks Beyond Additive Perturbations. In: 2020 IEEE International Conference on Data Mining (ICDM). S. 1412–1417, 2020.