https://doi.org/10.2352/EI.2024.36.4.MWSF-332

One Detector to Rule Them All? On the Robustness and Generalizability of Current State-of-the-Art Deepfake Detection Methods

Raphael Antonius Frick & Martin Steinebach; Fraunhofer SIT|ATHENE Center; Darmstadt, Germany

Abstract

With the advancements made in the field of artificial intelligence (AI) in recent years, it has become more accessible to create facial forgeries in images and videos. In particular, face swapping deepfakes allow for convincing manipulations where a person's facial texture can be replaced with an arbitrary facial texture with the help of AI. Since such face swapping manipulations are nowadays commonly used for creating and spreading fake news and impersonation with the aim of defamation and fraud, it is of great importance to distinguish between authentic and manipulated content. In the past, several methods have been proposed to detect deepfakes. At the same time, new synthesis methods have also been introduced. In this work, we analyze whether the current state-of-the-art detection methods can detect modern deepfake methods that were not part of the training set. The experiments showed, that while many of the current detection methods are robust to common post-processing operations, they most often do not generalize well to unseen data.

Introduction

In recent years, the rapid advancement of deep learning techniques, particularly Generative Adversarial Networks (GANs) [1], Variational Auto-encoders (VAEs) [2] and Diffusion Models [3], has led to the emergence of highly realistic and believable fabricated content. Among these technologies, deepfakes stand out as a powerful tool for creating synthetic media that can deceive the human eye.

The term *deepfake* is a fusion of the terms *deep* and *fake*. Hereby, deep refers to the application of deep learning, whereas fake refers to the synthesis and manipulation of media. It encompasses various modalities, such as images [4], videos [5], audio [6], and even text [7]. Deepfakes have both positive applications, such as enhancing creative arts and film production, and negative implications, including the spread of misleading information. Images and videos are frequently shared on the internet alongside written text to provide additional context and by this can convey a certain trust to an underlying story of an article or a social media post.

Face swaps are a specific type of deepfake, where the face inside an image or video can be exchanged with any desired face with the help of a neural network. Thus, they are especially at risk of being utilized with malicious intent. During the Ukraine-War, deepfakes have been used to produce fake videos showcasing the presidents of both nations surrendering. Thus, images and videos shared on social media or by unverified sources should not be trusted blindly. Further, it shows the need for methods that allow to verify the authenticity of multimedia.

Over the past years, several approaches to detecting deepfakes have been proposed. These can be divided into data-driven [8, 9, 10] and model-based approaches [11, 12]. While the former try to extract all the relevant information to distinguish between authentic and non-authentic content automatically from given raw input data, the latter aim to take advantage of handcrafted features. While these methods are able to achieve a high accuracy on scientific data sets, they often seem to fail to generalize well to unseen data in the wild. In Facebook's Deepfake Detection Challenge from 2020[8] the best performing approach was able to achieve 82.56% accuracy on the public test data set, yet only an accuracy of 65.18% during the evaluation on the private test data set. Another challenge is, that many classifiers are not robust against subtle image post-processing, such as the introduction of random noise or heavy compression.

In this paper, we evaluate a set of model-based and data-driven state-of-the-art methods to assess, whether they can generalize to unseen data and are at the same time robust to various common post-processing operations. For this, we augment the FaceForensics++ dataset [13] to undergo various post-processing steps. Moreover, we create a new test dataset containing videos from current-state-of-the-art one-shot deepfake methods and deepfake models used for real-time face swaps.

Face Swapping Deepfakes

Introduced in 2017, deepfakes involving face swaps were the first type of deepfakes to be developed. Although various new face swapping techniques have been proposed since then, the architecture presented in the original implementation is still very relevant today, as it forms the basis for many popular deepfake algorithms [5].

It consists of a customized autoencoder architecture that consists of a single encoder network and a set of two decoder networks. During model training, the encoder network creates embeddings from the input images that contains information on (a) the identity A present in the authentic footage and (b) the identity B to be inserted into the target image and video. For each of the identities, there is an identity-specific decoder network that attempts to reconstruct the facial images based on the embeddings provided by the encoder network. By swapping the decoder networks during inference, a face image with identity A fed into the decoder trained on identity B leads to the synthesis of an image with identity B, whereby the facial expression corresponds to that of the input image. After the synthesis, the generated face texture is inserted into the target medium. Hereby, facial landmarks [14] are used to determine the position of the face and its components. Optionally, color grading and adjustments can be made to the

mask, which determines which parts of the face texture are to be transferred, so that the generated face can be better blended into the target image frame.

By incorporating several optimizations to the trained deepfake models, this architecture can also be used in real-time [1], e.g., during online meetings. However, one major disadvantage of the architecture is, that it requires a vast number of images during training. In general, over 3000–5000 images are required for each of the identities displayed. These pictures should show the people depicted with different facial expressions, in different head positions and in different lighting conditions. Obtaining such images may not always be feasible. As such, over the past years, methods have been proposed that allow generating face swapping deepfakes using a single image [15][2].

Deepfake Detection

In this paper, we compare various state-of-the-art deepfake detection models regarding their robustness against common post-processing and their generalizability towards unseen data. For this, we consider a set of model-based and data-driven approaches. The selection of approaches was based on the availability of the source code, the date published, and the objectives (e.g., achieving cross-dataset generalizability or robustness).

Model-based Approaches

Model-based approaches take advantage of handcrafted features in order to distinguish between authentic and non-authentic content. These are usually derived from signal processing and consist of artifacts that are absent, modified or introduced within the manipulated material due to the deepfake creation process. In this paper, we focus on a technique analyzing geometric features derived from facial landmarks as well as a detection method that takes advantage of biophysical signals.

Improving the Efficiency and Robustness of Deepfakes Detection through Precise Geometric Features

The paper *Improving the Efficiency and Robustness of Deepfakes Detection through Precise Geometric Features* [11] proposes LRNet, a framework for detecting deepfakes videos using geometric features. In the past, most deepfakes video detection techniques analyzed the appearance of the displayed faces using data-driven models. However, these classification models can become subject to adversarial attacks which result in forced misclassifications [16]. By this, the detection can be bypassed. The paper addresses this by proposing a model-based approach taking advantage of the analysis of the temporal consistency of geometric features. In particular, it aims at exposing manipulated faces by detecting abnormal facial movement patterns and time discontinuities.

The framework consists of three components. The *face preprocessing* module is used to detect faces within a video, extract their facial landmarks, and to align the found faces. The authors identified, that the facial landmarks are prone to jitter in consecutive frames, even in cases where no head movement is involved. Thus, the *calibration* module is used to refine the found facial landmarks. By feeding the computed optical flows into a Kalman filter [17], the landmark coordinates can be denoised. The calibrated facial features and their difference along successive images are then fed into a two-stream recurrent neural network (RNN). The model is then trained to distinguish between authentic and forged faces.

In experiments, the model achieved 0.999 AUC on the FaceForensics++ dataset [13]. In addition, it also performed well on videos from an unseen dataset without retraining. However, the model's performance declined on videos that were part of the Celeb-DF dataset [18] with an AUC score of 0.569. However, it showed overall robustness towards noise and strong video compression.

DeepFakesON-Phys: DeepFakes Detection based on Heart Rate Estimation

The paper *DeepFakesON-Phys: DeepFakes Detection based on Heart Rate Estimation* [12] introduces a deepfake detection framework based on physiological measurement. As face swaps are generated for each frame separately, they are prone to artifacts with regard to temporal consistency. In this case, the paper leverages from remote photoplethysmography (rPPG) to analyze video sequences for subtle color changes in the human skin, revealing the presence of human blood under the tissues. Faces featuring irregular blood flow are thereby classified as fake.

DeepFakesON-Phys employs a convolutional attention network (CAN) to extract spatial and temporal information from video frames. By this, the model not only captures motion information, but also information about the facial appearance. These are then used to determine the overall blood flow and whether the video containing faces has been subject to a manipulation.

The model was able to achieve AUC scores of 0.98 and beyond on the Celeb-DF [18] and DFDC datasets [19], and thereby outperformed other methods tested.

Data-driven Approaches

Unlike model-based approaches, data-driven models can derive all the relevant information required for detecting deepfakes automatically during model training. They usually outperform model-based approaches, but often suffer from various issues, including generalizability, robustness and absence of explainability.

DeepFake Detection Challenge (DFDC) Solution

In 2020, Facebook held a deepfake detection competition [8] which was based on the data of the DFDC dataset [19]. In the competition, the solution proposed by Selim Seferbekov was able to place first. The proposed solution conducts a frame-per-frame detection that is based on the EfficientNet-B7 [20] network architecture that has been fine-tuned for the classification task. During training, a variety of augmentations on the data were applied, including cutouts, the introduction of blur and noise. In addition, an ensemble classifier consisting of seven weak models was used to further enhance the detection performance.

AltFreezing for More General Video Face Forgery Detection

Wang et al. [9] propose architectural modifications to data-driven detection methods so that they can generalize better to unseen data. Based on a 3D-convolutional neural network, which not only assesses spatial but also temporal information, they in-

[1] DeepFaceLive: https://github.com/iperov/DeepFaceLive

[2] FaceFusion: https://github.com/facefusion/facefusion

troduce a method called AltFreezing. It is used to ensure that more artifacts are considered in the classification process, thereby achieving higher generalizability. Weights are divided into two groups, which are alternately frozen during the training process. To focus the learning process on both, spatial and temporal features, the groups are built in such a way that those usually concerned about spatial-related information are grouped together, whereas those concerning temporal feature build another group. Since data augmentation are usually only applied on a per-frame basis, they solely affect the spatial-related features. Thus, the authors present a selection of data augmentation techniques on video level so that temporal feature extraction can be enhanced.

In their experiments, their model was able to achieve AUC scores between 0.895 to 0.993 on various datasets including Celeb-DF v2 [18] and Facebook's DFDC dataset [8], even when solely trained on data of the FaceForensics++ dataset [13]. Due to their data augmentation techniques, their model was able to be more robust against post-processing operations than the related work tested against during evaluation.

Detecting Deepfakes with Self-Blended Images

The paper *Detecting Deepfakes with Self-Blended Images* [10] presents a novel synthetic training data called self-blended images (SBIs) for detecting deepfakes. It takes advantage of the fact, that the generated face texture needs to be merged with the target frame after its synthesis. As such, synthetic training material is generated by blending pseudo source and target images from single pristine images. By this forgery artifacts such as blending boundaries and statistical inconsistencies between source and target images are simulated.

By not using specific deepfake models to create a training set, SBIs consist of more general and hardly recognizable deepfakes as they are derived from pristine images. The key idea is that this encourages classifiers to learn generic and robust representations without overfitting to manipulation-specific artifacts.

During evaluation, it was shown that their method provides good performance on unseen data. The model, that was trained on SBIs based on FaceForensics++, achieved an AUC of 0.7242 on the DFDC dataset and an AUC score of 0.9318 on the Celeb-DF dataset.

Evaluation

In this section, the results of the robustness and generalizability tests are presented. To carry out the experiments, a set of datasets have been specially crafted. Further, some algorithms required some adjustments to be made to its source. These changes alongside the process for creating the datasets will be presented in the following.

Datasets

For evaluating the robustness of the models, a dataset has been created based on the FaceForensics++ dataset [13]. Sine the dataset has been released in 2018, the data within the dataset consists of deepfakes following the original implementation. The resulting face swaps are of low quality, i.e., the face textures are limited to a low resolution of 64x64 pixels, feature visual artifacts such as blending artifacts, and temporal jitter. Therefore, these videos are most likely to be classified with high confidence by various detection methods. As such, the data can be used to measure the influences of post-processing operations on deepfake detection methods.

For the robustness tests, the following operations have been applied on the 1000 videos of the FaceForensics++ c23 deepfake dataset, which consist of videos of moderate compression.

- *Blur:* Blur can be applied to mitigate some of the blending artifacts occurring in deepfaked content. For the experiments, box-blur with kernel sizes of 3 and 5 were chosen as well as a median filter using a kernel size of 5.
- *Noise:* The introduction of artificial noise can lead to misclassifications in cases where the noise follows a particular pattern. In this paper, we investigate whether random noise can lead to similar results. Here, the Poisson as well as the salt and pepper noise was used. While Poisson noise only introduces minor changes into the image, the salt and pepper noise introduces visible artifacts and as such degrades the image quality.
- *Color Correction:* A color mismatch between the authentic and forged regions of an image may lead to the detection of a deepfake. In addition, if the brightness is too high or too low, some artifacts may disappear. Thus, we consider changes in brightness and contrast of 50% as possible post-processing operations that mail conceal a deepfake.

To analyze whether deepfake detection models can correctly classify videos created with unseen models, another dataset was created using real-time and one-shot deepfake algorithms. The dataset consists of 10 authentic stock videos, that have been manipulated using *DeepFaceLive*, *Inswapper*, *BlendSwap*, and *SimSwap*. For one-shot deepfakes based on *Inswapper*, *BlendSwap*, and *SimSwap*, FaceFusion was used. Examples are showcased in Figure 1.

Implementation

During evaluation, the open-source implementations and the provided pre-trained models of the aforementioned methods were used[3][4][5][6][7].

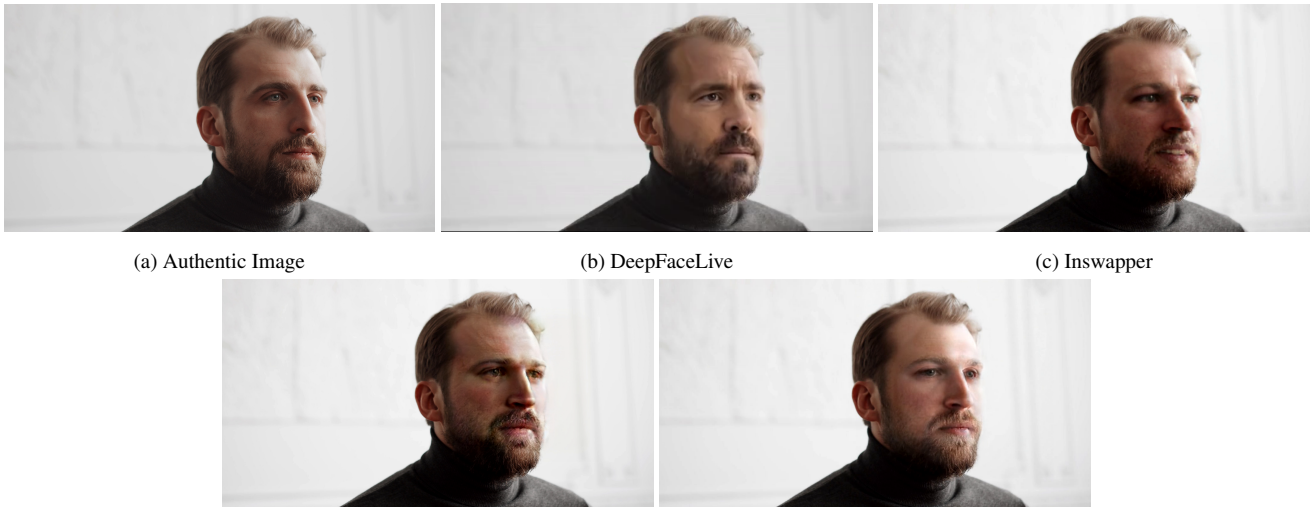
In the case of *LRNet*, the new checkpoints and scripts from 2024 were used that are supposed to provide better generalizability towards unseen data by improving the facial landmark calibration.

DeepFakesON-Phys provides a log for each of the classified videos indicating which frames are likely to be modified. During experiments, it was revealed that the model performs less well when averaging the scores. Instead, computing the standard deviation resulted in an improved performance of $2-8\%$.

Robustness Test

The results of the robustness test are displayed in Table 1. As it can be seen, AltFreezing performs best alongside the model trained on Self-Blended Images. It showcases well that data-driven models outperform model-based techniques, but also

[3]DFDC Solution: https://github.com/selimsef/dfdc_deepfake_challenge
[4]Self-Blended Images: https://github.com/mapooon/SelfBlendedImages
[5]AltFreezing: https://github.com/ZhendongWang6/AltFreezing
[6]LRNet: https://github.com/frederickszk/LRNet
[7]DeepFakesON-Phys: https://github.com/BiDAlab/DeepFakesON-Phys



(a) Authentic Image (b) DeepFaceLive (c) Inswapper

(d) SimSwap (e) BlendSwap

Figure 1: Examples of the videos used during the generalizability tests.

	FF++ C23	Blur 3	Blur 5	Median 5	Poisson Noise	Salt & Pepper	Brightness 0.5	Brightness 1.5	Contrast 0.5	Contrast 1.5	All
DFDC Solution	0.9280	0.9232	0.9144	0.9512	0.9224	**0.9040**	0.9152	0.9007	0.9104	0.8976	0.9013
AltFreezing	**1.0**	**1.0**	**1.0**	**1.0**	**1.0**	0.6864	**1.0**	**1.0**	**1.0**	**1.0**	**0.9790**
Self-Blended Images	**1.0**	0.9184	0.8736	0.8879	0.9744	0.8976	0.9792	0.9551	0.9840	0.9552	0.9226
LRNet	0.7976	0.8288	0.8056	0.7672	0.8992	0.6872	0.8152	0.7504	0.8064	0.8064	0.7888
DeepFakesON-Phys	0.5232	0.5200	0.5264	0.4640	0.5392	0.5312	0.4320	0.5120	0.4960	0.4832	0.5005

Table 1: Results of the robustness test. All values displayed represent the AUC-score measured on each of the modified datasets.

	DeepFaceLive	Inswapper	SimSwap	BlendSwap	All
DFDC Solution	**0.96**	**0.88**	**0.88**	**0.88**	**0.9000**
AltFreezing	0.1666	0.3999	0.36	0.16	0.3769
Self-Blended Images	0.6799	0.72	**0.88**	0.84	0.78
LRNet	0.56	0.5399	0.52	0.56	0.545
DeepFakesON-Phys	0.44	0.44	0.48	0.6799	0.51

Table 2: Results of the generalizability test. All values displayed represent the AUC-score measured on each of the datasets.

that not all the results presented in the respective papers could be reproduced using the provided checkpoints. Although training models using heavily augmented data did help to stabilize the classification performance even when trying to classify post-processed videos, some augmentations still have a strong influence on the performance. This is especially the case for salt and pepper noise. Unfortunately, DeepFakesON-Phys did not perform well on either dataset.

Generalizability Test

Table 2 showcases how the selected methods perform when classifying unseen data from newer deepfake creation methods. Especially Altfreezing showed a drastic decline in performance when classifying unseen data. The DFDC solution still performs best with an AUC-score of 0.9. However, it cannot classify one-shot deepfakes with the same confidence as real-time deepfakes generated using DeepFaceLive. As in the previous experiment, the model-based approaches performed worst on the crafted dataset.

Conclusion

In this paper, we evaluated various state-of-the-art methods for detecting face swapping deepfakes. We thereby mainly focussed on assessing the robustness against common post-processing operations as well as the generalizability towards unseen data. During experiments, it was revealed that data-driven approaches performed best across all test scenarios, especially when it comes to the robustness of the models. However, videos that had salt and pepper applied could not be classified with high confidence. Regarding generalizability, only the winning solution of the DFDC deepfake challenge was able to classify unseen data with high accuracies along with a model trained on Self-Blended Images. The experiments indicated that there does not exist a single detector that could detect all deepfakes with very high confidence. Thus, future work could revolve around combining several data-driven models to not only improve robustness, but also the generalizability at the same time.

Acknowledgments

This work was supported by the German Federal Ministry of Education and Research (BMBF) and the Hessian Ministry of Higher Education, Research, Science and the Arts within their joint support of “ATHENE – DREAM” and ”Lernlabor Cybersicherheit” (LLCS) as well as by the ”Bundesamt für Sicherheit in der Informationstechnik” as part of the project ”SecMedID”.

References

[1] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, “Generative adversarial networks,” *Communications of the ACM*, vol. 63, no. 11, pp. 139–144, 2020.

[2] D. P. Kingma and M. Welling, *An Introduction to Variational Autoencoders*, vol. 12. Now Publishers, 2019.

[3] L. Yang, Z. Zhang, Y. Song, S. Hong, R. Xu, Y. Zhao, W. Zhang, B. Cui, and M.-H. Yang, “Diffusion models: A comprehensive survey of methods and applications,” *ACM Computing Surveys*, vol. 56, no. 4, pp. 1–39, 2023.

[4] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, “High-resolution image synthesis with latent diffusion models,” 2022.

[5] 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: Integrated, flexible and extensible face-swapping framework,” 2021.

[6] K. Qian, Y. Zhang, S. Chang, X. Yang, and M. Hasegawa-Johnson, “Autovc: Zero-shot voice style transfer with only autoencoder loss,” 2019.

[7] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. M. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, and D. Amodei, “Language models are few-shot learners,” 2020.

[8] B. Dolhansky, J. Bitton, B. Pflaum, J. Lu, R. Howes, M. Wang, and C. Canton-Ferrer, “The deepfake detection challenge dataset,” *ArXiv*, vol. abs/2006.07397, 2020.

[9] Z. Wang, J. Bao, W. gang Zhou, W. Wang, and H. Li, “Altfreezing for more general video face forgery detection,” *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4129–4138, 2023.

[10] K. Shiohara and T. Yamasaki, “Detecting deepfakes with self-blended images,” *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 18699–18708, 2022.

[11] Z. Sun, Y. Han, Z. Hua, N. Ruan, and W. Jia, “Improving the efficiency and robustness of deepfakes detection through precise geometric features,” *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3608–3617, 2021.

[12] J. Hernandez-Ortega, R. Tolosana, J. Fierrez, and A. Morales, “Deepfakeson-phys: Deepfakes detection based on heart rate estimation,” *ArXiv*, vol. abs/2010.00400, 2020.

[13] A. Rössler, D. Cozzolino, L. Verdoliva, C. Riess, J. Thies, and M. Nießner, “Faceforensics++: Learning to detect manipulated facial images,” *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 1–11, 2019.

[14] S. Zhang, X. Zhu, Z. Lei, H. Shi, X. Wang, and S. Z. Li, “S3fd: Single shot scale-invariant face detector,” in *Proceedings of the IEEE international conference on computer vision*, pp. 192–201, 2017.

[15] R. Chen, X. Chen, B. Ni, and Y. Ge, “Simswap: An efficient framework for high fidelity face swapping,” in *MM ’20: The 28th ACM International Conference on Multimedia*, 2020.

[16] I. J. Goodfellow, J. Shlens, and C. Szegedy, “Explaining and harnessing adversarial examples,” 2015.

[17] G. Welch and G. Bishop, “An introduction to the kalman filter,” Tech. Rep. 95-041, University of North Carolina at Chapel Hill, Chapel Hill, NC, USA, 1995.

[18] Y. Li, X. Yang, P. Sun, H. Qi, and S. Lyu, “Celeb-df: A large-scale challenging dataset for deepfake forensics,” *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3204–3213, 2019.

[19] B. Dolhansky, R. Howes, B. Pflaum, N. Baram, and C. C. Ferrer, “The deepfake detection challenge (dfdc) preview dataset,” 2019.

[20] Q. Xie, M.-T. Luong, E. Hovy, and Q. V. Le, “Self-training with noisy student improves imagenet classification,” 2020.

Author Biography

Raphael Antonius Frick is a researcher at the Media Security and IT

Forensics division at Fraunhofer SIT. His current research is dedicated to detecting artificially generated multimedia with special focus on detecting deepfakes.

Prof. Dr. Martin Steinebach is the manager of the Media Security and IT Forensics division at Fraunhofer SIT. In 2003, he received his PhD at the Technical University of Darmstadt for this work on digital audio watermarking. In 2016, he became honorary professor at the TU Darmstadt.