https://doi.org/10.2352/EI.2026.38.4.MWSF-311

Face Swap Source Attribution

Martin Steinebach,Julian Götzinger; Fraunhofer SIT|ATHENE: Darmstadt, Germany

Abstract

Face swapping techniques enable the realistic manipulation of facial identities across images, posing significant challenges for digital forensics. While existing research has focused on detecting manipulated media, little attention has been given to identifying the specific source image used in a face swap. In this work, we investigate whether it is possible to reliably trace a manipulated image back to the exact photo that served as the source for the swapped face. We propose a comparison-based method that generates candidate face swaps using known source images and compares them to the target manipulation. Experiments data set demonstrate that our method identifies the correct source image based. We further evaluate robustness against common image distortions, such as JPEG compression and down-scaling, and find that the identification process remains reliable. Our findings highlight the potential of image-level forensic analysis to support source attribution in face-swapped media, with important implications for legal and investigative contexts.

Motivation

Face swapping [1, 2] is a process in which a face in a given image *ImA* is replaced with a face from another image *ImB*, resulting in a synthetic image *ImC*. In forensic investigations, the ability to determine the origin of such manipulated images is essential for verifying authenticity and detecting tampered content. Face swapping presents particular challenges in this regard. Identifying the source image used in the manipulation can serve as crucial evidence in cases involving identity fraud, misinformation, or the creation of illicit content. The forensic objective is thus to analyze the manipulated image *ImC* and trace the embedded face back to its original source *ImB*, enabling both the verification of the manipulation and contextualization of the swapped face. This is particularly significant when *ImB* is a private image, accessible only to a limited number of individuals. Demonstrating that *ImC* was generated using *ImB*—as opposed to a publicly available image—substantially narrows down the set of potential perpetrators. The primary objective of this study is to investigate whether it is feasible to reliably demonstrate that a specific photograph of a person was used as the source for a face swap in a manipulated image. This question is especially relevant in forensic scenarios such as cyber harassment or the creation of fake erotic imagery [3]. To this end, we explore methods capable of accurately comparing facial regions between candidate source images and the manipulated result to establish a definitive match. Furthermore, we assess the extent to which common image alterations—such as lossy compression and scaling—impact the reliability of this matching process. By systematically evaluating the effects of these transformations, we aim to understand their influence on the accuracy and robustness of source image identification in digital forensic investigations.

Figure 1. *Example source face set: 16 portraits of the same person*

Figure 2. *Example face swap: left - Target image, mid and right: face swaps using photos of the same person*

Challenge

Identifying a specific photograph of a person as the source of a face swap is a significant forensic challenge.Since the face belongs to the same person, all face-swap results from different photographs of that person may appear visually similar. This is due to the inherent consistency of facial features across images of the same subject, such as geometry, skin tone and expression as shown in figure 1. As a result, it becomes non-trivial to distinguish between images generated from different source photos. Image 2 shows a target and two face swaps using photos from figure 1.

However, face swapping is not a simple identity-preserving process; it involves a complex transformation that transfers not only high-level identity traits but also low-level visual cues from the source image to the target. These cues may include lighting,

resolution, facial expression, pose, and camera-specific artifacts. Therefore, even subtle differences between candidate source images can result in measurable variations in the output image. This subtle fingerprinting provides a potential approach for distinguishing between source candidates, provided that the detection and comparison mechanisms are sensitive and robust enough to capture these differences.

This task would be considerably more difficult if, instead of using a single image as the source, a generative model were trained on a set of photographs of the same individual. In that case, the model would produce a kind of averaged or generalized representation of the person's face, effectively abstracting away image-specific characteristics. The resulting face swap would then no longer reflect any single photograph in its original form, eliminating the possibility of direct source identification through image-level comparison.

Thus, the core scientific challenge lies in developing methods that can detect and quantify the subtle, image-specific traces that are preserved during the face swapping process, despite the high visual similarity between outputs derived from different images of the same person. Achieving this requires precise comparison strategies that are robust to facial similarity yet sensitive to the nuanced artifacts left by the manipulation pipeline.

State of the Art

Face swapping detection has emerged as a critical research domain in response to the proliferation of deepfake technologies that manipulate facial images and videos. To address this growing concern, researchers have been developing increasingly sophisticated methods to detect and localize such manipulations. A variety of algorithms have been proposed to determine whether a given piece of media has been generated or altered by artificial intelligence [4, 5, 6, 7].

To the best of our knowledge, no existing approaches besides our own [8] have focused on identifying the specific source image used in a face swap. Our earlier work required access to high quality images. This is the main difference to this paper besides more a detailed evaluation.

Substantial research has been devoted to the detection of face swapping and other deepfake generation techniques. As early as 2017, Zhang et al. employed Speeded-Up Robust Features (SURF) in combination with machine learning for face swap detection [9]. More recent approaches have leveraged deep learning: Ding et al. [10] developed a model to classify real versus swapped faces using convolutional architectures. Huang et al. [11] introduced a framework that incorporates explicit identity contrast loss and an implicit identity exploration (IIE) loss within a convolutional neural network. Yu et al. [12] conducted an extensive comparative analysis involving over 100 deep learning-based methods for detecting facial manipulations. Dang et al. [13] provided a comprehensive survey covering various manipulation techniques and corresponding detection strategies. Ghasemzadeh et al. [14] proposed a generalized approach for manipulation detection aimed at improving cross-domain robustness.

We also previously discussed a number of deepfake detection strategies for image [15], video [16, 17, 18] as well as audio [19, 19, 20]. As in the research listed above, their goal was to indentify wheter a deepfake occured or not.

Despite these advances, significant challenges persist in creating detection methods that generalize effectively across different datasets and manipulation techniques. Ongoing research is directed toward increasing the robustness and accuracy of detection models by incorporating both spatial and frequency domain features. Moreover, there is growing interest in developing universal detection tools capable of identifying a wide spectrum of face manipulation methods without relying on prior knowledge of specific algorithms.

Concept

The proposed method begins with the identification of an image *ImC* suspected to contain a swapped face. This image serves as the target image for a new round of face swapping. This means that *ImC* becomes *ImA'*.

Subsequently, we compile a set of candidate images $ImB_{1...n}$ that may have served as the source for the swapped face. Using a face swapping algorithm, we generate a series of synthetic images $ImC'_{1...n}$ by combining *ImA'* with each candidate face image. We then compute the difference between the original manipulated image *ImC* and each generated image $ImC'_{1...n}$. The candidate image yielding the smallest difference is identified as the most likely source of the swapped face.

To further evaluate the robustness of this method, we analyze the effects of common image transformations, such as lossy compression and scaling, applied to *ImC*. We assess whether these alterations significantly influence the difference metrics, which could affect the reliability of source identification.
The algorithm proceeds through the following steps:

- Selection of the face-swapped image to be analyzed.
- Replacing (*ImA*) with *ImC*
- Collection of potential source images for the swapped face ($ImB_{1...n}$).
- Re-creation of candidate face-swapped images ($ImC'_{1...n}$) using a face swapping algorithm.
- Computation of the difference between the original face-swapped image (*ImC*) and each candidate image:
  - Localization of the face region in *ImC*.
  - Cropping of *ImC* and candidate images to isolate the facial region.
  - Pixel-wise subtraction to compute visual difference.
  - Visualization of differences through inversion and gamma correction.
  - Calculation of the average pixel difference between *ImC* and each candidate image.
- Selection of the most probable source image based on the lowest average difference, with an optional threshold to account for ambiguous cases.

Implementation

We implemented the algorithm described above for evaluation. It calculates the difference of the pixel values between the image under investigation *ImC* and the newly generated candidates $ImC'_{1...n}$).

Bevore the comparison, we change all images to greyscale. Then we identify the face region of *ImC* with the help of the cv2 CascadeClassifier[1] haarcascades option. We compare it to the

[1] https://docs.opencv.org

identical image regions of the candidates $ImC'_{1...n}$) by the Pillow[2] ImageChops.difference function. The actual difference value is the average of the provided difference image. For better visualization, we invert the difference image after calculating the average.

For creating the face swap images, we used face fusion 3.0 with the parameters stated in table 1.

Relevant face fusion 3.0 arguments and their corresponding values.

Argument	Value
face_detector_model	scrfd
face_detector_angles	0
face_detector_size	640x640
face_detector_score	0.5
face_landmarker_model	2dfan4
face_landmarker_score	0.5
face_selector_mode	reference
face_selector_order	large-small
face_selector_gender	null
face_selector_race	null
face_selector_age_start	null
face_selector_age_end	null
reference_face_position	0
reference_face_distance	0.6
reference_frame_number	0
face_mask_types	box
face_mask_blur	0.3
face_mask_padding	0, 0, 0, 0
trim_frame_start	null
trim_frame_end	null
temp_frame_format	png
keep_temp	null
output_image_quality	80
output_image_resolution	852x1280
output_audio_encoder	aac
output_video_encoder	libx264
output_video_preset	veryfast
output_video_quality	80
output_video_resolution	null
output_video_fps	null
skip_audio	null
processors	face_swapper

Experiments

For the face sources, we used 14 photo sets from the "Portrait and 26 Photos Re-identification" dataset[3]. The target images were chosen from pixabay, with the only requirement that the images needed to show persons. For creating the test cases of *ImC* for investigation, we randomly selected one photo from the pixaby set as the target image and one image of the face source set as the source of the face swap. Figure 3 shows an example face swap.

Figure 4 shows an example of the differences per face. We randomly selected one face of a person as the true soruce and

[2]https://pillow.readthedocs.io

[3]https://www.kaggle.com/datasets/trainingdatapro/portrait-and-30-photos-test?resource=download

Figure 3. *Generation of face swap examples: ImA, ImB, ImC*

compared it to 10 other photos as candidate sources. Onc can see that the differes are less for the true source image.

Figure 7 shows the results for 10 runs of the experiment. Each column of the figure is one set of true source and candidates. One can see that in all cases the difference is lowest for the true source. But the overall differnce and the distances differ between the sets. There are cases where the difference to the true source is larger than the differences of all candidates in another set. One good examples is 10-11-9 (person 10, face 11, target image 9): the true source difference is greater than all differences in e.g. 10-13-2. Table 2 shows the results in detail.

Example of difference for pairs of original and new faces. The first pair for each set (original face of person and target) is the one to be expected to have the smallest difference

Person	Org.Face#	New Face#	Target	Difference
1	21	21	2	4,5734
1	21	15	2	5,0173
1	21	12	2	5,1306
1	21	26	2	5,2386
1	21	19	2	5,3937
1	21	4	2	6,3843
1	9	9	8	3,7203
1	9	21	8	5,6744
1	9	14	8	5,7426
1	9	12	8	6,0749
1	9	23	8	6,2108
10	11	11	9	6,8657
10	11	20	9	7,2290
10	11	19	9	7,4653
10	11	18	9	7,7640
10	11	2	9	8,2505
10	11	4	9	8,3605
10	13	13	2	4,6777
10	13	2	2	5,5893
10	13	17	2	6,0304
10	13	18	2	6,1353
10	13	22	2	6,2452

A more detailed visualization of the results is given by figure 6. One can see that both graphs are similar, but the original face has fewer pixels with high differences.

We also evaluated to impact of JPEG lossy compression and image scaling by 50% as examples for changes *ImC* could be subject to before is is examined. Only minimal impact is obeserved here. Figure 5 shows an example. The original face source is

Figure 4. *Differences of the face regions. First example top row: True source image.*

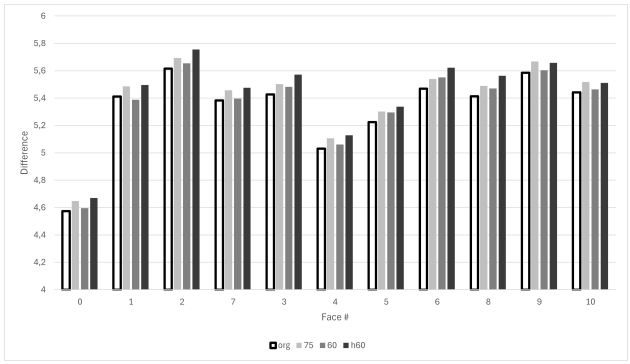


Figure 5. *Differences after JPG75, JPG60, and downscaling to 50%.*

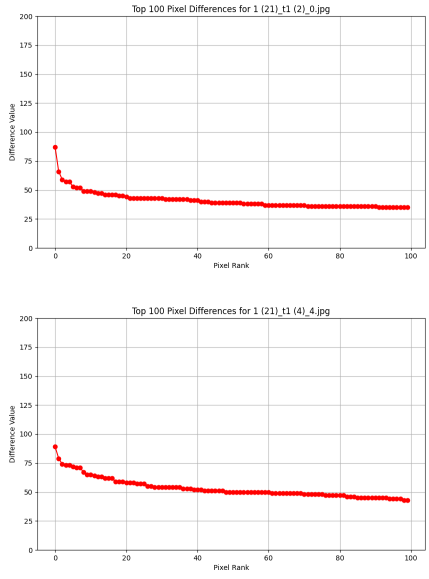


Figure 6. *Differences plots of the face regions. Difference values of pixels sorted from large to small. Top: Origincal face, bottom: Face candidate with smallest difference.*

a siginifcantly smaller difference value than the rest of the candidates. The ”attacks” do not change this. There seems to be only little influence on the results. It may be of interest that the stronger JPG compression with quality factor 60 produces smaller differences than JPEG quality 75. Table 3 shows the details.

Differences for original output and after jpg75, jpg 60 and downscaling 50%

Face#	org	JPG75	JPG60	Scaling
0	4,573	4,648	4,595	4,669
1	5,411	5,486	5,388	5,497
2	5,616	5,693	5,654	5,755
7	5,383	5,458	5,397	5,476
3	5,427	5,502	5,483	5,572
4	5,030	5,106	5,061	5,128
5	5,225	5,303	5,296	5,337
6	5,468	5,540	5,552	5,622
8	5,413	5,489	5,471	5,563
9	5,585	5,668	5,603	5,657
10	5,442	5,517	5,463	5,510

Discussion

The conducted experiments demonstrate the feasibility of identifying the specific source photograph used in a face-swapped image, even among visually similar candidates. We evaluated the ability of our method to distinguish the true source image from a pool of potential candidates based on pixel-level differences.

In a the evaluation, one photo of a person was randomly designated as the true source and compared to up to ten other photos of the same individual, treated as candidate sources. The computed difference between the image under investigation and each candidate consistently revealed a lower value for the true source compared to others. This supports the hypothesis that even subtle visual cues from the original photo are preserved during the face swap process and can be detected.

However, the magnitude of the differences varied substantially between experiments. For instance, in case 10-11-9 (person 10, face 11, target image 9), the difference between the manipulated image and the true source was greater than all differences in another test case (e.g., 10-13-2). .

To evaluate robustness, we subjected *ImC* to common image transformations, specifically JPEG compression and downscaling to 50% of the original size. Results indicate that these modifications had minimal impact on the outcome. The true source consistently exhibited the lowest difference, regardless of the applied distortion. Interestingly, stronger JPEG compression (quality factor 60) sometimes resulted in lower differences compared to moderate compression (quality 75). While counterintuitive, this may be attributed to the smoothing effect of compression, which reduces overall pixel variance and thus diminishes detectable differences. These findings suggest that the method remains effective under realistic forensic conditions where image quality may be degraded.

In summary, the results confirm that image-specific traces from a face source photo can persist through the face swapping process and can be leveraged to identify the original source with high reliability. Nevertheless, variability in absolute difference values between test sets suggests that contextual factors, such as

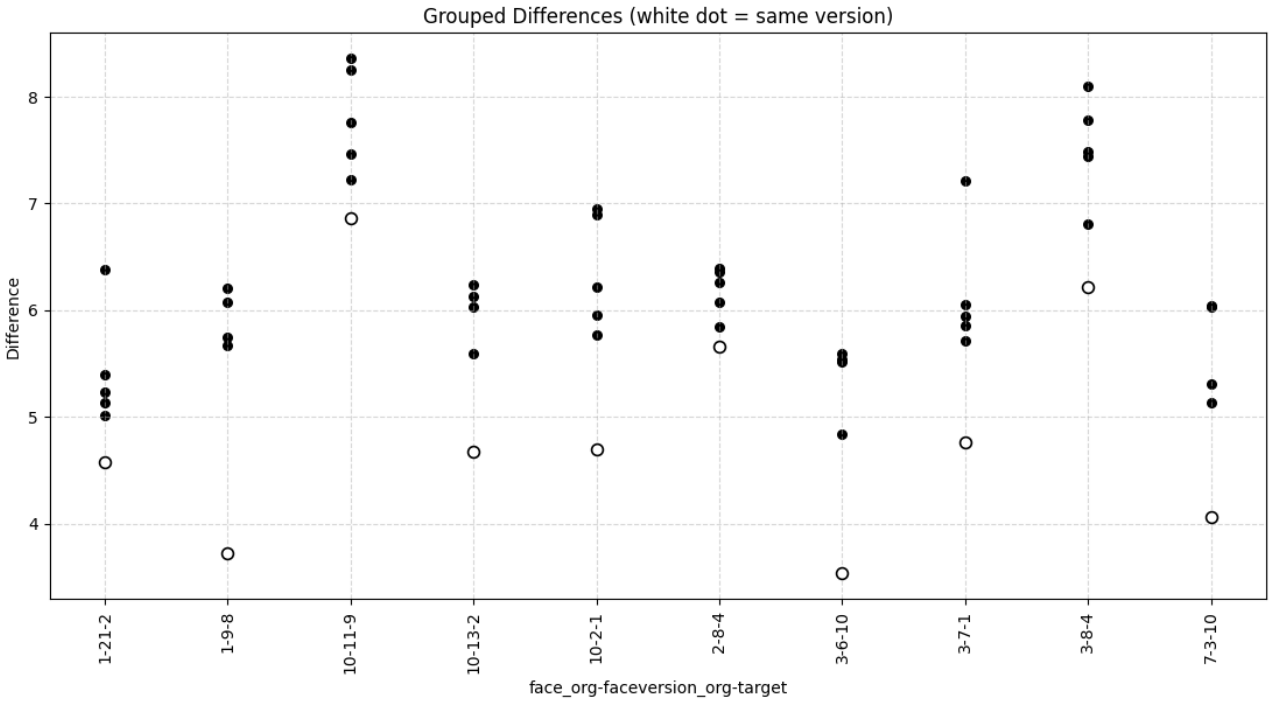


Figure 7. *The original faces in all ten test cases produce a lower difference than the other face photos of the same person*

lighting conditions and facial expressions, play a non-negligible role in the matching performance.

Summary and Future Work

In this work we introduce the concept of face photo attribution in face swap attacks. We show that is possible to identify a specific photo of a person that was used in a face swap scenario. This can be helpful for forensic investigations if one can show that a photo has been used that was only availble to the suspect.

It is likely that face swapping will continue to develop. Future research is needed to determine whether the findings presented here can be applied to new methods of face swapping.

As the expermiments in this paper are limited, future work could also address more test sets and various attacks. Currently the approach does not discuss evasion attacks trying to earse the traces of the photo used. It would also be interesting to know if this is possible and if there are any countermeasures.

Acknowledgments

This research work has been funded by BMBF and the Hessen State Ministry for Higher Education, Research and the Arts within their joint support of the National Research Center for Applied Cybersecurity ATHENE.

References

[1] S. Baliah, Q. Lin, S. Liao, X. Liang, and M. H. Khan, "Realistic and efficient face swapping: A unified approach with diffusion models," *arXiv preprint arXiv:2409.07269*, 2024.

[2] Y. Zhu, W. Zhao, Y. Tang, Y. Rao, J. Zhou, and J. Lu, "Stableswap: Stable face swapping in a shared and controllable latent space," *IEEE Transactions on Multimedia*, 2024.

[3] K. Hao, "Deepfake porn is ruining women's lives. now the law may finally ban it," *MIT Technology Review*, 2021.

[4] J. Frank, T. Eisenhofer, L. Schönherr, A. Fischer, D. Kolossa, and T. Holz, "Leveraging frequency analysis for deep fake image recognition," in *International conference on machine learning*, pp. 3247–3258, PMLR, 2020.

[5] D. C. Epstein, I. Jain, O. Wang, and R. Zhang, "Online detection of ai-generated images," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 382–392, 2023.

[6] S. Volkova and A. Bogdanov, "A deep learning approach to face swap detection," *International Journal of Open Information Technologies*, vol. 9, no. 10, pp. 16–20, 2021.

[7] J. Jiang, B. Wang, B. Li, and W. Hu, "Practical face swapping detection based on identity spatial constraints," in *2021 IEEE international joint conference on biometrics (IJCB)*, pp. 1–8, IEEE, 2021.

[8] M. Steinebach and M. Frühwein, "Face swap forensics," *Electronic Imaging*, vol. 37, pp. 1–8, 2025.

[9] Y. Zhang, L. Zheng, and V. L. Thing, "Automated face swapping and its detection," in *2017 IEEE 2nd international conference on signal and image processing (ICSIP)*, pp. 15–19, IEEE, 2017.

[10] X. Ding, Z. Raziei, E. C. Larson, E. V. Olinick, P. Krueger, and M. Hahsler, "Swapped face detection using deep learning and subjective assessment," *EURASIP Journal on Information Security*, vol. 2020, pp. 1–12, 2020.

[11] B. Huang, Z. Wang, J. Yang, J. Ai, Q. Zou, Q. Wang, and D. Ye, "Implicit identity driven deepfake face swapping detection," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 4490–4499, 2023.

[12] Z. Yu, Y. Qin, X. Li, C. Zhao, Z. Lei, and G. Zhao, "Deep learning for face anti-spoofing: A survey," *IEEE transactions on pattern analysis and machine intelligence*, vol. 45, no. 5, pp. 5609–5631,

2022.
[13] M. Dang and T. N. Nguyen, "Digital face manipulation creation and detection: A systematic review," *Electronics*, vol. 12, no. 16, p. 3407, 2023.
[14] F. Ghasemzadeh, T. Moghaddam, J. Dai, J. Yun, and D. D. Kim, "Towards generalized detection of face-swap deepfake images," in *Proceedings of the 3rd ACM Workshop on the Security Implications of Deepfakes and Cheapfakes*, pp. 8–13, 2024.
[15] R. A. Frick and M. Steinebach, "A photo and a few spoken words is all it needs?! on the challenges of targeted deepfake attacks and their detection," in *Proceedings of the 3rd ACM Workshop on the Security Implications of Deepfakes and Cheapfakes*, pp. 25–28, 2024.
[16] R. A. Frick, S. Zmudzinski, and M. Steinebach, "Detecting "deepfakes" in h. 264 video data using compression ghost artifacts," *Electronic Imaging*, vol. 32, pp. 1–7, 2020.
[17] R. A. Frick and M. Steinebach, "One detector to rule them all? on the robustness and generalizability of current state-of-the-art deepfake detection methods," *Electronic Imaging*, vol. 36, pp. 1–6, 2024.
[18] R. A. Frick, S. Zmudzinski, and M. Steinebach, "Detecting deepfakes with haralicks texture properties," in *International Symposium on Electronic Imaging Science and Technology (IS&T) 2021*, 2021.
[19] K. Schäfer and M. Steinebach, "Generalization in audio deepfake detection: Evaluating asr encoder-based feature extraction," in *2025 IEEE 12th International Conference on Data Science and Advanced Analytics (DSAA)*, pp. 1–5, IEEE, 2025.
[20] K. Schäfer and M. Steinebach, "Mfcc vs. lfcc for audio deepfake detection: The role of delta features and input length," in *2025 33rd European Signal Processing Conference (EUSIPCO)*, pp. 576–580, IEEE, 2025.

Author Biography

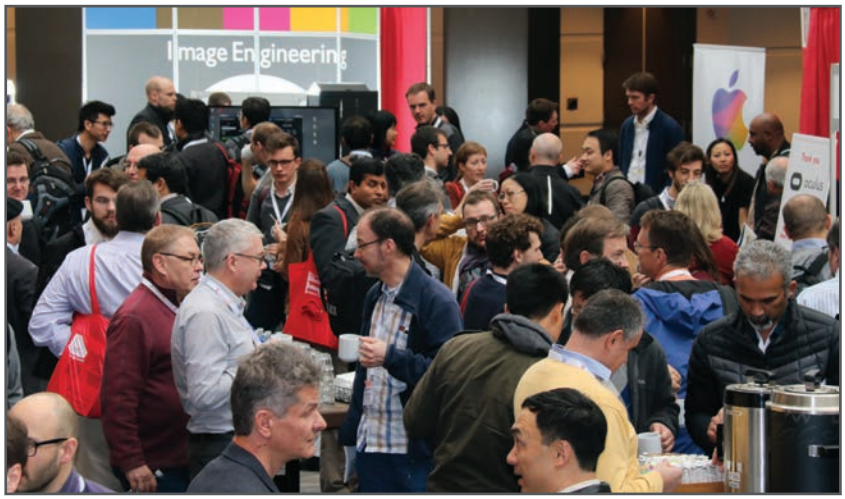
Martin Steinebach is the manager of the Media Security and IT Forensics division at Fraunhofer SIT. From 2003 to 2007 he managed the Media Security in IT division at Fraunhofer IPSI. He studied computer science at the Technical University of Darmstadt and finished his diploma thesis on copyright protection for digital audio in 1999. 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. He gives lectures on Multimedia Security as well as Civil Security. He is Principal Investigator at ATHENE and represents IT Forensics and AI security. Before he was Principal Investigator at CASED with the topics Multimedia Security and IT Forensics.

Julian Götzinger is a research associate at Fraunhofer SIT / ATHENE in Darmstadt. He obtained his Master's degrees in Computer Science and IT Security from TU Darmstadt in 2024. In the Media Security and IT Forensics department, he researches deep learning-based face verification methods, with a particular focus on applications for the protection of minors. His broader research interests include fairness and robustness in biometric systems, explainability and interpretability of deep learning models, as well as synthetic data generation.

JOIN US AT THE NEXT EI!

electronic IMAGING

Imaging across applications . . . Where industry and academia meet!



• SHORT COURSES • EXHIBITS • DEMONSTRATION SESSION • PLENARY TALKS •
• INTERACTIVE PAPER SESSION • SPECIAL EVENTS • TECHNICAL SESSIONS •

www.electronicimaging.org

ADVANCING IMAGING SCIENCE 75 1947 IS&T 2022 imaging.org