https://doi.org/10.2352/EI.2025.37.4.MWSF-300

Face Swap Forensics

Martin Steinebach, Marco Frühwein; Fraunhofer SIT|ATHENE: Darmstadt, Germany

Abstract

Multimedia forensics is an important field addressing the increasing misuse of digital content, such as deepfakes and face-swapping technologies. This paper focuses on detecting face swapping. Our goal is not to decide whether face swapping has occurred. We assume that we execute a forensic investigation in which it needs to be learned which photo of a person's face has been used for the face swap. We take a number of potential source face photographs and compare their behavior when reproducing the face swap. We show that the photo used for the face swap can be identified even after lossy compression and scaling.

Motivation

Face swapping [1, 13] allows taking an image *ImA* and replace a face shown in that image with a face from a second image *ImB*. This creates a fake image *ImC*. In forensic investigations, the ability to determine the origin of manipulated images is critical to verifying authenticity and identifying tampered content. Face-swapping poses unique challenges in this context. Identifying the source photo used in the face swap can provide crucial evidence in cases of identity fraud, misinformation, or illegal content creation. The forensic goal is therefore to analyze and trace the face in the manipulated photo *ImC* back to its original source *ImB*, allowing verification of both the manipulation and the original context of the swapped face. This can be important if *ImB* is a photo that only one or a few people have access to. If we can show that *ImC* was created using *ImB* and not another publicly available photo, this significantly reduces the number of potential creators.

The primary goal of this work is to analyze whether it is possible to reliably prove that a specific photo of a person has been used as the source for a face swap in a manipulated image. This may be particularly relevant for forensic investigations of cybermobbing, for example, when erotic images have been faked[7].

This requires methods that can accurately compare the effect of the facial region between potential source images to establish a clear match. In addition, the work aims to assess whether common image modifications, such as lossy compression and scaling, affect the reliability of this matching process. By evaluating the effects of compression artifacts and image resizing, we seek to understand how these modifications may affect the accuracy and robustness of source photo identification in forensic contexts.

State of the Art

Face swapping detection has become a critical area of research due to the rise of deepfake technologies that manipulate facial images and videos. Researchers are developing advanced methods to identify and localize these manipulations effectively. There are a number of algorithms aiming for deciding whether a given media is created or modified by AI [5, 4, 10, 9]

To our knowledge, there are so far no approaches identifying the actual photo used for face swapping. On the other hand, there are many works on the detection of face swapping and other deepfake methods.

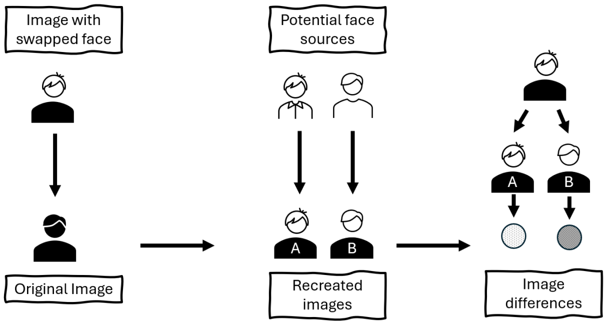

Figure 1. *Concept of approach*

Already in 2017 Zhang et a. utilized SURF features combined with machine learning for face swap detection [12]. Ding et al. [3] apply deep learning for classifying real and swapped faces. Huang et al. [8] use explicit identity contrast loss and implicit identity exploration (IIE) loss combined with a CNN. Yu et al. [11] compare a wide set (over 100) of deep learning methods for face manipulation detection. Dang et al. [2] provide a survey with different manipulation and detection approaches. Ghasemzadeh et al.[6] aim at a generalized detection approach.

Despite these advancements, challenges remain in developing detection methods that generalize well across diverse datasets and manipulation techniques. Future research is focusing on enhancing the robustness and accuracy of detection models, integrating spatial and frequency domain features, and creating tools that can identify a wide range of face manipulation methods without prior knowledge of specific techniques.

Approach

The method begins with the identification of an image *ImC* containing a swapped face. This is the object to be analyzed in a forensic examination. The first objective is to determine the original image *ImA* in which the face was replaced. This can be done by advanced inverse image searching. Then we gather a set of candidate images $\{ImB\}_{1...n}$ that may have served as potential sources for the swapped face. Using a face swapping algorithm, we generate new face-swapped images $\{ImC'\}_{1...n}$ by combining the original image *ImA* with each of the candidate face sources. We then compute the difference between the original face-swapped image *ImC* and each of the newly generated images $\{ImC'\}_{1...n}$. The candidate image that produces the smallest difference, falling below a predefined threshold, is identified as the likely source of the face swap. See also figure 1 for an illustration of the concept.

In addition, we assess the influence of image alterations on this process by applying lossy compression and scaling to *ImC*. We then evaluate whether these modifications significantly impact the calculated difference, potentially affecting the reliability of the source identification.

The algorithm can be summarized in the following steps:

- Selection of face swapped image to analyze
- Identification of image source by robust inverse image search
- Collection of potential face source images
- Re-creation of candidate face swapped images by face swapping algorithm
- Calculation of difference between face swapped image and candidate images
  - Identification of face area in face swapped image
  - Cropping of face swapped image and candidate images to focus on face region
  - Calculation of difference by subtraction
  - Visualization by inversion of face image difference and gamma correction
  - Calculation of average pixel difference between face swapped image and candidate images
- Selection of face swap source image by lowest difference and threshold in case all image candidates show similar distance

Evaluation

To verify our approach, we executed a number of experiments with image sets of varying quality. The section about manual high-quality experiments uses portrait photos from Pixabay. Here we find very detailed and focussed shots of portraits, as we would expect in an actual case of a quality fake. In addition, we looked at a publicly available data set, comparing the performance of 14 image sets. Here, due to the lower quality and background objects, we also show the potential failure of the approaches due to incorrect face detection. This will not be a problem in a real-world manual investigation. But it shows how error-prone fully approaches still are.

Manual High Quality Test

First experiments indicate that the approach can robustly identify the facial image used for face swapping event for similar images of the same person. As an example, we used faces of one Pixabay artist taken from one model shown in figure 3. In figure 5 we show the resulting face swap images in the first row. For the test face D from image 3 was used for the creation of the face swap. We can see that the differences between the found and the recreated images help to identify D. Compared to faces A to C the difference of D is small. This is also shown by the average pixel distances in figure 2. For the first run we only compressed the found image by JPEG quality factor 50, a strong lossy compression. For the second run, we also added downsizing by 20%.

The behavior of the method could depend on the combination of the manipulated image *ImA* and the face source *ImB*. We therefore created and analyzed face swap images with four face sources and four target images. As shown in figure 6, the behavior appears to be independent of the target image.

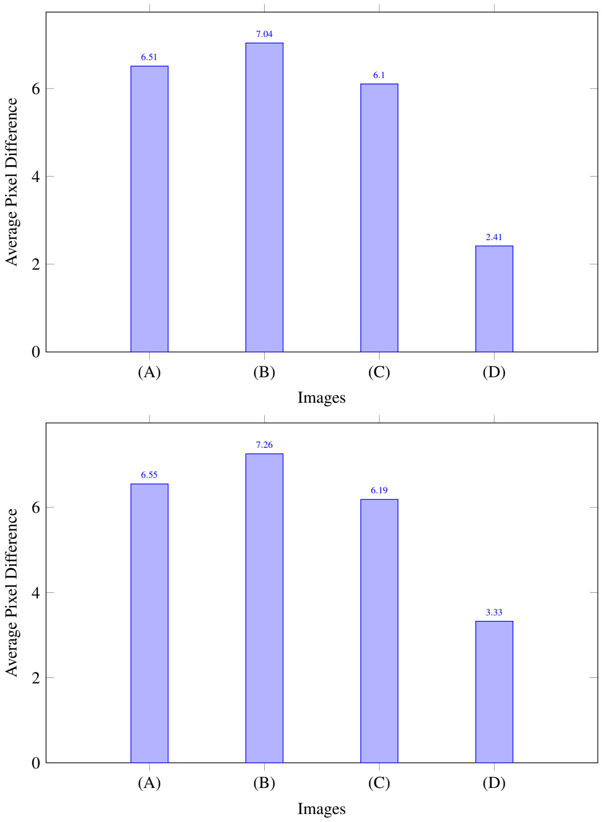

Figure 2. *Average pixel difference of face regions. D was the face used. Top: After JPEG 50 compression, Bottom: After JPEG 50 compression and downscaling by 20%*

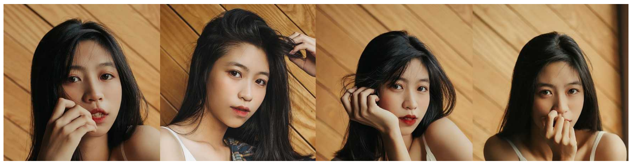

Figure 3. *Four (A to D) source faces used for swapping, CC0 Pixabay, KemDauArt*

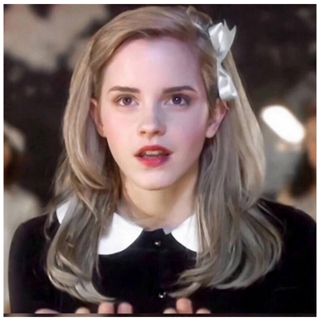

Figure 4. *Source for the face swapping image, example by seaart.ai, original: BBC, "Ballet Shoes"*

A note about the face swapping method used for the experiments in this subsection: we use the seaart.ai online free face swapping method. There is no detailed information about which

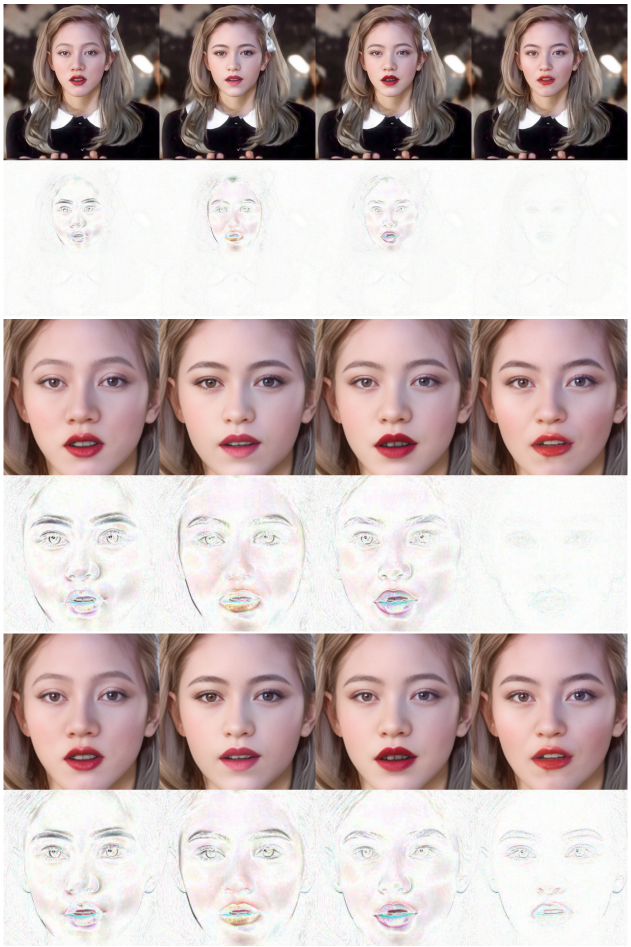

Figure 5. *Example for full image, image cropped to face region, and robustness after 20% down-scaling, each with JPEG 50 lossy compression. Resulting images as well as differences to test images are shown.*

AI method is used by searart.ai. But since most of their methods are based on Stable Diffusion, it is likely that inswapper-128[1] is used.

Figure 8 shows the performance of the method when also face sources of different persons are used. As to be expected, the difference of the original face source C is significantly lower than that of all other face sources (see also figure 7). The face sources of the different persons E to H all create a high difference, two face sources of the correct person show a lower difference than that wrong persons. This could be another research question: Is it possible to prove that an image was created with a specific persons face by using face sources of that person and of other persons and showing that the average difference of the person is significantly lower than that of other persons.

We checked the performance of the detection with additional online face swapping tools, remaker.ai[2] and aifaceswap.io[3], see figures 9 10. In both cases, the results were similar to those of seart.ai. Thus, the proposed method does not depend on a specific face swapping method.

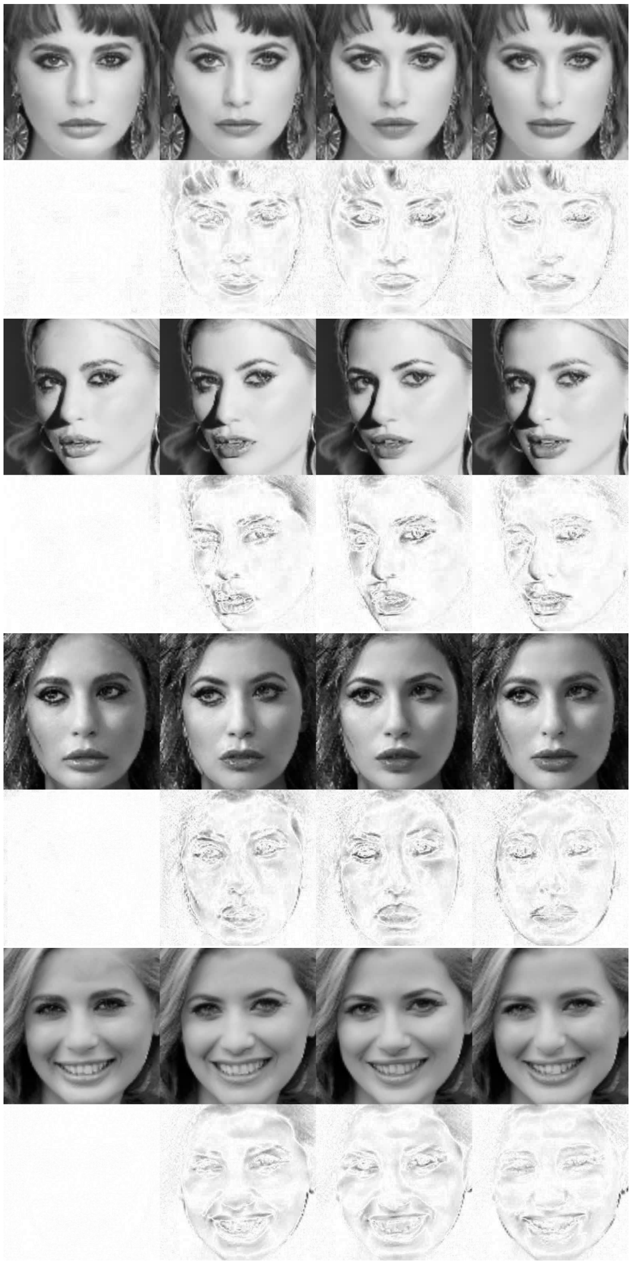

Figure 6. *The identical four face sources applied to four different targets. The face source is always face 1 on the left, image under analysis was stored with JPEG 60 lossy compression. Face Swap: inswapper 128 fp16*

Another question is whether it is necessary to use the same face swapping method for creation of the image under investigation *ImC* and re-creation by ImA and ImA in an investigation. Otherwise, it would be necessary to identify the face swapping method before the proposed method, requiring an additional analysis. Therefore, we executed one experiment where we used different face swapping methods for creation and analysis. In figure 11 we can see that while the difference between *ImC* and the newly created image with the correct face source is greater than when using the same face swapping method, there is still a sig-

[1] https://github.com/haofanwang/inswapper

[2] https://remaker.ai/

[3] https://aifaceswap.io

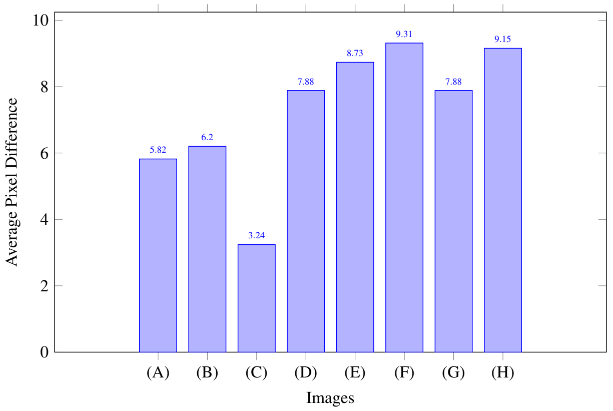

Figure 7. *Average-d pixel difference of face regions.*

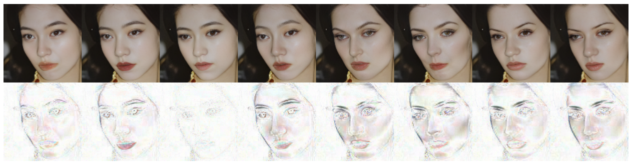

Figure 8. *A to D are faces from the same person, C is the face source, E to H are two additional persons*

nificant difference between the correct source and the other face sources.

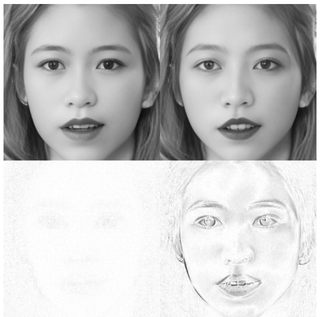

Figure 9. *Remaker AI example as cross-verification. Left=source face. Strong differences can be observed between the source face and the other face.*

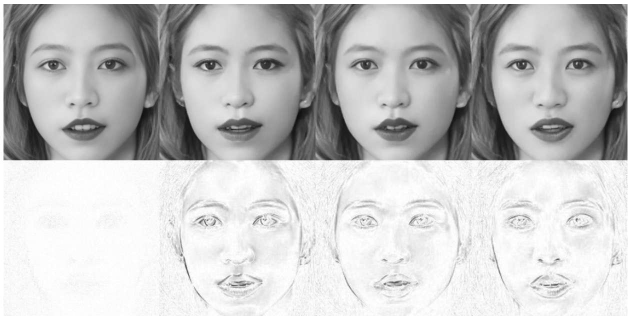

Figure 10. *AIfaceswap.io example as cross-verification. Left=source face. Strong differences can be observed between the face image used and the rest of the face sources. Same sources as in figures 3 and 4*

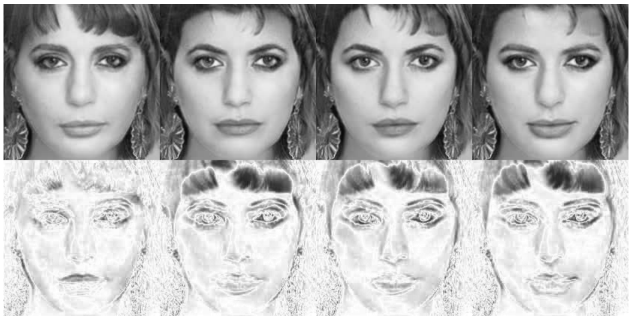

Figure 11. *Cross-Swapping. The image under alaysis was created with inswapper 128 fp16, the comparison images were created with ghost 1 256.*

Automated data set

For a broader evaluation, we used 14 photo sets from the ”Portrait and 26 Photos Re-identification” dataset[4]. For creating the fakes, we used face fusion 3.0 with the parameters stated in table 2. We randomly selected one photo from the photo sets as target images and three photos from a different set as the sources of the face swap. We then selected one of the three resulting face swap images, scaled them down by 20% and stored them with JPEG quality 60.

The results can be seen in figure 12 and in table 1. One can see that in most cases the identification of the original source is possible. There are also exceptions: sets 01055155 and 10dc56f2 produce identical differences for all three candidates. The reason here is a failed face recognition in the image under investigation. This causes an identical background part to be compared, so the difference is only the noise caused by scaling and lossy compression. 0efc5174 has an identical issue, and in addition in one of the candidate images the face region could not be identified, causing a missing difference value. Examples are shown in figure 13

We also conducted a small experiment to test whether some false face sources produce significantly different results than others. We used 20 faces from one set of photos and swapped them with a target image from another set. The image to be investigated (image 1) was down scaled from 3880 pixels to 2000 pixels (almost 50%) width and saved with JPEG quality 60. The results are shown in figure 14: The first (correct) face source has a significantly lower distance than all the other 19 faces. Still, the range is from 7.4 to 14.5 times more average pixel difference.

Discussion

Our approach addresses a different challenge than known approaches in deepfake and GenAI detection: our research goal was to determine which specific pair of images (source and target, *ImA* and *ImB*) was used to create the image under test. This is relevant in cases where it is obvious that image manipulation has occurred, but it is necessary to better understand and prove how that image was manipulated. We show that in the case where both the source image and the image from which the face for a face swap has been taken, we can show that a particular face image is the most likely face source, even when very similar photos of the same face are also candidates.

Our work brings a new perspective to the ongoing research

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

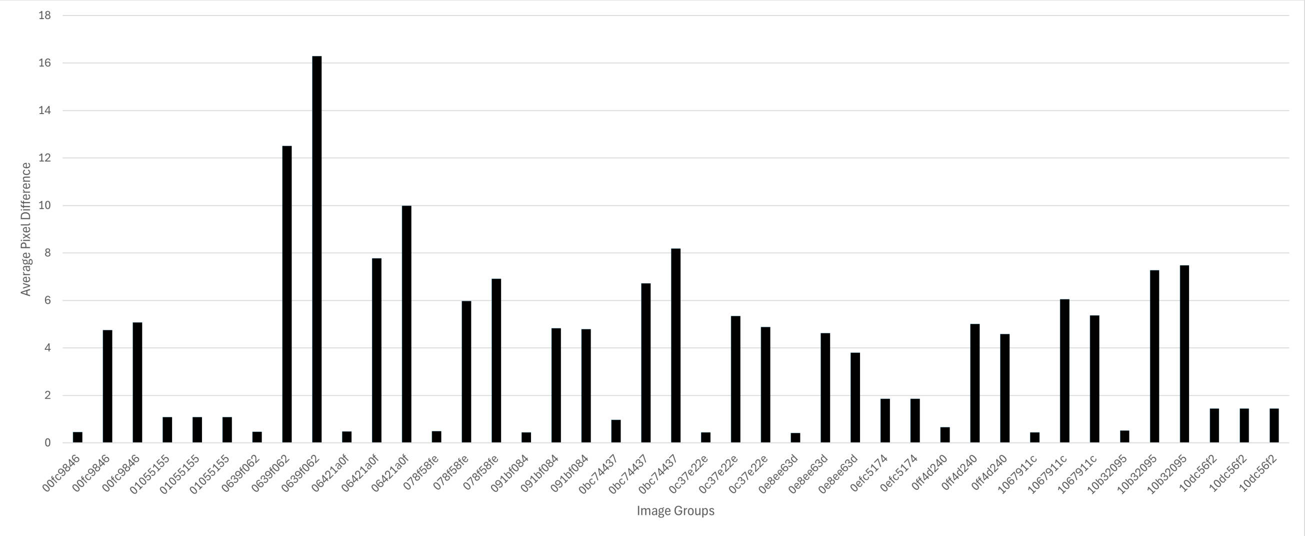

Figure 12. *Comparison of differences in sets of three.*

Table 1: Detailed results

Set	Inv	Can	Diff
00fc9846	12	12	0,46
00fc9846	12	14	4,75
00fc9846	12	9	5,07
01055155	18	18	1,09
01055155	18	3	1,09
01055155	18	5	1,09
0639f062	20	20	0,47
0639f062	20	21	12,51
0639f062	20	23	16,29
06421a0f	24	24	0,48
06421a0f	24	3	7,78
06421a0f	24	4	9,99
078f58fe	11	11	0,5
078f58fe	11	18	5,98
078f58fe	11	6	6,92
091bf084	18	18	0,45
091bf084	18	5	4,83
091bf084	18	7	4,79
0bc74437	11	11	0,97
0bc74437	11	26	6,72
0bc74437	11	9	8,19
0c37e22e	11	11	0,45
0c37e22e	11	15	5,34
0c37e22e	11	23	4,88
0e8ee63d	12	12	0,42
0e8ee63d	12	16	4,63
0e8ee63d	12	20	3,8
0efc5174	15	20	1,86
0efc5174	15	6	1,86
0efc5174	15		
0ff4d240	10	10	0,66
0ff4d240	10	24	5,01
0ff4d240	10	6	4,59
1067911c	10	10	0,45
1067911c	10	14	6,05
1067911c	10	4	5,37
10b32095	19	19	0,52
10b32095	19	25	7,27
10b32095	19	26	7,48
10dc56f2	17	17	1,45
10dc56f2	17	2	1,45
10dc56f2	17	9	1,45

in face swapping detection. The results presented demonstrate a promising ability to reliably identify the specific image used in the face-swapping process. In particular, the study highlights the importance of having access to the original source and target images and shows that re-performing the face-swapping process yields the best detection results. However, our approach is flexible in that it also allows detection using the manipulated image itself as a starting point when the original images are unavailable.

An important result of our work is the resilience of the detection process to common post-processing transformations. In particular, lossy compression and scaling, both of which are expected in practical scenarios following the creation of manipulated images, were shown to have only a minor impact on detection accuracy. This robustness underscores the potential applicability of our method in real-world settings where images may undergo such modifications.

Table 2: 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

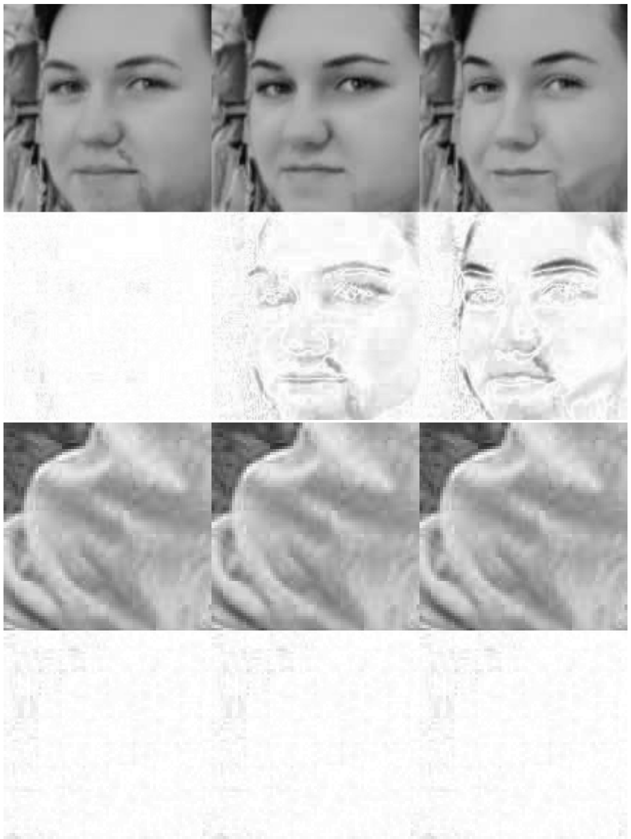

Figure 13. Examples for face fusion 3.0 results. Top: Successful example 0bc74437, Bottom: Wrongly detected face region of set 10dc56f2 leads to background region comparison.

As this is the first work to address this specific aspect of face swapping detection, further research is crucial to expand its scope and robustness. Future studies should investigate the impact of additional post-processing attacks, such as blurring or stronger compression, on detection accuracy. Exploring these scenarios will provide deeper insights into the limitations of the current approach and guide improvements that will increase its reliability under diverse and challenging conditions.

In summary, while our work lays a solid foundation for source identification of face-swapped images, it invites further experimentation and development to refine and strengthen detection methods in the face of evolving manipulation techniques and processing challenges. The approach is straightforward and easy to implement, which can be an advantage in forensic scenarios as the results are easy to explain and to reproduce.

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

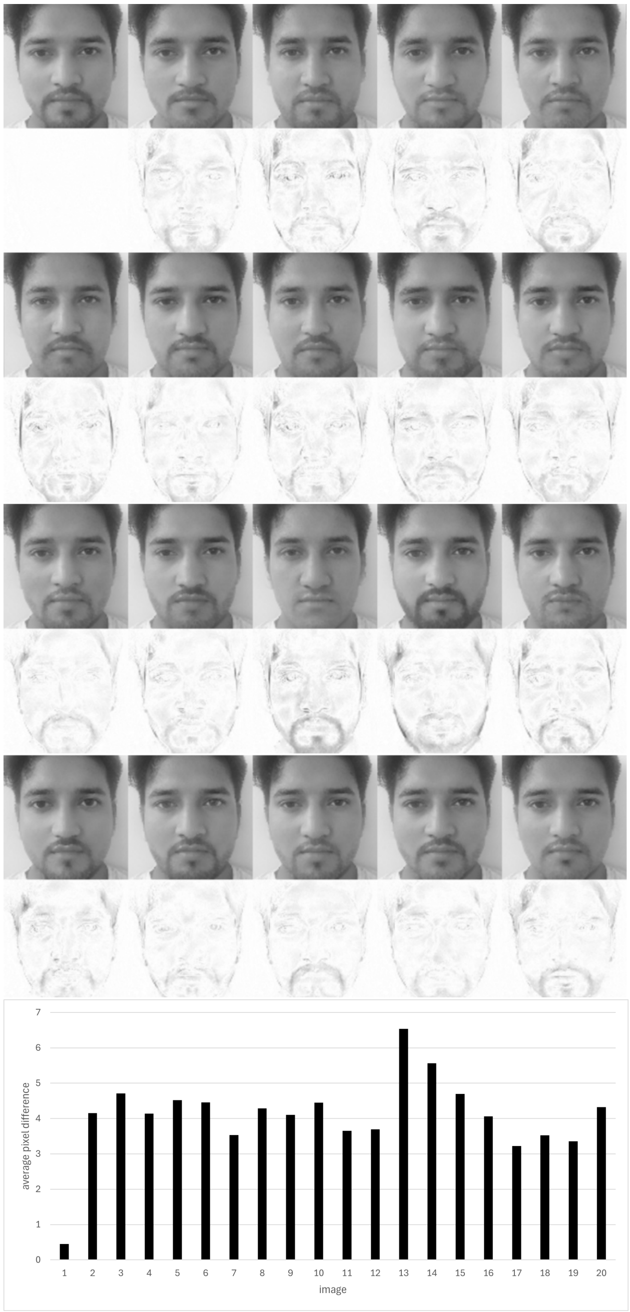

Figure 14. Comparison of 20 face versions.

[1] Sanoojan Baliah, Qinliang Lin, Shengcai Liao, Xiaodan Liang, and Muhammad Haris Khan. Realistic and efficient face swapping: A unified approach with diffusion models. *arXiv preprint arXiv:2409.07269*, 2024.

[2] Minh Dang and Tan N Nguyen. Digital face manipulation creation and detection: A systematic review. *Electronics*, 12(16):3407, 2023.

[3] Xinyi Ding, Zohreh Raziei, Eric C Larson, Eli V Olinick, Paul Krueger, and Michael Hahsler. Swapped face detection using deep learning and subjective assessment. *EURASIP Journal on Information Security*, 2020:1–12, 2020.

[4] David C Epstein, Ishan Jain, Oliver Wang, and Richard Zhang. Online detection of ai-generated images. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages

382–392, 2023.

[5] Joel Frank, Thorsten Eisenhofer, Lea Schönherr, Asja Fischer, Dorothea Kolossa, and Thorsten Holz. Leveraging frequency analysis for deep fake image recognition. In *International conference on machine learning*, pages 3247–3258. PMLR, 2020.

[6] Faraz Ghasemzadeh, Tina Moghaddam, Jingming Dai, Joobeom Yun, and Dan Dongseong Kim. Towards generalized detection of face-swap deepfake images. In *Proceedings of the 3rd ACM Workshop on the Security Implications of Deepfakes and Cheapfakes*, pages 8–13, 2024.

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

[8] Baojin Huang, Zhongyuan Wang, Jifan Yang, Jiaxin Ai, Qin Zou, Qian Wang, and Dengpan Ye. Implicit identity driven deepfake face swapping detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4490–4499, 2023.

[9] Jun Jiang, Bo Wang, Bing Li, and Weiming Hu. Practical face swapping detection based on identity spatial constraints. In *2021 IEEE international joint conference on biometrics (IJCB)*, pages 1–8. IEEE, 2021.

[10] SS Volkova and AS Bogdanov. A deep learning approach to face swap detection. *International Journal of Open Information Technologies*, 9(10):16–20, 2021.

[11] Zitong Yu, Yunxiao Qin, Xiaobai Li, Chenxu Zhao, Zhen Lei, and Guoying Zhao. Deep learning for face anti-spoofing: A survey. *IEEE transactions on pattern analysis and machine intelligence*, 45(5):5609–5631, 2022.

[12] Ying Zhang, Lilei Zheng, and Vrizlynn LL Thing. Automated face swapping and its detection. In *2017 IEEE 2nd international conference on signal and image processing (ICSIP)*, pages 15–19. IEEE, 2017.

[13] Yixuan Zhu, Wenliang Zhao, Yansong Tang, Yongming Rao, Jie Zhou, and Jiwen Lu. Stableswap: Stable face swapping in a shared and controllable latent space. *IEEE Transactions on Multimedia*, 2024.

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 Principle Investigator at ATHENE and represents IT Forensics and AI security. Before he was Principle Investigator at CASED with the topics Multimedia Security and IT Forensics.

Marco Frühwein has been a research associate at Fraunhofer SIT since 2021. He obtained his Bachelor's degree in 2017 and subsequently his Master's degree in 2018 at Darmstadt University of Applied Sciences in Optotechnology and Image Processing. During his studies, he focused on digital image processing. He is working on projects for forgery-proof barcode systems and deeplearning-based recognition of archaeological objects from camera images. He is also involved in the detection of AI-based content and has developed his own diagnostic tools for this purpose. He regularly works on IT forensic reports for the forensic evaluation and authentication of image and video data.

electronic
IMAGING

Image Engineering

STUDENT/YP SHOWCASE
PURDUE