https://doi.org/10.2352/EI.2025.37.4.MWSF-303

Robustness of Reverse Image Search Engines in the Era of AI-Generated Media

Raphael Frick, Fraunhofer SIT — ATHENE Center (Germany); Felix Stein, Technische Universität Darmstadt (Germany); Katharina Wallrabenstein, Technische Universität Darmstadt (Germany); Sascha Zmudzinski, Fraunhofer SIT — ATHENE Center (Germany)

Abstract

Reverse Image Search engines play a critical role in Open Source Intelligence (OSINT) and media forensics. They help uncover the origins of images, restoring original contexts and verifying authenticity. This capability is essential for distinguishing benign images from manipulated content. However, the robustness of these search engines is increasingly challenged by the application of post-processing operations and advancements in generative AI, which can produce highly realistic synthetic images. This paper analyzes the robustness of current reverse image search engines and highlights several shortcomings, emphasizing the need for enhanced resilience against AI-generated content to maintain their effectiveness in OSINT and media forensics.

Introduction

In today's modern, interconnected digital world, the sharing of images, videos, and audio has become ubiquitous. While some of this media originates directly from cameras or recording devices, a significant portion undergoes editing or enhancement through various software tools. For instance, filters are often used to blur facial features or refine skin texture. Earlier, these adjustments relied on traditional signal processing techniques, but the advent of artificial intelligence has drastically improved the speed and precision to create such modifications. Today, AI can not only be used to manipulate existing media but also enables the creation of entirely new content from scratch.

As these technologies become more sophisticated and harder to detect with the human eye, they bring about complex challenges. Manipulated media is increasingly used to harm reputations, facilitate fraud, and spread false information on a wide scale. Addressing these threats requires reliable detection methods, yet this task is exceptionally challenging. Many existing solutions based on deep learning fail to adapt effectively to the rapidly advancing generative techniques. Furthermore, the ability of detection systems to remain robust against post-processing operations, such as the artificial introduction of noise, has been shown to significantly diminish their effectiveness. Thus, detection methods often struggle to accurately identify face swaps or other manipulations when subjected to such transformations.

Many manipulated images and videos, however, originate from content that has already been shared online, making them more accessible for misuse. In a case study analyzing the dissemination of multi-modal disinformation during the Covid-19 pandemic in Germany between 2021 and 2022 as part of the ATHENE project Disco[1], we identified that much of the deceptive media was not generated by AI but rather consisted of existing images repurposed out of their original context. Frequently, photographs taken during unrelated events were reintroduced in misleading contexts to spread false narratives.

To address such issues, reverse image search engines are a valuable tool for tracing media back to its original context or identifying unaltered versions before manipulation. However, these tools face limitations similar to "blind" detection techniques that rely solely on analyzing the image under test [3, 6]. The application of strong post-processing operations, such as resizing or heavy filtering, can greatly reduce the success rate of these systems, making the detection of forgeries even more difficult. The Image Similarity Challenge organized by Meta in 2021 focused on the development of techniques capable of retrieving query images that had undergone extensive post-processing [4]. The findings from the challenge highlighted that the proposed methods frequently encountered challenges when dealing with image compositions, as well as images that had been cropped or blurred. Moreover, the results underscored the necessity for continued advancements, as even the most successful methods demonstrated limitations in accuracy across various sub-tasks [11, 10].

In this paper, we delve into the impact that various types of post-processing operations have on the performance of contemporary openly accessible reverse image search engines. Furthermore, we explore their applicability in detecting deepfakes, with the goal of gaining deeper insights into the obstacles presented by the advancements of AI-supported media manipulation technologies.

Retrieval of Reliable Information

Given the exponential proliferation of information shared online, and particularly through social networking platforms, the identification of reliable content has become increasingly complex. To mitigate the effort required for evaluating manipulated or synthetically generated media, the integration of analytic filtering methodologies during the data acquisition phase emerges as a viable solution (refer to Figure 1).

One technique that has demonstrated significant promise in recent years is the identification of fact-check-worthy content on social media [8, 1]. Rather than undertaking the exhaustive verification of all uploaded material for authenticity, this approach focuses on filtering content based on its degree of factuality. Factuality may be inferred from textual descriptions accompanying visual media (e.g., images or videos) as well as from the intrinsic attributes of the media itself. Empirical studies, however, indicate that textual information alone often serves as a robust indicator for determining whether a post necessitates further examination [5].

The evaluation process can be conducted via two principal methodologies: manual expert consultation or automated foren-

[1] DisCo – Disinformation and Corona

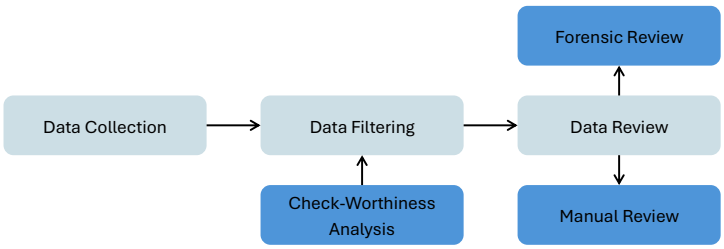


Figure 1: Process of obtaining reliable information.

sic techniques. Each of these approaches presents inherent challenges. Identifying domain-specific expertise can be intricate, while automated forensic methods, although capable of detecting manipulations, frequently yield outputs such as heatmaps that are challenging to interpret for non-specialist users.

Nevertheless, semi-automated approaches, such as reverse image search engines, offer practical utility. These systems facilitate the identification of altered images by tracing them back to their authentic sources, thereby assisting users in assessing the credibility of posts and supporting the retrieval of verifiable information.

Reverse Image Search Engines

Reverse image search engines represent advanced tools designed for querying information using images as input, instead of text-based queries. They analyze the visual attributes of an image to locate comparable or associated media on the internet.

Over recent years, multiple search engine providers have incorporated reverse image search functionalities into their platforms alongside conventional options for searching web-pages and images through textual queries. Among the most prominent reverse image search engines available today are Google, Microsoft Bing, and Yandex. While all of these enable users to conduct image-based searches, each engine offers distinct supplementary features that can add further precision to the search results.

Microsoft Bing - "Visual Search"

Microsoft Bing's visual search functionality was initially launched in 2009 [2]. At the time of its introduction, its capabilities were limited to identifying images that closely resembled the input image.

In its current iteration, however, the search engine has significantly expanded its functionality. It is now equipped to identify individuals within images and to extract and translate textual content from query images through the application of optical character recognition (OCR) technology. Moreover, Microsoft Bing's visual search functionality includes the capability to focus on specific regions within an image by allowing users to crop it. This enables a more refined and targeted search, enhancing the precision of the results obtained from the query image.

Google - "Search by Image"

Several years after the introduction of Microsoft Bing's visual search functionality, Google unveiled its own reverse image search engine in 2011 [3]. Although initially offering limited functionality, its capabilities have been significantly expanded over time.

In 2022, the integration of Google Lens into the platform marked a significant advancement. This enhancement enabled the translation of textual content identified within images, along with the ability to crop images for increased search precision. Beyond identifying visually similar images, the tool can also locate webpages that feature the exact same image.

Yandex - "Sibir / Cibir"

Among the three search engines, Yandex was the last to introduce reverse image search capabilities. Launched in 2013 [4], it was able to find similar images by splitting the images into so-called visual phrases. This technology converts user-submitted images into numerical "visual phrases," representing their key features, which are then matched with images found online.

Comparable to its competitors, the system has progressively incorporated a range of advanced functionalities over the years. These include OCR-based text extraction and translation, identification of websites hosting the same image, and retrieval of visually similar images. Additionally, when individuals are depicted within an image, the system identifies and provides information about the person. Furthermore, it offers an image-cropping feature to enhance the precision of search results.

Benchmarking Framework

To evaluate the robustness against post-processing techniques and the effectiveness in identifying face-swapping deepfakes of current state-of-the-art reverse image search engines, a benchmarking framework has been developed. The overall overview of the framework is displayed in Figure 2 and it consists of various components, which will be explained in the following.

Augmentations

Prior to conducting tests, each image within the test dataset is subjected to a series of augmentation processes. Augmentations, in this context, encompass typical image processing operations that may be employed to subtly modify genuine images or introduce malicious alterations. A total of eight distinct augmentation techniques were utilized, which are outlined below.

- **Compression Addition:** Reduces image quality using JPEG compression, with the level of compression specified as a percentage. A quality factor of 10 was used in our experiments to simulate high image compression.
- **Noise Addition:** Introduces noise to images, with options like Gaussian or salt-and-pepper noise types, and adjustable intensity levels.

[2] Microsoft Bing - "Visual Search"
[3] Google - "Search by Image"
[4] Yandex - "Sibir / Cibir"

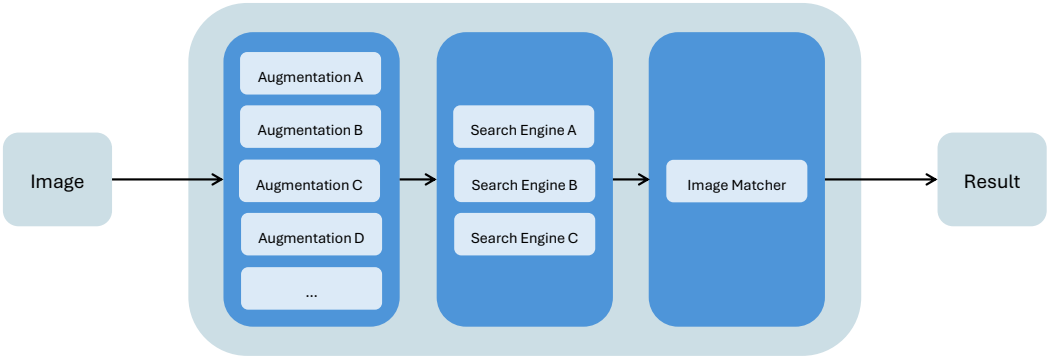


Figure 2: Overview of the benchmarking framework

- **Black-and-White Conversion:** Transforms the image into gray-scale.
- **Blocky Overlay:** Adds black squares to the image, where intensity controls the number of squares and size specifies their dimensions. To minimize disruption, a total of 45 blocks, each measuring 25 pixels, were inserted. This approach ensures a balance between augmenting the image and maintaining its overall coherence.
- **Image Blurring:** Uses Gaussian blur to soften the image, with configurable kernel size and intensity values. In our experiments, a kernel size of 25 was used alongside a sigma value of 10.
- **Cropping:** Removes portions of the image from the top, bottom, left, or right based on defined percentages. Here 25% were used to halve the dimensions of the image.
- **Mirroring:** Flips the image along the horizontal axis.

Examples of the applied augmentations can be viewed in Figure 3.

Beyond the post-processing augmentations, the evaluation also incorporated face swaps [2] and facial reenactment [7] syntheses (Figure 4). The first technique involves substituting the original facial texture with an arbitrary texture, thereby completely altering the appearance of the face. In contrast, the latter approach focuses exclusively on modifying head and lip movements, predominantly influencing facial expressions without replacing the underlying texture. These tests were introduced to further examine the ability of the search engines to identify and detect AI-manipulated content effectively.

Reverse Image Search Engines

Following the augmentation of the input images, each image is processed using the three specified reverse image search engines. While certain search engines, such as Google and Yandex, offer functionality to locate web-pages hosting the exact same image, this study exclusively employed similarity-based search across all three engines. Given the extensive augmentations applied to the images, their inclusion in the ”exact match” search results was deemed improbable. For each search engine, the top 10 search results were identified and subsequently evaluated.

Image Matching

To evaluate the ability of the search engines to accurately identify the original, unaltered source image, the retrieved search results were compared to the non-augmented image. Cryptographic hash functions, which generate divergent results for files subjected to variations in compression, scaling, or cropping, were deemed unsuitable for this comparison. Instead, robust image hashing techniques were utilized. Specifically, the pHash algorithm [12] was employed, enhanced with an extension to improve its resilience against cropping, ensuring reliable matching under such conditions [9].

Experiments

The subsequent sections provide a description of the experimental setup used for the benchmark test, followed by a comprehensive analysis and discussion of the results obtained.

Setup

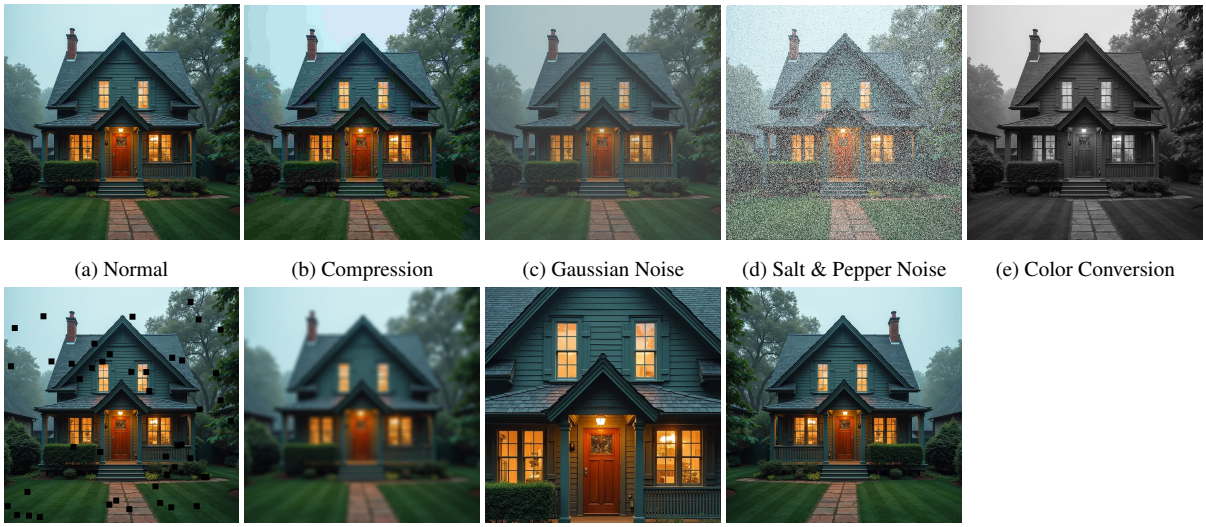
The test set comprises a total of 90 images, systematically categorized into three distinct topics: political affairs, the Covid-19 pandemic, and miscellaneous subjects, with 30 images allocated to each category. By incorporating both older images and those that have been widely shared across the internet, the dataset is designed to minimize potential biases arising from load balancing or outdated indexing within the database. In addition, 20 face swapping and facial reenactment videos were created with the help of SimSwap [2] and LivePortrait [7] to assess the engine’s capabilities of identifying deepfakes.

Results

The results of the robustness evaluation are presented in Table 1. It is important to note that this assessment is intended solely for the purpose of identifying limitations in current state-of-the-art search engine algorithms and to facilitate a discussion on potential directions for further improvement. Accordingly, the identities of the search engines have been anonymized.

The performance of the search engines exhibits notable variation. While Search Engine A and Search Engine B demonstrate strong capabilities in identifying most images, Search Engine C achieves success in only approximately half of the instances, highlighting a significant disparity.

During the testing phase, an interesting observation was made regarding one of the search engines, which leveraged its OCR capabilities to identify similar images. While the engine occasionally failed to retrieve the exact original or augmented image, e.g., in cases involving protests against Covid-19 restrictions, it demonstrated the ability to locate other images depicting protests on the same thematic subject. This observation highlights that certain search engines incorporate the extraction of diverse



(a) Normal (b) Compression (c) Gaussian Noise (d) Salt & Pepper Noise (e) Color Conversion

(f) Blocky Overlay (g) Blur (h) Cropping (i) Mirroring

Figure 3: Examples of the augmentations used for benchmarking the robustness of reverse image search engines.

features beyond visual content to generate their search results as textual data recognized through OCR, appear to play a significant role in enhancing their capability to retrieve contextually relevant images.

In the context of robustness analysis, the findings indicate that JPEG compression represents the least challenging augmentation for classification. This is evidenced by consistently high precision scores across all three search engines: Search Engine A achieves a precision of 0.8442, Search Engine B records a precision of 0.7532, and Search Engine C achieves a precision of 0.5714. These results suggest that the features introduced by JPEG compression are effectively recognized and classified by the algorithms with a high degree of accuracy.

Conversely, mirror augmentation emerges as the most difficult transformation to classify. Search Engine A records an exceptionally low precision score of 0.0130, reflecting substantial difficulty in detecting this specific augmentation. In contrast, Search Engine B achieves a superior precision of 0.8442, while Search Engine C attains a moderate precision score of 0.5714. These significant disparities underscore the variability in the algorithms' capabilities when handling this transformation. Images augmented with salt-and-pepper noise posed significant challenges for the search engines during retrieval. In many cases, the noise was misinterpreted as snow, contrasting with augmentations such as the black blocky overlay, leading to a notable reduction in performance. Specifically, precision values dropped to 0.1299 for Search Engine A, 0.2078 for Search Engine B, and 0.4286 for Search Engine C, underscoring the difficulty of accurately detecting such augmentations across different algorithms.

Overall, while the classification performance across most augmentations exhibits general consistency among the search engines, certain augmentations reveal notable differences. For example, Search Engine A faces pronounced challenges with mirror transformations but excels in classifying JPEG compression and Gaussian noise. Conversely, Search Engine C demonstrates more uniform performance across augmentations, albeit with comparatively lower precision scores.

Beyond the robustness tests, an additional evaluation was conducted to determine the capability of state-of-the-art reverse image search engines in identifying the source material of deepfake manipulations, specifically focusing on face swaps and facial reenactments. This analysis aimed to assess their effectiveness in detecting and tracing these advanced AI-driven content alterations. The results are displayed in Table 2.

As evidenced by the results, Search Engine C, while performing poorly in identifying previous images, exhibited notable success in detecting face swaps. In contrast, Search Engine B struggled significantly, failing to detect approximately half of the face swaps, whereas Search Engine A demonstrated slightly better performance in comparison.

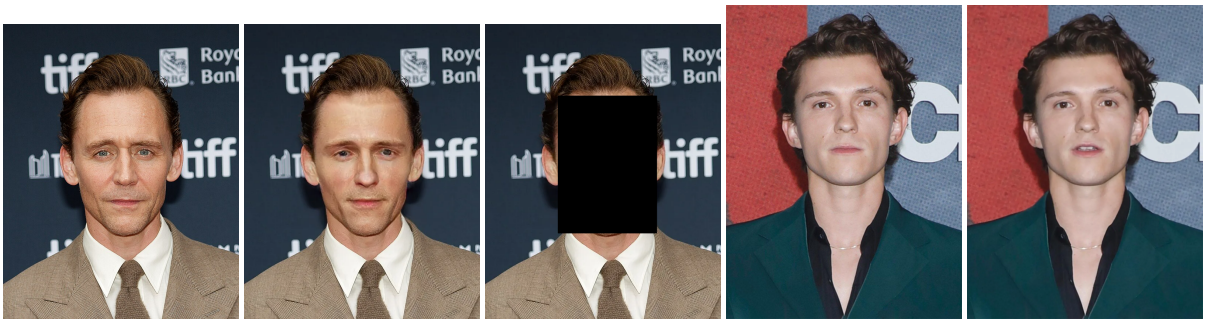
In cases where the authentic source image could not be identified, the deepfaked individual was often matched to the person depicted in the retrieved image. Although blacking out the face improved results in certain cases, it generally did not encourage the search engines to focus on the background. Instead, blurred facial features were frequently observed in the retrieval results, ultimately reducing overall precision.

Facial reenactment methods, which primarily alter facial expressions without significantly modifying the face itself, posed minimal challenges for the search engines. None of the search engines exhibited difficulties in identifying these alterations. This suggests that while current search engines are proficient at handling minor modifications to images, they encounter greater challenges when faced with substantial changes, such as those introduced by face swaps.

Conclusion

In this paper, we present a benchmarking framework aimed at evaluating the robustness of state-of-the-art reverse image search engines. Since forensic detection methods often lack generalizability or are too complex for use by the general public, reverse image search engines frequently serve as the most practical solution.

The robustness evaluation, conducted using a dataset of 90 images subjected to 8 different augmentations, identified that certain augmentations, such as blocky overlays, JPEG compression,



(a) Original Image (b) Face Swap (c) Pre-Processed Image (d) Original (e) Reenacted Image

Figure 4: Examples of face swaps and facial reenactments.

Search Engine	Normal	Overlay	Blur	Color Conversion	JPEG	Crop	Mirror	Gaussian Noise	Salt & Pepper Noise
A	0.8182	0.8182	0.4675	0.5714	0.8442	0.2078	0.0130	0.7792	0.1299
B	0.7403	0.7403	0.4156	0.6104	0.7532	0.3896	0.8442	0.7792	0.2078
C	0.5844	0.5065	0.4545	0.6104	0.5714	0.4805	0.5714	0.5714	0.4286

Table 1: Comparison of precision values for various processing operations across Search Engines A, B, and C.

Search	Face Swap	Facial Reenactment
A	0.5714	1.0
B	0.4280	1.0
C	0.8571	1.0

Table 2: Comparison of precision values for detecting deepfakes using Search Engines A, B, and C.

and Gaussian noise, had minimal impact on search engine performance. However, augmentations including blur, cropping, and salt-and-pepper noise posed significant challenges, often distracting the search engines and resulting in degraded performance. These findings underscore the need for further research to improve robustness against geometric transformations and strong artificial noise and blur.

Furthermore, while the detection of facial reenacted videos was achieved with high precision, identifying face swaps presented greater challenges. In certain cases, pre-processing of images was required, which involved obscuring the swapped face to direct the matching process toward the background. However, this approach failed to yield improved results in most instances. Despite these limitations, the findings demonstrate that current search engines possess a degree of effectiveness in detecting deepfakes, even without specialized forensic knowledge. Nonetheless, their applicability remains constrained, particularly when significant alterations have been applied to an image.

Future research directions could involve assessing long-term robustness and incorporating a broader range of AI-supported manipulation techniques, such as inpainting.

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] Firoj Alam et al. “Overview of the clef-2023 checkthat! lab task 1 on check-worthiness of multimodal and multigenre content”. In: (2023).

[2] Renwang Chen et al. “SimSwap: An Efficient Framework For High Fidelity Face Swapping”. In: *MM ’20: The 28th ACM International Conference on Multimedia*. 2020.

[3] Brian Dolhansky et al. *The DeepFake Detection Challenge (DFDC) Dataset*. 2020. arXiv: 2006.07397 [cs.CV]. URL: https://arxiv.org/abs/2006.07397.

[4] Matthijs Douze et al. *The 2021 Image Similarity Dataset and Challenge*. 2022. arXiv: 2106.09672 [cs.CV]. URL: https://arxiv.org/abs/2106.09672.

[5] Raphael Antonius Frick and Martin Steinebach. “Improved identification of check-worthiness in social media data through multimodal analyses.” In: *ROMCIR@ ECIR*. 2024, pp. 31–40.

[6] Raphael Antonius Frick and Martin Steinebach. “One Detector to Rule Them All? On the Robustness and Generalizability of Current State-of-the-Art Deepfake Detection Methods”. In: *Electronic Imaging* 36.4 (2024), pp. 332–1–332–1. DOI: 10.2352/EI.2024.36.4.MWSF-332. URL: https://library.imaging.org/ei/articles/36/4/MWSF-332.

[7] Jianzhu Guo et al. “LivePortrait: Efficient Portrait Animation with Stitching and Retargeting Control”. In: *arXiv preprint arXiv:2407.03168* (2024).

[8] Maram Hasanain et al. “Overview of the CLEF-2024 CheckThat! lab task 1 on check-worthiness estimation of multigenre content”. In: *25th Working Notes of the Conference and Labs of the Evaluation Forum, CLEF 2024*. CEUR Workshop Proceedings. 2024, pp. 276–286.

[9] Martin Steinebach, Huajian Liu, and York Yannikos. "Efficient Cropping-Resistant Robust Image Hashing". In: *2014 Ninth International Conference on Availability, Reliability and Security*. 2014, pp. 579–585. DOI: 10.1109/ARES.2014.85.

[10] Wenhao Wang et al. *D^2LV: A Data-Driven and Local-Verification Approach for Image Copy Detection*. 2021. arXiv: 2111.07090 [cs.CV]. URL: https://arxiv.org/abs/2111.07090.

[11] Shuhei Yokoo. *Contrastive Learning with Large Memory Bank and Negative Embedding Subtraction for Accurate Copy Detection*. 2021. arXiv: 2112.04323 [cs.CV]. URL: https://arxiv.org/abs/2112.04323.

[12] Christoph Zauner. "Implementation and Benchmarking of Perceptual Image Hash Functions". Master's thesis. Hagenberg, Austria: Upper Austria University of Applied Sciences, Hagenberg Campus, 2010.

Author Biography

Raphael Antonius Frick is a research associate at Fraunhofer SIT / ATHENE Center in Darmstadt and a PhD candidate at Technische Universität Darmstadt. In the "Media Security & IT Forensics" department, he researches new detection methods for AI-generated and manipulated audiovisual data, as well as techniques for identifying disinformation in social media across multiple modalities. Additionally, he investigates the possibilities of synthetic data to achieve other security goals such as anonymity and to improve the robustness and performance of deep learning classifiers.

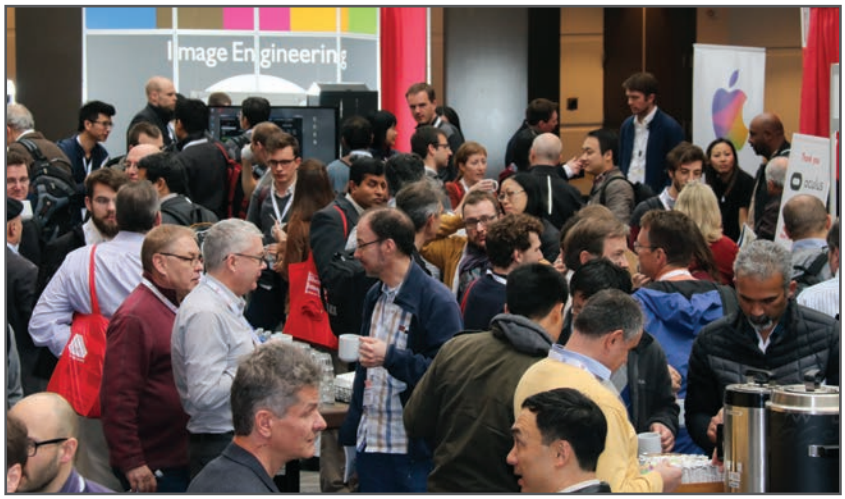
Felix Stein and Katharina Wallrabenstein pursued their studies at Technische Universität Darmstadt, where they each earned a master's degree in computer engineering.

Sascha Zmudzinski is a senior researcher at Fraunhofer SIT in the Media Security and IT Forensics division. He received his PhD at the TU Darmstadt in 2017 for his work on authentication audio watermarking.

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