Report from Dagstuhl Seminar 23021

# Media Forensics and the Challenge of Big Data

**Irene Amerini***[1], **Anderson Rocha***[2], **Paul L. Rosin***[3], and **Xianfang Sun***[4]

1 **Sapienza University of Rome, IT.** amerini@diag.uniroma1.it
2 **State University – Campinas, BR.** arrocha@unicamp.br
3 **Cardiff University, GB.** paul.rosin@cs.cf.ac.uk
4 **Cardiff University, GB.** sunx2@cardiff.ac.uk

## Abstract

With demanding and sophisticated crimes and terrorist threats becoming more pervasive, allied with the advent and widespread of fake news, it becomes paramount to design and develop objective and scientific-based criteria to identify the characteristics of investigated materials associated with potential criminal activities. We need effective approaches to help us answer the four most important questions in forensics regarding an event: "who," "in what circumstances," "why," and "how." In recent years, the rise of social media has resulted in a flood of media content. As well as providing a challenge due to the increase in data that needs fact-checking, it also allows leveraging big-data techniques for forensic analysis.

The seminar included sessions on traditional, deep learning-based methods, big data, benchmark and performance evaluation, applications, and future directions. It aimed to orchestrate the research community's efforts in such a way that we harness different tools to fight misinformation and the spread of fake content.

**Seminar** January 8–13, 2023 – https://www.dagstuhl.de/23021
**2012 ACM Subject Classification** Applied computing → Computer forensics; Computing methodologies → Image manipulation
**Keywords and phrases** Digital forensics, Image and video authentication, Image and video forensics, Image and video forgery detection, Tampering detection
**Digital Object Identifier** 10.4230/DagRep.13.1.1

## 1 Executive Summary

*Anderson Rocha (State University – Campinas, BR)*

This summary summarizes the outcomes of our Dagstuhl Seminar. The seminar focused on

- important issues,
- relevant problems, and
- adequate solutions.

In the end, we provide a panorama of the last 20 years of the area, its main advances, and its challenges ahead. We go through several key aspects regarding research and development, the translational gap between academia and industry, and what we need to fill this gap. We also highlight key areas and decisions we must focus on in the years ahead. Digital Forensics is part of our lives, and we need to bring together the best minds to tackle its open problems and challenges.

* Editor / Organizer

Media Forensics and the Challenge of Big Data, *Dagstuhl Reports*, Vol. 13, Issue 1, pp. 1–35
Editors: Irene Amerini, Anderson Rocha, Paul L. Rosin, and Xianfang Sun
DAGSTUHL REPORTS Dagstuhl Reports
Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

In our discussions, we confront traditional techniques with a range of new data-driven solutions, clearly pointing out the advantages and disadvantages of each kind of formulation. We also discuss their needs regarding scaling up to deal with ever-growing data sets.

We bring to bear aspects related to the development of fair, accountable, unbiased, and explainable solutions respecting directives such as the General Data Protection Regulation.

Finally, we point out that one of the biggest challenges nowadays in the presence of big data is the emergence of artificial intelligence generative techniques that easily allow the creation of never-seen-before content at unprecedented scale and speed, giving rise to what we have been referring to as synthetic realities. Only an orchestrated effort taking advantage of all different techniques from various formulations will allow us to fight back against such synthesized realities.

## 2 Table of Contents

**Executive Summary**

**Overview of Talks**

**Round Table Discussions**

# 3 Overview of Talks

## 3.1 Traditional Methods in Forensics

*Mauro Barni (University of Siena, IT)*

The dawn of multimedia forensics traces back to some seminal works published in the early 2000s by researchers previously working on steganalysis. Such works focused mostly on camera identification and detection of double JPEG compression. Since then, a large number of techniques have been developed dealing with a wide variety of forensic problems, including[1] detection of image resizing, color correction, detection of copy-move editing, detection of geometric and illumination inconsistencies etc... The methods developed in the first decade of multimedia research were based on the intuition that every step in the life of a multimedia document leaves within it a specific trace, often referred to as fingerprint or footprint, whose presence (or absence) can be used to derive some useful information about the past history of the document. Most methods developed in that period were adopting a model-based approach, according to which the process leading to the generation of the footprint was carefully modeled (by means of geometric or statistical tools), and the model used to develop sound footprint detection and/or localization techniques. In some cases, the forensic models were quite accurate allowing the development of extremely powerful tools. This was the case, for instance, of source camera identification based on PRNU (Photo-Response-Non-Uniformity) and detection of copy-move forgeries. This approach contrasts with more recent data-driven techniques based on deep neural network architectures, which base their success on the availability of massive amounts of training data. It is the goal of this talk to review the early history of multimedia forensics techniques and compare them with the most recent developments in the field, by paying particular attention to discuss the pros and cons of model-based and data-driven solutions, eventually advocating a synergistic use of both approached so to leverage on their complementary strengths.

## 3.2 Deep Learning in Multimedia Forensics

*Christian Riess (Universität Erlangen-Nürnberg, DE)*

Deep learning drives the development of new methods in Multimedia Forensics. Since deep learning derives decision rules from examples, it not only improves traditional model-based forensic tasks, but it also enables entirely new forensic tasks where analytic models cannot be constructed. However, after harvesting the immediate benefits of deep learning in forensics, we are now entering a period where its challenges become more visible.

In this talk, we discuss the most pressing challenges, and we raise the question for future directions of research. We hypothesize that a combination of the virtues of traditional methods with the power of deep learning can move the field significantly forward. The talk reviews four recent examples for such combinations, namely GAN fingerprints, image self-consistency, NoisePrint, and Bayesian learning.

[1] Here and afterwards we focus mainly on image forensics.

## 3.3 Compliance Challenges in Forensic Image Analysis Under the Artificial Intelligence Act

*Benedikt Lorch (Universität Innsbruck, AT)*

**Joint work of** Benedikt Lorch, Nicole Scheler, Christian Riess

In many applications of forensic image analysis, state-of-the-art results are nowadays achieved with AI methods. However, concerns about their reliability and opacity raise the question whether such methods can be used in criminal investigations from a legal perspective. In April 2021, the European Commission proposed the Artificial Intelligence Act, a regulatory framework for the trustworthy use of AI. Under the draft AI Act, high-risk AI systems for use in law enforcement are permitted but subject to compliance with mandatory requirements. In this paper, we summarize the mandatory requirements for high-risk AI systems and discuss these requirements in light of two forensic applications, license plate recognition and deep fake detection. The goal of this talk is to raise awareness of the upcoming legal requirements and to point out avenues for future research. For full details, see: [1].

### References

1 Benedikt Lorch, Nicole Scheler, and Christian Riess. Compliance Challenges in Forensic Image Analysis Under the Artificial Intelligence Act. In *30th European Signal Processing Conference (EUSIPCO)*, pages 613–617. IEEE, 2022.

# 4 Round Table Discussions

## 4.1 Day 1 – Initial Introductory Discussions

*Christian Riess (Universität Erlangen-Nürnberg, DE) – recorder of the session*

Thorsten Beck introduces his work and background. He works on scientific integrity education. He reports about a database of images that was compiled by researchers at the Humboldt-Elsevier Advanced Data + Text Centre (HEADT Centre), supported by publishers such as Elsevier, PLOS, Frontiers and others. The images stem from retracted papers. From the point of view of a journal reviewer, he is interested in solutions to detecting the (very diverse) types of manipulations to support the reviewing process with an automated screening for image-based scientific fraud.

A discussion emerges on the challenges of analyzing such images. Concerns are raised that even though the database consists of about 500 papers (which may seem to be a lot from some point of view), the individual cases are too diverse to think about a "universal" forensic tool. HEADT Centre also came to this conclusion, which is why they work with major publishers to collect enough data for creating a training set for machine learning approaches to specific types of tampering, and to develop specific tools for scientific reviewers. Such a tool might inform a reviewer for example whether an image has been previously used in a publication (image repurposing), or whether there are indications for a copy-move forgery in an image. It is clear that such tools cannot cover all cases of fraud and cannot replace humans in the decision-making process. On the other hand, the Dagstuhl participants agree that such well-defined computational tasks are feasible goals to achieve, and may help to catch some cases of scientific fraud.

The discussion shifts towards the different roles of images in different scientific fields. In biomedical imaging, an image sometimes constitutes the actual contribution of a paper, as a proof for some type of (expected or unexpected) behavior. Similarly, for imaging or image generation tasks, the image is the "proof of work", and hence an integral part of the contribution. In contrast, in other fields of computer science, images oftentimes only serve as illustrations, and are hence less of a priority for forensic verification. It is also noted that in various fields (biology, computer graphics, computer vision) images often serve the purpose of advertising a work. It is also pointed out that a single image of a successful experiment may in most cases not be sufficient scientific proof per se, since it does not indicate anything about error probabilities. An analogy to COVID tests is made, which may be positive, but to get a satisfying statement, one should actually present a number of different tests and a confidence value associated with their accuracy.

The discussion then shifts towards the difficulty of realistically, conservatively assessing the performance of tools. Scientific results are often too optimistic. One notorious issue are evaluation setups that are too simple and do not cover the diversity of real-world data. Another prevalent issue are side channels in the evaluation dataset that greatly simplify the classification task. Several participants report first-hand experiences with such side channels across various application fields.

The discussion further shifts towards comparisons in scientific works. It is raised that one issue in the community are unfair comparisons due to a lack of care in fully tuning the competing algorithms for a comparative evaluation. Martin Steinebach mentions the "Plagiarism Analysis, Authorship Identification, and Near-Duplicate Detection" (PAN) challenges at the workshop of the CLEF initiative (Conference and Labs of the Evaluation Forum). Here, instead of performing a self-evaluation, the workshop is centered around challenges where participants submit a docker image and all code is evaluated at a central site, to ensure a fairer comparison of scientific results. Another example is SHREC, a shape retrieval contest. Here, a list of results is computed by each participant, and sent to the organizers for comparison to the ground truth. A criticism is that the participants can look at the test set, which is not possible in the docker approach.

The discussion returns to the challenges that Thorsten Beck initially raised. In particular, participants address the question what forensic algorithms can be considered to work robustly. The participants agree that copy-move is quite mature, and up until a couple of years ago Photo-Response Non-Uniformity (PRNU) was also a go-to forensic cue. A conversation around copy-move emerges. Two possible use cases for copy-move forgeries are to either cover something (e.g. an airplane in the sky, or a car by wood) or to emphasize something (e.g., a crowd of people). For low-texture content, block-based detectors work better, but they are quite expensive to compute. For high-texture content, keypoint-based descriptors work quite well. Aerial images are a good use case for copy-move forgeries, since there are fewer perspective constraints. It was pointed out that creating a large-scale copy-move dataset is a challenge: if done manually, it takes a large amount of time. If done automatically, it exhibits typically hard edges at the cuts which put the usefulness of the data into question. An experience is reported that one can splice semi-automatically foreground and background objects, and thereby create a larger dataset.

One challenge in the transition from academic research to practice is that in practice the priors are greatly skewed. In academia, classification tasks are often set up such that there is a 50-50 chance to be correct when guessing. In practice, for example in CSAM detection or steganalysis, the odds are skewed to a prior probability of 1 out of $10^6$. Hence, even a low false positive rate overwhelms an analyst if she/he has to skim through all of these cases.

A short detour to video analysis. A case is reported where a Ph.D. student achieves better results on real data than on minimal, clean, academic data. It is acknowledged that this may be possible depending on the exploited cue. However, it may also be the case that the data preparation is just flawed, and a side channel is opened.

The discussion returns to the question on a characterization of forensic tools and their practical use. It is agreed among the group that forensic tools for proof in court are different from forensic tools to fight disinformation. Image reuse detection can be a good tool to fight disinformation.

The discussion then turns towards the broad family of detection or localization of synthetically generated visual content. First, how big is the actual thread from so-called DeepFakes? Maybe the actual threat vector is relatively narrow. A counter-example could be the Zelenskiy video ("drop your guns, surrender"), even though this was debunked after publication. However, variations of this task could bear realistic threat vectors in the future, for example to generate a synthetic image from a line of text. Hence detecting such synthetically generated data can be quite relevant. There are now also advanced possibilities for image retouching, e.g., by asking a model to replace a logo from a truck. With respect to the practical applicability, the networks are currently not good at creating interaction, e.g.: "draw a picture where Biden chokes Trump". From the perspective of image creation, it is better to take an image of someone choking another person, and to replace the two involved persons by Biden and Trump.

Finally, the discussion turns towards the role of deep learning in multimedia forensics research (this is an anticipation of the following seminar days). Deep learning papers are highly cited and are taking over many communities that would in principle also be interested in other approaches. For example, "traditional SIGGRAPH" people might also be happy about other methods, but deep learning dominates the conference. Deep learning also highly affects the funding landscape, and it is difficult to get a grant through without deep learning. Also, it impacts the culture of evaluation, in the sense that much more empirical comparisons are required and it is difficult to get something published without a demonstrated improvement over related work.

## 4.2 Day 1 – Traditional Methods in Forensics

*Christian Riess (Universität Erlangen-Nürnberg, DE) – recorder of the session*

**Opening.** Two statements are made to enter the conversation on traditional methods in forensics. First, a thought on traditional copy-move forgery detection (CMFD) algorithms is raised. They are tedious to parameterize, and a good strategy for overcoming that is unclear. However, it is appreciated that this is a classic, explainable image processing pipeline. Second, a thought on traditional methods versus deep learning methods is raised: it would be interesting to see hybrid approaches that make the best use of both paradigms.

**Remarks on Explainability.** An extended block of the discussion then focuses on explainability, which is oftentimes attributed as a key advantage of traditional forensic methods. The conversation is very lively, every seminar member contributes his or her perspective.

Why is explainability important? One could also do a controlled experiment to modify something and then check how good such a modification is detected, in order to convince for example a court of law of the workings of a method. It is a problem to base a court decision on an empirical evaluation without even any chance of understanding what is going on in the ML model. To illustrate this statement with an example from the US: parole decisions based on machine learning achieve the same performance as decisions that are reduced to three simple features: age, sex, and prior convictions [1]. However, the three features are much better understandable, and based on that understanding one can then discuss whether these features are a agreeable basis for the decision.

Hence, explainability is a critical asset in forensic investigations. It is noted that some classes of model-based methods are indeed well explainable, at least the main intuition behind them. Examples are physics-based geometric cues, like shadows and lighting conditions. One practical example from Brasil are video recordings that allegedly document a case of bribery. A forensic expert showed that there is a 1 in a million chance that the video is a forgery, otherwise it is real. This straw was used by the defendant, and only explainable systems can add further trust in the analysis.

There are several remarks that question the advantage of explainability in traditional forensic methods. It is noted that the claim that traditional forensic methods are inherently explainable comes with limitations, in particular when interacting with representatives from law enforcement without technical background. In this case, forensic cues that would otherwise be considered to be quite elementary from an information theoretic point of view, for example JPEG artifacts. This is even exacerbated in court, where the lawyer from the other party acts as an opponent. It happened in the past that expert witnesses failed to even explain linear interpolation in a satisfying way. On the other hand, law enforcement officers arguably also do not need to understand every detail of a method (who understands DNA analysis? Raise your hand!). From that point of view, input modifications and tracking of the change of output or an associated heatmap is the closest to the needs of the police. Hence, what you can explain to a non-technical audience is the ability of the tool, and the false positives, but you cannot explain the method itself. As a side note, judges then treat traditional methods and deep learning methods the same, since both are not explainable from their point of view. That doesn't negate the difference between traditional methods, whose functioning can be explained to suitably trained professionals, and deep learning tools, whose decision process is often obscure. However, the "level of obscurity" for AI-based methods differs with the type of task that AI fulfills. A binary classifier might indeed be unpredictable in its results. However, one could think of hybrid methods that use traditional elements and AI elements (e.g., AI for denoising the image, traditional methods for extracting hand-crafted features) which can be expected to satisfy these requirements very well. To conclude, it is important that our community develops more awareness to the other stakeholders (lawyers, judges) that are supposed to use our methods.

Regarding explainability in the context of the combination of traditional and deep learning techniques: a relatively easy scenario is when an image region is locally manipulated. In this case, deep learning and traditional methods can be cascaded. The deep learning approach can be used to find the relevant region, and traditional tools can be used in a manual analysis to verify this finding. The explainability comes in this case from the manual analysis. Such an approach is pursued in the analysis of fraud in scientific papers. Another option for combining traditional and AI methods is to use (AI-)learned filters and to re-inject them into a traditional method, e.g., by training a random forest.

It is noted that deep networks are also not entirely opaque. Instead, one can aim to get an impression about their behavior and some confidence that the correct functions are learned 1) by modifying the input and observing how the output behaves (e.g., noisier input should lead to less crisp results), 2) by backpropagating the decision scores to understand which parts of the input were most relevant for that decision, as it is done in gradCAM, and 3) by manually checking the learned filters. However, while this three-element list is not questioned per se, several participants note that these tools do not fit well to some multimedia forensics tasks. For example, heatmaps are usually not quantitative, oftentimes hard to interpret ("messy"), they are only useful for artifacts that coincide with certain locations in the image. For example, a sensor fingerprint (PRNU) can not really be visualized, so how can it be explained? Another example for a lack of possibilities for explainability is authorship attribution of a post at a social media platform. To make this example even more difficult, how can such an attribution be distinguished from a spoofing attempt?

**Compression of the next Generation – an opportunity for traditional forensic methods?** One remark is that HEIF images have not been forensically investigated. One problem here could be the lack of data. Researchers at Florence studied HEIF images and collected a small HEIF dataset. The analysis showed that, although the sensor pattern noise is still present on HEIF images, it is much more attenuated than in JPEG images, posing serious limitations to its effectiveness in realistic scenarios [2]. It is also noted that creating forgeries in HEIF data requires particularly high effort. On the other hand, it is not clear whether there is a forensic use case for such manipulations.

### Standardization of Forensic Methods

Regarding generalizability, existing forensic methods are doing quite good on known attacks and known processing chains, but we fail on generalization of social network laundered data and unknown new generators. So, generalization, explainability, certification/standardization are central issues. If a method is standardized, then it does not need to be explainable anymore. For example, DNA testing is standardized because at some point in the past scientist have proven that it works. However, how could possibly a deep neural network be standardized? And what if someone then demonstrates an adversarial example attack on the network? Will this not immediately destroy the validity of the proof and destroy the standardization, because a judge sees two images that look the same, but they create different predictions? Against this concerns one can argue that in Western countries, certification is usually done for the operator of a method, not the method itself.

### Evaluation of forensic methods: too far away from practice?

However, we are not quite in the situation to standardize methods, and one issue towards this lies in the evaluation.

One critique of traditional physics-based methods (e.g., methods that assess inconsistencies in shadows, perspective, or lighting) is that they only work in very controlled scenarios, whereas they are typically too constrained to be used in real world examples.

However, to be fair, this limitation to methods that work in lab environments is not only limited to physics-based methods. In practice, strong laundering of forensic traces happens when sharing images over social media. The sharing introduces recompression artifacts and geometrical modification on the uploaded visual content that degrades or erases the traces previously left by a manipulation, thus hindering the analysis. One specific example is that there is to our knowledge no paper on deep fake detection in the web.

In research evaluation we often make the simplifying assumption that we only need to decide whether one specific attack did occur or not. For example, we check for copy/paste, scaling or double compression. In real-world scenarios the challenge is more open. One often is tasked with stating whether *something* happened to the image, resulting in a manipulation of its perceived content. In practice, one strategy could be to run several detectors on the image, like double compression detection, inverse image search, stitching detection and more, and then to aggregate the different results in a graphical interface with an alert function that is fed back to the analyst who needs to decide about the evidence. The potential usefulness of such an approach is also reported in a TIP paper by Anderson Rocha's group. There, the output of many detectors is combined in a Bayesian way into a probability map. The success of this method may indicate that one needs multiple complementary methods.

### Public Code

It is acknowledged that today more code is made publicly available than "back in the days" for traditional forensic methods. However, it would be good to have more efforts to collect code and to benchmark existing approaches. Meanwhile, code is available from various groups. There are also some benchmarks.

However, there is no grander community work to publish code. Biometrics has good practices by conducting challenges. In forensics, there was the 1st IEEE IFS Challenge, and there were some other minor events (ICASSP 2017, NIST/DARPA MFC, deepfake challenges). Maybe it would be good to do a challenge with a) synthetic generators e.g. based on stable diffusion, b) photoshop, and c) synthetic generators and photoshop. The evaluation should then be done in a way that the generators are not known.

### Acceptance of Various Types of Evidence in Court

A generated piece of data, like a synthetic license plate, can not be accepted as proof at a court. However, this situation changes if an algorithm enhances an existing license plate, and an expert witness reads what he can decipher from this enhanced license plate. Besides the scenario of a court case, the second scenario is to read a license plate as an investigative cue. In this case, also machine learning classification is admissible (which would be impossible in court, due to the unknown error probability). As always, there are exceptions to this rule: in a case from the US, a person was sent to jail because his/her face was matched with a database, even though the person was innocent [4].

Then, it is discussed what national regulations exist for using a photograph or social media images in court. In a court case in the US, a social media picture was used to establish a link between the person and a gang. The photographer was asked whether the image/scene is real, and the photographer confirmed it. Hence, it does not necessarily need a technical method to authenticate images, there are also other ways. In Italy, it depends on a case-by-case basis. If the opposing lawyer does not challenge an image, then it should be admissible. Amped had a case where they challenged an image that was allegedly transported through WhatsApp, but in general, the judge can decide what is accepted as evidence.

From a technical point of view, it can be interesting to look into confidences for decisions. For example, a neural network can provide a confidence, and this can be a real benefit over traditional forensic methods, for example in super-resolution. This can also be a reason for revisiting AI methods in court cases: If I trust a network better than some flawed assumptions about a Gaussian distribution in a traditional forensic method, then it is probably better to use that network, argue why the method is better trusted, and provide its empirical accuracy to the court.

The threat of adversarial attacks should probably not be too much overstated for multimedia forensics. Adversarial attacks also exist for example for face detection. However, face recognition is a widely accepted technique, maybe because it is a visible cue. In our case, we are dealing with invisible cues, which could be the reason why it is more difficult for us to argue against adversarial attacks. However, in principle the threat assessment from adversarial attacks should in both cases be equal.

**GoF versus DL Forensics.** Traditional methods are maybe better suited for looking at one individual object, e.g., whether the shadow is fitting. However, in order to establish context between objects, then maybe machine learning methods can learn correlations that are otherwise inaccessible.

It is important to note that even our strongest traditional methods are limited in their generalizability in the field. For example, traditional CMFD detectors have a recall of about 20% on sufficiently difficult data (like scientific papers that are screened for fraud). This leaves a lot of work open.

From the perspective of a researcher: did we stop to do research in traditional methods because everything was done, or did we move to AI because we had no other choice due to the overall "AI wave"?

The rise of AI methods has also brought more datasets. Is there a way that we can benefit from these datasets with traditional methods? Arguably, the low-hanging fruits of traditional methods are taken, and the deep learning fruits were much easier to reach. For traditional methods, it could also be a selling point that the method only needs 10 images to calibrate, or that the method can generalize better than deep learning methods. But in any case, it is necessary to compare novel methods that follow the traditional paradigm also to AI methods. Such a comparison is difficult to do in a reasonable way, since traditional cues only pick up isolated aspects oftentimes, but nevertheless it has to be done.

**References**

1 Elaine Angelino, Nicholas Larus-Stone, Daniel Alabi, Margo Seltzer, and Cynthia Rudin. Learning certifiably optimal rule lists. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 35–44, 2017.

2 Daniele Baracchi, Massimo Iuliani, Andrea G. Nencini, and Alessandro Piva. Facing image source attribution on iphone x. In Xianfeng Zhao, Yun-Qing Shi, Alessandro Piva, and Hyoung Joong Kim, editors, *Digital Forensics and Watermarking*, pages 196–207, Cham, 2021. Springer International Publishing.

3 Jonathan W. Hak. The admissibility of video and photographs posted to social media: Inconsistent court rulings. `https://tinyurl.com/yc3eymcc`, 2020.

4 K. Hill. Another arrest, and jail time, due to a bad facial recognition match. The New York Times, Dec. 29, 2020, available: https://www.nytimes.com/2020/12/29/technology/facial-recognition-misidentify-jail.html.

5 ICCV. 1st workshop on traditional computer vision in the age of deep learning (TradiCV). `https://sites.google.com/view/tradicv`, 2021.

6 Wikipedia. Phantom of Heilbronn: Example issues of DNA forensics analysis. `https://en.wikipedia.org/wiki/Phantom_of_Heilbronn`, 2022.

## 4.3 Day 2 – Discussion on the AI Act

*Christian Riess (Universität Erlangen-Nürnberg, DE) – recorder of the session*

Benedikt Lorch presents challenges for the use of AI in criminal investigations that arise from the draft Artificial Intelligence Act. The presentation is followed by a discussion. It is clarified that the AI Act aims at companies/providers of AI solutions, not on AI methods per se. For example, it is not a DeepFake detector that is 'high risk' per se, but instead it is the application of a DeepFake detector in a court case, where the fundamental rights of the defendant are at stake.

The following discussion touches a number of concerns. One concern is that GDPR is preventing research on faces, because all data/models that is created in a non-GDPR conforming way is tainted, and (strictly speaking) it can not be used for research. Another concern is that the AI act will create obstacles not only to companies using AI for commercial use, but also to researchers. All the more that it is not clear if the restrictions and obligations also extend to the models used as initial point for fine tuning and transfer learning. Another question that is raised is whether the transparency requirements for companies in the AI Act should also be extended to research? Stating the limitations of the system is a good practice in papers, but not everyone does it, and some people write pseudo justifications.

## 4.4 Day 2 – Deep Learning Based Methods

*Irene Amerini (Sapienza University of Rome, IT) – recorder of the session*

### 4.4.1 Christian Riess: Deep Learning in Multimedia Forensics

Christian Riess opens the session with a stimulating presentation on the advantages and challenges of deep learning methods in multimedia forensics. As a sidenote, Teddy Furon's WIFS 2021 keynote is mentioned which highlights the analogies between ML security and the typical goals in information forensics and security [2].

He cites three works:

- GAN fingerprint (Marra et al) depends on upsampling in GAN. However, this trace is easily removed by compression [3]
- Self-consistency (Efron et al), they didn't do any assumption on the kind of the attacks
- NoisePrint (Verdoliva et al) [1]
- Detection of out-of-distribution samples (cases in multimedia forensics of out-of-distribution samples are an huge amount)
  - Supervised approach calibration: needs another dataset
  - Bayesian methods that model weights as probability distributions
  - Bayesian approach

The talk ends with some final questions:

- Tangible benefits of DL?
- Are we just replacing models by dataset?
- Other interesting DL methods?

**References**

1 Davide Cozzolino and Luisa Verdoliva. NoisePrint: A CNN-Based Camera Model Fingerprint. *IEEE Transactions on Information Forensics and Security*, 15:144–159, 2019.

2 Teddy Furon. WIFS 2021 keynote. https://www.youtube.com/watch?v=Gh2_tR-hgyU, 2021.

3 Francesco Marra, Diego Gragnaniello, Luisa Verdoliva, and Giovanni Poggi. Do GANs Leave Artificial Fingerprints? In *IEEE Conference on Multimedia Information Processing and Retrieval (MIPR)*, pages 506–511. IEEE, 2019.

### 4.4.2 Deep learning – Discussion

**Paul Rosin:** There are also limitations in deep learning, and it is unsettling that the choice of architecture feels somewhat random: e.g. is tanh activation better than sigmoid? You don't know unless you try it out empirically.

**Luca Cuccovillo:** Neural networks should learn forensically useful properties. For example, features to describe the reverberation properties of the room in which the recording took place.

**Martin Steinebach:** Adding transformed input to a neural network, e.g., additional frequency information, really helps.

**Christian Riess:** Agnostic about the attack. I would like a method that generalizes.

**Luca Cuccovillo:** Algorithms for audio synthesis are meant to create speech which sounds plausible, and produced by the voices the network was trained upon – not to create audio meant to overcome a forensics analysis. Rather than looking *only* for synthesis traces, we should also look if the distributions of (meaningful) features inside the evidence about, e.g., speaker identity, recording device, room acoustics fit the allegation or not. If not, then something is off with the recording.

**Paul Rosin:** My experience with deep learning is that although the results are often good in general, if we look in detail there can be a lot of flaws. I found this when we had to compare our non-DL approach to colorization with competing DL approaches; the latter were not as good as I expected from an initial superficial view. Sometimes, in an attempt to achieve good results it seems that you rather then spend time on hand crafting features, instead you had craft loss functions. But it can be difficult to control the output of these deep learning models. In comparison, with the traditional methods, to do what you want is trivial.

**Luca Cuccovillo:** When you want to deal with a lot of complexity you should use deep learning to cover this complexity. This can be done directly – e.g., to perform end-to-end single/double encoding detection – or indirectly, – e.g. to perform microphone identification in presence of strong background noise, using a network to remove the noise while preserving the colour of the microphone.

**Martin Steinbach:** In detection CSAM or fake news deep learning methods are working. Manipulation detection is not working well with deep learning methods of the box. We added spectral transformation as a second channel to the input data and the performance improved a lot.

**Paul Rosin:** An interesting topic is neurosymbolic AI, which combines neural and symbolic AI in order to better capture prior information than purely using machine learning.

**Benedikt Lorch:** In the past few years, deep learning has been applied to almost any application in multimedia forensics. In light of all the success stories, little attention has been given to the limitations of deep learning. Only now are the failure cases of deep learning receiving increasing attention.

**Isao Echizen:** Benefit of DL, data. For a Deepfake detector for a company you should vary the dataset. Provide simple models to companies and companies improve the model, continuing to train the model. For rolling out a deep neural network in a company, then the data is often quite limited.

**Thorsten Beck:** What are the implications of the lack of sufficiently large datasets for the development of DeepLearning models and resulting tools? Are artificially generated datasets able to contribute to the development of effective tools?

**Tiziano Bianchi:** Deep Learning for analyzing robustness of deep learning, but not used a lot. Maybe one of the tools that we need is on the explanation of out-of-distribution samples.

**Mauro Barni:** My impression is that with DL we are just replacing models with datasets. The limits of model-based methods is that they cannot be used in the absence of good models and they cannot be used outside the precise limits used to build the models. The limit of data-driven methods, conversely, is that they cannot be used in the absence of representative and vast datasets, and they cannot be used with data which is not coherent with the datasets used for training. One may argue, though, that datasets are easier to build, while good models describing the complexity of real life may simply not exist. On the other hand, model-based methods seems to generalize a bit better to situations deviating from the models.

**Mauro Barni:** Maybe it is the right thing to replace models with datasets: if you want to describe real life then data are better than models. So maybe models are more robust. Confidence is the key. If I want to describe images why shouldn't I use as many images as possible as examples. Dataset mismatch: is the same as model mismatch.

**Christian Riess:** Unsupervised confidence measure. Bayesian neural network – I like the paradigm.

**Marco Fontani:** Dempster Shafer theory to measure the confidence measure [2]. Law enforcement 5% authentication, 95% enhancement (AI dangerous for enhancement). Paper by Boato and Pasquini [9]: more real than real (AI-generated images are considered more real than real images by humans). AMPED also published a paper where celebrities faces were upsampled with bicubic interpolation and with deep learning, and the recognition rate was not really affected [3]. With AI super-resolution, you create an average face, but real faces may contain strange artifacts (e.g., scars, moles) that the network tends to neglect; these artifacts are the most valuable for law enforcement when doing face recognition.

**Irene Amerini:** The work proposed by Mayer at al [12] is an interesting DL-based method. The authors introduce a digital image forensics approach called forensic similarity, which determines whether two image patches contain the same forensic trace or different forensic traces. The system is evaluated determining whether two image patches were captured by the same or different camera model and manipulated by the same or a different editing operation and the same or a different manipulation parameter, given a particular editing operation. Regarding Deepfake detection many different DL-methods exist in the literature. Those methods suffer from a number of shortcomings some of which are particularly relevant for their applications, so to say, in the wild, where strictly controlled laboratory conditions do not hold. Another point that should be addressed is the detection of Deepfakes in real-time such as recognizing the fake contents in a video-call on a device like a smartphone. For this purpose it is necessary to design models with low inference time and a small number of parameters, able to run on hardware with limited memory but able to recognize the fake with an high accuracy.

**Alessandro Piva:** Farid had another paper with similar results to Boato and Pasquini [13]. What are your experiences of Continual Learning?

**Christian Riess:** You add training data on the fly without going to catastrophic forgetting. Good paper but I don't know if I want to use it in forensics. It is autonomous in general you have a plan when you decide to retrain. So I think it is difficult to apply in real forensics scenario. It is used in network intrusion detection and in biometrics.

**Lakshmanan Nataraj:** We have a couple of papers on seam carving, most recently in the CVPR media forensics workshop. Our experience in deep learning methods in video forensics: training and test data should be the same; changing model changes a little bit the accuracy

**Roberto Caldelli:** In our experiments, deep learning works super great for specific tasks, but generalization and vulnerability to adversarial attacks are a problem. Ablations are important to understand the impact of certain design decisions. Confidences are also important. Input perturbations are fundamental to understand what is happening inside of the network.

**Xianfang Sun:** image segmentation, super resolution. Data hungry not only forensic application but also other areas. The results should be scalable. Weak supervised learning is not so popular in forensics.

**Christian Riess:** the community is sometimes a bit slow to absorb insights from the ML and vision communities. For example, we used shallow networks for a while. The vision community extensively explored self-supervised learning to mitigate the data bottleneck. This is probably something that we should be paying more attention to.

**Law Ngai Fong:** extract noise pattern Siamese network, forensic similarity, metrics

**Roberto Caldelli:** 1. When you do a good training, with a sufficient number of data, the performances that deep neural networks can achieve are amazing but what about generalization, black box scenario, adversarial point of view? All these kinds of issue should be put on the table and be analyzed. 2. Methods that look inside the box (looks for activations and so on). A paper, we gave at ICIP 2019 analyzes the confidence, the internal layers and tries to understand what it is inside the black box (explainability); it considers the point of view of adversarial.

**Anderson Rocha:** Bot classification can be done either by content with a language model (this is our community) or based on connectivity (which is done in the field of network analysis)

**Anderson Rocha:** Smartphone authentication with multimodal: image, video, and audio reflection(!) Fusion is an important topic in forensics, because sometimes one signal is not strong enough. Example: we record biosignals with smart watches, then do anxiety classification, because the person e.g. is sweating, heart rate is increasing, but the person is standing. This needs to be validated with medical insights.

**Anderson Rocha:** What do you think are the biggest challenge in cross-modal algorithm design?

**Paul Rosin:** Are there datasets available?

**Anderson Rocha:** Yes, for various tasks.

## References

1 Thierry Denoeux. A k-nearest neighbor classification rule based on Dempster-Shafer theory. *IEEE Transactions on Systems, Man, and Cybernetics*, 25(5):804–813, 1995.

2 Velan Erik, Marco Fontani, Sergio Carrato, Jerian Martino, et al. Does deep learning-based super-resolution help humans with face recognition? *Frontiers in Signal Processing*, 2, 2022.

3 Federica Lago, Cecilia Pasquini, Rainer Böhme, Hélène Dumont, Valérie Goffaux, and Giulia Boato. More real than real: A study on human visual perception of synthetic faces [applications corner]. *IEEE Signal Processing Magazine*, 39(1):109–116, 2021.

4 Sophie Nightingale and Hany Farid. Synthetic faces are more trustworthy than real faces. *Journal of Vision*, 22(14):3068–3068, 2022.

## 4.5 Day 2 – Cross-Media Approaches for Multimedia Forensics

*Irene Amerini (Sapienza University of Rome, IT) – recorder of the session*

**Luca Cuccovillo:** It is a problem to analyse image and audio over time.

**Martin Steinbach:** Lipsynch movement of lips compare to voice. Synthetic tools that do that are good. Can be applicable

**Isao Echizen:** detection of fake news, inconsistency

**Martin Steinbach:** disinformation detection, take image out of the original content and reuse it, image search and take the text

**Luca Cuccovillo:** Finding duplicates is a problem, figure out how to do cross in social media, Next media to consider: the metaverse

**Paul Rosin:** Could GPS tracking be considered a new media?

**Thorsten Beck:** What about using video codecs?

**Christian Riess:** You can use image forensic tools that analyze key frames.

**Mauro Barni:** It is pretty obvious that a video sequence provides more information than its single frames taken in isolation, yet the current state of the art in video forensics shows that in most cases Working at the frame level is enough to get very good accuracy. Problems like lack of generalization are not easily solved by passing from frames to video sequences. Of course, I am not saying that working at the sequence level does not provide any advantage, this looks more like a limitation of currently available techniques.

Possibly, temporal based analysis of deepfake based on LSTM is a little bit less more prone to adversarial attacks in terms of transferability [1], still I do not know if this small advantage makes temporal analysis worth.

**Martin Steinbach:** Fraunhofer study on a tool for fake news detection

**Roberto Caldelli:** We have studied how to improve source identification by using different sensors on-board of a smartphone (e.g. accelerometer, gyroscope). Not necessarily adding different media improve the identification.

**Tiziano Bianchi:** useful for disinformation detection: text

**Irene Amerini:** Multi-modal approach is useful in the context of fact-checking. The general idea is to do topic mining on tweets to identify facts, e.g., the first tweets about covid at the time when it was not yet well known what it was. So the goal is to work on a system that knows how to map tweets and the images associated with it into a multi-modal embedding in which images and text pertaining to the same facts are close to each other. Why is this useful? Imagine that we find a tweet about a new fact, but we do not have enough elements in the tweet to tell whether it is true or false. With this system we can do retrieval of all tweets similar to the one I am considering, and through these I can get more information about that fact. Another example on the use of different media is related to social media provenance where images and videos data are considered together. The main reason behind such choice is that collecting datasets large enough to train neural networks for the task has become difficult because of the privacy regulations that have been enacted in recent years. To mitigate this limitation, in [10] authors propose two different solutions based on transfer learning and multitask learning to determine whether a video has been uploaded from or downloaded to a specific social platform through the use of shared features with images trained on the same task. Moreover they introduce a model based on multitask learning, which learns from both tasks simultaneously.According to our knowledge, this is the first work that addresses the problem of social media platform identification of videos through the use of shared features.

**Anderson Rocha:** authorship attribution. Connectivity graph (Facebook), authorship (emoji are important)

**Mauro Barni:** Multi-modal approaches surely make sense yet it is important that the various modalities are fused properly to avoid inheriting the weaknesses of the various modes rather than their strengths.

**Anderson Rocha:** Cross-modality parental control in real time. Images, videos, caption, audio. Process them in real time to block the video. It is a classification problem but you don't have time coherence or series of classifier that we combine over time. How to combine different modality over time → fusion. Doing this real time with no deep learning in our case. Sometimes audio says one thing, but the image says something different.

**Paul Rosin:** Data fusion is a common topic in computer vision, and there are many different approaches. Perhaps we can use some of these in forensics.

**Anderson Rocha:** For recent papers this is true. Jointly optimizing different modalities. Early fusion or decision fusion if you don't have a network. Which one is better depends.

**Irene Amerini:** Most of the methods for Deepfake detection rely on extracting salient features from RGB images to detect through a binary classifier if the image is fake or real. In [11] is proposed DepthFake, a study on how to improve classical RGB-based approaches with depth-maps. The depth information is extracted from RGB images with monocular depth estimation techniques. Using multi-modal information can help increase the performance of the detectors and in generalization capacity of these features with respect to deepfake generation techniques that have not been seen in training.

**Anderson Rocha:** How to combine temporal information. One of the challenges when dealing with multiple detectors across time is how to combine the different responses overtime so that temporal information is incorporated. This was, for instance, discussed in the paper "Multimodal data fusion for sensitive scene localization"[10] in which the authors propose a novel multimodal fusion approach to sensitive scene localization. The solution can be applied to diverse types of sensitive content, without the need for step modifications. Such solutions are key to deal with the ever-changing scenario of forensics in which actors keep proposing new ways of defeating detectors.

**Alessandro Piva:** Our experience on data fusion concerns the exploitation of both content-based features and file structure-based features for the identification of the source of video content (e.g. which brand of the source device, or in which social network was the content uploaded). The idea is to extend the work to exploit both audio and video features.

**Anderson Rocha:** Fusion is important and promising path in forensics. We are using smartwatches to capture biosignals. With different data we are able to understand what it is doing.

**Anderson Rocha:** What prevents you using a cross modality?

**Paul Rosin:** Lack of datasets.

**Mauro Barni:** Video and audio lip synchronization is quite popular, still frame by frame analysis seems to work better.

**Anderson Rocha:** This is a dataset bias

**Benedetta Tondi:** Maybe we need a bigger dataset.

**Anderson Rocha:** Generalization could help solve working with more modalities.

**Mauro Barni:** Maybe the current networks do not exploit well the availability of more than one single modality. For sure we need larger datasets, which are not easy to build in the multimodal case.

**Anderson Rocha:** Example of a work by Christian Riess on his PhD on reflectance for forgery detection. So do not exploit well the availability of more than one single modality. For sure it is important to transform the input.

**Mauro Barni:** I really think the way to go is to fuse results from GOF and AI-based methods.

**Alessandro Piva:** Continual learning: we investigate the potential of continual learning techniques to build an extensible social network identification neural network where multiple new tasks, each one comprising multiple new social platforms, are considered, in order to simulate the possibility that new social media can appear.

**Marco Fontani:** Reproducibility of the methods found in the literature is often impossible. The results are quite different. A problem can be a different dataset.

**Benedetta Tondi:** We should do all make efforts to release the code including the trained models, and also all the instructions for methods' training. Without that, reproducing results turns out to be a hard task in deep learning. Also, research advances so fast that we need to be able to run comparisons in a fast way.

**Anderson Rocha:** and if you can publish because of that?

**Anderson Rocha:** In video you cannot do cross-validation and this often happens.

**Mauro Barni:** Often the problem is the way you test your algorithm.

**Benedetta Tondi:** It helps in reproducibility (for company and for us to compare our results).

**Anderson Rocha:** In a Nature paper [14] they analyze 62 algorithm COVID Xray image detection. None working! When you submit a paper reviewer ask to compare with arXiv

**Christian Riess:** What do you do?

**Mauro Barni:** If the AE is not responding or insist that you should consider arXiv papers as state of the art, then you should talk to the EIC. IEEE, for instance, has a clear policy stating that arXiv papers CANNOT be considered state of the art and asking a comparison against arXiv papers is not allowed.

**Anderson Rocha:** And you have to compare with published papers not arxiv!

### References

1 Dongdong Lin, Benedetta Tondi, Bin Li, Mauro Barni, Exploiting temporal information to prevent the transferability of adversarial examples against deep fake detectors. *2022 IEEE International Joint Conference on Biometrics (IJCB).*

2 Thierry Denoeux. A k-nearest neighbor classification rule based on Dempster-Shafer theory. *IEEE Transactions on Systems, Man, and Cybernetics*, 25(5):804–813, 1995.

3 Velan Erik, Fontani Marco, Sergio Carrato, Jerian Martino, et al. Does deep learning-based super-resolution help humans with face recognition? *Frontiers in Signal Processing*, 2, 2022.

4 Marco Fontani, Enrique Argones-Rúa, Carmela Troncoso, and Mauro Barni. The watchful forensic analyst: Multi-clue information fusion with background knowledge. In *IEEE International Workshop on Information Forensics and Security (WIFS)*, pages 120–125. IEEE, 2013.

5 Information Processing Fraunhofer Institute for Communication and Ergonomics. Software that can automatically detect fake news. `https://www.fraunhofer.de/en/press/research-news/2019/february/software-that-can-automatically-detect-fake-news.html`, 2019.

6 Chandrakanth Gudavalli, Erik Rosten, Lakshmanan Nataraj, Shivkumar Chandrasekaran, and BS Manjunath. SeeTheSeams: Localized detection of seam carving based image forgery in satellite imagery. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1–11, 2022.

7 Oren Halvani and Philipp Marquardt. An unsophisticated neural bots and gender profiling system. In *CLEF (Working Notes)*, 2019.

8 Muhammad Rifki Kurniawan. Catastrophic forgetting in neural networks explained. `https://mrifkikurniawan.github.io/blog-posts/Catastrophic_Forgetting/`, 2021.

9 Federica Lago, Cecilia Pasquini, Rainer Böhme, Hélène Dumont, Valérie Goffaux, and Giulia Boato. More real than real: A study on human visual perception of synthetic faces [applications corner]. *IEEE Signal Processing Magazine*, 39(1):109–116, 2021.

10 Luca Maiano, Irene Amerini, Lorenzo Ricciardi Celsi, and Aris Anagnostopoulos. Identification of social-media platform of videos through the use of shared features. *Journal of Imaging*, 7(8), 2021.

11 Luca Maiano, Lorenzo Papa, Ketbjano Vocaj, and Irene Amerini. Depthfake: a depth-based strategy for detecting deepfake videos, 2022.

12 Owen Mayer and Matthew C Stamm. Forensic similarity for digital images. *IEEE Transactions on Information Forensics and Security*, 15:1331–1346, 2019.

13 Sophie Nightingale and Hany Farid. Synthetic faces are more trustworthy than real faces. *Journal of Vision*, 22(14):3068–3068, 2022.

14 Michael Roberts, Derek Driggs, Matthew Thorpe, Julian Gilbey, Michael Yeung, Stephan Ursprung, Angelica I Aviles-Rivero, Christian Etmann, Cathal McCague, Lucian Beer, et al. Common pitfalls and recommendations for using machine learning to detect and prognosticate for COVID-19 using chest radiographs and CT scans. *Nature Machine Intelligence*, 3(3):199–217, 2021.

15 Inna Vogel and Peter Jiang. Bot and gender identification in twitter using word and character n-grams. In *CLEF (Working Notes)*, 2019.

16 Kyra Wittorf, Martin Steinebach, and Huajian Liu. Automated image metadata verification. *Electronic Imaging*, 33:1–6, 2021.

## 4.6 Day 3 – Big Data Challenges

*Tiziano Bianchi (Polytechnic University of Turin, IT) – recorder of the session*

The participants were asked to report their experience with big data.

**Tiziano Bianchi:** Large scale PRNU search, dataset of about 25 million images. The problem is that for this kind of data search does not scale sub-linearly (e.g, log) with the size of dataset. Data is noise-like, standard indexing techniques (e.g LSH) are unstable.

**Isao Echizen:** Problems with bias on datasets. Construction of large dataset by starting with reference dataset and doing preprocessing and augmentation. Promote construction of large datasets involving more communities.

**Luca Cuccovillo:** Experience with speech matching (a sort of specialized Shazam). Problem with scalability, e.g., the need to replicate pairwise correlations for aggregation of similar speech. Problem mainly related to engineering, e.g., how to design short fingerprint with enough quality. One challenge is doing audio phylogeny, i.e, finding relation graphs of audio signals. Including synthetic audio. Need of collaboration, common understanding. Need of many different tools for dataset generation, different community should provide them to have scalability, single institutions cannot do this. Some datasets in challenges may have biases (e.g ASVspoof) [3]

**Martin Steinebach:** Working with real datasets has many issues not found in scientific research (transcoding, etc.). Research does not often consider efficiency on large scale.

**Luca Cuccovillo:** Agrees on additional engineering for managing speed required for big data.

**Irene Amerini:** Dealing with big data is a huge challenges due to many issues in order to have access to it. One of them is the time needed to collect dataset since big datasets are not always already available and, secondly, the data storage if you are a small institution. Furthermore all of the collected data need to be filtered and pre-processed in order to be used. Multimodal is even a bigger problem if should scale to big data.

**Paul Rosin:** A challenge with 3D datasets can be the large amount of data required to be stored. The feasibility of large scale digitization has been demonstrated for museum artifacts where companies have captured millions of images. As part of a project we were working on automatic segmentation, and needed to manually segment a large amount of data ($>$ 1000 images) for ground truth.

**Mauro Barni:** There is a lack of real big data due to problems in gathering them.

I have two experiences in this sense.

In a first case we were trying to develop a print and scan attack against a detector of synthetic images. The original detector does not work after print and scan, so we had to retrain it on printed and scanned images. To do that we had to build a dataset of printed and scanned images. The required effort was huge, and were able to get *only* 20,000 images, obtained with one single printer and one single scanner. Generalization to other devices was out of reach due to the lack of equipment. As a result, at test time the detector does not work well with different printers and scanners. [4]

As a second example, we collected 4 million outdoor images to classify geographic provenance (country recognition) by relying on the cultural features of urban architecture, social habits, etc ... The country was determined from the GPS position of the image. We got a huge improvement when the dataset grew from 0.5M to 4M images. Yet, gathering the images was not easy. We had to filter the images based on their content, to retain only urban scenes, remove persons etc ... We also had to ensure enough diversity gathering images from more source, including, street view, Flickr and Mapillary). Diversity and representativeness are big problems in large-scale image collection. For example, in our dataset there are many more images from the US & Europe than from some small countries (e.g. in Africa). We tried to solve this bias by balancing the dataset, i.e., building macro classes with the same size, but in this way our classifier was less discriminative. We also tried by weighting underrepresented countries, however the overall accuracy decreased. [1]

Similar problems are surely present in other application domains. For instance, how can we gather face images from small ethnic groups?

**Benedikt Lorch:** We gained practical experience with big data on an image retrieval task where the image database was growing every day. The concern was that the search would slow down with size of the database. We were able to address this concern with an approximate search method. Looking ahead, larger machine learning models create a demand for larger datasets. However, it becomes increasingly difficult to screen larger datasets and assert data quality properties, which is also required by the Artificial Intelligence Act. To this end, an interesting direction for future research would be quality metrics for datasets and automatic methods to assess data quality.

**Luca Cuccovillo:** An example of dataset quality assessment is to classify degraded training data, to get what most representative data are. [9, 13]

**Christian Riess:** experience in building dataset for image superresolution. One problem is the exponential number of combinations of parameters in dataset, to be done manually. [10]

License plate recognition project with police, mix of real and synthetic to ease annotation. Use of augmentation, and post-processing. Built a rack of different cameras to automate acquisition of renderings on screen. Real acquisition of cars can be done but is very time consuming (700 labeled so far). Difficult to add realistic features like weather effects, lighting, etc. [16]

**Thorsten Beck:** Mentions dataset of images from retracted articles from Elsevier. Annotation is manually done from article retraction notices. The dataset is not really large. Compiling such a dataset comes with significant legal challenges, e.g. when results are published (e.g., only for research use). The dataset does not cover all forms of manipulation, consequently representativity is another issue, larger collection of images exist only for few categories (duplicates). The dataset comprises of multiple kinds of manipulations (since it is build of real-world data). Automatically generated manipulation datasets (e.g., copy-move) are not very realistic. Getting enough publishers to the table is a demanding task, since they are not necessarily ready to invest resources and man power. Still, the problem of inacceptable image manipulation in scholarly works will hardly be resolved without a contribution from the side of the publishers. [17]

**Roberto Caldelli:** not much experience on big data for forensics. We have been gathering data for testing image provenance from social media (Facebook, Twitter) and we developed automatic tools for crawling and downloading. Problems are the interaction with social network API, which can be time consuming and complying with their policies for gathering data. We also experimented with a copy-move detection tool on print and scan images by testing with different devices. A comment is that limited availability of very big datasets for everybody sometimes makes research less democratic.

**Lakshmanan Nataraj:** Detection of GAN generated images. Collection of datasets from different GAN tools (6 types of GAN). Millions of images. Classification of GAN types. [5, 11]

**Alessandro Piva:** In PRNU estimation for video there is the need to process multiple frames, but this process is hindered by the presence of video stabilization, requiring the synchronization of ech frame. No efficient methods found in the state of the art when research was done. Managing crops, resize, rotation. Analyzing large datasets of videos requires huge computational effort. Experience in building dataset (VISION, multimedia forensic challenge), one of the problems is organizing the dataset before starting the collection of data. For recent datasets this is complicated by multiple acquisition settings and resolutions available. Usually only few settings/resolution are considered. For some published datasets there is not enough information on video settings used during acquisition, or inconsistent setting were adopted. Care must be taken on these aspects when building new datasets, such that in our opinion a single research group is not enough for the task.

**Marco Fontani:** Our products are for case works (mainly police), not many big data cases. We've been testing automatic analysis of images for insurance companies, there's a problem with the complexity of real cases (acquisition pipeline, etc.), and unclear definitions of authenticity in some scenarios. In video surveillance, a large amount of data is collected and must be stored for possible later use as evidence. Some storage and evidence management systems do not preserve the integrity of data (e.g., they systematically use transcoding of the original footage); this is a problem for forensics. Also, it is expensive to use commercial storage systems. Some police forces try to revert to local storage lately.

**Benedetta Tondi:** country recognition task, joint work with Mauro. Satellite images, and the detection of manipulated satellite images. Problem of datasets of satellite images, especially large scale datasets. Different sources are different domains. Tools trained on one sensor do not generalize to other sensor (e.g., Google Maps, other satellites). Need to include images from multiple sensors in the dataset.

**Anderson Rocha:** scientific retraction papers (DARPA 2017). 5000 papers with retraction notes. System receives pdf and extract images from pdf. Analyze images for forgery. Compare all images from papers of same authors. Analysis of images in suspicious scientific papers. Compare all papers from Scholar profile of authors, build a graph with similarities. The system produces a report to help human expert. No automatic decision should be allowed according to rules. Library to create copy move and forgery with different tools, to generate data for training. Completely annotated since synthetically produced. Freely available. (DARPA semaphore project). Only can detect about 20-25% of forgery right now. (Papers in biology and medicine). Detector should be improved. Right now only images are used, no content or text from papers. [12]

Detection of pornographic images/videos. 200 hours of pornography in dataset, problems with authorization from University for storing them. For illegal material (child pornography) training should be done on virtual machine by police. Multiple levels of training: Imagenet, fine tuning on generic pornography, fine tuning on police virtual machine for child pornography. Should have very low false positives. 40000 child pornography cases in police dataset. 40000 normal (including non pornography and regular pornography). 35% detection, less than 5% FP. No decision, only filtering. Usually run on suspect's harddrive, the tool gives the most likely files, manual inspection is required. You should reduce the number of hours used by manual expert inspection, so low FP is required in this application. One video is enough for prosecution, so even if few videos are detected over the total is perfectly fine. Right now we are collecting everything form social media (whatsapp, telegram, tiktok, facebook, twitter) on attacks in Brasil. Billions of data, most is garbage, should be filtered.

**Mauro Barni:** What are you looking for?

**Anderson Rocha:** To localize faces and identify spreaders, i.e.,most frequent faces seen in videos. Collecting related text.

**Roberto Caldelli:** what kind of real images did you use during training for pornography detection? Common images such as objects, landscapes, cars and so on, or did you select specific cases of presence of normal nudity?

**Anderson Rocha:** We selected difficult cases, for example we use images selected by skin detector (beach, swimming pools, etc.).

Then, the discussion turned on discussing challenges and opportunities offered by big data. The following challenges are identified:

- Copyright
- How to manage storage requirements
- How to distinguish what is useful in collected data
- How to generate synthetic data
- How to guarantee diversity and representativity.
- Computational power to collect all required data.
- Versioning.
- For university is difficult to have storage and computation capabilities.
- Problems of privacy when collecting some kind of data (e.g. faces).

Then discussion follows:

**Mauro Barni:** You get outstanding performance if and only if you have enough data. You cannot use AI without enough data. Someone claims that with big data and enough computational power you can explain everything? I do not quite agree with this view, understanding is more than just finding patterns in data.

**Anderson Rocha:** Most of the correlations are spurious, but how to separate useful from spurious? There are three levels for acquiring knowledge:

1. Find correlations, machines are very good at this
2. Find possibilities, like cause – effect relations, AI is usually bad at this
3. Analyse past decisions, project alternative future based on different choices, machines cannot do this.

**Marco Fontani:** Are machines accountable? Who is accountable? Producers will say this is just a help for human decision, the expert operating the system should be accountable.

**Benedetta Tondi:** A theoretical challenge is represented by the security of networks in an adversarial environment. Data should be representative of possible attacks. This turns out to be a very big challenge for forensic tools.

**Mauro Barni:** With big data it is easier to hide poisoned samples, and more difficult to spot them.

**Anderson Rocha:** Attacks exploiting triggers. How can we inspect a network to see whether we have a backdoor. This can be a forensic problem.

**Mauro Barni:** A possible solution is to inspect the datasets used during training, not only the trained network. Attacks can be carried out at different levels. Backdoor can be used also to watermark a network. We have a good experience in checking datasets. In [7] we used cluster analysis in the latent space to detect poisoned samples.

**Martin Steinebach:** This is important for autonomous driving. Training robust classifier. More machine learning security. But also forensic if you analyze the dataset for anomalies.

**Luca Cuccovillo:** Opportunities of federated learning in big data (privacy, complexity, but also vulnerabilities to attacks).

## 4.7 Day 3 – Benchmark and Performance Evaluation

*Tiziano Bianchi (Polytechnic University of Turin, IT) – recorder of the session*

**Anderson Rocha:** There is a need of a validation protocol, for comparisons among different tools. Problems to be solved are how to access to data, how to choose test and training data, how to choose proper metrics.

**Martin Steinebach:** Huge datasets sometimes are prone to overfitting, if not diverse enough. A black-box evaluation protocol could be more fair. Give a blind test set to prevent overfitting over it.

**Anderson Rocha:** : We need to be responsible for this black box.

**Martin Steinebach:** A public body could be the standardization body. Blind virtual machine for evaluation of security of tools, including AI tools.

**Paul Rosin:** What about the feedback, will this be useful for improving tools?

**Marco Fontani:** The main issue for the practitioner is explainability. Heat maps are often not enough.

**Paul Rosin:** When benchmarking for image standardization, often a benchmark dataset is used both for training and testing, with a random split. It is better to use a separate benchmark dataset for only testing purposes. Collecting different data for training can be left up to the developer. I advocate a structured benchmark where different levels of difficulty are provided in the testing set (depends on application) A small testing dataset can be more curated. [15]

**Anderson Rocha:** : We have good dataset for deepfake detection, however the performance when performing intra-dataset evaluation is saturated. This is observed also for spoofing detection and copy-move detection. Cross-dataset evaluation is needed as next step. Difficult to have different levels for testing in forensics.

**Marco Fontani:** confirm the experience of cross-dataset evaluation, performance drops in this case.

**Paul Rosin:** One issue is the diversity of types of images in forensic datasets.

**Anderson Rocha:** most of datasets include natural images. Experience with separation of specific images (biomedical) from natural during dataset preparation.

**Mauro Barni:** Most work is done fine tuning network trained on natural images. In some fields, e.g. GAN generated images, few architectures are available. Better cross-validation by training on images produced by one architecture and testing on another one is needed.

**Anderson Rocha:** : there is a shift from the real-fake detection problem to fingerprinting of GAN generation algorithm. Maybe in the future we will shift to fingerprinting, which is a more challenging problem and requires training on all available tools.

**Paul Rosin:** I recently came across ForgeryNet dataset for benchmarking [8], which contains a lot of data: 3 millions images, 200000 videos. This dataset should be considered a useful resource.

## References

1 Omran Alamayreh, Giovanna Maria Dimitri, Jun Wang, Benedetta Tondi, and Mauro Barni. Which country is this picture from? New data and methods for DNN-based country recognition. *arXiv preprint arXiv:2209.02429*, 2022.

2 João P Cardenuto and Anderson Rocha. Benchmarking scientific image forgery detectors. *Science and Engineering Ethics*, 28(4):35, 2022.

3 Luca Cuccovillo, Christoforos Papastergiopoulos, Anastasios Vafeiadis, Artem Yaroshchuk, Patrick Aichroth, Konstantinos Votis, and Dimitrios Tzovaras. Open challenges in synthetic speech detection. In *IEEE International Workshop on Information Forensics and Security (WIFS)*, pages 1–6. IEEE, 2022.

4 Anselmo Ferreira, Ehsan Nowroozi, and Mauro Barni. VIPPrint: A large scale dataset of printed and scanned images for synthetic face images detection and source linking. *arXiv preprint arXiv:2102.06792*, 2021.

5 Michael Goebel, Lakshmanan Nataraj, Tejaswi Nanjundaswamy, Tajuddin Manhar Mohammed, Shivkumar Chandrasekaran, and BS Manjunath. Detection, attribution and localization of GAN generated images. *arXiv preprint arXiv:2007.10466*, 2020.

6 Chandrakanth Gudavalli, Erik Rosten, Lakshmanan Nataraj, Shivkumar Chandrasekaran, and BS Manjunath. SeeTheSeams: Localized detection of seam carving based image forgery in satellite imagery. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1–11, 2022.

7 Wei Guo, Benedetta Tondi, and Mauro Barni. A master key backdoor for universal impersonation attack against DNN-based face verification. *Pattern Recognition Letters*, 144:61–67, 2021.

8 Yinan He, Bei Gan, Siyu Chen, Yichun Zhou, Guojun Yin, Luchuan Song, Lu Sheng, Jing Shao, and Ziwei Liu. ForgeryNet: A versatile benchmark for comprehensive forgery analysis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4360–4369, 2021.

9 Zohaib Amjad Khan, Giuseppe Valenzise, Aladine Chetouani, and Frédéric Dufaux. Towards an image utility assessment framework for machine perception. In *30th European Signal Processing Conference (EUSIPCO)*, pages 568–572. IEEE, 2022.

10 Thomas Köhler, Michel Bätz, Farzad Naderi, André Kaup, Andreas Maier, and Christian Riess. Toward bridging the simulated-to-real gap: Benchmarking super-resolution on real data. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 42(11):2944–2959, 2019.

11 Tajuddin Manhar Mohammed, Jason Bunk, Lakshmanan Nataraj, Jawadul H Bappy, Arjuna Flenner, BS Manjunath, Shivkumar Chandrasekaran, Amit K Roy-Chowdhury, and Lawrence Peterson. Boosting image forgery detection using resampling features and copy-move analysis. *arXiv preprint arXiv:1802.03154*, 2018.

12 Daniel Moreira, João Phillipe Cardenuto, Ruiting Shao, Sriram Baireddy, Davide Cozzolino, Diego Gragnaniello, Wael Abd-Almageed, Paolo Bestagini, Stefano Tubaro, Anderson Rocha, et al. SILA: A system for scientific image analysis. *Scientific Reports*, 12(1):18306, 2022.

13 Curtis G Northcutt, Anish Athalye, and Jonas Mueller. Pervasive label errors in test sets destabilize machine learning benchmarks. *arXiv preprint arXiv:2103.14749*, 2021.

14 Judea Pearl and Dana Mackenzie. *The Book of Why: The New Science of Cause and Effect*. Hachette UK, 2018.

15 Paul L Rosin, Yu-Kun Lai, David Mould, Ran Yi, Itamar Berger, Lars Doyle, Seungyong Lee, Chuan Li, Yong-Jin Liu, Amir Semmo, et al. NPRportrait 1.0: A three-level benchmark for non-photorealistic rendering of portraits. *Computational Visual Media*, 8(3):445–465, 2022.

16 Andreas Spruck, Maximilane Gruber, Anatol Maier, Denise Moussa, Jürgen Seiler, Christian Riess, and André Kaup. Synthesizing annotated image and video data using a rendering-based pipeline for improved license plate recognition. *arXiv preprint arXiv:2209.14448*, 2022.

17 Humboldt-Universität zu Berlin. Image integrity database. `https://rs.cms.hu-berlin.de/iidb/pages/home.php`, 2023.

## 4.8 Day 4 – Morning Discussion

*Benedikt Lorch (Universität Innsbruck, AT) – recorder of the session*

Luca Cuccovillo presents open challenges in synthetic speech detection based on his talk from IEEE WIFS 2022 [1]. The goal of the WIFS paper was to review limitations of current datasets and discuss requirements for good synthetic speech datasets.

Neural speech synthesis: Ground-breaking applications vs. unprecedented forms of misuse

Synthetic speech detection:

- Potential: Plenty of room for research and development
- Danger: Lack of common planning/directions
  - Unclear technical requirements for datasets
  - No interpretability of model outputs
  - Lack of robustness/generalization
  - Lack of exchange between research and potential end users

Datasets: Large number of datasets available, but all of them have problems: Undisclosed synthesis algorithms, synthesized voices do not have real counterparts (speaker recognition would solve the task), single female speaker, not redistributable in original/derivative form, single text-to-speech pipeline

Detection algorithms: Many excellent proposals with hand-crafted features and deep neural networks. But they also have some issues: Unseen synthesis methods are problematic, unseen speaker/recording conditions, methods based on flawed dataset, lack of interpretability and explainability, unclear functional/non-functional experiments

How to do data collection right?

- Curating the data: Balance the speakers, gender, age, languages, accents
- Has to have transcriptions, enough data for training/fine-tuning
- Adhere to legal constraints.

Requirements for the creation of synthetic data:

- High linguistic and voice variability
- Diverse vocoding qualities
- Diverse feature extraction qualities
- Maximum expressiveness

Efforts and costs should be shared:

- Data collection and storage requirements
- What about federated learning (FL), leaving data on-premise?
- Is federated learning feasible for non-IID audio data?

Explainability is more than nice-to-have:

- Right of explanation prompted by the EU
- Current AI Act proposal considers forensic algorithms “high-risk”
- Journalists and forensic analysts have strong demand for explainability

- Question: Should we rely on XAI methods from image domain, or go further?
- Question: Are saliency maps on spectrograms understandable to end users, or only to researchers? Useful as debugging tools but not really explainable

Discussion: How many of these challenges are related to synthetic image detection?

- Image datasets can also contain biases
- Difficult not to inject any side channels in speech
- Possession asymmetry: A few companies possess the most amount of speech data, which gives them an advantage. In speech, this asymmetry arose earlier than in vision.
- The general problems are the same across application domains: dataset diversity, dataset size, explainability. The way these problems manifest themselves are different, calling for different mitigation strategies. Visualization maps can be more difficult to interpret for audio. In other words, audio and image forensics share the same general problems, but solutions can be very different.

**Mauro Barni:** Research in AI (and AI-based forensics) proceeds in a chaotic way. Everyone is somehow steered by their own goals. But we can do small things to advance our field: serving on the editorial board of a journal allows you to some extent direct the community. Similarly, competition steers the communities for the next years.

**Martin Steinebach:** There are many parallel, duplicate efforts, just using other taxonomies and not knowing about each other.

**Mauro Barni:** The newly proposed AI-based watermarking methods are rarely compared to the traditional watermarking techniques. Yet, classical watermarking provides satisfactory, sometimes excellent, solutions to many problem, so a comparison would be really needed.

## 4.9 Day 4 – Current and Future Applications

*Benedikt Lorch (Universität Innsbruck, AT) – recorder of the session*

There were eight short talks on current and future applications.

### 4.9.1 Marco Fontani about Amped Software

About Amped software:

- Mission: Provide customers with reliable algorithms based on scientific papers
- More than 100 users around the world
- Quest to provide good support, provide a complete product, forensically sound, widely adopted and accepted worldwide, deeply involved in the scientific community

Amped ecosystem:

- Amped Five, the top tool with all filters (Swiss knife)
- Amped Replay (simplified Swiss knife): an advanced player with streamlined processing and basic enhancement
- DvrConv: CCTV systems use proprietary video formats, and this software allows batch conversion of such formats in a forensically sound manner
- Amped Authenticate: Authenticate images; since recently Authenticate includes a DL detector for deepfake detection, but with all the necessary warning messages.

Survey on video forensics state of the art based on user survey [2]

- Main issues: Low image quality, proprietary CCTV/DVR video files, amount of cases/data, interpretation of video evidence; budget is not an issue
- Increase of video casework in recent years
- Increase of crime, change on image and video quality, pandemic made an increase of casework
- Evidence used to solve a crime: CCTV, mobile device data, images and videos from other sources
- Training: Vendor training, self-learning, job training,...

Should AI be used in forensics:

- Only 2% say “avoid AI”
- Majority said the use of AI should be limited to cases when proven reliable
- A good percentage said “to be used for investigative leads only”

Question: What tool would Amped like to develop?

- Users want Amped FIVE to be faster
- Functionality: Image enhancement, e.g., improve denoising
- Authenticate: Need for video authentication tools; users request tools for deepfake detection, although they are to date not very relevant in practice yet

Amped also contributed to the ENFSI best practice guidelines for audio authentication [9] and image authentication [8].

### 4.9.2 Isao Echizen: Fake media detection and its practical application

- Examples of where fake faces have been seen recently
- Five types of face synthesis methods
- Detection approaches: MesoNet, Capsule network, joint facial video detection and segmentation
- AIaaS for automatic detection of fake facial videos: AI-based web service accessible via web API
- License status of SYNTHETIQ VISION: Will be used by several companies, including CyberAgent Inc (advertising company in Japan).

Common issues for fake media detection:

- Performance degrades when images are redistributed via social networks (item Detection of unseen types of fakes. Periodic updating of training data and model training are necessary
- Users do not necessarily need a generic detector. Accurate detection of a specific kind of fake media is acceptable, e.g. digital twin: faceswap / eKYC: facial reenactment

Question about compliance with AI regulations: Are there similar restrictions and laws about privacy in Japan as in Europe? Data comes from companies.

Question: What is the most important face synthesis technique to detect for companies? Facial reenactment used in KYC.

### 4.9.3 Martin Steinebach: KIKU: Utilizing AI for the protection of cultural property

KIKU = Künstliche Intelligenz für den Kulturgutschutz (Artificial intelligence for the protection of cultural assets)

The project KIKu[2] (for a video demo please see [11]) is a follow-up project of the BMBF project Illicid[3]. In Illicid, various technical methods for the protection of cultural assets were developed, including a machine-learning based app that classifies robbery excavations and thus helps, for example, customs officers to detect illegal imports. This part was considered so relevant by users that KIKu was designed, with several stages to further develop the application. The core continues to be the detection of robbery excavations through machine learning [6] [4]. Deep learning will be used both to classify and to recognize similar objects. However, the project will also address issues such as the detection of forgeries of cultural goods.

The project is relevant to security because illegal excavations and lootings serve, among other things, to finance terrorist groups. The excavation sites are occupied, cultural goods are looted and then smuggled to third countries, where they are sold. The proceeds then flow back to the terrorists. To identify artifacts that come from looted excavations, the knowledge of experts is necessary. But these are not available where the objects are brought across the border or offered for sale. It is not possible for the customs officers to verify the information on the objects, for example regarding origin or age. This is where the KIKu tool comes in: Items assessed by experts become training data with images and metadata. The trained

[2] https://www.sit.fraunhofer.de/de/kiku/

[3] https://www.sifo.de/sifo/de/projekte/schutz-vor-kriminalitaet-und-terrorismus/schutz-vor-organisierter-kriminalitaet/illicid/illicid-verfahren-zur-erhellun-beispiel-antiker-kulturgueter.html

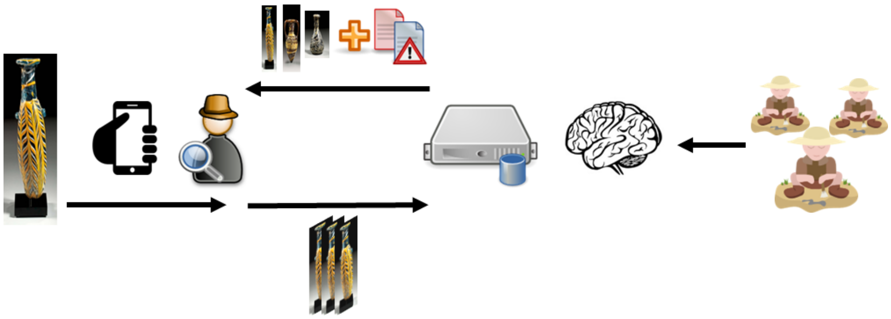

**Figure 1** KIKu workflow.

network can then be accessed with an app from a smartphone with a photo of an item to be examined. The customs officer can now compare the information on the object with KIKu's assessment and take further steps in the event of discrepancies (see figure 1).

As common in machine learing, training data is an important issue. In the first project Illicid, 2-3 archaeologists provided 3,000 hand-labelled datasets, which did show promising results. The strategy here was not to aim for a generic recognition of cultural goods, but learning only items of a narrow area and epoch. In KIKu, crawling of museum data was utilized, with currently 140,000 training sets available, increasing the potential of generalization.

For cultural good recognition there are already applications in Poland for recognizing stolen paintings. However, the KIKU project also includes similar paintings and other objects. The core goal is classification and not re-identification.

Discussion about collecting more images from museums, whether users can prefer texture or shape features, whether the network prefers any particular features

Discussion about maturity of technology: Retrieval tasks seem to work quite well, checking constraints for pixel-level differences is error prone to do at scale in practice

### 4.9.4 Lakshmanan Nataraj: Current and future applications in media forensics

- Seam carving and seam insertion
- Seam carving detection with a CNN
- Object removal examples with and without heatmaps
- Satellite image object removal with heatmaps [12]
- Seam carving for object displacements
- Potential future applications:
  - Satellite image forensics
  - Different domains: image, video, audio, metaverse, NERFs, diffusion, etc.

Discussion how to do object displacement using seam carving.

## 4.10 Day 4 – Challenges Ahead

*Benedikt Lorch (Universität Innsbruck, AT) – recorder of the session*

### 4.10.1 Jane Wang: Convergence of signal processing to machine learning

- Signal processing (SP) and image processing (IP) plays a key role in the *preprocessing and transformation and feature extraction*, before the DL design. SP-based processing is critical in digital media security and forensics research.
- Relationship between SP/IP operations and DL components
- The SP/ML boundary is getting blurred

Future: signal/model-driven DL

- Challenges: data-driven (lack of generalizibility in out-of-sample scenarios); limited/noisy training samples; interpretablity/explainability; DL security/trustfulness; robustness to noise/attacks; uncertainty in deep learning
  - Potential direction: combine domain knowledge and the DL's learning capabilities to mitigate deficiencies of traditional SP/IP and black-box DNN approaches
  - Bring DSL(?) in statistical SP into DL, e.g. statistical DL, Bayesian DL
  - Bring DL into SP, e.g. deep unfolding
- Combining physics-based modeling and DL
- Perspective: Seeing will no longer be believing

Adversarial ML:

- scrutinize potential security vulnerabilities of DL models by (virtually) attack them
- requires proper threat model

Analogies to forensics, anti-forensics, and counter anti-forensics: Both digital images and DL models are vulnerable to manipulations and attacks, intentionally or unintentionally, posing critical challenges in trusting digital images

Potential directions:

- combine both SP and IP with DL
- leverage domain knowledge in signal/image processing
  - investigate interpretation for DL-based digital image forensics problems
  - focusing on the vulnerability of digital images themselves
  - focusing on vulnerability of current DL models

Both attack and defense side will improve. It is harder to fool the traditional image processing features

**Paul Rosin:** There is no guarantee that combining learning- and model-based techniques can gain the benefits of the two. In fact, how can we be sure that the combination does not inherit the weaknesses of the two?!

Discussion about interpretability: Use domain knowledge where possible.

### 4.10.2 Sebastien Marcel: Biometrics Security and Privacy

Biometrics security and privacy (BSP) research group: signal processing and ML applied to BSP, e.g. biometric recognition, security, privacy, multi-modal fusion AI and responsible datasets: fairness, Trojan/backdoors, ethics and synthetic datasets

- BATL: Create face anti-spoofing technology with a multi-spectral sensor. Created a multi-spectral PAD dataset
- FairFace: Metric to measure fairness in biometric systems (fairness discrepancy rate), now working on fairness mitigation strategy [3, 7]
- SAFER: Generate synthetic datasets for training and testing
- Media forensics challenges ahead:
  - Hyper-realistic and real-time audio-visual fakes
    - Detection: generalization to unseen attacks
    - Attribution: identification of the source of the attacks
  - Fairness and transparency compliance (e.g. EU Artificial Intelligence Act)
    - bias assessment and mitigation (biometrics and forensics)
    - synthetic datasets (e.g. face datasets) for training/testing classifiers to circumvent data protection issues
    - certification labs: push for AI certification scheme

What is bias? When you consider the error rate for the general population, you have a low error rate. As soon as the population is broken down into groups, the performance of the subdistribution is diverse. Same errors for everybody.

Bias mitigation strategies:

1. post-processing of the scores
2. if you have access to the model: regularization in order to balance the errors
3. fix the dataset

### 4.10.3 Anderson Rocha: Key challenges ahead

1. Synthetic realities: People with their own view of the world, fabrication of views and images, fake news, deep fakes: How to deal with synthetic generators for faces/images/videos?
2. How to generalize, dealing with the openness and unseen scenarios, e.g., in spoofing or deepfake detection? Try to devise methods that adaptively train themselves, i.e. self-supervised learning.
3. Fusion: Combining different sensors and modalities for solving a particular problem.
4. Solutions to compliance problems with privacy laws: Federated learning, self-supervised learning with access to some information only

### 4.10.4 Irene Amerini: Multimedia forensics: Challenges ahead

Research objectives:

- Design multimedia forensics techniques able to detect manipulated contents
- Scale forensic investigations to real-world applications: deepfake detection, social network provenance

Future trends:

- Forgery detection and source identification on internet-style data, not only on lab datasets. Semantic forensics on multimedia/multimodal assets

- Defense solutions against disinformation attacks, e.g. images generated with text-to-image techniques
- Adversarial deep learning: understanding the robustness and security of developed techniques. Build platforms and procedures to test robustness of models.

Future trends in deep fake detection:

- Continual deepfake detection (continual learning)
- Multimodal approach for deepfake detection (or generative models)
- Generalization issues
- Real-time deepfake detection
- Certifying authorship (even of deepfakes) via blockchain. Back to watermarking?

Future trends:

- Datasets are huge but not huge enough. Potential solutions are self-supervised learning, generating synthetic training images, augmenting datasets with generative models
- Problems: Biased datasets
- Computational cost for training and hardware resources. Potential solution: Creating lightweight models that require less hardware resources but without sacrificing much of performance

Common themes in all 8 talks: Self-supervised learning, combining different ways (e.g. model-based and learning-based techniques, different modalities), and synthetic generators pose a pressing problem.

### References

1 Luca Cuccovillo, Christoforos Papastergiopoulos, Anastasios Vafeiadis, Artem Yaroshchuk, Patrick Aichroth, Konstantinos Votis, and Dimitrios Tzovaras. Open challenges in synthetic speech detection. In *IEEE International Workshop on Information Forensics and Security (WIFS)*, pages 1–6. IEEE, 2022.

2 Amped Software. Survey results: The state of video forensics 2022. `https://blog.ampedsoftware.com/2022/12/20/survey-results-the-state-of-video-forensics-2022/`, 2022.

3 Kimmo Karkkainen and Jungseock Joo. FairFace: Face attribute dataset for balanced race, gender, and age for bias measurement and mitigation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1548–1558, 2021.

4 Waldemar Berchtold, Huajian Liu, Simon Bugert, York Yannikos, Jingcun Wang, Julian Heeger, Martin Steinebach, and Marco Frühwein. Recognition of objects from looted excavations by smartphone app and deep learning. *Electronic Imaging*, 34:1–4, 2022.

5 Gabriel Bertocco, Antônio Theófilo, Fernanda Andaló, and Anderson Rocha. Reasoning for complex data through ensemble-based self-supervised learning. *arXiv preprint arXiv:2202.03126*, 2202.

6 Simon Bugert, Huajian Liu, Waldemar Berchtold, and Martin Steinebach. Cultural assets identification using transfer learning. *Electronic Imaging*, 34:1–4, 2022.

7 Tiago de Freitas Pereira and Sébastien Marcel. Fairness in biometrics: A figure of merit to assess biometric verification systems. *IEEE Transactions on Biometrics, Behavior, and Identity Science*, 4(1):19–29, 2022.

8 European Network of Forensics Science Institutes – Digital Imaging Working Group. Best practice manual for digital image authentication. `https://enfsi.eu/wp-content/uploads/2021/10/BPM_Image-Authentication_ENFSI-BPM-DI-003-1.pdf`, 2021. Issue No. 001.

9 European Network of Forensics Science Institutes – Forensic Speech and Audio Analysis Working Group. Best practice manual for digital audio authenticity analysis. https://enfsi.eu/wp-content/uploads/2022/12/FSA-BPM-002_BPM-for-Digital-Audio-Authenticity-Analysis.pdf, 2022. Issue No. 001.

10 Moreira, Daniel and Avila, Sandra and Perez, Mauricio and Moraes, Daniel and Testoni, Vanessa and Valle, Eduardo and Goldenstein, Siome and Rocha, Anderson. Multimodal data fusion for sensitive scene localization. Elsevier Information Fusion, v45, pp 307–323, 2019

11 Fraunhofer SIT. The KiKu-App: Using artificial intelligence to automatically recognize cultural assets. https://youtu.be/un4ED05Ag_I.

12 Chandrakanth Gudavalli, Erik Rosten, Lakshmanan Nataraj, Shivkumar Chandrasekaran, and BS Manjunath. SeeTheSeams: Localized detection of seam carving based image forgery in satellite imagery. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1–11, 2022.

## 4.11 Day 5 – Forensics Questions and the Future of the Field

*Thorsten Beck (HU Berlin, DE) – recorder of the session*

### 4.11.1 Discussion questions for the report

- How can we advance the field?
- How has the field changed in the past 5 years?
- What do you predict will happen in the next to 5-10 years?
- What is the biggest challenge in the field at the moment?
- What are the most critical changes that we must make to face the future effectively?
- What effect has deep learning made on the field?
- Who is making the greatest advancements in the field, and what are they doing?

### 4.11.2 The Future

10 year perspective:

- improved quality of synthetic media
- pervasiveness of synthetic media
- link between physical world and digital content will be broken – then crypto?
  and forensics may not help to reinforce trustworthiness/authenticity
- few generators will emerge possibly watermarked
- integration of AI and GoF (AI comes first)
- self-supervised DL

### 4.11.3 Research Challenges in the Field of Media Forensics

Core Challenges

- Generalization (if I know how to identify one deep fake, how do I know to detect different ones?)

- Distribution mismatches/distributional shift (how can we handle out of distribution samples?)

  (above items boil down to lack of realistic models in GoF MMF)
- Modeling (various kinds of) uncertainty/dealing with uncertainty

see also: limitations due to amount of training data

Data-related problems (different twist for GoF)

- representativity (number of variables considered)
- privacy / copyright and legal restrictions (* see security)
- bias (exists as variable, but does not necessarily consider real-world distribution of age/gender etc.)
- for A.I. forensic approaches, big data is required (sometimes one might be confronted with one-shot problems, that require GoF approaches, see §What speaks for A.I.?)

**Marco Fontani:** from the point of view of COURTS and JUDGES, it is generally not plausible to make decision about an individual by data derived from other sources.

Explainability (resp. Interpretability?) – for "AI eyes" only?

- Check for correct behavior
- for forensic use (how do machines "see the invisible")

**Marco Fontani:** Research papers ought to distinguish between explainability and interpretability.

**Martin Steinebach:** Interpretation of forensic results in court and trials must generally represent not only the perspective of the prosecution, but also the perspective of the defense (neutrality).

**Luca Cuccovillo:** It should be considered that explainability in the literature is discussed as subset of trustworthiness.

Security

- enlarged attack surface wrt GoF (also because of training) – see also adversarial examples
- develop suitable threat models
- cat and mouse loop

What speaks for the application of A.I.?

- lack of good models for GoF fitting the complexity of real life
- coping with dynamic changes (e.g. software updates for cameras)
- benefits from pre-training/immediate benefit from available standard computer vision models (?)
- less domain knowledge needed (?)

**Christian Riess:** greybox/blackbox examples cannot always be sufficiently addressed via GoF

**Martin Steinebach:** problem with A.I. – in real world cases: one needs maybe 10 photos to identify a camera, but with A.I. you need thousands of images to train models. Real-life scenarios may require GoF approaches. It may be hard to explain criminal investigators or the police that large amounts of data is required to make A.I. work.

## Participants

- Irene Amerini
Sapienza University of Rome, IT
- Mauro Barni
University of Siena, IT
- Thorsten Beck
Humboldt Universität zu Berlin, DE
- Tiziano Bianchi
Polytechnic University of Turin, IT
- Luca Cuccovillo
Fraunhofer IDMT – Ilmenau, DE
- Isao Echizen
National Institute of Informatics – Tokyo, JP
- Benedikt Lorch
Universität Innsbruck, AT
- Christian Riess
Friedrich-Alexander-Universität Erlangen-Nürnberg, DE
- Paul Rosin
Cardiff University, UK
- Martin Steinebach
Fraunhofer SIT – Darmstadt, DE

## Remote Participants

- Roberto Caldelli
CNIT – Florence and Mercatorum University, IT
- Marco Fontani
Amped Software – Trieste, IT
- Haiying Guan
NIST – Gaithersburg, US
- Zulfiqar Habib
Comsats University – Lahore, PK
- Lakshmanan Nataraj
Trimble Inc. – Chennai, IN
- Ngai Fong Law
The Hong Kong Polytechnic University, HK
- Sebastian Marcel
Idiap Research Institute – Martigny, CH
- Alessandro Piva
University of Florence, IT
- Anderson Rocha
State University – Campinas, BR
- Xianfang Sun
Cardiff University, UK
- Benedetta Tondi
University of Siena, IT
- Z. Jane Wang
University of British Columbia – Vancouver, CA