PhD in Information Technology and Electrical Engineering

Università degli Studi di Napoli Federico II

PhD Student: Fabrizio Guillaro

Cycle: XXXVII

Training and Research Activities Report

Academic year: 2022-23 - PhD Year: Second

Fabrizio Guillaro

Tutor: prof. Luisa Verdoliva Luisa Verdoliva

Co-Tutor: Giovanni Poggi

Date: October 23, 2023

1. Information:

- **PhD student: Fabrizio Guillaro** **PhD Cycle: XXXVII**
- **DR number: DR995863**
- **Date of birth: 06/01/1996**
- **Master Science degree: Computer Engineering**
- **University: Università degli Studi di Napoli Federico II**
- **Scholarship type: UNINA - DII, DISCOVER project, funded by DARPA under the SEMAFOR program**
- **Tutor: Prof. Luisa Verdoliva**
- **Co-tutor: Prof. Giovanni Poggi**

2. Study and training activities:

Activity	Type[1]	Hours	Credits	Dates	Organizer	Certificate[2]
From Handcrafted to End-to-End Learning, and Back: a Journey for Multi-Object Tracking	Seminar	2	0.4	02/12/22	International AI Doctoral Academy, University of Modena and Reggio Emilia	Y
Digital Forensics	Seminar	2	0.4	06/12/22	DIETI – UNINA	Y
Face Presentation Attack Detection	Seminar	1.5	0.3	07/12/22	IEEE Biometrics Council	Y
Advances on Multimodal Machine Learning Solutions for Speech Processing Tasks and Emotion Recognition	Seminar	1	0.2	19/01/23	IEEE Signal Processing Society	Y
The Super Neuron Model - A new generation of ANN-based Machine Learning and Applications	Seminar	1	0.2	09/02/23	EURASIP	Y
How to boost your phd	Course	16	4	11/01/23 – 01/03/23	DIETI ITEE-ICTH - CQB PhD programs	Y

What’s Up with Image and Video Forensics?	Seminar	1	0.2	02/03/23	EURASIP	Y
Unleashing the Power of LLMs: a Historical perspective on Generative AI	Seminar	1	0.2	02/03/23	DIETI - UNINA	Y
Statistical Multimedia Security and Forensics	Course	20	4	08/05/23 – 12/05/23	University of Trento – IECS Doctoral School	Y
Computational Disinformation Symposium	Seminar	7	1.4	06/06/23	NYU Tandon School of Engineering	Y
Prompting in Vision	Seminar	3	0.6	19/06/23	IEEE/CVF (CVPR 2023)	Y
International Computer Vision Summer School (ICVSS)	Doctoral School	30	6	09/07/23 – 15/07/23	University of Catania	Y
Scientific Integrity Verification Through Image Forensics	Seminar	1	0.2	06/07/23	IEEE Signal Processing Society - Information Forensics and Security Technical Committee	Y
Elaborazione di Segnali Multimediali	Tutorship	7	0.28	01/05/23 – 30/06/23	Luisa Verdoliva	
Preparation of the conference paper “TruFor: Leveraging all-round clues for trustworthy image forgery detection and localization” for CVPR 2023 Attendance to weekly technical meetings. Participation to International Workshop on Information Forensics	Research		8.9	01/11/22 – 31/12/22		

(WIFS) (13-16 /12/2022)						
Study of state of the art methods for image forgery detection and localization. Study of state of the art methods for synthetic image generation. Implementation of an end-to-end approach for image forgery detection and localization that relies on the confidence map. Benchmarking several SOTA methods on different splicing and copymove datasets. Generation of a synthetic splicing dataset using diffusion models. Attendance to weekly technical meetings. Preparation of the camera ready version of the conference paper “TruFor: Leveraging all-round clues for trustworthy image forgery detection and localization” for CVPR 2023			8.6	01/01/23 – 28/02/23		
Study of state of the art methods for image forgery detection and localization. Study of state of the art methods for Language Models. Generation of synthetic splicing datasets using image generation methods.			5.6	01/03/23 – 30/04/23		

Attendance to weekly technical meetings. Experiments regarding an improved version of the method described in our conference paper “TruFor: Leveraging all-round clues for trustworthy image forgery detection and localization” Experiments using text-guided methods for image forgery detection.						
Participation to Computer Vision and Pattern Recognition Conference (CVPR), Vancouver, Canada (18-22 /06/23) and presentation of our paper “TruFor: Leveraging all-round clues for trustworthy image forgery detection and localization” Generation of synthetic splicing datasets using image generation methods. Attendance to weekly technical meetings. Experiments regarding an improved version of our image forgery detection method “TruFor”. Implementation of methods for image forgery detection using text descriptions of forgeries.			4	01/05/23 – 30/06/23		
Attendance to weekly technical meetings. Experiments regarding an improved version of our image forgery detection method “TruFor”. Experiments regarding	Research		4	01/07/23 – 31/08/23		

image forgery detection using text descriptions.						
Experiments regarding the influence of the training dataset on the overall performance of our method. Creation of locally manipulated images using recent synthetic image generation methods. Experiments regarding a more efficient version of our image forgery detection method "TruFor". Attendance to weekly technical meetings. Presentation of our paper "TruFor: Leveraging all-round clues for trustworthy image forgery detection and localization" at the Research Working Group organized under the SemaFor project (DARPA).	Research		10	01/09/23 – 31/10/23		

1) Courses, Seminar, Doctoral School, Research, Tutorship
2) Choose: Y or N

2.1. Study and training activities - credits earned

	Courses	Seminars	Research	Tutorship	Total
Bimonth 1	**-**	**1.1**	**8.9**	**-**	**10**
Bimonth 2	**-**	**0.4**	**8.6**	**-**	**9**
Bimonth 3	**4**	**0.4**	**5.6**	**-**	**10**
Bimonth 4	**4**	**2.0**	**4**	**0.28**	**10.3**
Bimonth 5	**6**	**0.2**	**4**	**-**	**10.2**
Bimonth 6	**-**	**-**	**10**	**-**	**10**
Total	**14**	**4.1**	**41.1**	**0.28**	**59.5**
Expected	**30 - 70**	**10 - 30**	**80 - 140**	**0 – 4.8**	

3. Research activity:

Manipulating images has never been easier, with new powerful editing tools appearing by the day. These new possibilities stimulate the creativity of benign and malicious users alike. Previously, crafting a multimedia disinformation campaign required advanced skills, and attackers were limited to copy, replicate or remove objects in an image, classic forms of image manipulations also known as "cheapfakes". With the explosive growth of deep learning, image manipulation tools have become incredibly easier to use and much more powerful, allowing users to generate on-the-fly images of persons that do not exist or to produce realistic deepfakes, without the need for any editing skill. Diffusion models [1, 2] enable the creation of realistic edits using natural language prompts, photorealistically adapting the inserted manipulation to the style and lighting of the context. Such tools clearly represent a threat in the wrong hands. Indeed, in recent years there has been a rising interest on the part of governments and funding agencies in developing forensic tools capable of countering such attacks. In particular, a major focus in the field of **multimedia forensics** is on local image manipulations that change the semantics of the image.

In my second year of PhD I focused my research activities on the tasks of **image forgery detection**, which consists in providing the probability that the image has been manipulated, and **image forgery localization**, which aims at localize and highlight the manipulated regions.

Image forgeries leave some weak marks on the image: traces that invisible to the naked eye. Low-level artifacts are caused by the in-camera acquisition process, such as the sensor, the lens, the color filter array or the JPEG quantization tables. These traces can be enhanced by high-pass filtering or denoising, suppressing the image content as a result. When images are tampered, these traces may be corrupted, which allows one to carry out powerful forensic analyses. However, the uploading of images on social media makes the detection of manipulation more challenging, due to re-compression and other post-processing performed by the social network.

Some state-of-the-art detectors are built to exploit well-defined low-level features, such as traces of JPEG compression, demosaicking or interpolation [3, 4, 5], while others are typically developed to work well only on specific types of manipulations, like splicing [6, 7]. Despite considerable advances in the area, current SOTA detectors are not yet performant enough for in-the-wild deployment, due mainly to limited generalization or robustness and insufficient detection performance. Nevertheless, most methods perform image forgery localization, without a real focus on detection [8], deriving the image-level integrity score from the localization heatmap itself [9, 10, 11]. Few methods address the detection task directly [12, 13, 14, 15]. As a result, detection accuracy is poor, often with a high false positives rate.

My research focused on addressing such shortcomings, with a focus on the reliability of the detection and robustness under varied manipulations. The first component of my research product is the **Noiseprint++**. It is an extension of the work [16], in which a deep learning-based method to extract from an image its noise fingerprint, called Noiseprint, has been proposed. In the Noiseprint, traces related to in-camera processing steps are collected and emphasized, while the image content is suppressed. However, it shows limited robustness to image impairments induced by out-camera processes. The Noiseprint++, on the other hand, is an improved fingerprint which highlights traces related not only to the camera model but also to its **editing history**.

Training is aimed at obtaining the same noise-sensitive fingerprint for patches that share the same properties and different noise residuals for patches that are different under some respect. This means that images with different editing histories or coming from different camera models will have a different Noiseprint++.

In our method, which we called **TruFor** (Trustworthy Forensics) [P1], we feed the input RGB image and its Noiseprint++ in two networks that extract the anomaly localization map and an estimated confidence map. The information gathered in the localization map is then summarized in a compact descriptor, by means of a weighted pooling block, with weights depending on the confidence information. Finally, this descriptor is processed by a classifier which computes the integrity score. In particular, the **forgery localization** task is treated as a supervised binary segmentation problem, adopting the CMX architecture [17], a cross-modal fusion framework designed for multi-modal semantic segmentation. It consists in two parallel branches which have a shared encoder architecture based on SegFormer [18].

Another important aspect of my research concerns the **reliability** of the prediction. Designing reliable detectors is crucial in several computer vision applications, however, it is even more critical for our task, since forensic traces are often imperceptible to visual inspection. The problem is even more relevant when deep learning based methods are used, since image forensics tools are challenged by out-of-distribution data [19]. In the context of JPEG artifacts and resampling analysis, initial efforts to develop reliable forensics detectors are carried out in [20, 21], where it is proposed to use Bayesian neural networks that provide an uncertainty range with every prediction. In this way, the user can quantify trust on the final prediction. Inspired by [22], our work aims at making a further step in this direction and proposes a method using external uncertainty quantification [23] to design a **confidence map** from the anomaly localization heatmap. We propose a method to compute a per-pixel confidence estimate of the predicted anomaly map, based on the concept of True Class Probability (TCP) [22]. The TCP map is close to 1 for a good prediction and close to 0 for a bad prediction. After training our localization network, we train our confidence network, which learns to estimate the TCP, providing an estimated confidence map. This process can also be considered as a form of failure prediction, where a network is trained to predict where another trained network fails. Finally, to maximize the system reliability, the final binary classifier, responsible for the **forgery detection** task, is trained jointly with the confidence estimator. This detector gathers information from both the localization map and the confidence map, hence providing a more reliable integrity score, less affected by false positives in the localization map.

We benchmarked our model on seven publicly available datasets and one more dataset of local manipulations created by us using modern image synthetic generation models. We compared our performances to the ones of other SOTA methods, obtaining promising results. On average, our method provides the best localization performance, as well as detection performance, which testifies of a remarkable generalization ability across manipulations. We also carried out a robustness analysis on images impaired by compression and resizing. To this end, we use datasets uploaded on Facebook and Whatsapp, and TruFor performs consistently better than all competitors. Ablation studies have also been carried out, proving that the combination of Noiseprint++ and RGB, as well as the use of the confidence map, were the key for obtaining a competitive advantage.

References:

[1] Omri Avrahami, Ohad Fried, and Dani Lischinski. Blended Latent Diffusion. *ACM Trans. Graph.* 42, 4, Article 149, August 2023

[2] Alexander Quinn Nichol et al. GLIDE: Towards photorealistic image generation and editing with text-guided diffusion models. *In International Conference on Machine Learning*, volume 162, pages 16784–16804, July 2022

[3] Quentin Bammey et al. An adaptive neural network for unsupervised mosaic consistency analysis in image forensics. *In IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020

[4] Tiziano Bianchi, Alessia De Rosa, and Alessandro Piva. Improved DCT coefficient analysis for forgery localization in JPEG images. *In IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2444–2447, 2011

[5] Jinseok Park et al. Double JPEG Detection in Mixed JPEG Quality Factors using Deep Convolutional Neural Network. *In European Conference on Computer Vision (ECCV)*, September 2018

[6] Myung-Joon Kwon et al. CAT-Net: Compression Artifact Tracing Network for Detection and Localization of Image Splicing. *In IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 375–384, 2021

[7] Ronald Salloum, Yuzhuo Ren, and C-C Jay Kuo. Image splicing localization using a multi-task fully convolutional network (MFCN). *Journal of Visual Communication and Image Representation*, pages 201–209, 2018

[8] Davide Cozzolino, Giovanni Poggi, and Luisa Verdoliva. Data-Driven Digital Integrity Verification, pages 281–311. *Springer*, 2022

[9] Minyoung Huh et al. Fighting fake news: Image splice detection via learned self-consistency. *In European Conference on Computer Vision (ECCV)*, September 2018

[10] Yuan Rao, Jiangqun Ni, and Hao Xie. Multi-semantic CRFbased attention model for image forgery detection and localization. *Signal Processing*, 183:108051, 2021

[11] Yue Wu, Wael AbdAlmageed, and Premkumar Natarajan. ManTra-Net: Manipulation Tracing Network for Detection and Localization of Image Forgeries With Anomalous Features. *In IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019

[12] Xinru Chen et al. Image manipulation detection by multi-view multiscale supervision. *In IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2021

[13] Francesco Marra et al. A Full-Image Full-Resolution End-to-End Trainable CNN Framework for Image Forgery Detection. *IEEE Access*, 8:133488–133502, 2020

[14] Junke Wang et al. ObjectFormer for image manipulation detection and localization. *In IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2022

[15] Rongyu Zhang and Jiangqun Ni. A dense U-Net with crosslayer intersection for detection and localization of image forgery. *In IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2982–2986, 2020

[16] Davide Cozzolino and Luisa Verdoliva. Noiseprint: A CNNBased Camera Model Fingerprint. *IEEE Transactions on Information Forensics and Security*, 15:144–159, 2020

[17] Jiaming Zhang et al. CMX: Cross-Modal Fusion for RGB-X Semantic Segmentation With Transformers. *In IEEE Transactions on Intelligent Transportation Systems*, pages 1-16, 2023

[18] Enze Xie et al. SegFormer: Simple and efficient design for semantic segmentation with transformers. *Advances in Neural Information Processing Systems (NeurIPS)*, 34:12077–12090, 2021

[19] Luisa Verdoliva. Media forensics and deepfakes: An overview. *IEEE Journal of Selected Topics in Signal Processing*, 14(5):910–932, 2020

[20] Christian Riess, Benedikt Lorch, Anatol Maier. Reliable JPEG Forensics via Model Uncertainty. *In IEEE Workshop on Information Forensics and Security (WIFS),* 2020

[21] Anatol Maier, Benedikt Lorch, and Christian Riess. Toward reliable models for authenticating multimedia content: detecting resampling artifacts with Bayesian Neural Networks. *In IEEE International Conference on Image Processing (ICIP)*, pages 1251–1255, 2020

[22] Charles Corbière et al. Addressing failure prediction by learning model confidence. *In Advances in Neural Information Processing Systems (NeurIPS)*, pages 2902–2913, 2019

[23] Jakob Gawlikowski et al. A survey of uncertainty in deep neural networks. Artificial Intelligence Review, 2023

4. Research products:

Conference Papers:

[P1] **F. Guillaro**, D. Cozzolino, A. Sud, N. Dufour, and L. Verdoliva, "TruFor: Leveraging all-round clues for trustworthy image forgery detection and localization" in IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Vancouver, June 2023, pp. 20606-20615. *Published, NOT yet indexed in Scopus*

5. Conferences and seminars attended

IEEE International Workshop on Information Forensics (WIFS 2022)

- Dates: 13/12/2022 – 16/12/2022
- Place: online
- Presenter of the paper “TruFor: Leveraging all-round clues for trustworthy image forgery detection and localization”

IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR 2023)

- Dates: 18/06/2023 – 22/06/2023
- Place: Vancouver, Canada
- Presenter of the paper "TruFor: Leveraging all-round clues for trustworthy image forgery detection and localization"

6. Periods abroad and/or in international research institutions

The research period abroad starts on October 30th. The hosting institution is Google LLC in Mountain View, California, USA, under the supervision of Dr. Avneesh Sud, member of the Mint Team of Google Research. The activities at Google concern the development of methods for image manipulation detection, with a focus on AI generated images.

Since the start date is at the very end of the second year, the amount of time spent for studying abroad during the year is only 2 days.

7. Tutorship

- Teaching assistance for the course "Elaborazione di Segnali Multimediali" (7 hours)

8. Plan for year three

During my third PhD year, I will focus on improving our proposed image forgery detection methodology to reduce the false positives directly in the localization map. Given the proliferation of open source AI image generators, I will also explicitly target manipulations generated with more recent diffusion-based models. Moreover, I also intend to investigate the potential of the rising Large Language Models (LLMs) to understand if it is possible to exploit the power of textual prompts to get a better detection of image forgeries.

I will explore some of these topics during my period abroad. I will spend 3 months at Google in Mountain View (California), USA, starting from October 30th 2023 to January 29th 2024, under the supervision of Dr. Avneesh Sud.