PhD in Information Technology and Electrical Engineering
Università degli Studi di Napoli Federico II

PhD Student: Vincenzo De Rosa

Cycle: XL

Training and Research Activities Report

Year: First

Vincenzo De Rosa

Tutor: prof. Luisa Verdoliva

Luisa Verdoliva

Date: October 24, 2025

1. Information:

- **PhD student: Vincenzo De Rosa**
- **DR number: DR999874**
- **Date of birth: 06/05/1998**
- **Master Science degree: Computer Engineering** **University: Federico II**
- **Doctoral Cycle: XL**
- **Scholarship type: UNINA – DIETI, Google Gift project**
- **Tutor: Luisa Verdoliva**

2. Study and training activities:

Activity	Type[1]	Hours	Credits	Dates	Organizer	Certificate[2]
Study of adversarial attacks and defences for synthetic image detection Presentation of the paper "Exploring the Adversarial Robustness of CLIP for AI-generated Image Detection" at the IEEE Workshop on Information Forensics and Security, Rome 2-5 December 2024	Research		6	01/11/24-31/12/24		N
Using Deep Learning Properly	Course	10	4	03/02/25-19/02/25	DIETI-UNINA	Y
Can we Rely on AI? Reliability Issues in Artificial Neural Networks and Potential Solutions for Autonomous Vehicles	Seminar	1	0.2	16/01/25	DIETI-UNINA	Y
The Good, the Bad, and the Ugly in Quantum Computing: Computational Power, Intrinsic Noise, and Transient Faults	Seminar	1	0.2	17/01/25	DIETI-UNINA	Y
Study on defenses to adversarial attacks for Synthetic Image Detector	Research		6	01/01/25-28/02/25		N
How to boost your PhD	Course	16	5	08/01/25-12/02/25	DIETI-UNINA	Y

AI for applications in psychiatry	Seminar	1	0.2	12/03/25	EURASIP	Y
Multiphysics Simulation of Power Transformers	Seminar	1	0.2	20/03/25	IEEE	Y
Automatic Control in the Era of Artificial Intelligence	Seminar	1	0.2	24/03/25	DIETI-UNINA	Y
Multiuser MIMO Wideband Joint Communications and Sensing With Subcarrier Allocation	Seminar	1	0.2	26/03/25	SPS-IEEE	Y
Graph Fourier Transform for Directed Graphs	Seminar	1	0.2	28/03/25	SPS-IEEE	Y
Robot Autonomy among Decision-Making Agents	Seminar	1	0.2	15/04/25	DIETI-UNINA	Y
On the Security of Semantic Watermarking to Detect AI-Generated Content	Seminar	1	0.2	29/04/25	DIETI-UNINA	Y
Study on protection methods against style mimicry	Research		6	01/03/25-30/04/25		N
8th Advanced Course on Data Science & Machine Learning	Doctoral School	40	8	09/06/25-13/06/25	ACDL 2025	Y
Frequency Artefacts in Diffusion Models: an Achilles' Heel for Deepfakes?	Seminar	1	0.2	06/05/25	SPS-IEEE	Y
YDTR: Infrared and Visible Image Fusion via Y-Shape Dynamic Transformer	Seminar	1	0.2	23/05/25	SPS-IEEE	Y
PhD Survival Strategies	Seminar	1.5	0.3	30/05/25	DIETI-UNINA	Y
Robotic Manipulation @Vanvitelli Robotics Lab: A bird’s eye view on the last 5 years	Seminar	1	0.2	18/06/25	DIETI-UNINA	Y
Neural Acoustic Feedback Cancellation: From	Seminar	1.5	0.3	18/06/25	SPS-IEEE	Y

Signal Processing to Deep Learning						
Study on protection methods against style mimicry	Research		3	01/05/25-30/06/25		N
Directly Parameterized Neural Network Construction for Generalization and Robustness in Imaging Inverse Problems	Seminar	1	0.2	17/07/25	SPS-IEEE	Y
Multi-Scale Spectral Loss Revisited	Seminar	1	0.2	23/07/25	SPS-IEEE	Y
Distributed Signal Processing for Extremely Large-Scale Antenna Array Systems	Seminar	1	0.2	30/07/25	SPS-IEEE	Y
Machine Learning Methods for Trustworthy Autonomous Systems	Seminar	1	0.2	14/08/25	SPS-IEEE	Y
Foundational Speech Models and their Efficient Training with NVIDIA NeMo	Seminar	1.5	0.3	27/08/25	SPS-IEEE	Y
Trade-offs and Non-idealities in ISAC Systems: From Monostatic to Bistatic	Seminar	1	0.2	28/08/25	SPS-IEEE	Y
Strategies for improving innovation outcomes: How IP and R&D leaders turn patent data into business intelligence	Seminar	1	0.2	28/08/25	IEEE Spectrum	Y
NeuroAI: From HoloBrain to HoloGraph	Seminar	1	0.2	29/08/25	SPS-IEEE	Y
Study of protection methods against style mimicry	Research		8.3	01/07/25-31/08/25		N
11th IEEE - EURASIP Summer School on Signal Processing	Doctoral School	26	5.2	21/09/25-26/09/25	IEEE - EURASIP Summer School on Signal Processing	Y

IEEE Authorship and Open Access Symposium: Tips and Best Practices to Get Published from IEEE Editors	Seminar	1.5	0.3	15/10/25	IEEE Xplore	Y
Self-Supervised Coordinate Projection Network for Sparse-View Computed Tomography	Seminar	1	0.2	20/10/25	SPS-IEEE	Y
Study of protection methods against style mimicry	Research		5	01/09/25-31/10/25		N

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	-	-	6	-	6
Bimonth 2	4	0.4	6	-	10.4
Bimonth 3	5	1.4	6	-	12.4
Bimonth 4	8	1.2	3	-	12.2
Bimonth 5	-	1.7	8.3	-	10
Bimonth 6	5.2	0.5	5	-	10.7
Total	22.2	5.2	34.3	-	61.7
Expected	30 - 70	10 - 30	80 - 140	0 – 4.8	

3. Research activity:

The rapid advancement of generative AI, particularly through Diffusion Models, has created a paradigm shift in synthetic media generation. These powerful models can synthesize images with extraordinary realism and can be conditioned on textual descriptions, enabling a wide range of applications across entertainment, education, design, and healthcare. At the same time, such capabilities pose serious threats in the context of disinformation, copyright infringement, and digital forensics, motivating intense research on automated methods for detecting and mitigating the misuse of synthetic content. During my first year of PhD, my work focused on the robustness of forensic image detectors and the protection of visual content against style mimicry.

In recent years, the field of multimedia forensics has seen extensive research into detectors for AI-generated images, driven by the need to prevent their malicious use. For a long time, Convolutional Neural Networks (CNNs) have been the dominant architecture for this task and have been thoroughly investigated. However, newly proposed Transformer-based detectors, particularly those utilizing Contrastive Language-Image Pretraining (CLIP) with Visual Transformer (ViT) backbones, have demonstrated performance that matches or even surpasses CNNs, especially in their ability to

generalize to new types of synthetic images [1, 2, 3]. While the detection capabilities of these models are advancing, their security and resilience against deliberate attacks remain a critical concern. The field of adversarial machine learning has shown that deep learning models, including forensic detectors, are vulnerable to inputs that are subtly modified to cause misclassification.
Previous work has successfully attacked CNN-based deepfake detectors in both white-box and black-box settings, proving these threats are practical [4, 5]. However, these studies have primarily focused on CNNs, leaving a significant gap in understanding the adversarial robustness of the emerging CLIP-based Transformer detectors.
To address this research gap, [P1] presents a comprehensive study of the adversarial robustness of CLIP-based AI-generated image detectors, directly comparing them with established CNN-based methods. We evaluate their robustness against a variety of adversarial attacks and analyze the patterns of adversarial noise in the frequency domain. Our analysis reveals that while both types of detectors are highly vulnerable to white-box attacks, there is limited transferability of attacks between CNN-based and CLIP-based architectures. This behavior can be explained by the fundamental difference in the features they rely on, as evidenced by their adversarial noise spectra: CNN-based attacks perturb medium-high frequencies, whereas CLIP-based attacks concentrate on low frequencies.
Although these observations are specific to multimedia forensics, similar results have been reported in other computer vision fields as well [6, 7]. Overall, CLIP-based detectors are not intrinsically more robust than CNN-based ones, but their architectural dissimilarities significantly reduce attack transferability. This result is consistent with findings in image classification [8]. These insights deepen our understanding of how forensic detectors operate and provide a foundation for the development of more resilient and secure detection strategies.

Building upon these findings, my research also extends to the defensive side of adversarial machine learning, exploring how similar principles can be harnessed not only to attack but also to protect visual content. Modern diffusion models, such as Stable Diffusion, can be fine-tuned on a small set of an artist's publicly available works to replicate their distinctive style without consent. In this context, my research investigates the use of adversarial perturbations as protective tools, focusing on how subtle and imperceptible modifications can prevent diffusion models from replicating artistic styles. Several protection methods, such as Glaze [9], Photoguard [10], and Mist [11], have been developed to embed these perturbations into artworks, aiming to disrupt the fine-tuning process of generative models. Despite promising results, these defenses remain highly vulnerable to standard image processing operations [12]. Attackers can easily circumvent protections through common post-processing techniques such as JPEG compression, Gaussian noise injection, or image resizing, which severely limits their effectiveness and real-world robustness.
To address these limitations, my research will focus on developing a robust defense method that leverages adversarial techniques to protect artworks against style mimicry. The goal is to design perturbations that are not only imperceptible to human observers but also resilient to common image manipulations, thereby ensuring long-term and practical protection of artistic content.

References

[1] U. Ojha, Y. Li, and Y. J. Lee, "Towards universal fake image detectors that generalize across generative models," in CVPR, 2023.
[2] Z. Sha, Z. Li, N. Yu, and Y. Zhang, "DE-FAKE: Detection and Attribution of Fake Images Generated by Text-to-Image Generation Models," in ACM CCS, 2023.

[3] D. Cozzolino, G. Poggi, R. Corvi, M. Nießner, and L. Verdoliva, "Raising the Bar of AI-generated Image Detection with CLIP," in CVPR Workshop, 2024, pp. 4356–4366.
[4] N. Carlini and H. Farid, "Evading deepfake-image detectors with white and black-box attacks," in CVPR Workshop, 2020, pp. 658–659.
[5] S. Hussain, P. Neekhara, M. Jere, F. Koushanfar, and J. McAuley, "Adversarial Deepfakes: Evaluating Vulnerability of Deepfake Detectors to Adversarial Examples," in WACV, 2021, pp. 3348–3357.
[6] P. Benz, S. Ham, C. Zhang, A. Karjauv, and I. S. Kweon, "Adversarial Robustness Comparison of Vision Transformer and MLP-Mixer to CNNs," in BMVC, 2021.
[7] J. Bai, L. Yuan, S.-T. Xia, S. Yan, Z. Li, and W. Liu, "Improving Vision Transformers by Revisiting High-frequency Components," in ECCV, 2022, pp. 1–18.
[8] K. Mahmood, R. Mahmood, and M. Van Dijk, "On the Robustness of Vision Transformers to Adversarial Examples," in ICCV, 2021.
[9] S. Shan, J. Cryan, E. Wenger, H. Zheng, R. Hanocka, and B. Y Zhao, "Glaze: Protecting artists from style mimicry by {Text-to-Image} models" in 32nd USENIX Security Symposium (USENIX Security 23), 2023, pp. 2187–2204.
[10] H. Salman, A. Khaddaj, G. Leclerc, A. Ilyas, and A. Madry, "Raising the cost of malicious ai-powered image editing," arXiv preprint arXiv:2302.06588, 2023.
[11] C. Liang and X. Wu, "Mist: Towards improved adversarial examples for diffusion models," arXiv preprint arXiv:2305.12683, 2023.
[12] R. Honig, J. Rando, N. Carlini, and F. Tramèr, "Adversarial perturbations cannot reliably protect artists from generative ai," arXiv preprint arXiv:2406.12027, 2024.

4. Research products:

*[P1] **V. De Rosa**, F. Guillaro, G. Poggi, D. Cozzolino, and L. Verdoliva*
"Exploring the Adversarial Robustness of CLIP for AI-generated Image Detection", in 2024 IEEE International Workshop on Information Forensics and Security (WIFS), Rome, Italy, Published

5. Conferences and seminars attended:

2024 IEEE International Workshop on Information Forensics and Security (WIFS)
- Dates: 02/12/2024 – 05/12/2024
- Place: Rome, Italy
- Presentation of the paper "Exploring the Adversarial Robustness of CLIP for AI-generated Image Detection"

6. Activity abroad:

None

7. Activity in partner companies:

None

8. Tutorship:

None