Anxhelo Diko

Ph.D. Fellow, Department of Computer Science, Sapienza University of Rome

PERSONAL DETAILS

Birth	February 19, 1997
Address	Via Salaria 113, 00198 Rome, Italy
Phone	(+39) 3497921839
Mail	diko@di.uniroma1.it
Profile	www.linkedin.com/in/anxhelodiko97
Website	https://anxhelodiko.dev

PERSONAL STATEMENT

A highly motivated and results-oriented Computer Vision Ph.D. student with a deep passion for advancing the field of artificial intelligence. My research focuses on building multimodal representations and understanding human activities, addressing key challenges for autonomous agents and AI in general. I have extensive experience with multimodal large language models for video captioning and question answering and a keen interest in view-invariant video representation learning. I am particularly committed to exploring how to effectively bridge the gap between representations of different modalities while preserving their unique characteristics.

In addition to my research expertise, I have a strong engineering foundation honed through academic and industry experiences. Proficient in Python, C++, and CUDA, I excel at rapidly prototyping and implementing innovative ideas. I am eager to leverage my skills and knowledge to contribute to cutting-edge research and development in this dynamic field.

WORK EXPERIENCE

Applied Computer Vision Scientist, Huawei *2024-Ongoing*

Huawei, (7 months)

- Research on Multimodal LLMs focuses on Video Captioning and Video Question and Answering, targeting long-term videos (10+ minutes to hour-long) and hallucinations.
- References: Ioannis Patras, i.patras@qmul.ac.uk, Lead AI Scientist at Huawei and Full Professor at Queen Mary University of London

Contracted Machine Learning Researcher *2021-Ongoing*

Sapienza University of Rome (24 months)

- Conduct research on various computer vision problems like anomaly detection, action recognition, action anticipation, object detection, and vision-language models.

Machine Learning Engineer, LexmaTech *2020-2021*

LexmaTech, (16 months)

- Designed and deployed machine learning models at scale.
- Designed and implemented a parallel and scalable Ray-Tracing algorithm that runs on GPUs for discretizing 3D mesh representation of geometries into a volumetric representation. The implemented algorithm would cut the computational costs of the services offered by LexmaTech by 30% in the preparation phase.
- Reference: Simone Melchionna, simone.melchionna@gmail.com (Ex-Harvard Professor of Advanced Physics and CNR)

Applied Machine Learning Scientist, PaperClicks *2018-2018*

PaperClicks, Internship (6 months)

- Responsible for designing and implementing machine learning algorithms to optimize core business operations at PaperClicks.

EDUCATION

Ph.D. in Computer Science *2021-2025*
Sapienza University of Rome

- Research Area/s: Multimodal Video Representations.
- Advisor: Prof. Luigi Cinque

MSc. Computer Science, Sapienza University of Rome *2018-2020*

- Important Courses: Machine Learning, Computer Vision, Applied Artificial Intelligence, Multimodal Interaction, Cloud Computing, Advanced Software Engineering, Distributed Systems.
- Graduated as top 1% of the class.
- Finsl Grade: Cum laude

BSc. Business Computer Science, University of Tirana *2015-2018*

- Important Courses: Algorithms, Data Structures, Lnear Algebra, Calculus, C++, Java, Computer Architecture, Databases, Computer Networks, Information Security, Statistics.
- Thesis: Graduated as top 1% of the class.
- Final Grade: 10/10

HIGHLIGHTED PROJECTS

SEARCHER – Smart unmannEd AeRial vehiCles for Human likE monitoRing *2022-Ongoing*
Italian Ministry of Defense
Senior R&D Engineer

- Study and analysis about state-of-the-art anomaly detection, novelty detection, and UAV attention mechanism algorithms.
- Design and development of novel deep learning algorithms for UAV application in navigation and surveillance.
- Engineering the execution pipeline of the deep learning system on edge.

The project has raised funds from the Ministry of Defense.

ReWind: Understanding Long Videos with Instructed Learnable Memory. *2024-2024*
Huawei

- Designed and implemented a VLM for long-term video understanding equipped with a memory module.
- We introduce a parametrized memory module.
- A dynamic frames selection mechanism based on user instructions is introduced for selecting the most important frames related to the question.

Semantically Guided Representation Learning for Action Anticipation *2023-2024*
Individual Project, part of PhD research

- Designed and implemented neural networks that could anticipate future human action from RGB videos.
- Introduced two novel attention mechanisms that model the causality between video events happening in different time steps and preserve temporal order.
- Introduced a novel learning approach for action anticipation which bridges the gap between visual interpretation of the future and text.
- Achieved competitive results on three benchmarks, namely EpicKtchens-100/55 and EGTEA++.

View-Invariant Video Understanding exploiting relative similarities from RGB videos *2023-Ongoing*
Collaboration with IPVLab supervised by Prof. Giovanni Maria Farinella, part of PhD research

- Align the understanding of actions from multiple different synchronized views.
- Exploit the relative similarities between different actions in latent space.
- Exploit differences between POV to create disentangled view-dependent latent spaces.

ReViT – Enhancing Vision Transformers With Residual Attention (Submitted for publication at Pattern Recognition) *2022-2023*

Individual Project, part of PhD research

- Designed and implemented a novel residual connection between transformer blocks that propagates and accumulates knowledge from shallow to deeper layers.
- Enhanced Vision Transformers performance by improving their feature diversity in deeper layers.
- Obtained an improvement of +4% on ImageNet1K

Enabling Smart Assistants Communication with Deaf and Mute People Through Sign-Language Detection and Text-To-Speech Multimodal Interaction *2019-2020*

Individual Project

- Designed and implemented a solution that could read sign languages from an RGB camera, translate the sign language into text, transform the text into speech to communicate with a smart assistant, capture the smart assistant's answer, and convert it into written text.
- The implemented solution could help deaf and mute people communicate with smart assistants controlled by voice commands.

SKILLS

Skills Summary Hard Skills

- **Languages:** English, Albanian (mother tongue), Italian.
- **Professional Competences:** Algorithms, computer vision, MMLLM (Multimodal Large Language Models), LLMs, image processing, video processing, machine learning, deep learning, representation learning, visual recognition, action anticipation, action classification, object detection, semantic segmentation, homography, programming, parallel computing, transformers, convolutional neural networks, Unit testing, CI/CD, debugging, Prompting.
- **Programming Languages:** Python 3 (6+ years), C++ (2+ year), CUDA (1+ year), SQL (1+ year).
- **Tools and Frameworks (Expert):** PyTorch, HuggingFace api, OpenCV, NumPy, SciPy, Scikit-learn, Pandas, PlotLy, Docker, Bash, git, AWS.
- **Tools and Frameworks (Proficient):** PyTorch-Lightning, Tensorflow, Detectron2, MMDetector, DLlib, CuPy, PlotLy, Scikit-Image, MongoDB, MySQL, VTK, ITK, PoreSpy, MPI, OpenMP, Jenkins, Kubernetes.
- **Operating System:** Debian-based Linux (Ubuntu, Mint).

Soft Skills

- Communication, teamwork, attention to detail, problem-solving, adaptability, time management, work ethic, perseverance, consistency, and persistence.

PUBLICATIONS

Research works

1. "ReWind: Understanding Long Videos with Instructed Learnable Memory." - (Submitted at CVPR25) - First Author. Link: `https://arxiv.org/abs/2411.15556`
2. "LAGUNA: LAnguage Guided UNsupervised Adaptation with structured spaces" - (Submitted at CVPR25) - First Author. Link: `https://arxiv.org/abs/2411.15557`
3. "Semantically Representation Learning for Action Anticipation" - (Accepted at ECCV24) - First Author.
4. "ReViT: Enhancing Vision Transformers with Residual Attention." - (Pattern Recognition, 2024) - First Author. link: `https://github.com/ADiko1997/Vision.ai-PhD/tree/main/ReViT`.
5. "MS-Faster R-CNN: Multi-Stream Backbone for Improved Faster R-CNN Object Detection and Aerial Tracking from UAV Images." - Remote Sens. 2021, 13, 1670. `https://doi.org/10.3390/rs13091670` (Co-Author, Software implementation, Methodology and writing).
6. "Low-Altitude Aerial Video Surveillance via One-Class SVM Anomaly Detection from Textural Features in UAV Images." - Information 2022, 13, 2. `https://doi.org/10.3390/info13010002` (Co-Author, Methodology).

AWARDS AND FELLOWSHIPS

1. Research Fellowship, Sapienza University of Rome, IT (2021).
2. Research Fellowship, Sapienza University of Rome, IT (2022).
3. Research Fellowship, Sapienza University of Rome, IT (2023).
4. Research Grand "Avvio alla ricerca", Sapienza University of Rome, IT (2023).
5. LazioDisco Scolarship, Rome, IT (2018 - 2020).

CONFERENCES AND PUBLIC SPEAKING

1. ECCV 2024, Author
2. ICCV 2023, Volunteer Student
3. RomeRehab 2023, Invited Speaker, Computer Vision and Video Understanding Applied to GAIT Analysis for Rehabilitation