ellis unit | TURIN Talk

Politecnico di Torino

ARTIFICIAL INTELLIGENCE HUB
POLITECNICO DI TORINO

VANDAL
http://vandal.polito.it

www.ellis.eu

Mattia Segù
ETH Zurich

May 21st, 2025
starting at 14:30 AM cet
Link:
https://tinyurl.com/yc296vdp

Advancing Instance-Level Perception: End-to-End Sequence Modeling for Tracking and Efficient Multi-Modal Segmentation

Instance-level perception - the ability to localize, segment, and classify individual objects over time - is fundamental to systems that interact with the physical world. Recent advances in model architectures and data quality have enabled unified models capable of detecting and segmenting objects across both closed-set categories and free-form referring expressions. However, existing approaches struggle to scale to end-to-end instance tracking and efficiently adapt to edge deployment, posing key challenges for real-world applications. In this talk, I will present two recent works - SambaMOTR and MOBIUS - that push the boundaries of instance-level perception by addressing multi-object tracking and efficient segmentation. SambaMOTR enables end-to-end multi-object tracking by leveraging Samba, a set-of-sequences model that captures long-range dependencies, tracklet interactions, and temporal occlusions, improving robustness in dynamic environments with complex motion. MOBIUS makes vision-language instance segmentation scalable through a bottleneck encoder for efficient scale and modality fusion, and a language-guided calibration loss for adaptive decoder pruning, reducing inference time by up to 75% while maintaining state-of-the-art performance across both high-end and mobile devices. Through the lens of these two works, I will explore how sequence modeling and efficient multi-modal perception can be leveraged to develop scalable, real-time object perception models, enabling robust tracking and segmentation in complex environments.

Mattia Segù is a PhD candidate at the Computer Vision Lab at ETH Zurich, co-supervised by Prof. Luc Van Gool and Prof. Bernt Schiele. As a member of the Max Planck ETH Center for Learning Systems, he visited the Computer Vision and Machine Learning Department at the Max Planck Institute for Informatics, collaborating with Prof. Bernt Schiele. He has also worked as a Student Researcher at Google, contributing to Federico Tombari's team. His research focuses on instance-level perception, advancing multi-object tracking methods that can learn end-to-end from long video sequences, dynamically adapt, and leverage limited annotations in a self-supervised manner. His recent work extends to vision-language instance perception, developing models that can detect, track, and segment objects based on free-form referring expressions. In the past, he has also explored fundamental challenges in deep learning, including domain generalization and uncertainty estimation.