

Institute of Artificial Intelligence

UNIVERSITY OF CENTRAL FLORIDA

FINAL ORAL EXAMINATION

OF

Akash Kumar
B.Tech., Delhi Technological University, 2019

FOR THE DEGREE OF

DOCTOR OF PHILOSOPHY
(COMPUTER SCIENCE)

13, October, 2025, 2 P.M.
UCF Global-101

DISSERTATION COMMITTEE

Professor Yogesh S Rawat, *Chair*	*yogesh@crcv.ucf.edu*
Professor Mubarak Shah	*shah@crcv.ucf.edu*
Professor Chen Chen	*chen.chen@crcv.ucf.edu*
Professor Shaurya Agarwal	*shaurya.agarwal@ucf.edu*

DISSERTATION RESEARCH IMPACT

Deep learning has significantly advanced visual understanding tasks like object detection, tracking, and spatio-temporal grounding, benefiting fields such as autonomous vehicles, surveillance systems, and robotics. While large-scale labeled datasets have been crucial to this success, the process of spatio-temporally labeling video requires immense human effort, creating a major bottleneck for scaling this technology. Furthermore, progress has largely been confined to closed-world settings, hindering the ability of models to handle free-form queries and new concepts in unconstrained, real-world environments.

This dissertation addresses these fundamental challenges through several key contributions. First, to reduce annotation dependency, we introduce semi-supervised learning frameworks for dense tasks like video action detection. By incorporating spatio-temporal coherence constraints and student-teacher refinement, our methods achieve performance competitive with fully supervised models while using only a fraction of the labeled data. Second, to bridge the gap to open-world understanding, we propose novel methods for weakly-supervised learning. We introduce foundation models that ground natural language queries by employing contextual and progressive learning paradigms. These models learn to interpret compositional actions and navigate complex scenes without bounding box annotations, significantly enhancing their ability to generalize to unconstrained, real-world scenarios.

SELECTED PUBLICATIONS (h-index: 10, Total Citations: 376)

1. **End-to-End Semi-Supervised Learning for Video Action Detection,** Akash Kumar, Yogesh Singh Rawat, *in IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022*.

2. **Semi-supervised Active Learning for Video Action Detection,** Ayush Singh, Aayush J Rana, Akash Kumar, Shruti Vyas, Yogesh Singh Rawat. *in Association for the Advancement of Artificial Intelligence (AAAI), 2024*.

3. **Stable Mean Teacher for Semi-Supervised Video Action Detection,** Akash Kumar, Sirshapan Mitra, Yogesh Singh Rawat. *in Association for the Advancement of Artificial Intelligence (AAAI), 2025*.

4. **Contextual Self-paced Learning for Weakly Supervised Spatio-Temporal Video Grounding,** Akash Kumar, Zsolt Kira, Yogesh Singh Rawat. *in International Conference on Learning Representations (ICLR), 2025*.

5. **STPro: Spatial and Temporal Progressive Learning for Weakly Supervised Spatio-Temporal Grounding,** Aaryan Garg, Akash Kumar, Yogesh Singh Rawat. *in IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2025*.

DISSERTATION

TOWARDS LABEL-EFFICIENT APPROACHES FOR DENSE VIDEO TASKS

Although deep learning has advanced video understanding, its deployment is hindered by two major challenges: reliance on manual annotations and limited ability to operate beyond closed-world settings. Spatio-temporal labeling is costly, while real-world applications demand models capable of interpreting novel, open-ended queries. This dissertation develops methods to improve label efficiency and enable flexible, open-world video analysis.

To alleviate the annotation bottleneck, we establish a semi-supervised framework for video action detection based on consistency regularization, where a model learns to produce stable predictions across augmented views of the same unlabeled video. However, this is challenging - generic augmentations often affect only static regions, providing limited information about dynamic actions. We introduce spatio-temporal consistency constraints modeling motion dynamics. Two regularizers: temporal coherency and gradient smoothness - leverage the continuous nature of human actions to ensure predicted localizations remain coherent.

Building on this foundation, we refine our approach through an improved student-teacher framework. Traditional student–teacher models can accumulate errors when the teacher produces noisy pseudo-labels. We mitigate this with an error recovery module that learns from the student's mistakes on labeled data and transfers corrective knowledge to improve teacher predictions on unlabeled videos. This feedback progressively enhances pseudo-label quality. Additionally, a pixel-level difference loss enforces temporal consistency, enabling strong performance with only a fraction of labeled data.

While effective within known categories, these semi-supervised methods are insufficient for real-world scenarios requiring understanding beyond predefined labels. Our research extends to the open-world challenge of grounding free-form textual queries in video without bounding box supervision. We study localizing subjects spatio-temporally using only textual queries. Inspired by vision-language foundation models, we explore adapting their broad visual understanding. However, direct adaptation lacks essential spatio-temporal grounding and fine-grained alignment. To bridge this gap, we introduce a method linking textual queries to spatio-temporal predictions. Although effective, it struggles with compositional actions and crowded scenes. We address these limitations through a progressive learning framework that builds understanding gradually - first recognizing simple sub-actions, then adapting to dynamic contexts.

Extending this further, we propose a context-aware progressive learning approach that deepens understanding of nuanced scenes and language. It enriches grounding by dynamically using surrounding context to refine object identification and action interpretation. Instead of learning all concepts simultaneously, the model follows a self-paced curriculum - starting with coarse cues and advancing toward finer distinctions. This strategy strengthens temporal reasoning and contextual adaptation, enabling robust comprehension of complex, real-world narratives. Together, these advances form a cohesive path toward open-world video understanding, bridging the gap between closed-set perception and flexible, language-driven interpretation.

In summary, this dissertation advances two key frontiers in video understanding. By developing data-efficient semi-supervised methods for action detection and pioneering weakly supervised frameworks for open-world video grounding, this work moves the field toward scalable, flexible, and practical systems capable of interpreting complex visual narratives with minimal human supervision.



Akash Kumar

1997	Born in Patna, India
2015-2019	B.Tech., Delhi Technological University
2019-2020	Research Assistant, Indian Institute of Technology, Mandi
2024	Applied Scientist Intern, Amazon
2025	Applied Scientist Intern, Amazon
2020-2025	Ph.D., University of Central Florida