Efficient-data approaches for Dense Video Tasks

UCF Center for Research in Computer Vision

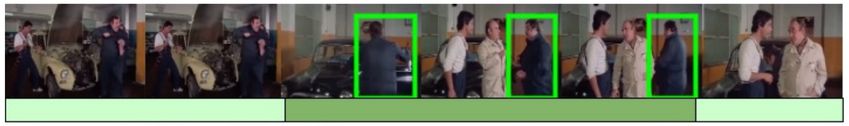
Akash Kumar, Yogesh Singh Rawat

WACV 2025 TUCSON, ARIZONA • FEB 28 - MAR 4

What comprises dense video tasks?

Video Action Detection: Action Classification + Spatio-Temporal Localization

Spatio-Temporal Video Grounding: Spatial + Temporal Referral Grounding

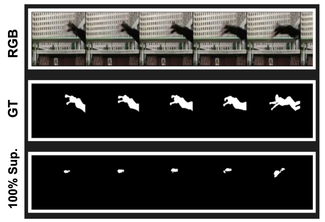


The man in blue shirt walks to car, speaks to man who gets out of car and walks behind him.

Why Data-efficient?

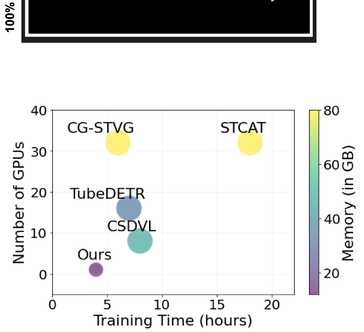
❖**Shortcomings of fully supervised approaches:**

- Costly Training + Laborious + **Expensive**
- Video Action Detection
  - **Lacks** precise spatio-temporal localization
  - **Weaker** temporal coherency
- Spatio-Temporal Video Grounding
  - **More** biasness
  - **Non-Scalability** to large-scale datasets



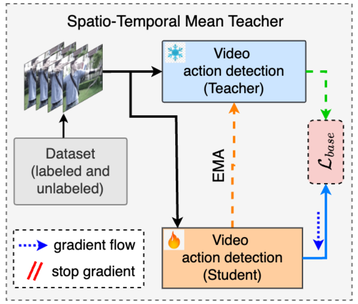
❖**Solutions:**

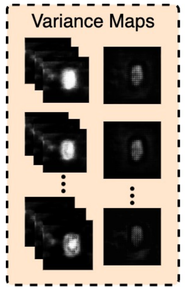
- Label **Efficient** Approach
- Video Action Detection
  - **Recover** fine-grained level localization
  - Enforce temporal **smooth** flow
- Spatio-Temporal Video Grounding
  - **Less** biasness (Frozen models)
  - **Scalability** for large-scale datasets

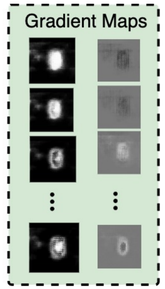


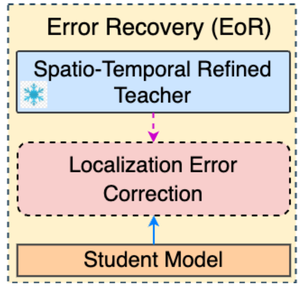
Closed-Set (Pre-defined set of actions): Semi-Supervised Video Action Detection

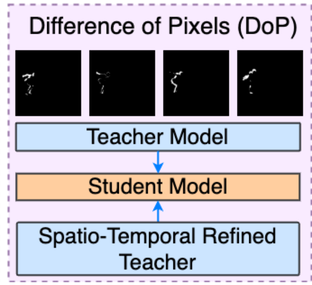
Baseline Semi-supervised Model | Solution: Temporal coherency [1] | Solution: Error Corrective fine-grained Localization [2,3]







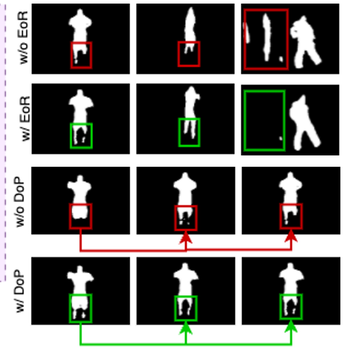




Limitations: Fine-grained localization + Temporal Coherency

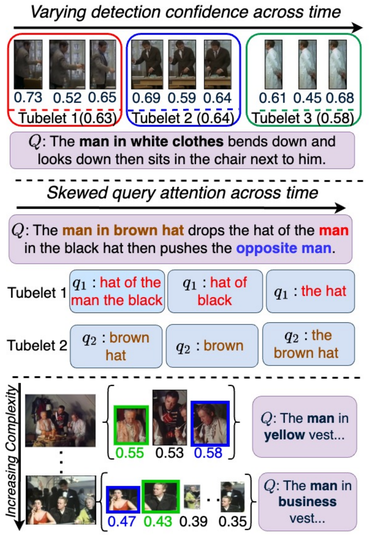
- Variance: Short-term fine-grained
- Gradient: Long-term smoothness
- Error Recovery: Class-agnostic Spatial boundary refinement
- Difference of Pixels: Spatio-Temporal Coherency Induction

Qualitative Analysis

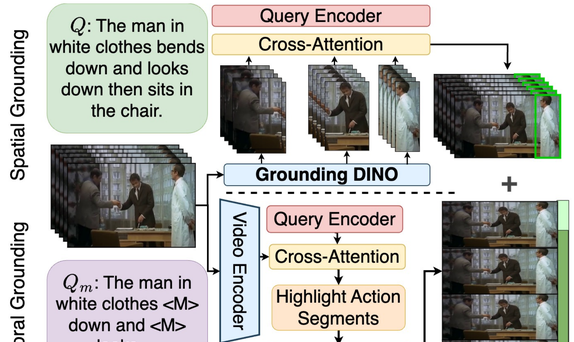


Open-Set (Free-form Query Understanding): Weakly-Supervised Spatio-Temporal Video Grounding

Limitations of **Grounding DINO**



First Step towards solution: **tubelet** ⟶ Tubelet Phrase Grounding | Solution: Fine-grained Attributes + Curriculum Learning[4]



Contextual Referral Grounding (CRG)

Noun	Adjective	Verb
Man	white	bends, looks, sits

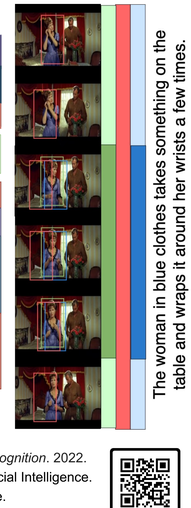
Global Query

Q: The man in white clothes bends down...

Local Query

Attribute	Verbs	Background
The man in white clothes	He bends down. He looks down. He sits.	next to him

The man in white clothes bends down
The man in white cloth looks down
The man in white cloth sits



Limitations: Free-form Query & Dense Scene complexity Understanding

[1] Kumar, Akash, and Yogesh S Rawat. "End-to-end semi-supervised learning for video action detection." *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2022.
[2] Singh, A., Rana, A.J., Kumar, A., Vyas, S. and Rawat, Y.S., 2024, March. Semi-supervised active learning for video action detection. In Proceedings of the AAAI Conference on Artificial Intelligence.
[3] Kumar, A., Mitra, S., and Rawat, Y.S., 2025, March. Stable Mean Teacher for Semi-supervised video action detection. In Proceedings of the AAAI Conference on Artificial Intelligence.
[4] Kumar, A., Kira, Z., and Rawat, Y.S., 2025, April. Contextual Self-paced Learning for Weakly Supervised Spatio-Temporal Video Grounding. In Proceedings of the ICLR.

SCAN ME