BUPT MCPRL

BUPT-MCPRL at TRECVID 2023 ActEV-SRL Challenge

Yang Song , HongPu Zhang , ZeLiang Ma , Zhe Cui, Yanyun Zhao
Beijing University of Posts and Telecommunications, China
{sy12138, zhp, mzl, cuizhe, zyy}@bupt.edu.cn

北京郵電大學
Beijing University of Posts and Telecommunications

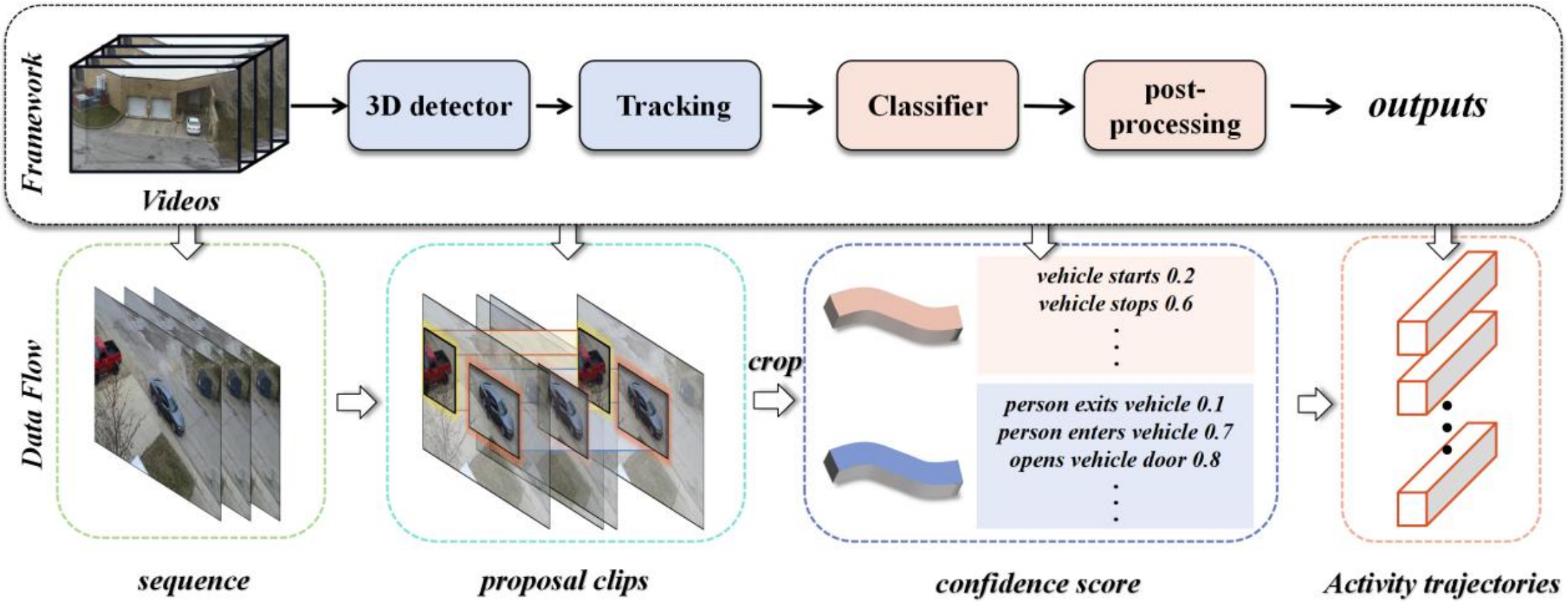
method

The MEVA dataset contains a total of 20 categories of activities, which we roughly divide into 5 activity groups and process them separately:

person-object	*person-specific object*	*vehicle-only*	*person-vehicle*	*scene-related and person-person*
person_reads_document, **person_texts_on_phone,** **person_picks_up,** **person_puts_down,** **person_sits_own,** **person_stands_up,** **person_transfers_object**	**person_interacts_ with_laptop**	**person_exits_vehicle,** **person_enters_vehicle,** **person_opens_vehicle _door,** **person_closes_vehicle _door**	**person_exits_vehicle,** **person_enters_vehicle,** **person_opens_vehicle _door,** **person_closes_vehicle _door**	**person_opens _facility_door,** **person_enters_scene _through_structure,** **person_exits_scene _through_structure,** **person_talks_to_person**

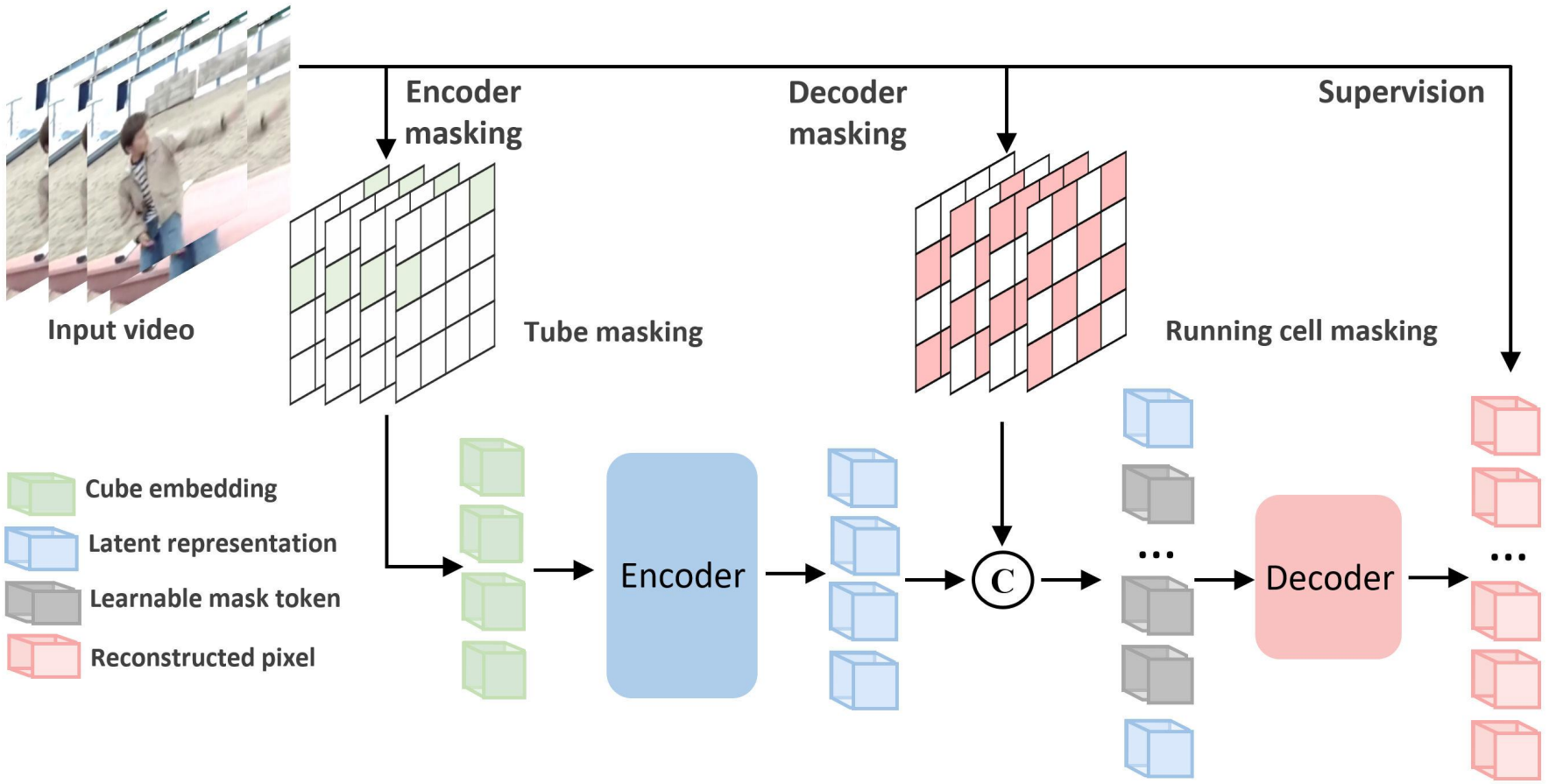
method - Framework

- 3D detector: Cascade R(2+1)D
- 5 different classification methods
- 3D classifier: VideoMAE V2



* R(2+1)D: https://arxiv.org/abs/1711.11248
* VideoMAEv2: https://arxiv.org/abs/2303.16727

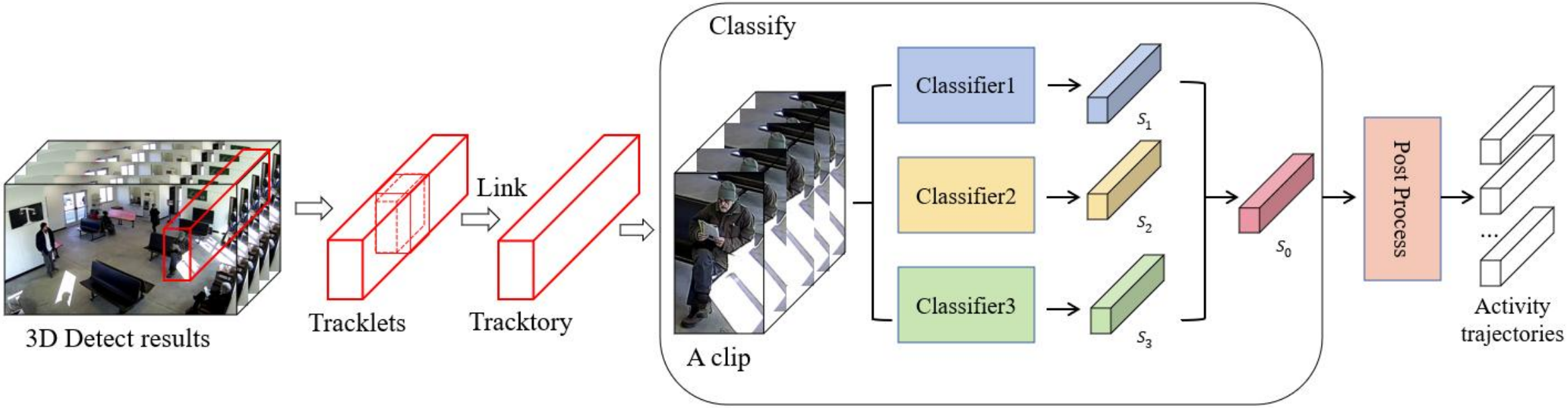
method - Classifier



- Model's Outstanding Performance
- Effective Video Feature Extraction
- Simplified Model Structure
- Similarity with Pretraining Dataset

person-object activity group detection

person_reads_document, person_texts_on_phone, person_picks_up, person_puts_down, person_sits_own, person_stands_up, person_transfers_object;



- 3D detector: Cascade RCNN
- 3D classifier: VideoMAEv2+ActionCLIP+Swin Transformer

* *ActionCLIP: https://arxiv.org/abs/2109.08472v1*

person-object activity group detection

❖ key issues and solutions

1. Sensitive to background information
Solution: Add more other activity categories in training stage.

2. Difficult to extract the feature information
Solution: Fine-tune the Large-scale model to extract more feature information.

3. Adopting a classifier score merge strategy
Solution: Adopt a classifier score merge strategy to synthesize the results of different classifiers to obtain a more representative score results.

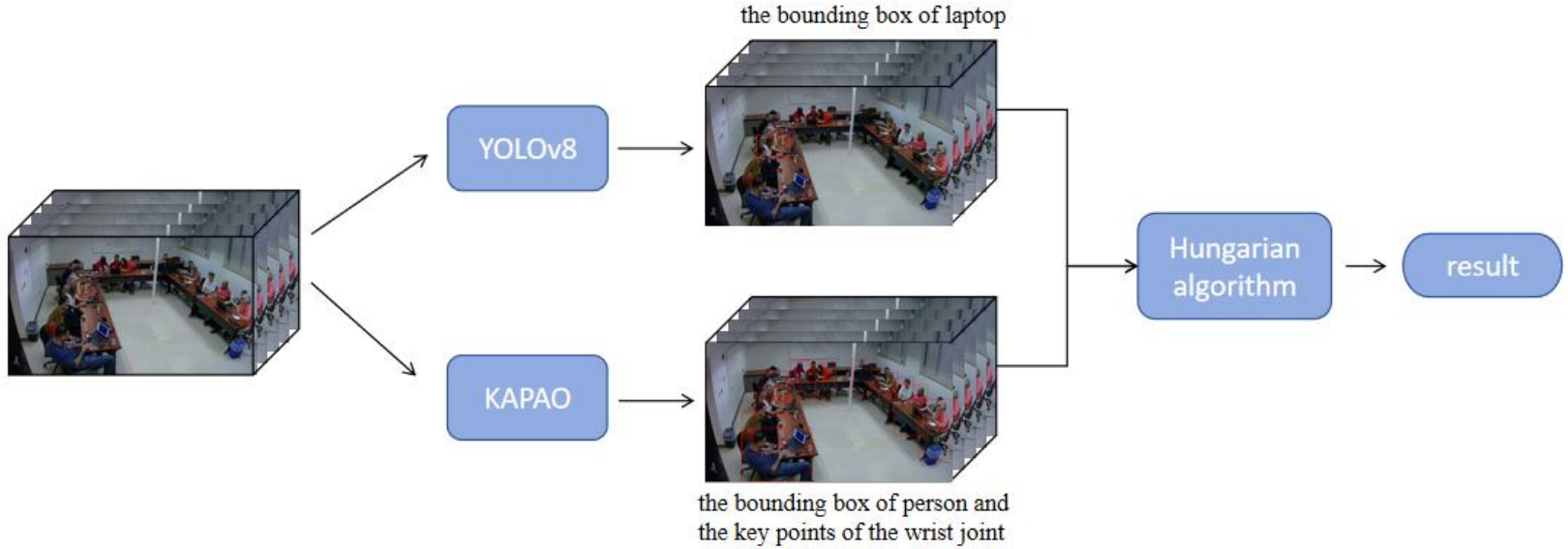
person-specific object activity group

person interacts with laptop;

- 3D detector: Cascade RCNN
- 2D detector: YOLOv8
- 2D pose estimation: KAPAO

❖ key issue

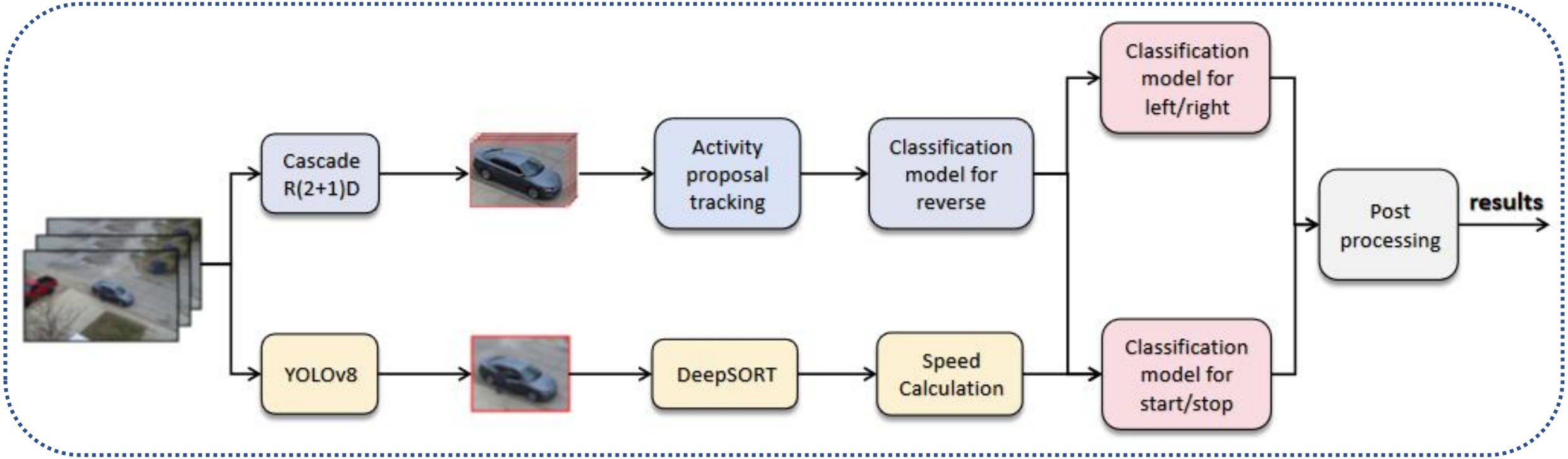
The generalization of the classifier is poor, resulting in poor results for scenes that do not exist in the training set and validation set on the test set.



** YOLOv8: https://github.com/ultralytics/ultralytics*
** KAPAO: https://arxiv.org/abs/2111.08557*

vehicle-only activity detection

vehicle starts, vehicle stops, vehicle turns left, vehicle turns right;



- 3D detector: Cascade R(2+1)D
- 3D classifier: Swin Transformer
- 2D detector: YOLOv8
- Tracker: DeepSORT

vehicle-only activity detection

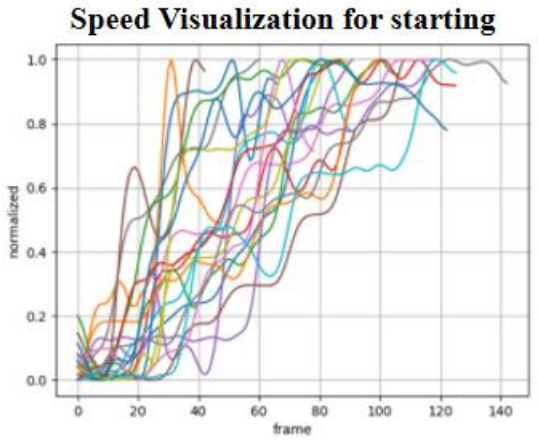
❖ key issues and solutions

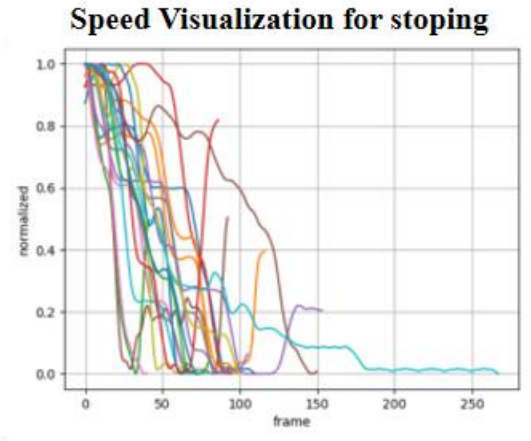
1. The Limited Dataset of Reverse Behavior

Solution: Reversing the samples for left and right turn behaviors to create synthetic reverse samples.

2. Method for Assisting Start-Stop Behavior Classification

Solution: Introducing 2D Detection and Tracking, Calculating the Speed of Each Frame in Start-Stop Category, and Assisting in Tracking Based on the Duration of Speed Increase/Decrease.





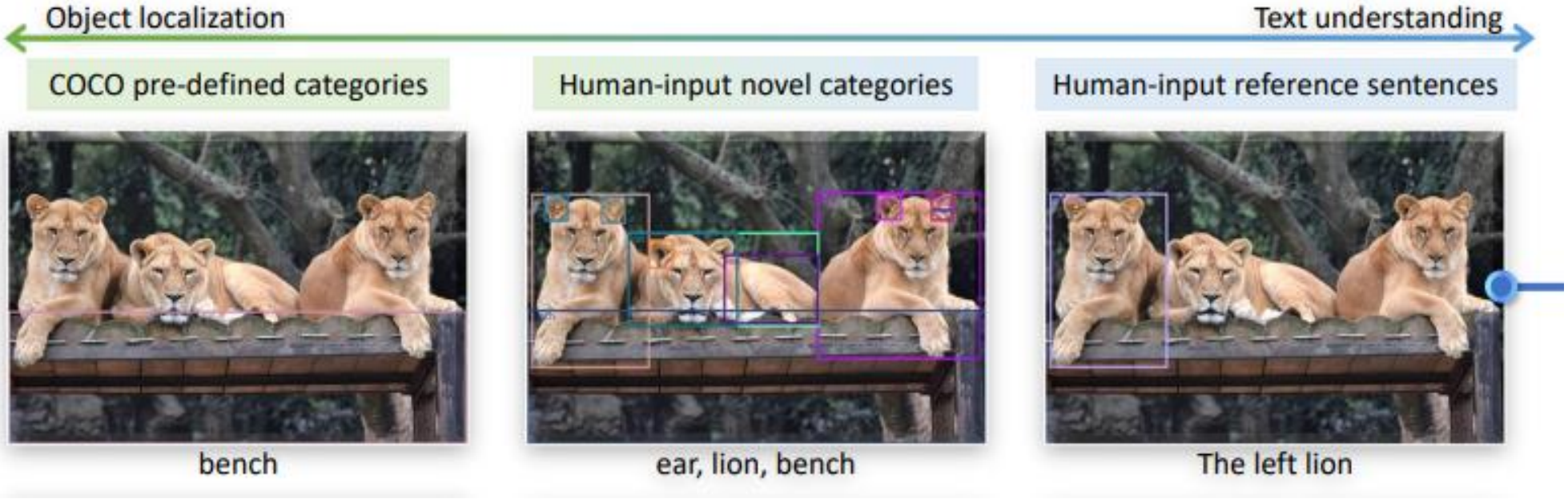
person-vehicle activity group

person exits vehicle, person enters vehicle, person opens vehicle door, person closes vehicle door;

❖ **key issues: A single action area contains multiple actions**



Solution:
Grounding DINO

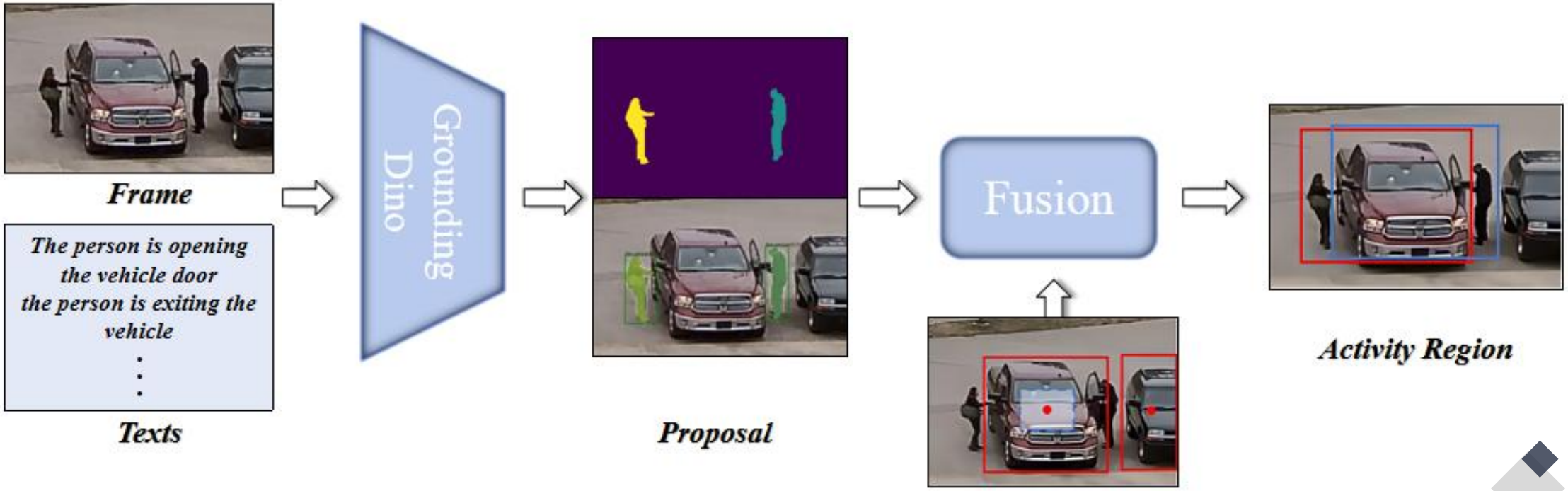


* *Grounding DINO:*
https://arxiv.org/pdf/2303.05499.pdf

person-vehicle activity group

Solution:

Utilizing Grounding DINO to detect individuals engaged in actions and incorporating the vehicle's location information to generate the Activity Region

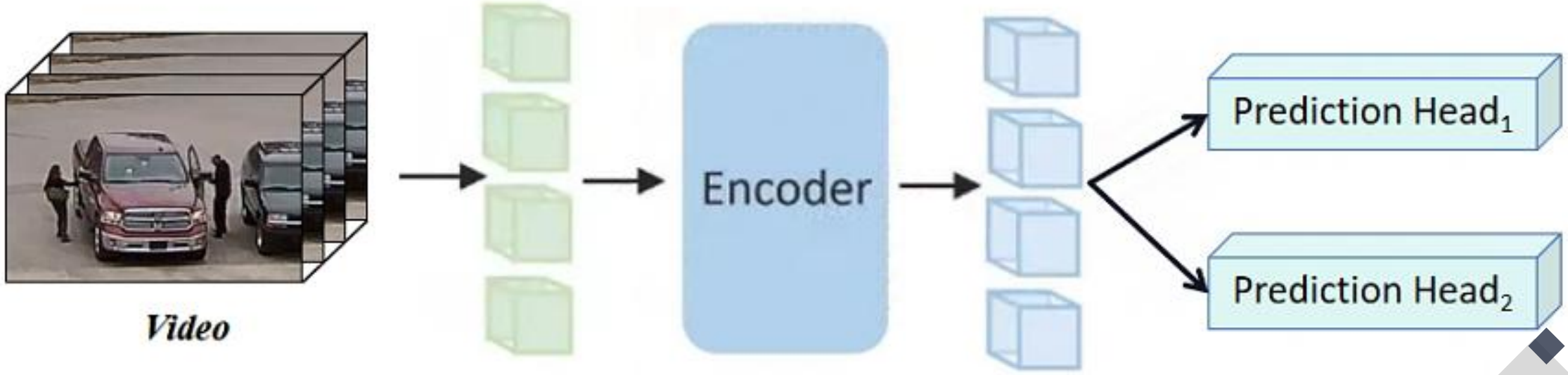


person-vehicle activity group

❖ key issues and solutions

1. Relevance between activities

Solution: In the training phase, we employ a multi-task training strategy. We use the same backbone with two separate prediction heads, each predicting one of the two mutually exclusive categories: opening and closing vehicle doors and entering and exiting the vehicle.



scene-related activity group

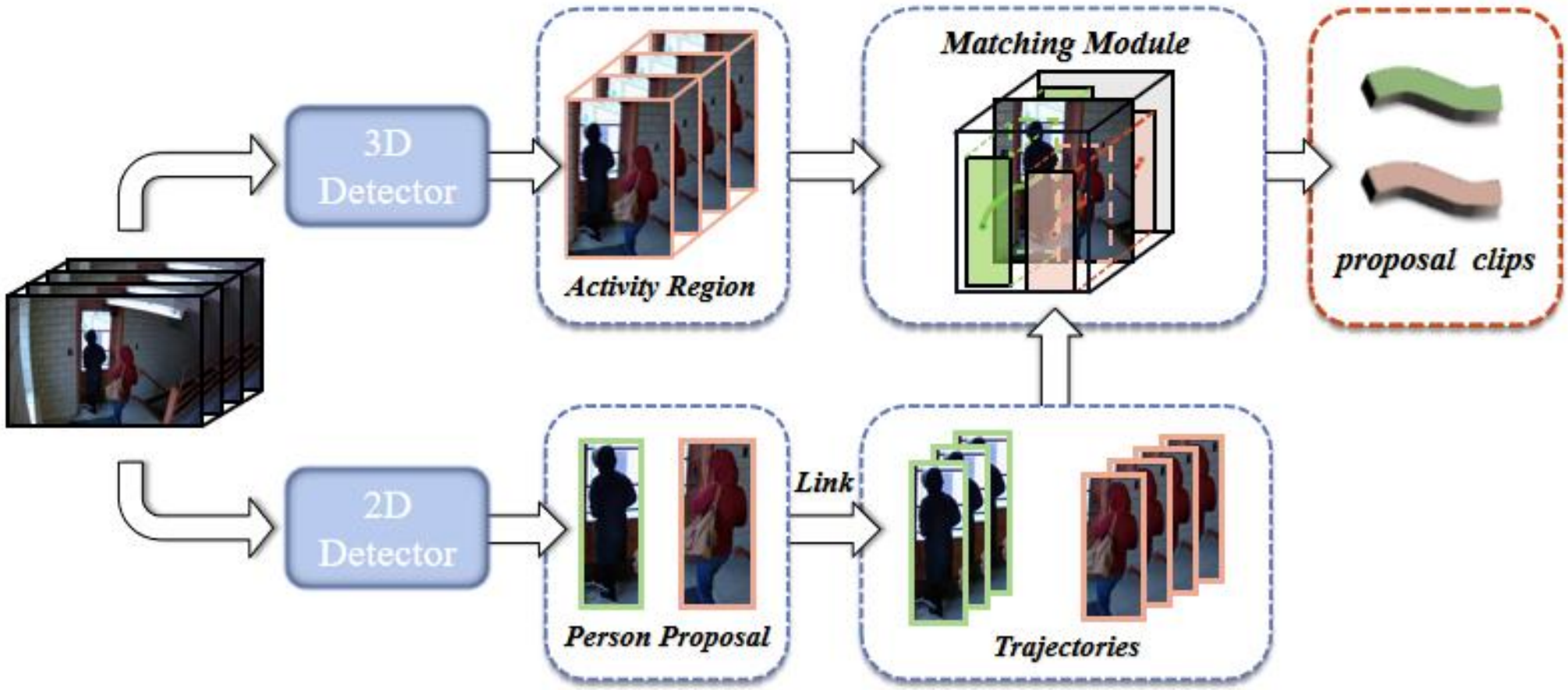
person opens facility door, person enters scene through structure, person exits scene through structure, person talks to person

❖ key issues and solutions

1. Unable to resolve the issue of following into and out of scenes

scene-related activity group

person opens facility door, person enters scene through structure, person exits scene through structure, person talks to person



- 3D detector: Cascade RCNN
- 2D detector: YOLOv8

Results

Results in TRECVID 2023 ActEV Self-Reported Leaderboard Challenge

Team	PMiss
BUPT-MCPRL	0.5781
mlvc_hdc	0.8952
WasedaMeiseiSoftbank	0.9985
FDU_AWS	0.9999
406	1
qwer1	1
hsmw	1

** Results from: https://actev.nist.gov/SRL#tab_leaderboard*

BUPT
MCPRL

THANKS