BUPT MCPRL

BUPT-MCPRL at TRECVID 2022 ActEV SRL Challenge

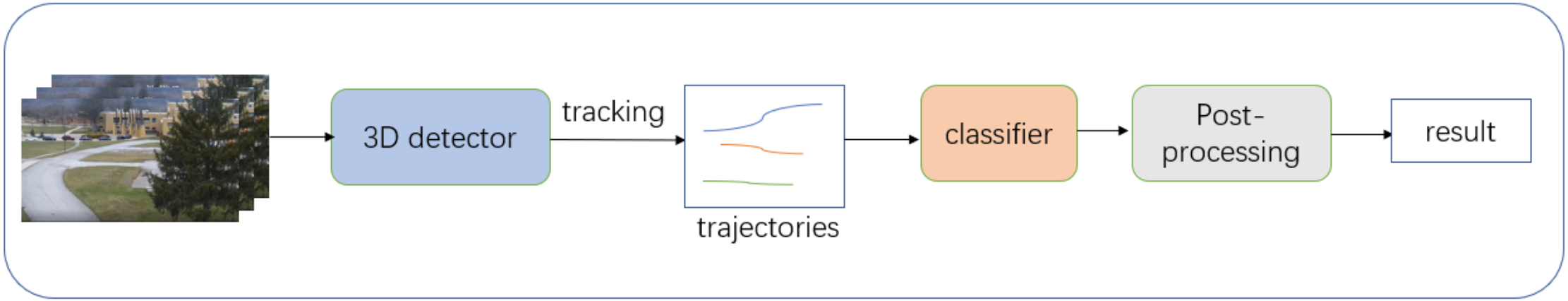
Hangyue Zhao,Zhihang Tong, Yuchao Xiao, Shuhao Qian, Song Li,
Zihan tian, Yanyun Zhao
Beijing University of Posts and Telecommunications, China
{zhaohy21315, tongzh, ycxiao, qiansh, ls0577,Tianzh, zyy}@bupt.edu.cn

北京郵電大學
Beijing University of Posts and Telecommunications

NIST National Institute of Standards and Technology
U.S. Department of Commerce

method

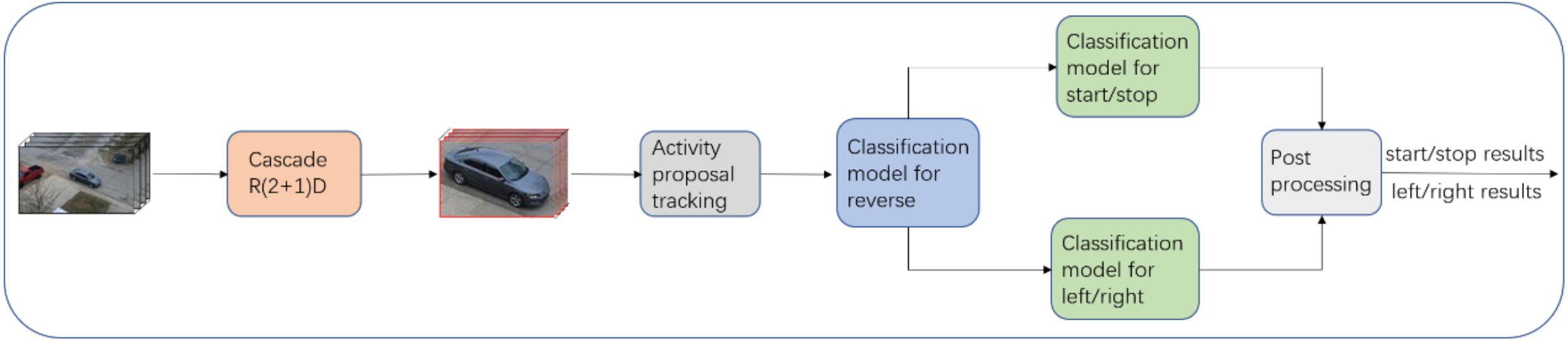
- 3D detectors
- 5 different classification methods



1. vehicle-only: vehicle starts, vehicle stops, vehicle turns left, vehicle turns right;
2. person-object: person picks up, person puts down, person sits down, person stands up, person transfers object;
3. person-specific object: person interacts with laptop, person reads document, person texts on phone;
4. person-vehicle: person exits vehicle, person enters vehicle, person opens vehicle door, person closes vehicle door;
5. scene-related and person-person: person opens facility door, person enters scene through structure, person exits scene through structure, person talks to person.

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

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

** Swin Transformer: https://arxiv.org/abs/2106.13230*

vehicle-only activity detection

❖ key issues and solutions

1. Multi label activities and mutual exclusion activities

Solution: Divide activities into two groups, train the classifiers respectively, and each classification only contains mutually exclusive activities. Convert the problem to single label classification.

2. Interference in reverse segment

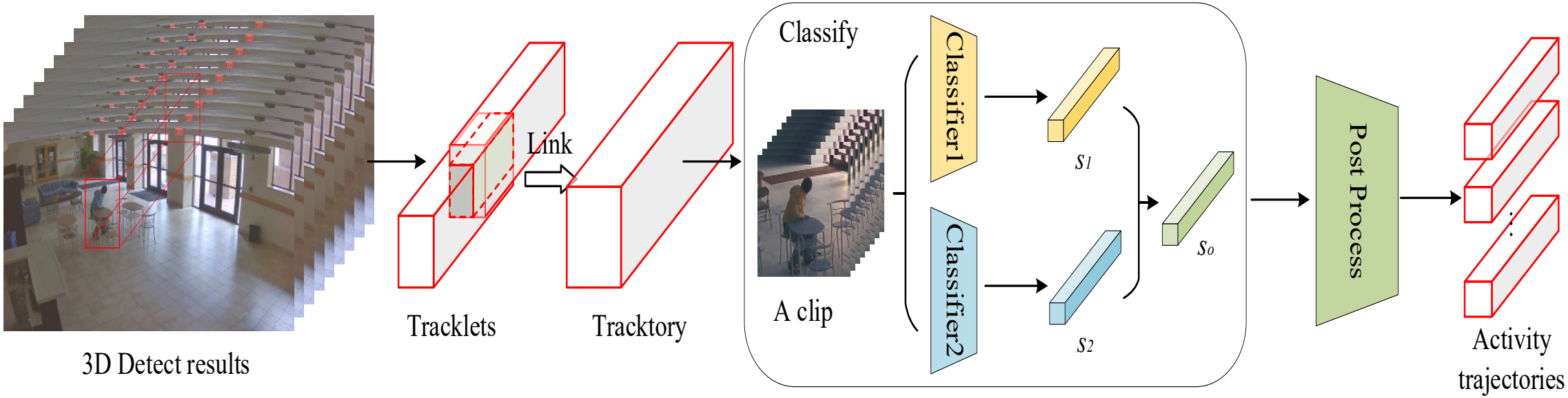
Solution: Filter the reverse segment by reversing classifier.

3. Temporal domain anchor selection

Solution: In inference stage, adopt short temporal anchor for the start and stop, and long temporal anchor for the turn left and right, so as to better accommodate both short and long time activities.

person-object activity group detection

person picks up, person puts down, person sits down, person stands up, person transfers object;



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

** Cascade R-CNN: https://arxiv.org/abs/1712.00726*
** 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 semantic information
Solution: Use Vision-Language-based model to modify the Vision-based model's result.

3. Using detector scores in post-processing
Solution: Construct the fusion formula for detection score and classification score, and make full use of the detection score through careful adjustment.

person-specific object activity group

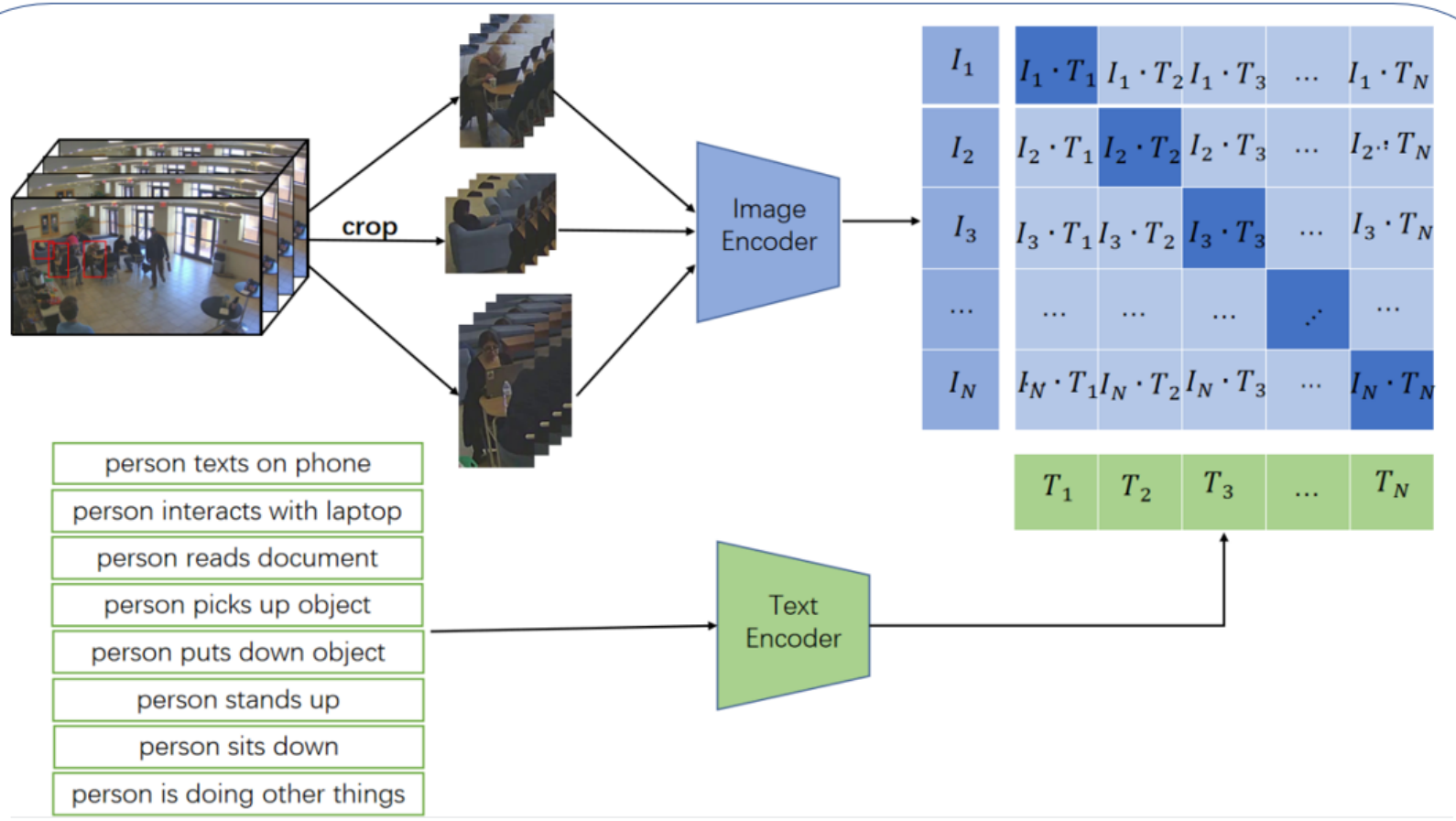
person interacts with laptop, person reads document, person texts on phone;

- 3D detector: Cascade RCNN
- 3D classifier: CLIP

* *CLIP: https://arxiv.org/abs/2103.00020*

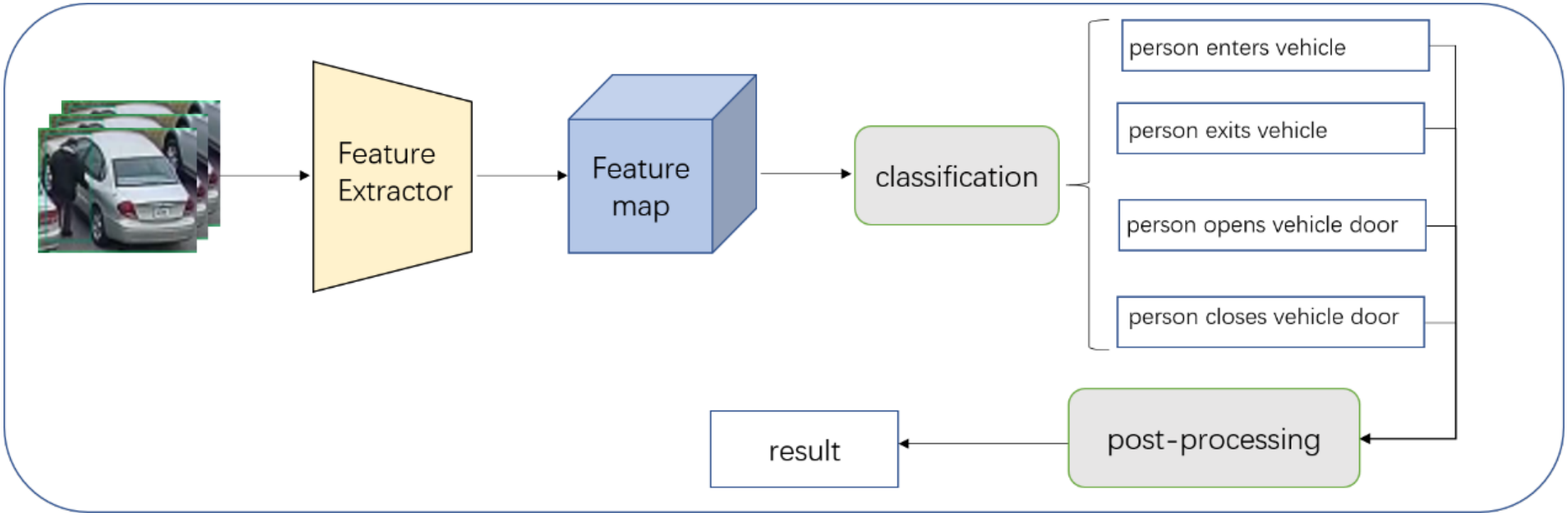
❖ **key issue**

the actor's posture and the interacting objects in single-frame



person-vehicle activity group

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



- 3D detector: Cascade RCNN
- 3D classifier: MViT

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

person-vehicle activity group

❖ key issues and solutions

1. Relatively long time-related activities

Solution: Use sliding windows with 64 frames

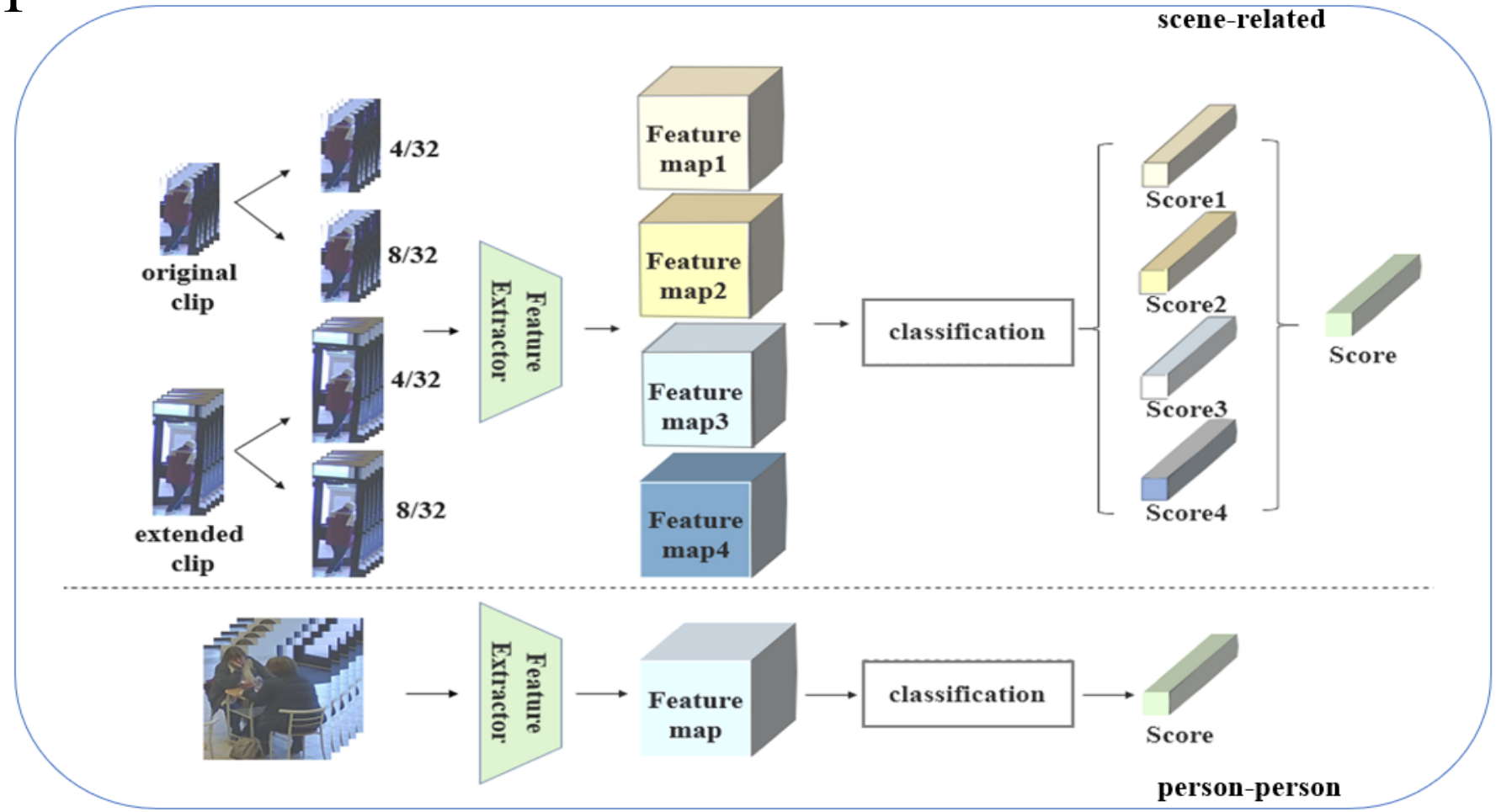
2. Relevance between activities

Solution: Constrain the activity result with the other related activity in post-processing

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
- 3D classifier: MViT



scene-related activity group

❖ key issues and solutions

1. Multi-label activity but single-label annotation
Solution: Fuse the space overlapped and time overlapped activities to make multi-label annotaions.

2. little scene information with person-only bounding box
Solution: Extand the bounding box to rich scene information for classifier training.

3. False Alarms.
Solution: We use frame-difference strategy to decrease the false detection of stationary target and filter too short tracks to mitigate false alarms.

Results

Results in TRECVID 2022 ActEV Self-Reported Leaderboard Challenge

Team	PMiss
BUPT-MCPRL	0.6309
UMD	0.8131
mlvc_hdc	0.9921
WasedaMeiseiSoftbank	0.9961

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

BUPT
MCPRL

THANKS