A proposal-based solution to spatio-temporal action detection in untrimmed videos

Ketul Shah[1], Joshua Gleason[2], Rama Chellappa[1]

[1]Johns Hopkins University

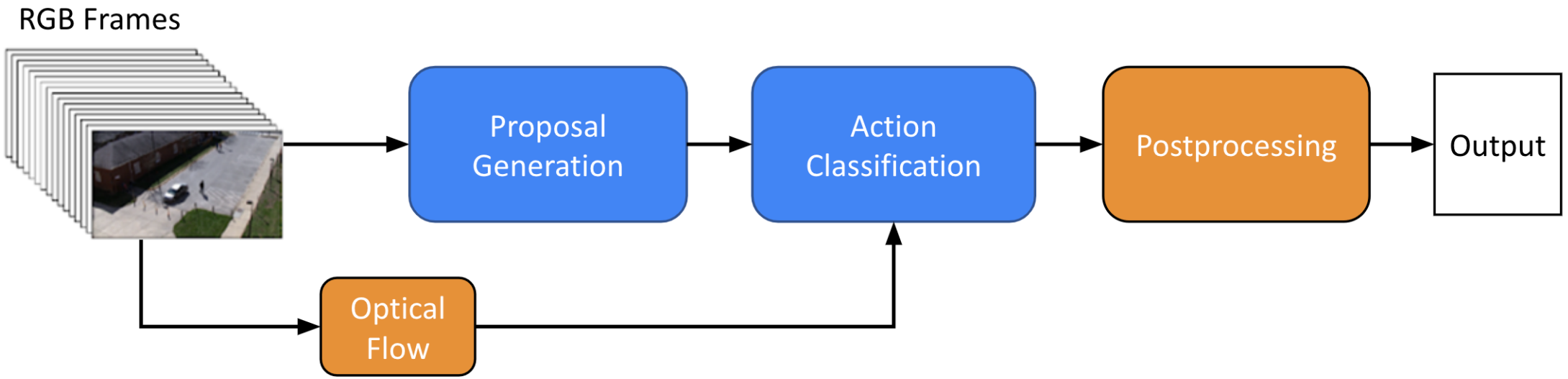
[2]University of Maryland, College Park

Problem Statement

- Spatio-temporal Action Detection

- Challenges
  - Large variations in scale (few pixels to recognize from)
  - Wide range of activity durations (e.g. talking, opening door, person laptop interaction)
  - Indoor and outdoor environments with clutter, occlusion, etc

Overall Pipeline

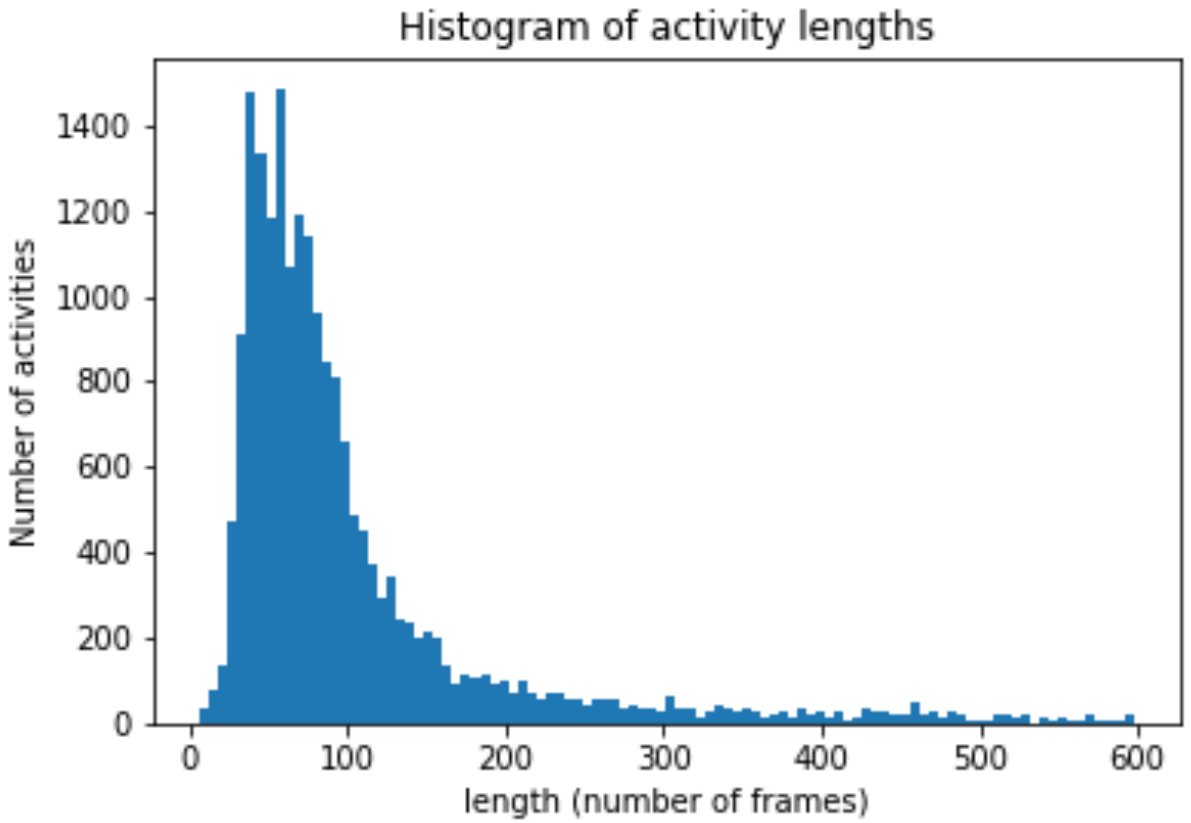


- Two-stage, modular
- Real-time system
- Re-trained and verified by multiple teams

Data & Annotations

- MEVA[1], VIRAT[2]
- 37 activities
- Number of videos: ~ 2230
- Total duration: ~ 7.7 days

Developed an AI assisted annotation tool which can be used for creating dense accurate annotations, quality assurance and fixing incorrect annotations.

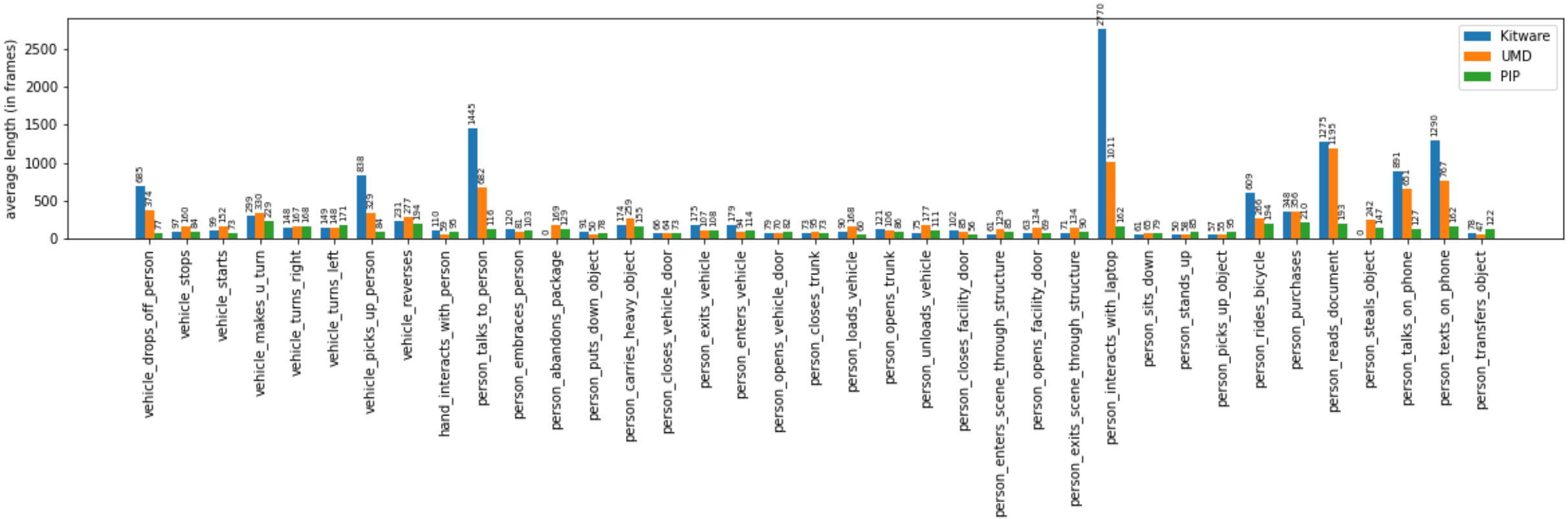


* From Kitware annotations

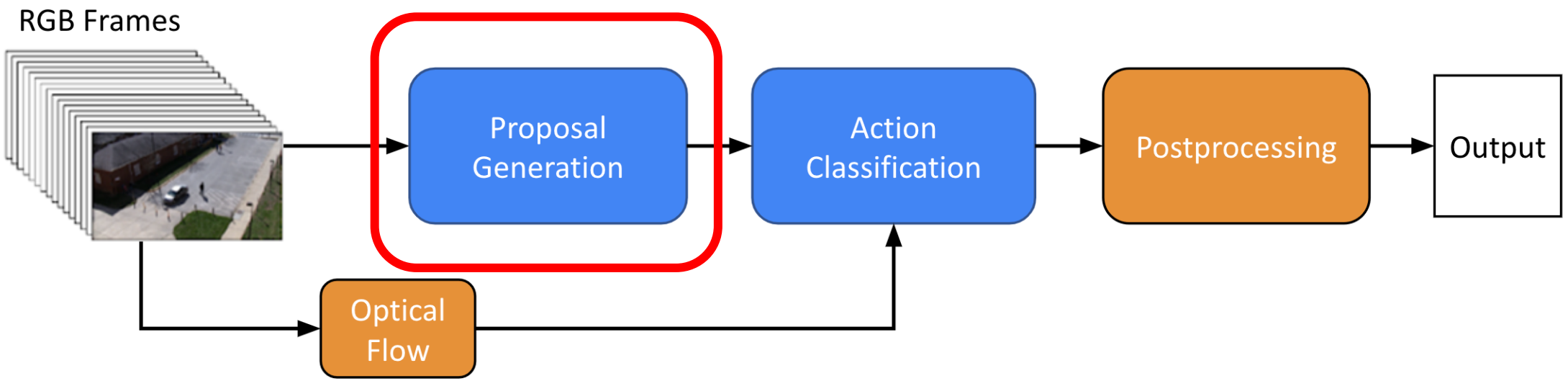
[1]: https://mevadata.org

[2]: Oh, Sangmin, et al. "A large-scale benchmark dataset for event recognition in surveillance video." *CVPR 2011*. IEEE, 2011.

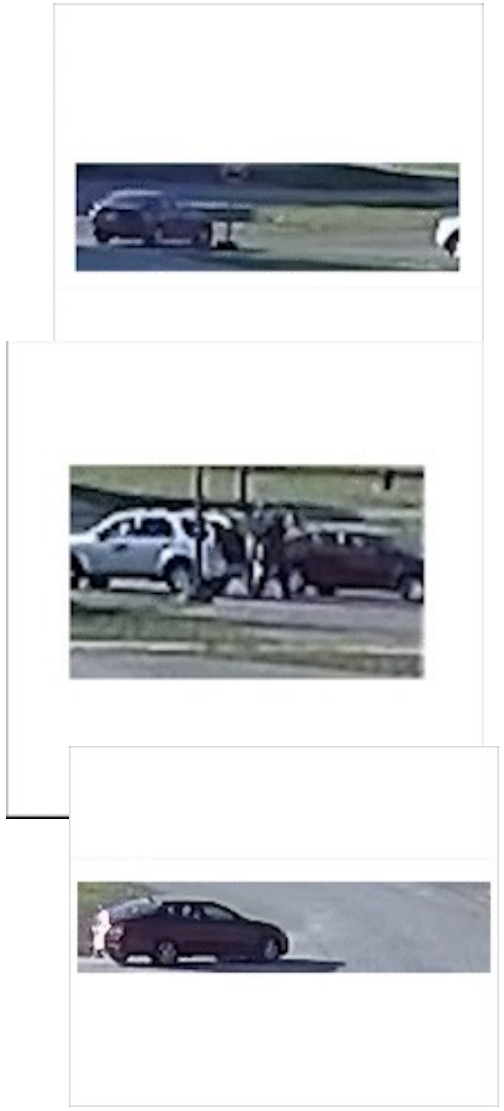
Average duration per activity



Proposal Generation


RGB Frames
Proposal Generation
Action Classification
Postprocessing
Output
Optical Flow

- Proposals = spatio-temporal cuboid of regions in the video where activities are potentially occurring
- (x_{min}, y_{min}, x_{max}, y_{max}, f_{start}, f_{end})

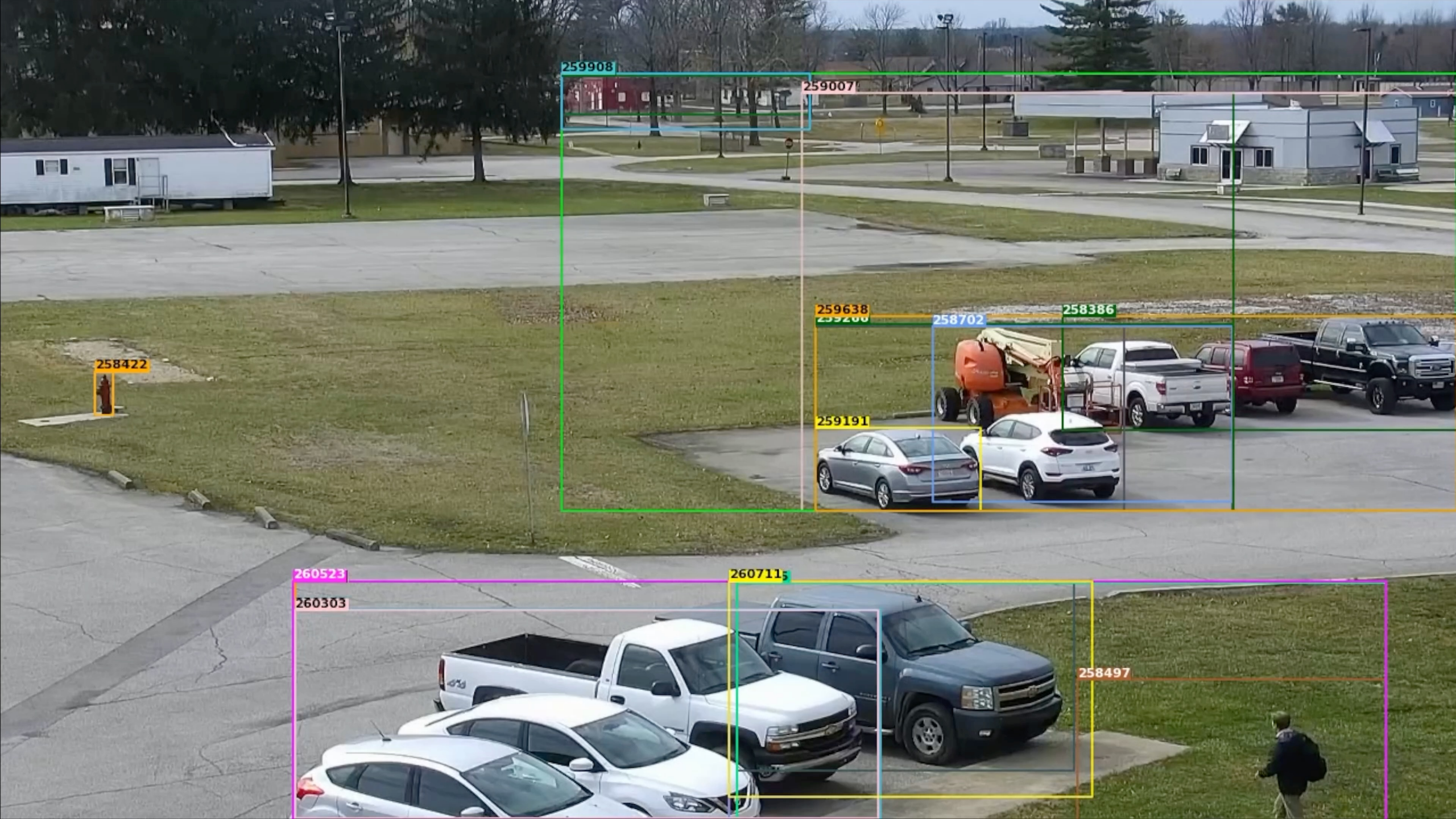


Training-time Proposals using Hierarchical Clustering

- Object Detection using Mask-RCNN[1] on every n-th frame
- Only keep person and vehicle detections
- Represent objects by a 3D feature vector (x, y, f)
  - (x, y) : Center of the object bounding box
  - f : Frame number
- Hierarchical clustering of these 3D features
- Split the resulting linkage tree at various levels to create k clusters
- Generate proposals as the max cuboid of all objects in a cluster
- $(x_{min}, y_{min}, x_{max}, y_{max}, f_{start}, f_{end})$

[1]: He, Kaiming, et al. "Mask R-CNN." *Proceedings of the IEEE international conference on computer vision*. 2017.


person
person
car
car
car
car
car
car
car
car
car
car


259908
259007
259638
258702
258386
258422
259191
260523
260711
260303
258497

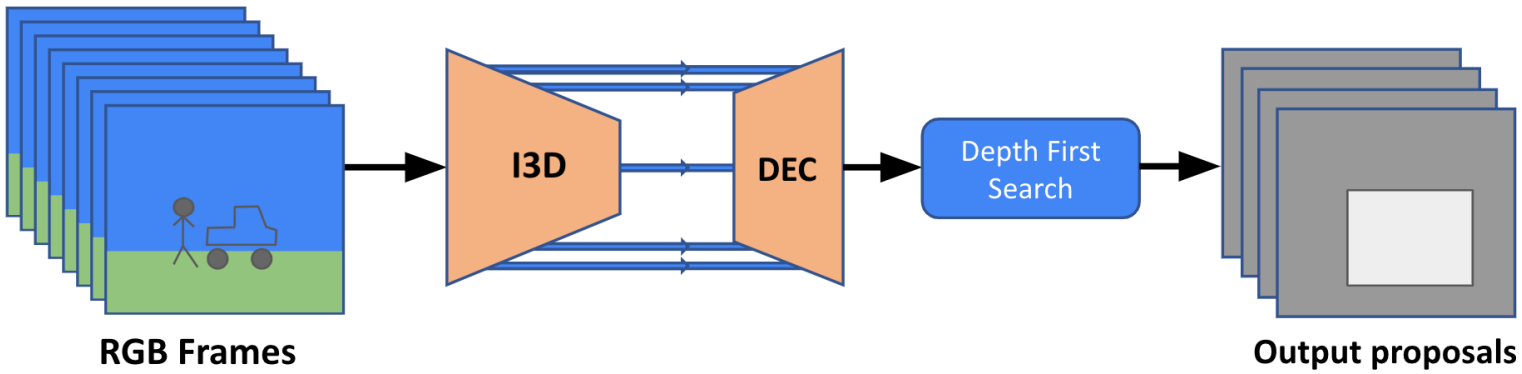
Proposal Jittering and Refinement

- Jitter proposals temporally to obtain dense proposals
  - Higher Recall
  - Data augmentation

- Next, each proposal is labeled as either:
  - non-action class (background): Easy/Hard Negatives
  - action classes (potentially multiple activities)

- Action classes determined based on spatio-temporal IoU overlap with ground truth annotations

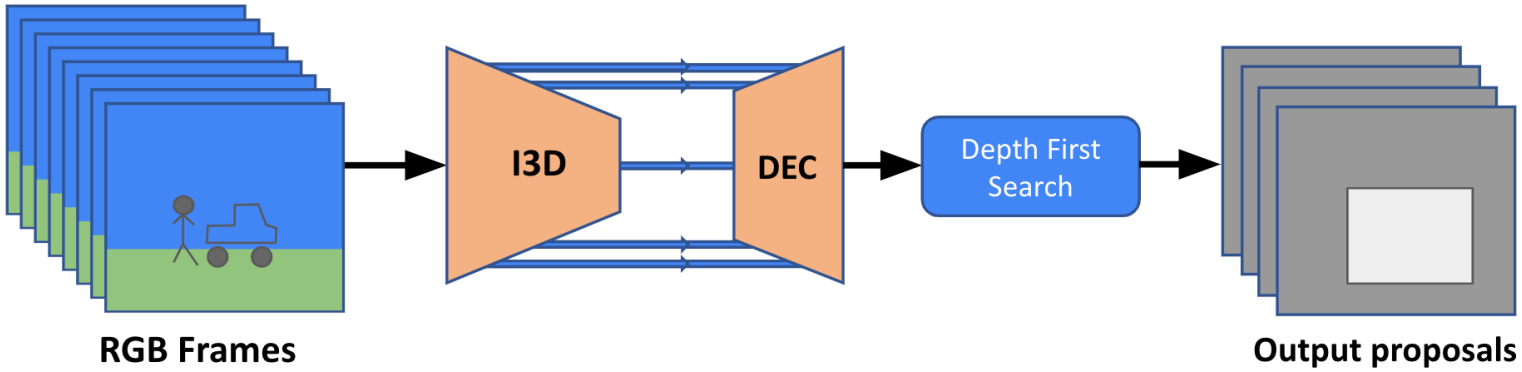
Data-driven proposals during inference



- At testing time, the system uses a data-driven proposal mechanism
- The proposal model uses ideas from 3D semantic segmentation

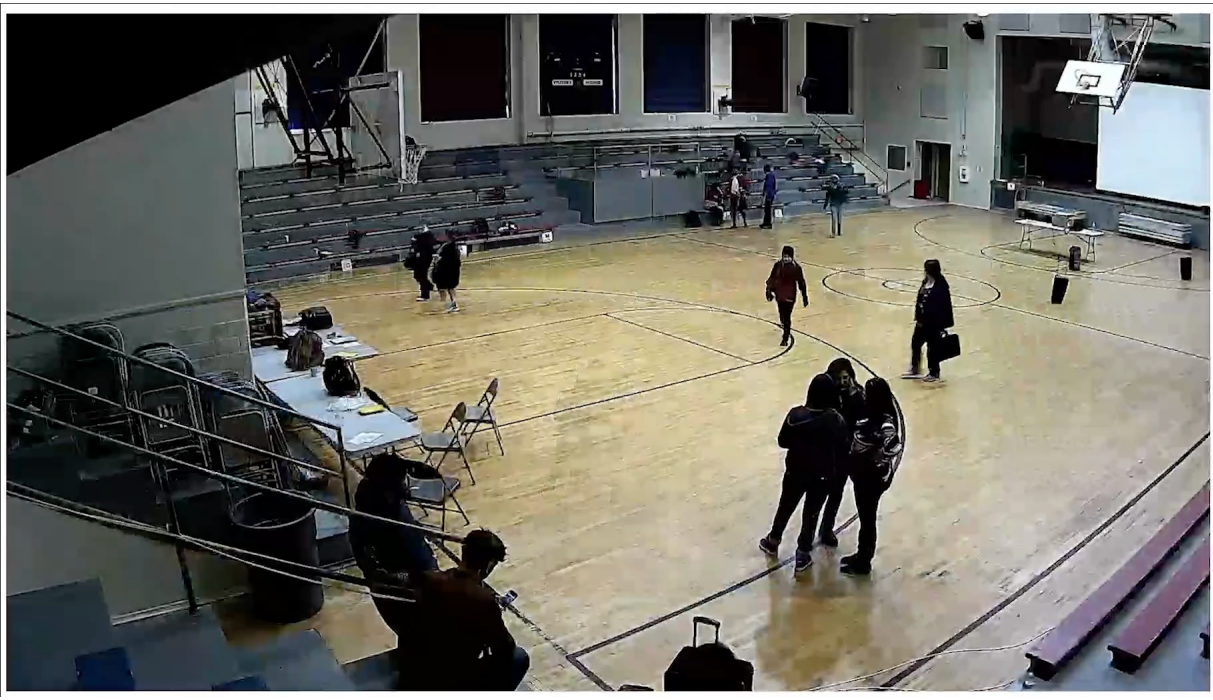
- Given a XYT volume, predict if each voxel is part of an activity

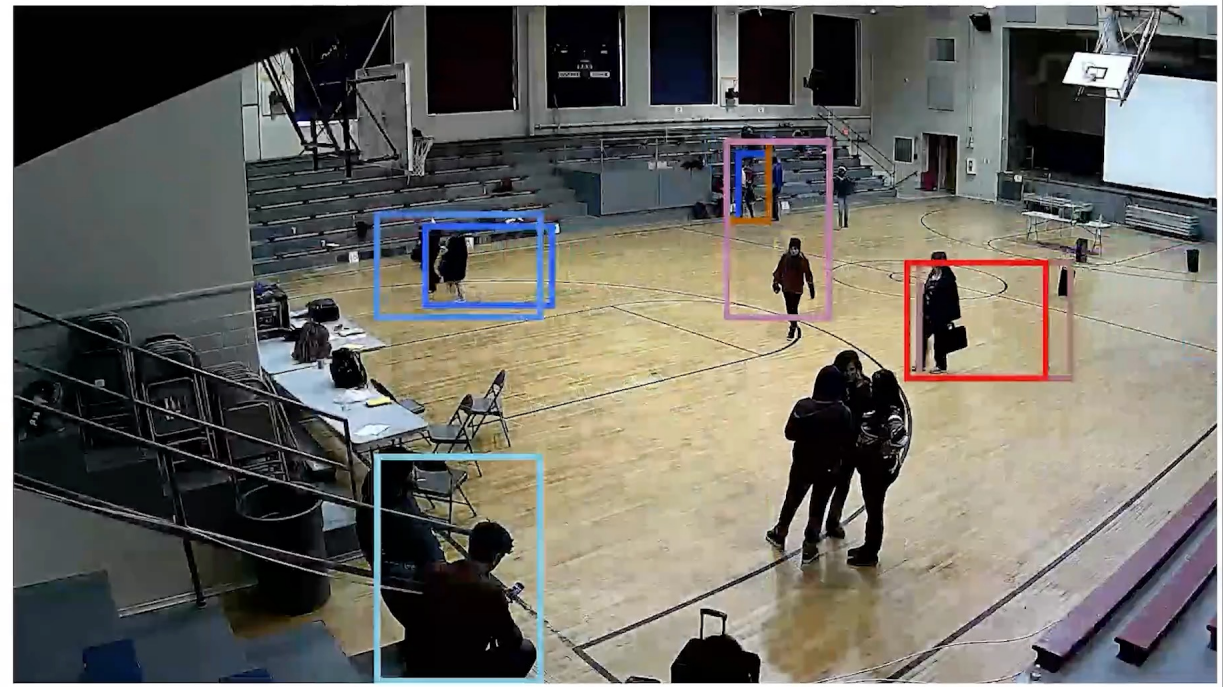
Data-driven proposal network



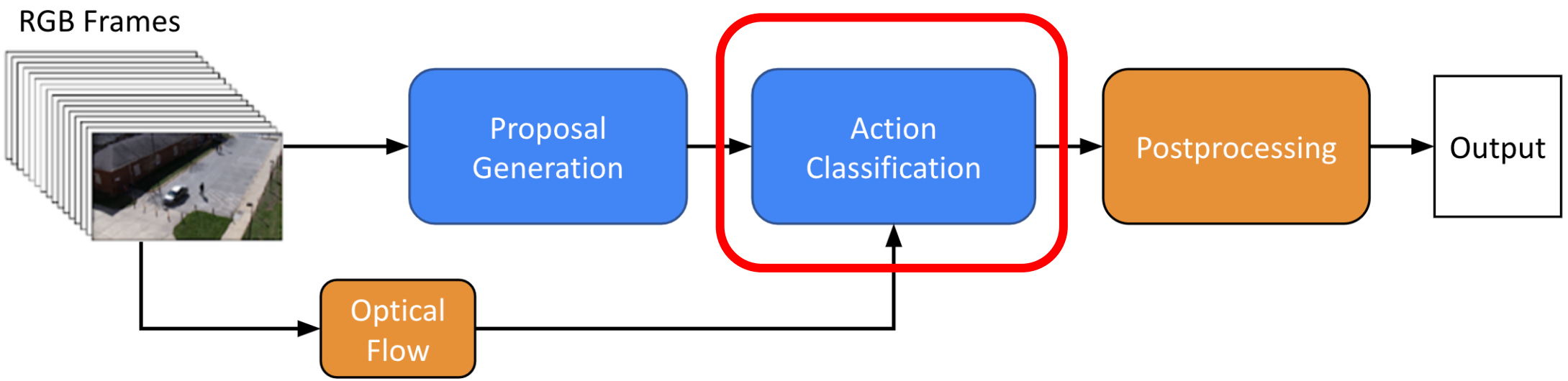
- 3D U-Net architecture using I3D
- Loss: Combination of BCE loss and Tversky Loss [1]
- Trained on a fixed number of strided uncropped frames
- Final proposals are produced by taking axis-aligned bounding boxes of connected components

[1]: S. S. M. Salehi, D. Erdogmus, and A. Gholipour, “Tversky loss function for image segmentation using 3d fully convolutional deep networks,” in International workshop on machine learning in medical imaging. Springer, 2017





Action Classification


RGB Frames
Proposal Generation
Action Classification
Postprocessing
Output
Optical Flow

Model and Input

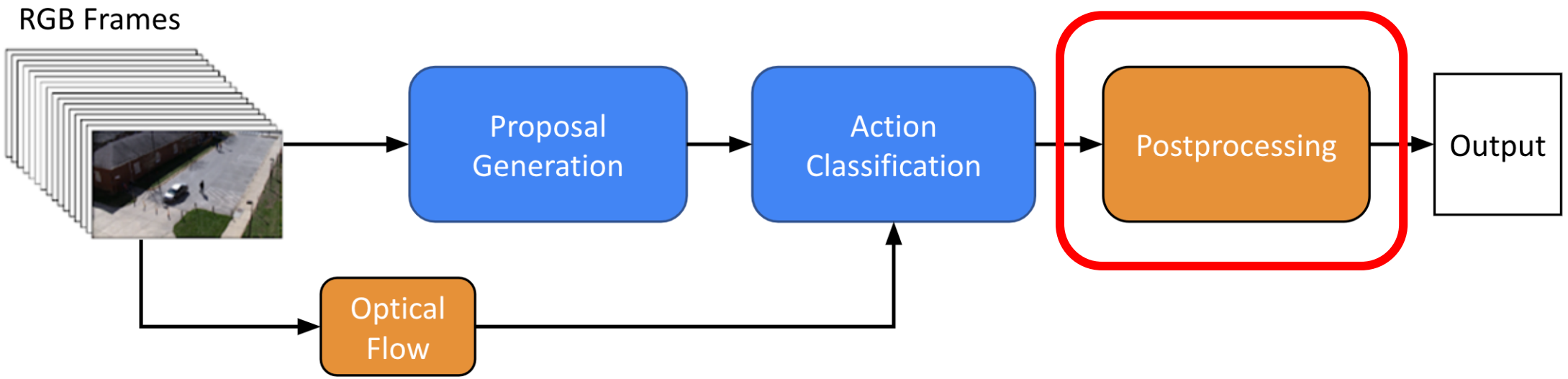
- I3D backbone
- Input modality: Optical Flow
- Input to the network: 64 x 224 x 224
- 64 frames sampled uniformly across temporal span of each proposal
- Videos are resized so that the smallest dim = 256
  - Random 224 x 224 crop during training
  - Horizontal flip (except for vehicle right/left/u turn)

Loss function

- Multi-label Classification (BCE Loss)
- Each proposal gets multiple labels (all overlapping GT activities)

$$\ell(\hat{y}, y) = -\frac{1}{C}\sum_{i=1}^{C}\left(y_i \log \frac{1}{1+\exp(-\hat{y}_i)} + (1-y_i)\log \frac{\exp(-\hat{y}_i)}{1+\exp(-\hat{y}_i)}\right) \qquad \hat{y} \in \mathbb{R}^C, y \in \{0,1\}^C$$

Post-processing


RGB Frames
Proposal Generation
Action Classification
Postprocessing
Output
Optical Flow

Post-processing

- Threshold for each action class
- 3D NMS
- Camera conditional Filtering
- Object conditional Filtering

Threshold + 3D-NMS

- For each proposal, we get a probability value of presence of each action class
- We set a low threshold to remove noisy predictions

- Our proposal generation method creates many highly-overlapping action proposals, many of which belonging to the same class
- Apply 3D-NMS to prune overlapping cuboids
  - Applied to each class separately

Camera Conditional Filtering

- We filter resulting predictions in additional post-processing based on the location of the camera, i.e. indoor vs outdoor
- If the camera is located indoors, we suppress all vehicle activities.
  - This could fail in certain cases, e.g. indoor parking lots
- Camera location is available at inference time in the provided metadata
- To be more flexible, we also perform object conditional filtering on predictions for each proposal

Object Conditional Filtering

- We filter predictions during post-processing based on consistency with object detections
- The set of activities is split into person-only, vehicle-only and person-vehicle activities
- Based on all the objects detected within the cuboid, we filter activity predictions by ensuring the following:
  - Person-only activities: have at-least one person detection
  - Vehicle-only activities: have at-least one vehicle detection
  - Person-Vehicle activities: have at least one person and vehicle detection

Results


Talking: 0.89

ActEV SRL Leaderboard

Rank	Team Name	Submission ID	Submission Date	System Name	AOD mean PMiss @0.1rfa	AOD mean nMODE @0.1rfa	AOD mean nAUDC @0.2rfa	AD mean PMiss @0.1rfa	AD mean nAUDC @0.2rfa
1	BUPT-MCPRL	27305	2022-11-02	MCPRL_S0	0.6309	0.0538	0.6705	0.5805	0.6231
2	UMD	27264	2022-06-16	UMD-JHU	0.8131	0.1620	0.8300	0.7789	0.7995
3	mlvc_hdu	27288	2022-10-28	mlvc_hdu_baseline	0.9921	0.0303	0.9922	0.9728	0.9732
4	WasedaMeiseiSoftbank	27279	2022-10-24	WasedaMeiseiSoftbank_P	0.9961	0.1080	0.9964	0.9829	0.9850
5	TokyoTech_AIST	27309	2022-11-23	p-merge	0.9965	0.1827	0.9961	0.9824	0.9830
6	M4D_team	27268	2022-10-18	baseline				0.9823	0.9819

Thank you!