Spatio-Temporal Action Detection in Untrimmed Videos

Rajeev Ranjan, Joshua Gleason, Steve Schwarcz, Carlos D. Castillo, Jun-Cheng Chen, Rama Chellappa

University of Maryland College Park

11/14/2018

Outline

- Introduction
- A Proposal-Based Solution to Spatio-Temporal Action Detection
- Experimental Results
- Conclusion

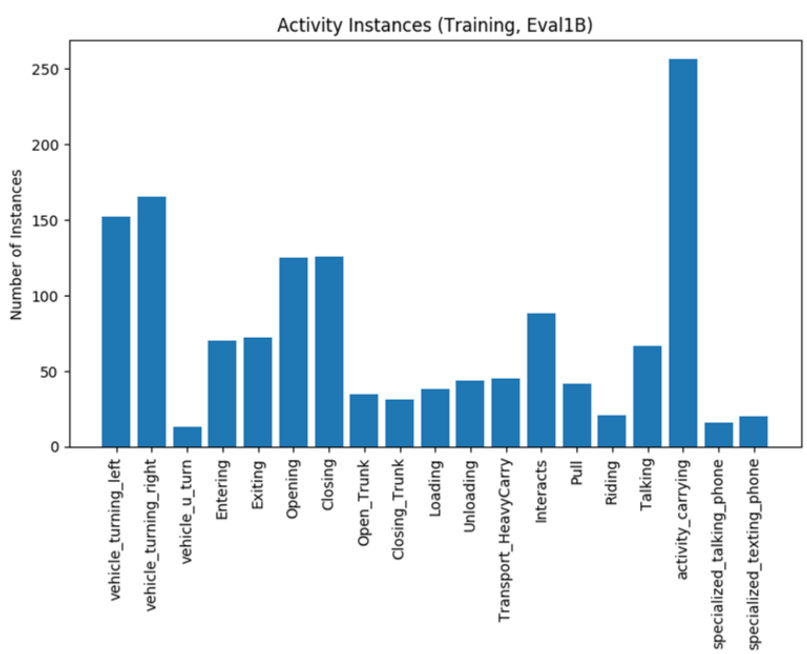
Challenges of DIVA - Sparsity

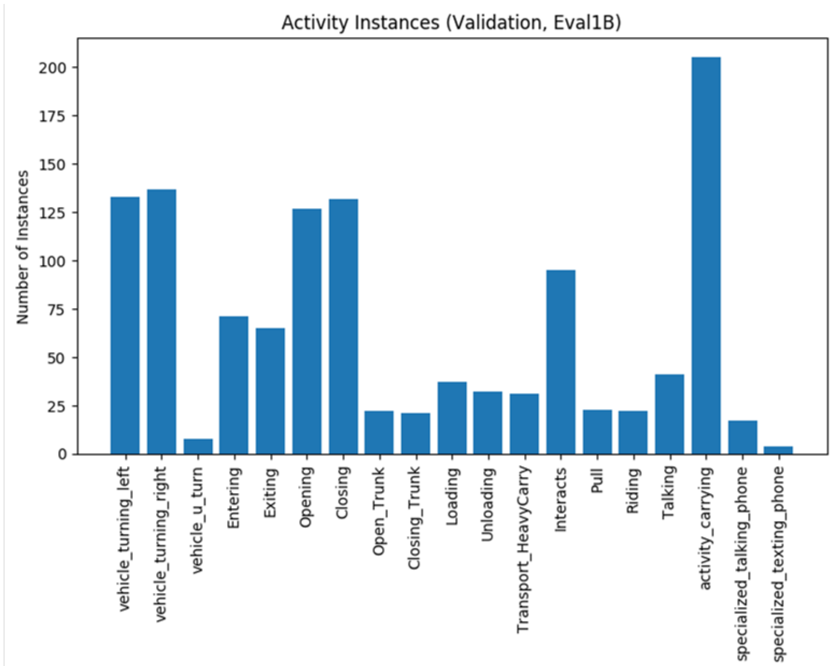
- DIVA actions are very small
  - The average activity is 150x300 resolution
  - Every video in ActEV dataset is either 1920x1080 or 1200x720
  - Most pixels in any given scene have no actions.

Spatial Sparsity Example

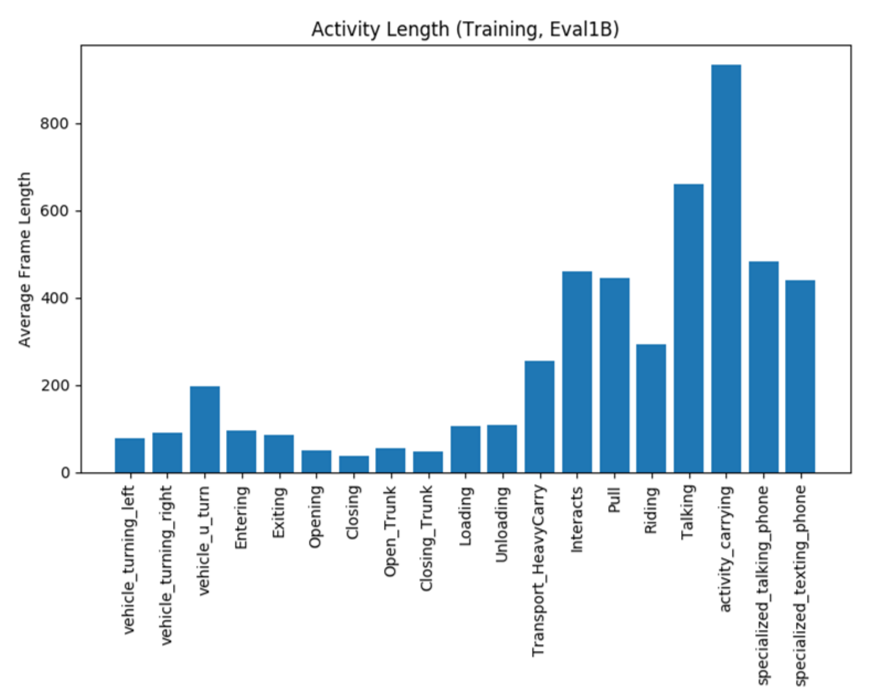
Figure 3. On the left, the DIVA action `Closing` makes up only a small portion of the image, and the surrounding context has no value for the action classification task. The THUMOS action `Cricket` on the right is much larger in the image, and the entire image's context is useful for classification.

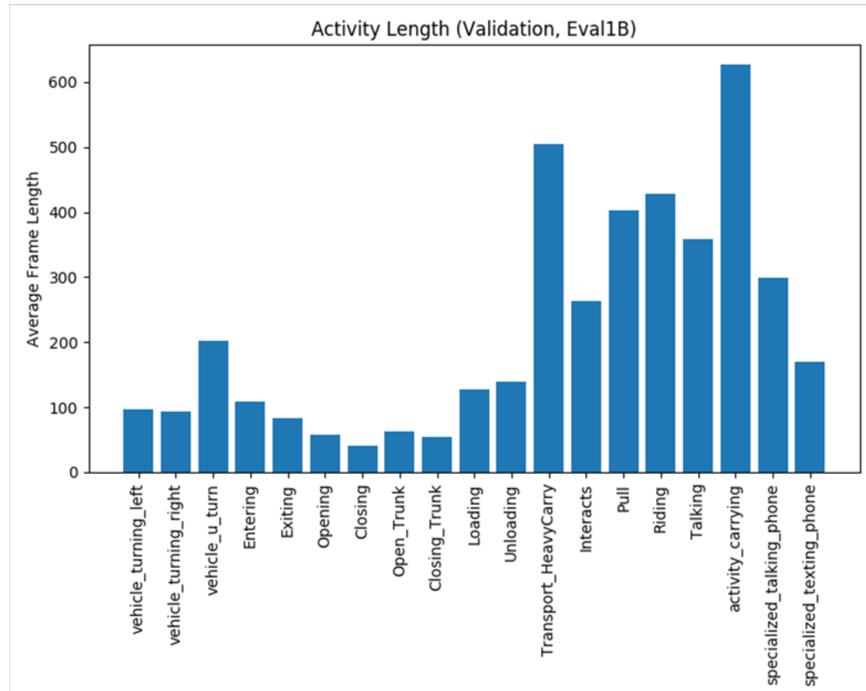
Challenges of DIVA - Limited Data





Challenges of DIVA - Variable Length Actions





Addressing Challenges - Sparsity

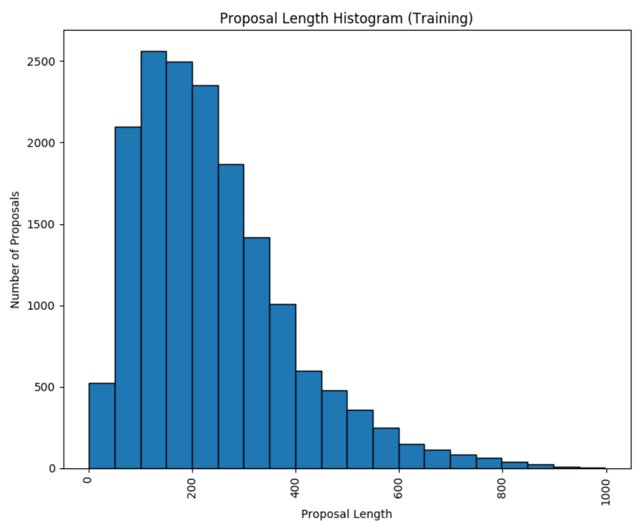
- Proposal based approach
  - Proposals are generated where people/vehicles are detected
  - Run classification on small sub-section of frame
  - Addresses sparsity by targeting where we look
  - Proposals can tightly bound regions of interest spatially
- Focus on High Recall
  - As long as proposals overlap a little, they can be refined later

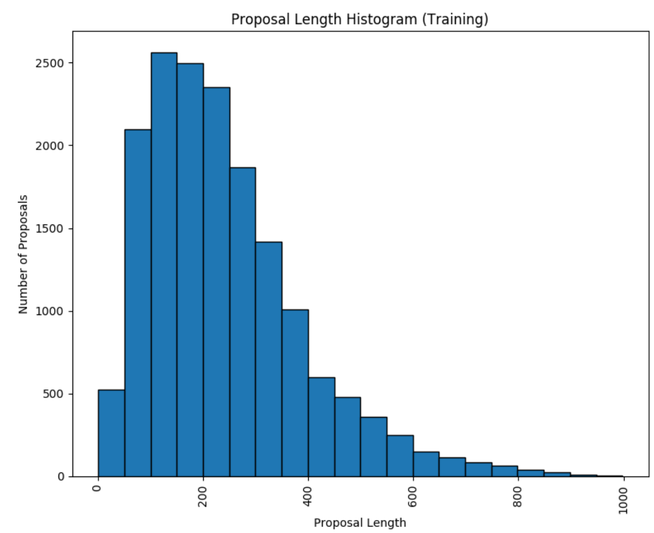
Addressing Challenges - Limited Data

- Utilize pre-trained classifier (I3D)
  - Trained on Kinetics-400 dataset (300k videos, 400 actions)
- Trained on proposals
  - Significantly more proposals than actions
  - Acts as implicit data-augmentation

Addressing Challenges - Variable Length Actions

- Proposals may have vastly different spans

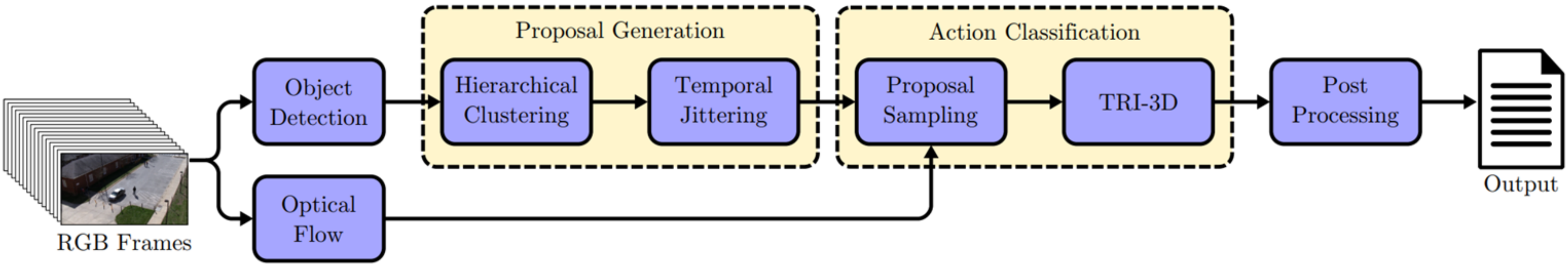




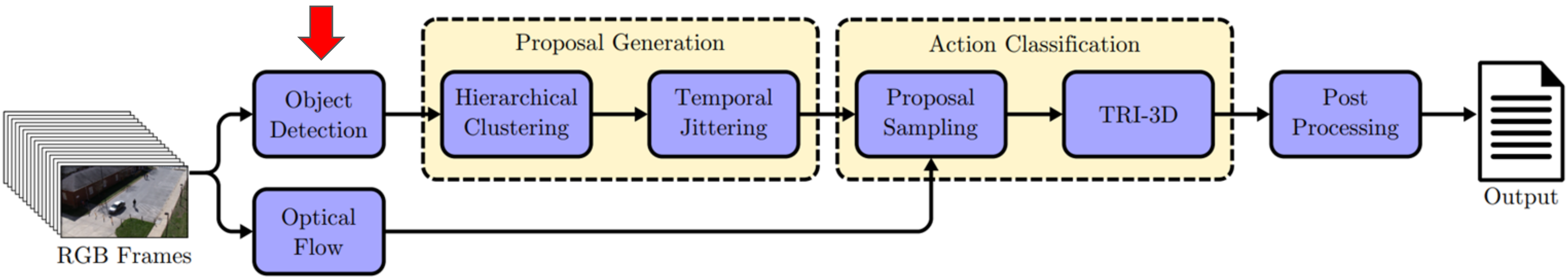
- Actions can often be accurately classified using a subset of frames
- Our solution is to classify using fixed number of frames from each proposal

System Overview

- Modular system design
  - Modules may be improved independently
  - Easily extendible pipeline

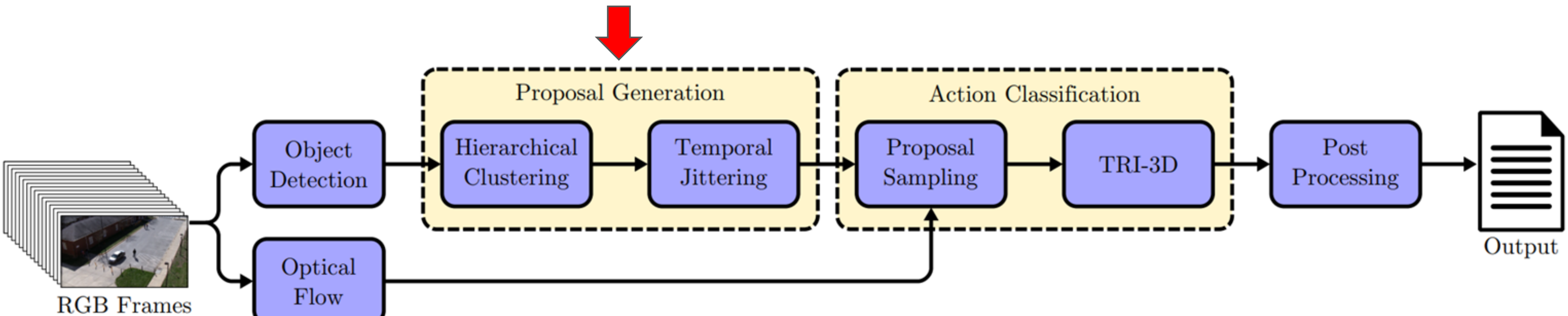


Object Detection



- Mask R-CNN
  - Trained on COCO
  - Accurate detection of humans and vehicles at different scales

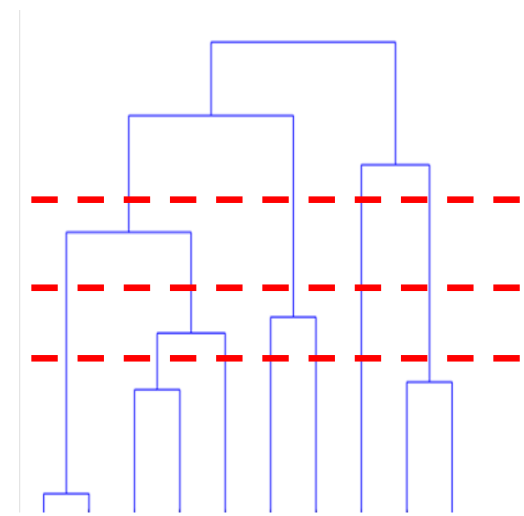
Proposal Generation



- Generate high-recall proposals
- Two step process
  - Cluster detections into proposal cuboids
  - Generate extra proposals via temporal jittering

Proposal Generation - Hierarchical Clustering

- Hierarchical Clustering for Proposal Generation
  a. For each detection let (x,y) be the center and f be the frame number
  b. Perform Divisive Hierarchical Clustering* on 3-d features (x,y,f)
  c. Dynamically split linkage tree at various levels to create k clusters
  d. Define cuboid from resulting clusters $(x_{min}, y_{min}, x_{max}, y_{max}, f_{st}, f_{end})$

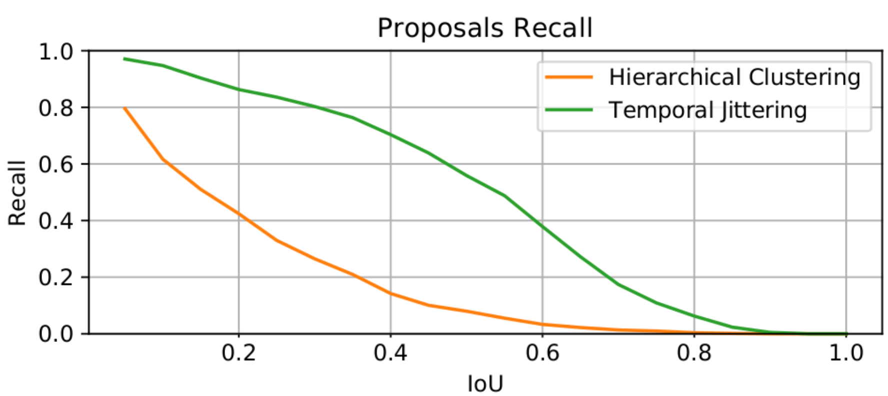


- Statistics on DIVA 1.A. validation
  - Approximately 250 proposals per video
  - Recall 42% at spatio-temporal IoU of 0.2

* Müllner, Daniel. "Modern hierarchical, agglomerative clustering algorithms." *arXiv preprint arXiv:1109.2378* (2011).

Proposal Generation - Temporal Jittering

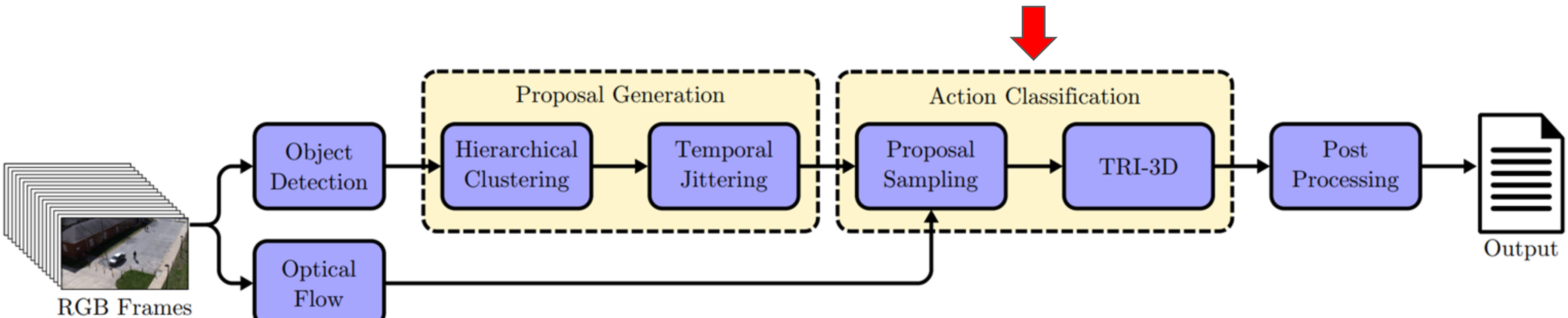
- Jittering to improve recall
  - Generate temporally jittered cuboids from each proposal
- Recall improvements after jittering
  - 42% → 86% at IoU of 0.2



Algorithm 1 Dense Proposal Generation

```
detections ← Mask − RCNN(video)
orig_proposals ← hierarchical_clustering(detections)
new_proposals ← orig_proposals
s ← 15
for proposal in orig_proposals do
    x0, y0, x1, y1 ← spatial_bounds(proposal)
    f_st, f_end ← temporal_bounds(proposal)
    for f from f_st to f_end step s do
        new_proposals.add(f − 16, f + 16)
        new_proposals.add(f − 32, f + 32)
        new_proposals.add(f − 64, f + 64)
        new_proposals.add(f − 128, f + 128)
final_dense_proposals ← new_proposals
```

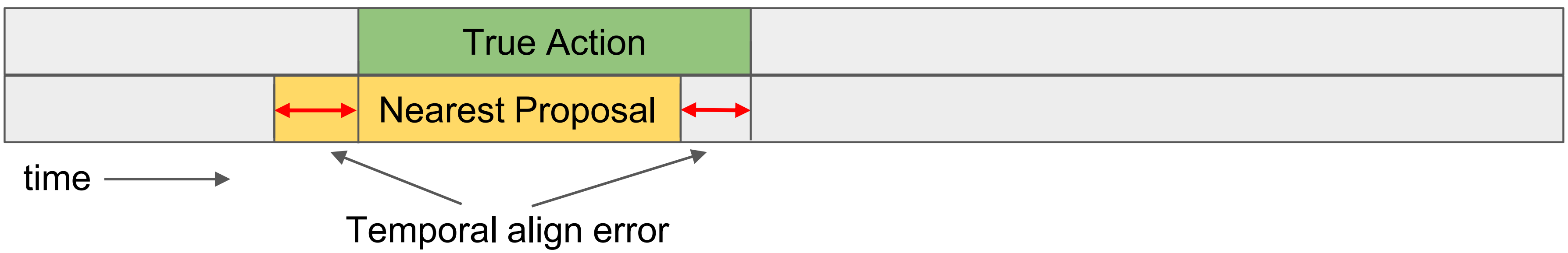
Action Classification



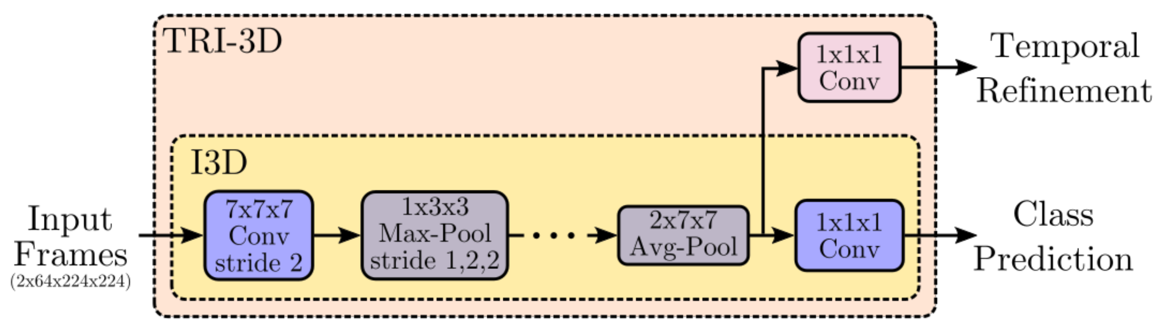
- Action Classification
  - Improves temporal localization of proposals
  - Rejects False Proposals
  - Classifies Valid Proposals

Temporal Refinement I3D (TRI-3D)

- Proposal temporal alignment to ground truth is imprecise

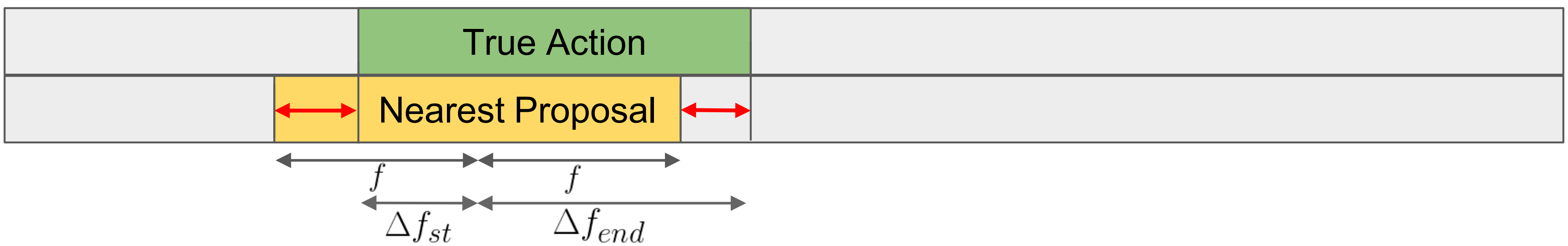


- TRI-3D network adds temporal refinement module



TRI-3D - Temporal Refinement

- Label proposal with extra temporal refinement



- Estimate how much adjustment is needed
  - Temporal Refinement labels $r_{st}, r_{end} = \left(-\frac{\Delta f_{st}}{f}, \frac{\Delta f_{end}}{f}\right)$

TRI-3D - Input Pre-processing

- Proposal Cuboids expanded to have 1-1 spatial aspect ratio
  - Padding improved results. Likely due to extra contextual information.
- Optical flow input
  - Each optical flow frame captures fast motions
- Uniformly sample 64 frames from cuboid
  - TRI-3D CNN infers high level action from multiple simultaneous frames

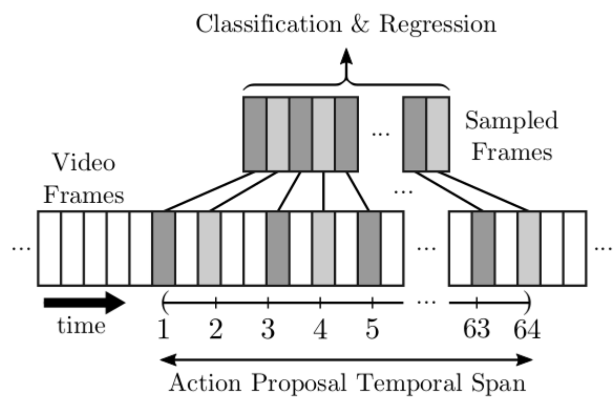


Figure. Uniform sampling of frames

Input Mode	Accuracy
RGB+Flow	0.704
RGB	0.585
Opt. Flow	**0.716**

Table. Preliminary Experiments on RGB vs optical flow by classifying ground truth validation proposals

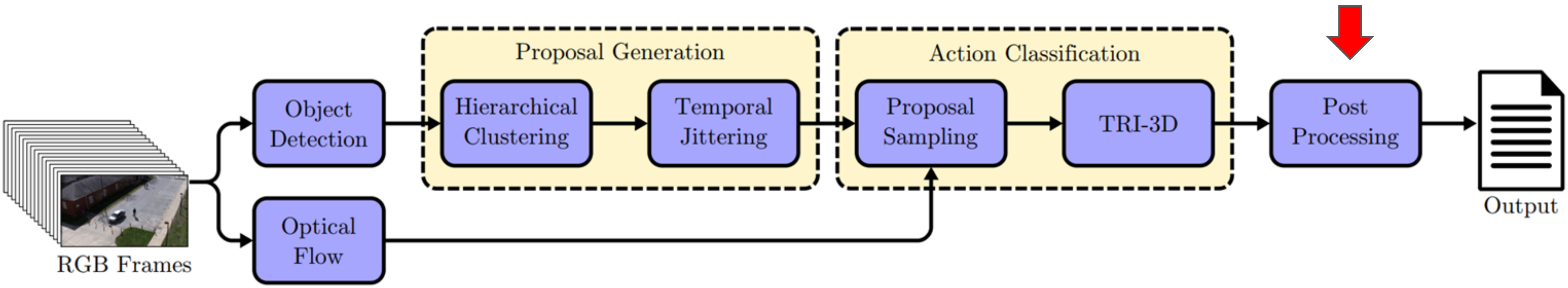
TRI-3D - Rejecting Negative Proposals

- Proposals with insufficient overlap with real action should be discarded
- Add an extra “negative” label during training
- Consider two types of negative proposals
  - Easy: Little to no overlap with true activity
  - Hard: Some overlap with true activity
- Strongly favor hard negatives during training
  - Makes classifier more robust (less false positives)

Designation	Count
Positive	**12,752**
Easy Negative	**9,525**
Hard Negative	**13,574**
Total Used	**35,851**

Table 1. Number of proposals used for training TRI-3D network.

Post Processing



- Spatio-temporal non-maximum suppression
- Select AODT objects

Post Processing - Non-maximum suppression

- Due to overlap in proposals a single action may have many overlaps
    a. Perform per-class non-maximum suppression on remaining proposal cuboids
- Selecting AOD(T) Objects
    a. Generate tracks for object detections through multi-target Kalman-filtering trackers
    b. Gather tracks with sufficient overlap with proposal cuboid
    c. Clip tracks to cuboid length
    d. Reject tracks that don’t make sense, e.g.
        - Stationary vehicles and people for turning actions
        - Vehicles in person only actions
    e. Remaining tracks make up AOD/AODT results

THUMOS’14 Results

- With minimal modification, our system outperforms many recently published results on the THUMOS’14 action dataset
- Two observations
  - @ 0.5 tIoU our system outperforms all but SoTA
  - The DIVA baseline algorithm (Xu et al.) is comparable to our system on THUMOS’14. However, we significantly outperform it on DIVA. This further emphasizes how much DIVA differs from other common action detection datasets.

	tIoU	0.1	0.2	0.3	0.4	0.5	0.6	0.7
	Karaman *et al.* [24]	4.6	3.4	2.4	1.4	0.9	-	-
	Oneata *et al.* [36]	36.6	33.6	27.0	20.8	14.4	-	-
	Wang *et al.* [45]	18.2	17.0	14.0	11.7	8.3	-	-
	Caba *et al.* [4]	-	-	-	-	13.5	-	-
	Richard *et al.* [38]	39.7	35.7	30.0	23.2	15.2	-	-
	Shou *et al.* [42]	47.7	43.5	36.3	28.7	19.0	10.3	5.3
	Yeung *et al.* [48]	48.9	44.0	36.0	26.4	17.1	-	-
	Yuan *et al.* [49]	51.4	42.6	33.6	26.1	18.8	-	-
	Escorcia *et al.* [9]	-	-	-	-	13.9	-	-
2017	Buch *et al.* [3]	-	-	37.8	-	23.0	-	-
	Shou *et al.* [40]	-	-	40.1	29.4	23.3	13.1	7.9
	Yuan *et al.* [50]	51.0	45.2	36.5	27.8	17.8	-	-
	Buch *et al.* [2]	-	-	45.7	-	29.2	-	9.6
	Gao *et al.* [12]	60.1	56.7	50.1	41.3	31.0	19.1	9.9
	Hou *et al.* [18]	51.3	-	43.7	-	22.0	-	-
	Dai *et al.* [8]	-	-	-	33.3	25.6	15.9	9.0
	Gao *et al.* [13]	54.0	50.9	44.1	34.9	25.6	-	-
	Xu *et al.* [46]	54.5	51.5	44.8	35.6	28.9	-	-
	Zhao *et al.* [53]	**60.3**	56.2	50.6	40.8	29.1	-	-
2018	Huang *et al.* [19]	-	-	-	-	27.7	-	-
	Yang *et al.* [47]	-	-	44.1	37.1	28.2	20.6	12.7
	Chao *et al.* [6]	59.8	**57.1**	53.2	**48.5**	**42.8**	**33.8**	**20.8**
	Nguyen *et al.* [32]	52.0	44.7	35.5	25.8	16.9	9.9	4.3
	Alwassel *et al.* [1]	-	-	51.8	42.4	30.8	20.2	11.1
	Gao *et al.* [11]	-	-	-	-	29.9	-	-
	Lin *et al.* [27]	-	-	**53.5**	45.0	36.9	28.4	20.0
	Shou *et al.* [41]	-	-	35.8	29.0	21.2	13.4	5.8
	Ours	52.1	51.4	49.7	46.1	37.4	26.2	15.2

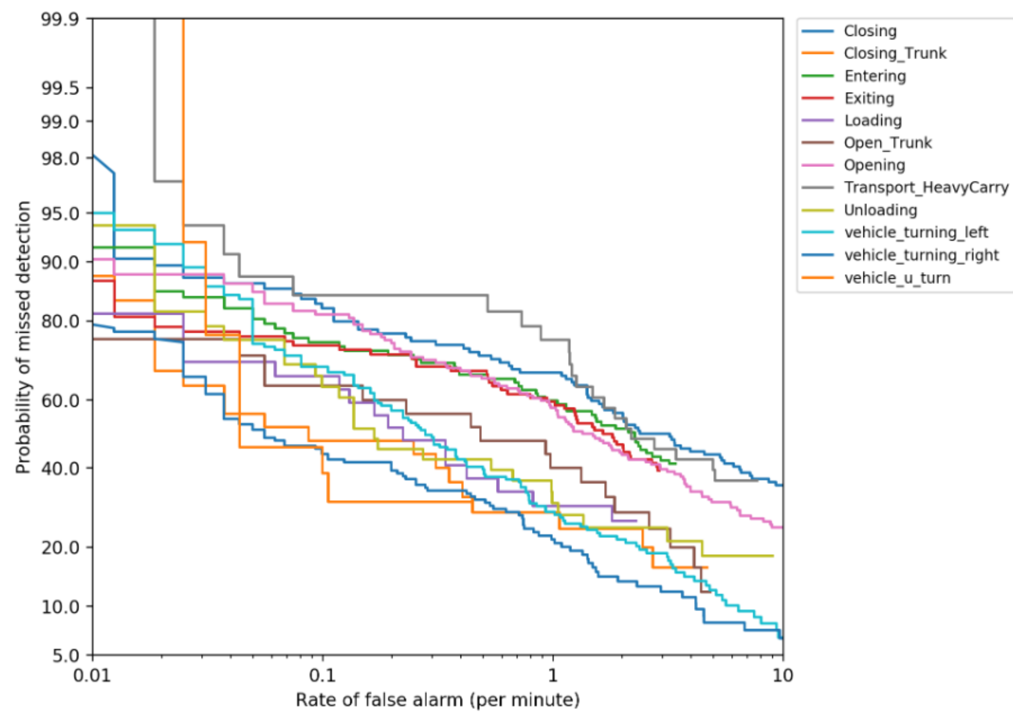
Table 4. Comparison to THUMOS’14 performers on the mAP metric at various temporal IoUs. Missing entries indicate that results are not available. We note that Xu *et al.* [46] is the same system used to compute the DIVA V1 baseline; see Table 3. Bold indicates best performance.

Results - DIVA Test 1.A. (AD)

Measure	**Value**
mean p_miss @ 0.15 rfa	0.6181246
mean p_miss @ 1 rfa	0.4405567
mean n_mide @ 0.15 rfa	0.2162213
mean n_mide @ 1 rfa	0.2231658

Results - DIVA Test 1.A (AD per class)

activity	p_missAtRfa.15	p_missAtRfa1	n_mideAtRfa.15	n_mideAtRfa1
vehicle_turning_right	0.4160000	0.2160000	0.1189398	0.1406151
Closing_Trunk	0.4800000	0.2800000	0.1952250	0.2005133
vehicle_turning_left	0.6330935	0.2805755	0.1041266	0.1455894
Loading	0.5925926	0.2962963	0.2320325	0.2185177
Unloading	0.5151515	0.3030303	0.1473277	0.1523250
vehicle_u_turn	0.3076923	0.3076923	0.1343847	0.1343847
Open_Trunk	0.6000000	0.4000000	0.1877447	0.2123635
Opening	0.7777778	0.5763889	0.2990197	0.2796126
Exiting	0.7340426	0.5957447	0.2144886	0.2502759
Entering	0.7319588	0.5979381	0.2417627	0.2779568
Closing	0.7807018	0.6754386	0.2838833	0.2275646
Transport_HeavyCarry	0.8484848	0.7575758	0.4357200	0.4382713

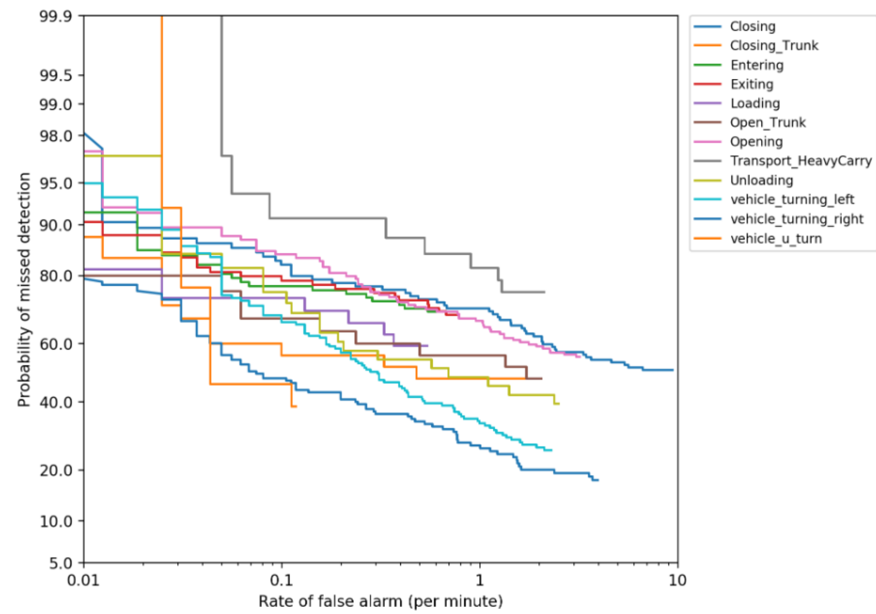


Results - DIVA Test 1.A (AOD)

Measure	**Value**
mean p_miss @ 0.15 rfa	0.6801261
mean p_miss @ 1 rfa	0.5576526
mean n_mide @ 0.15 rfa	0.2083421
mean n_mide @ 1 rfa	0.2198618
mean object p_miss @ 0.5 rfa	0.3063430

Results - DIVA Test 1.A (AOD per class)

activity	p_missAtRfa.15_AOD	p_missAtRfa1_AOD	n_mideAtRfa.15_AOD	n_mideAtRfa1_AOD	object_p_missAtRfa.50_AOD
vehicle_turning_right	0.4320000	0.2640000	0.1224045	0.1467868	0.0972335
vehicle_turning_left	0.6258993	0.3309353	0.1104631	0.1457217	0.0708223
vehicle_u_turn	0.3846154	0.3846154	0.0956272	0.0956272	0.0047506
Closing_Trunk	0.5600000	0.4800000	0.2053459	0.2281162	0.2839879
Unloading	0.6969697	0.4848485	0.0919419	0.1504894	0.7281739
Open_Trunk	0.6800000	0.5600000	0.1754733	0.1903678	0.2280419
Loading	0.7037037	0.5925926	0.2424638	0.2416688	0.3603478
Opening	0.8402778	0.6736111	0.3131711	0.2719526	0.3773797
Exiting	0.7765957	0.6914894	0.2240591	0.2680818	0.4643600
Entering	0.7628866	0.7010309	0.2242854	0.2373795	0.4214567
Closing	0.7894737	0.7105263	0.2750829	0.2148685	0.3389735
Transport_HeavyCarry	0.9090909	0.8181818	0.4197867	0.4472813	0.3005877

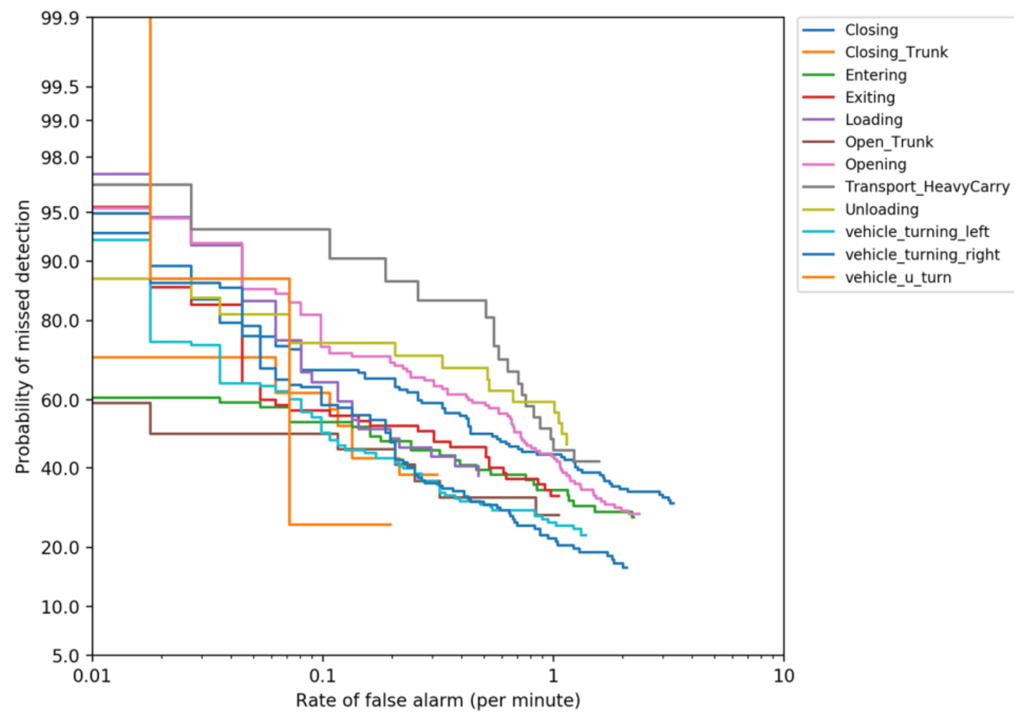


Results - DIVA Validation 1.A (AD)

Measure	**Value**
mean p_miss @ 0.15 rfa	0.5630079
mean p_miss @ 1 rfa	0.3613007
mean n_mide @ 0.15 rfa	0.2091128
mean n_mide @ 1 rfa	0.2279841

Results - DIVA Validation 1.A (AD per class)

activity	p_miss@0.15rfa	p_miss@1rfa	n-mide@0.15rfa	n-mide@0.1rfa
vehicle_turning_right	0.554745	0.218978	0.112873	0.127510
vehicle_u_turn	0.250000	0.250000	0.156234	0.156234
vehicle_turning_left	0.451128	0.255639	0.170556	0.175456
Open_Trunk	0.454545	0.272727	0.193195	0.233416
Exiting	0.538462	0.323077	0.206481	0.214736
Entering	0.521127	0.338028	0.193284	0.249608
Loading	0.513514	0.378378	0.264423	0.280165
Closing_Trunk	0.428571	0.380952	0.162767	0.155613
Opening	0.716535	0.433071	0.205391	0.235101
Closing	0.674242	0.439394	0.174487	0.176459
Transport_HeavyCarry	0.903226	0.451613	0.407549	0.476384
Unloading	0.750000	0.593750	0.262114	0.255128

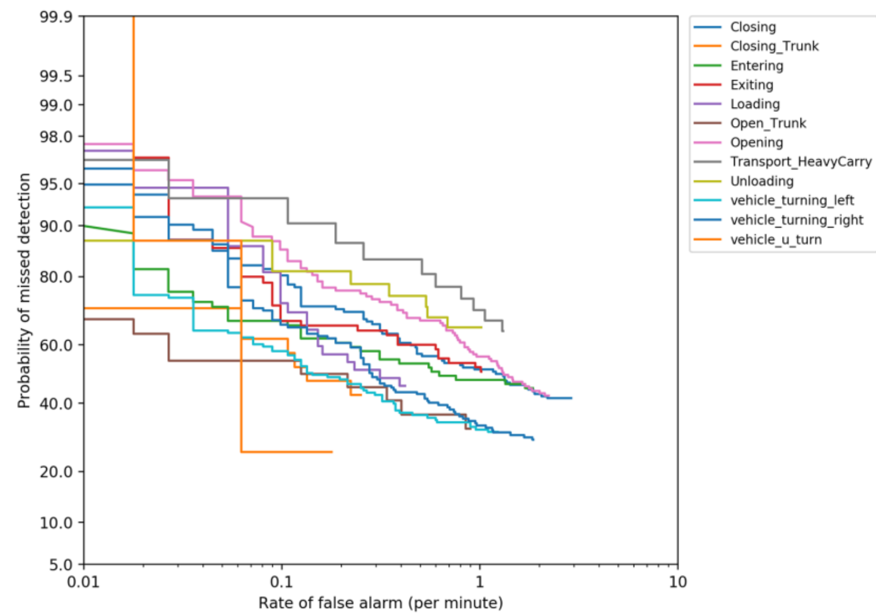


Results - DIVA Validation 1.A (AOD)

Measure	**Value**
mean p_miss @ 0.15 rfa	0.6271621
mean p_miss @ 1 rfa	0.4618795
mean n_mide @ 0.15 rfa	0.1994476
mean n_mide @ 1 rfa	0.2225540
mean object p_miss @ 0.5 rfa	0.2442836

Results - DIVA Validation 1.A (AOD per class)

activity	p_miss@0.15rfa	p_miss@1rfa	n-mide@0.15rfa	n-mide@0.1rfa	object-p_miss@0.5rfa
vehicle_u_turn	0.250000	0.250000	0.156234	0.156234	0.000000
vehicle_turning_left	0.496241	0.315789	0.167330	0.174516	0.057533
Open_Trunk	0.500000	0.318182	0.195864	0.252256	0.156534
vehicle_turning_right	0.635036	0.328467	0.107419	0.126918	0.063412
Closing_Trunk	0.476190	0.428571	0.173147	0.164532	0.184372
Loading	0.648649	0.459459	0.245856	0.272109	1.000000
Entering	0.619718	0.478873	0.153037	0.254387	0.301736
Closing	0.719697	0.515152	0.177939	0.188579	0.156172
Exiting	0.661538	0.523077	0.151852	0.169056	0.166017
Opening	0.803150	0.559055	0.227259	0.248732	0.072314
Unloading	0.812500	0.656250	0.229887	0.236279	0.244898
Transport_HeavyCarry	0.903226	0.709677	0.407549	0.427050	0.528414



Conclusion

- The dense proposals help increase the recall significantly.
- The proposed TRI-3D can effectively refine the temporal boundaries of the proposals.
- The modular design of the proposed system allows easy integration of better components.