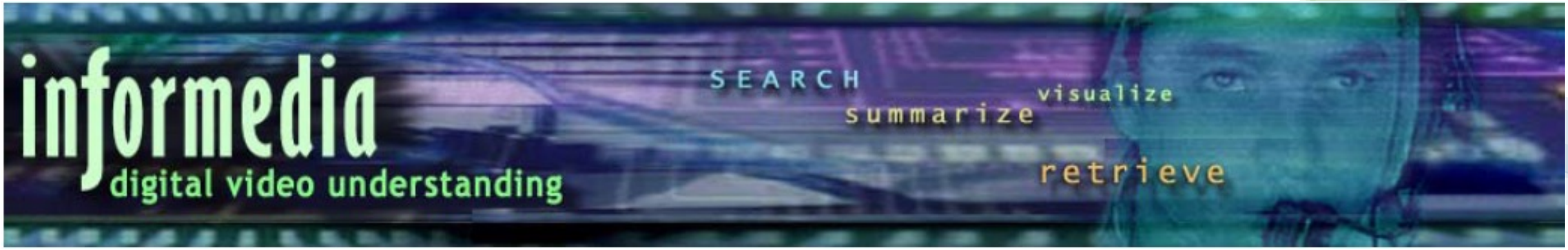


Argus++: Robust Real-time Activity Detection for Unconstrained Video Streams with Overlapping Cube Proposals

Lijun Yu, Yijun Qian, Wenhe Liu
and Alexander G. Hauptmann

Carnegie Mellon University
Language Technologies Institute

DIVA

Overview

Activity detection

- In ***unconstrained*** videos:
  untrimmed and with large field-of-views
- Three aspects
  - Temporal localization
  - Spatial localization
  - Action classification

Strict target

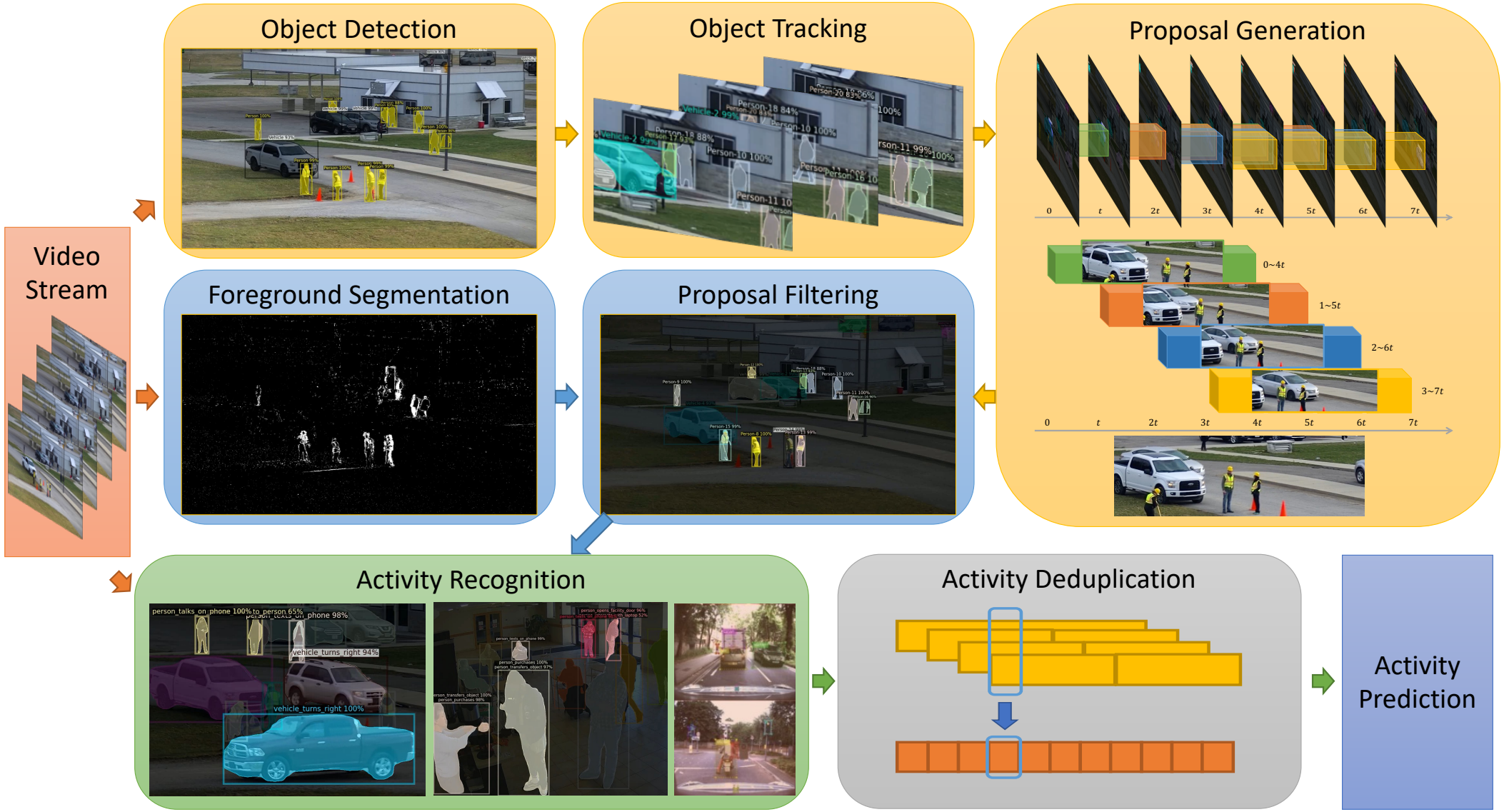
- Detect all atomic activities
- Bipartite match between predictions and ground truths

Loosened target

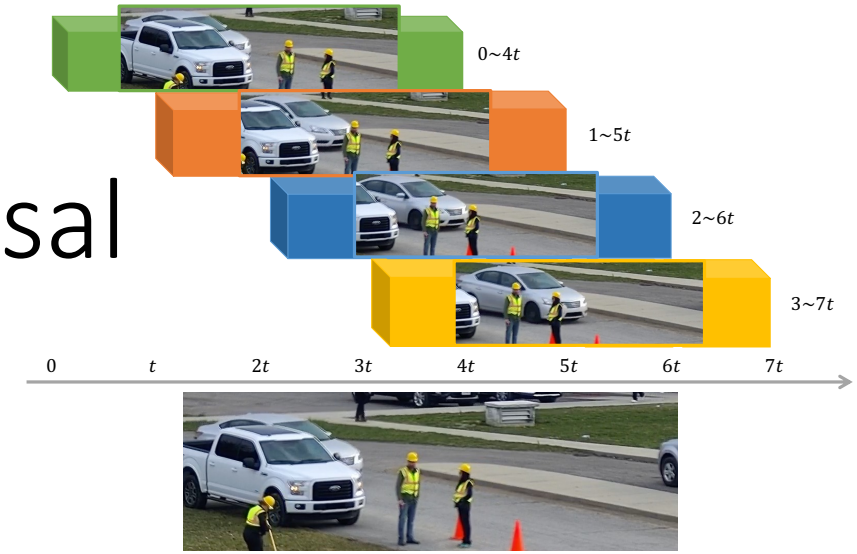
- Detect either atomic activities (e.g., standing up) or continuous repetitive activities (e.g., walking)
- Match multiple non-overlapping predictions to each ground truth

Argus++ Framework

Argus++ Architecture



Intermediate Concept: Cube Proposal



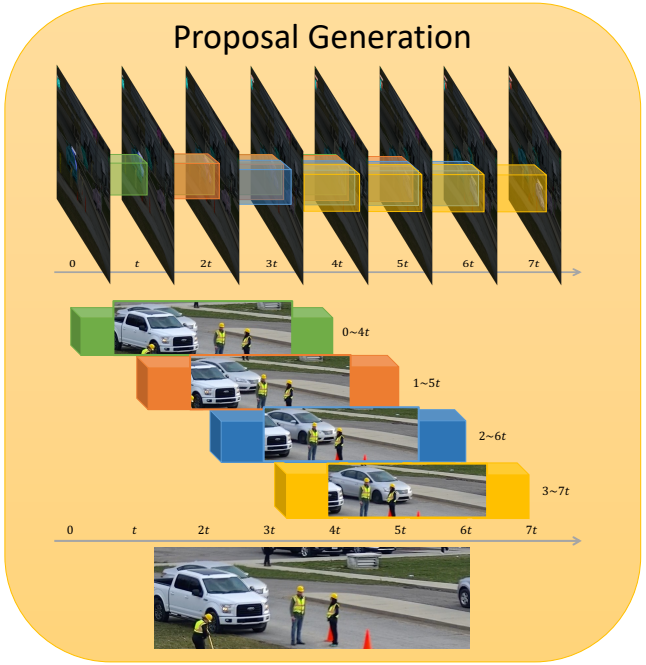
- Proposal
  - A candidate region where activity may occur
  - Processing element for activity recognition

- Spatio-temporal cube proposal
  - A simple six-tuple defining the boundaries in three dimensions

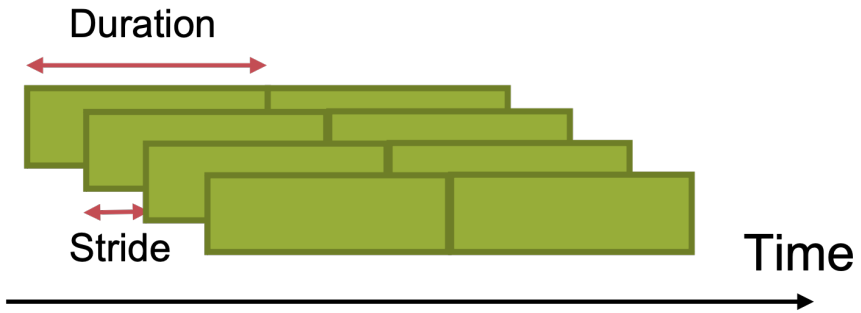
$$p_i = (x_0^i, x_1^i, y_0^i, y_1^i, t_0^i, t_1^i)$$

  - Fixed temporal duration when sampled
  - Much simpler than activity instances or tube proposals

Proposal Generation



- Proposal sampling
  - Dense overlapping sampling on untrimmed videos
  - Ensure completeness and coverage of any activity instance

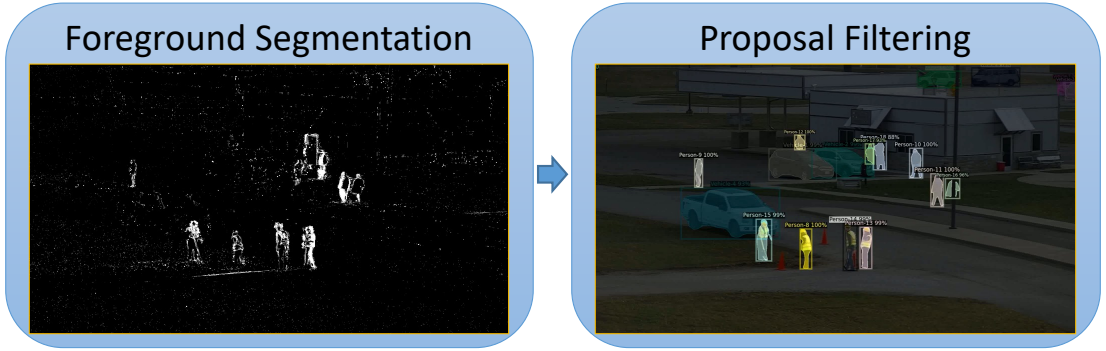


- Proposal refinement
  - Seed track ids from central from in each temporal window
  - Enlarge bounding boxes as union across the window

$$(x_0, x_1, y_0, y_1)_k = \bigcup(\{(x_0, x_1, y_0, y_1)_{i,j} \mid t_0 \le i \le t_1, tr_{i,j} = tr_{t_c,k}\})$$
$$k = 1, \cdots, n_{t_c}$$

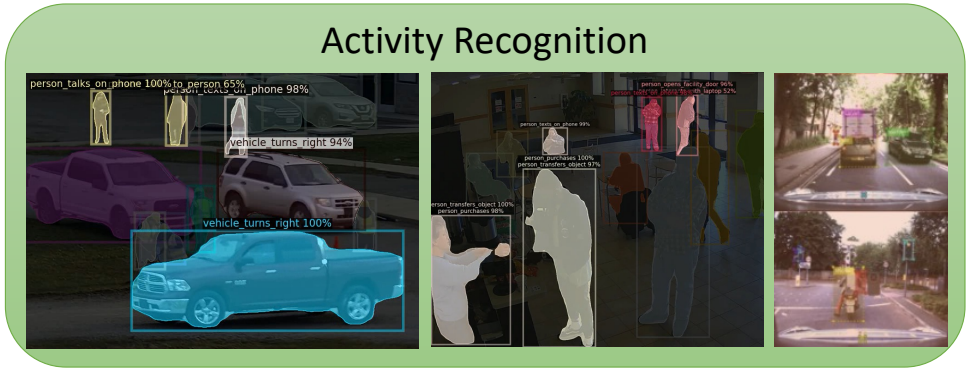
Proposal Generation: An Example

Proposal Filtering

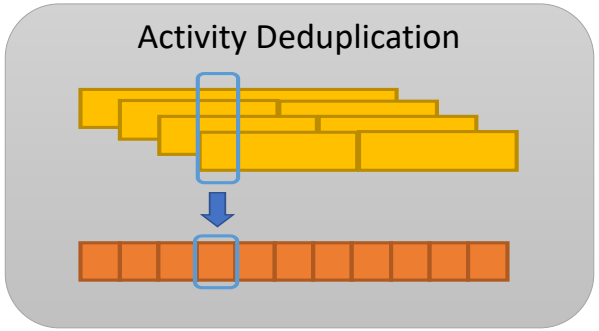


- Foreground segmentation
  - Frame-level binary mask for foreground pixels
  - Proposal foreground score as average value of pixel mask inside the cube
  - Learn a filtering threshold by allowing up to some sacrificed true positive
- Label assignment
  - Convert annotation into cube format by dense sampling
  - Estimate spatial IoU between proposal and ground truth cubes
  - Follow Faster R-CNN in selecting positive and negative samples
- Proposal evaluation
  - Assume perfect classifier by using assigned labels
  - Pass through following steps and use official metrics to estimate upper bound

Activity Recognition

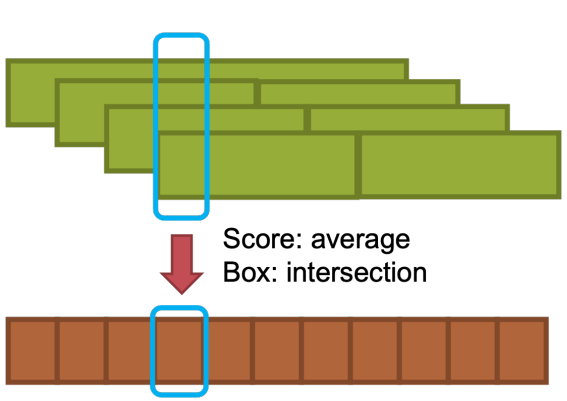


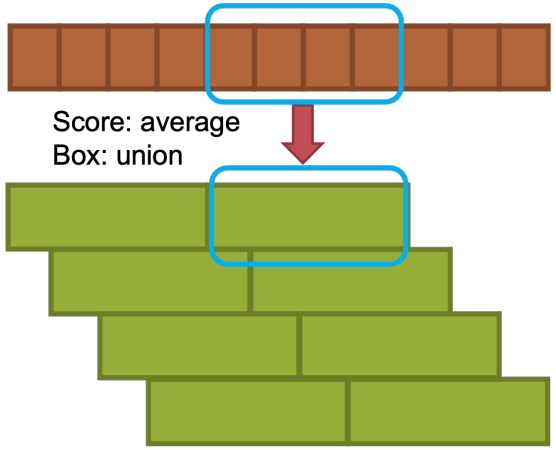
- Multi-label Classification
  - Binary cross entropy loss
  - Weighted by proposal scores
  - Balance activity-wise pos/neg samples
  - Balance samples of different activities
  - Balance samples of different datasets when used
- Action-wise late fusion

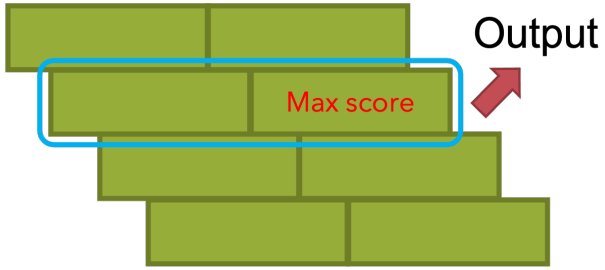


Activity Deduplication

- Overlapping instances







- Adjacent instances
  - Merge adjacent cubes above certain threshold, subject to a minimum duration

Experimental Results

Implementation Details

- Object detection: Mask R-CNN with Resnet-101 on COCO, stride 8
- Multi-object tracking: Towards-Realtime-MOT
- Foreground segmentation: HoG
- Proposal: duration 64, stride 16
- Classifiers: R(2+1)D, X3D, TRM

Evaluation Protocols

- NIST Activities in Extended Videos (ActEV)
  - Sequestered Data Leaderboards (SDL):
    MEVA Unknown Facility, MEVA Known Facility
  - Self-reported: VIRAT
  - Loosened target, P_{miss} @ T_{fa} based metrics
- ICCV 2021 ROAD Challenge
  - Road action detection in autonomous driving
  - Strict target, mAP @ $3D\ IoU$ based metrics

CVPR 2021 ActivityNet Challenge ActEV SDL Unknown Facility

System/Team	$nAUDC@0.2T_{fa} \downarrow$	Mean$P_{miss}@0.02T_{fa} \downarrow$	Relative Processing Time
Argus++ (Ours)	**0.3535**	**0.5747**	0.576
UMD_JHU	<u>0.4232</u>	0.6250	0.345
IBM-Purdue	0.4238	0.6286	0.530
UCF	0.4487	<u>0.5858</u>	0.615
Visym Labs	0.4906	0.6775	0.770
MINDS_JHU	0.6343	0.7791	0.898

NIST ActEV'21 SDL Known Facility

System/Team	$nAUDC@0.2T_{fa} \downarrow$	Mean$P_{miss}@0.02T_{fa} \downarrow$	Relative Processing Time
Argus++ (Ours)	**0.1635**	**0.3424**	0.413
UCF	0.2325	0.3793	0.751
UMD	0.2628	0.4544	0.380
IBM-Purdue	0.2817	0.4942	0.631
Visym Labs	0.2835	0.4620	0.721
UMD-Columbia	0.3055	0.4716	0.516
UMCMU	0.3236	0.5297	0.464
Purdue	0.3327	0.5853	0.131
MINDS_JHU	0.4834	0.6649	0.967
BUPT-MCPRL	0.7985	0.9281	0.123

NIST ActEV'21 SDL Unknown Facility

System/Team	$nAUDC@0.2T_{fa} \downarrow$	Mean$P_{miss}@0.02T_{fa} \downarrow$	Relative Processing Time
Argus++ (Ours)	**0.3330**	<u>0.5438</u>	0.776
UCF	<u>0.3518</u>	**0.5372**	0.684
IBM-Purdue	0.3533	0.5531	0.575
Visym Labs	0.3762	0.5559	1.027
UMD	0.3898	0.5938	0.515
UMD-Columbia	0.4002	0.5975	0.520
UMCMU	0.4922	0.6861	0.614
Purdue	0.4942	0.7294	0.239
MINDS_JHU	0.6343	0.7791	0.898

NIST TRECVID 2021 ActEV

System/Team	$nAUDC@0.2T_{fa} \downarrow$	Mean $P_{miss}@0.15T_{fa} \downarrow$	Mean $wP_{miss}@0.15R_{fa} \downarrow$
Argus++ (Ours)	**0.39607**	**0.30622**	<u>0.81080</u>
BUPT	<u>0.40853</u>	<u>0.32489</u>	**0.79798**
UCF	0.43059	0.34080	0.86431
M4D	0.84658	0.79410	0.88521
TokyoTech_AIST	0.85159	0.81970	0.94897
Team UEC	0.96405	0.95035	0.95670

NIST TRECVID 2020 ActEV

System/Team	$nAUDC@0.2T_{fa} \downarrow$	Mean $P_{miss}@0.15T_{fa} \downarrow$	Mean $wP_{miss}@0.15R_{fa} \downarrow$
Argus++ (Ours)	**0.42307**	**0.33241**	**0.80965**
UCF	<u>0.54830</u>	0.50285	<u>0.83621</u>
BUPT-MCPRL	0.55515	<u>0.48779</u>	0.84519
TokyoTech_AIST	0.79753	0.75502	0.87889
CERTH-ITI	0.86576	0.84454	0.88237
Team UEC	0.95168	0.95329	0.98300
Kindai_Kobe	0.96267	0.95204	0.93905

ICCV 2021 ROAD Action Detection

System/Team	Action@0.1 ↑	Action@0.2 ↑	Action@0.5 ↑	**Average ↑**
Argus++ (Ours)	**28.54**	**25.63**	6.98	**20.38**
THE IFY	28.15	20.97	6.58	18.57
YAAAHO	26.81	20.40	7.02	18.07
hyj	26.52	20.32	**7.05**	17.97
3D RetinaNet [21]	25.70	19.40	6.47	17.19
LeeC	13.64	9.89	2.23	8.59

Ablation Study

- Coverage of Proposal Formats

Table 8. Lower Bounds of $nAUDC@0.2T_{fa}$ on VIRAT Validation Set with different proposal formats. Italic values are non-overlapping proposals while the others are overlapping proposals. Duration and stride are in the unit of frames.

Duration / Stride	16	32	64	96
32	0.0705	*0.1208*	-	-
64	**0.0127**	0.0621	*0.0673*	-
96	0.0275	0.0504	-	*0.0688*

- Performance of Proposal Filtering

Table 9. Statistics of Proposals on VIRAT Validation Set

Name	Unfiltered	Filtered
Number of Proposals	211271	62831
Positive rate	0.1704	**0.5204**
Rate of unique label	0.4558	0.4415
Rate of two labels	0.4127	0.4252
Rate of three labels	0.1017	0.1060

Table 7. Proposal Quality Metrics on VIRAT Validation Set

$nAUDC@0.2T_{fa}$		IoU			Reference Coverage	
Threshold	Average	≥ 0	≥ 0.5	Average	≥ 0.5	≥ 0.9
Unfiltered Proposals	0.2358	0.0772	0.1518	0.1562	0.1125	0.4211
Filtered Proposals	0.2352	0.0772	0.1469	0.1563	0.1099	0.4280

Table 10. Proposal Filter on NIST ActEV'21 SDL Unknown Facility Micro Set

Proposal Filter	$nAUDC@0.2T_{fa} \downarrow$	Processing Time
Enabled	**0.4822**	0.582
Disabled	0.5176	0.925

Conclusion and Future Work

- Argus++: Robust Real-time Activity Detection
- Overlapping spatio-temporal cube proposals
- Superior performance in CVPR ActivityNet ActEV 2021, NIST ActEV SDL UF/KF, TRECVID ActEV 2020/2021, ICCV ROAD 2021

- Extending strict target into ActEV settings: bipartite matching with spatial localization
- Generalizing to more scenarios such as UAV videos
- Zero-shot or few-shot activity detection


informedia
digital video understanding
SEARCH
summarize
visualize
retrieve

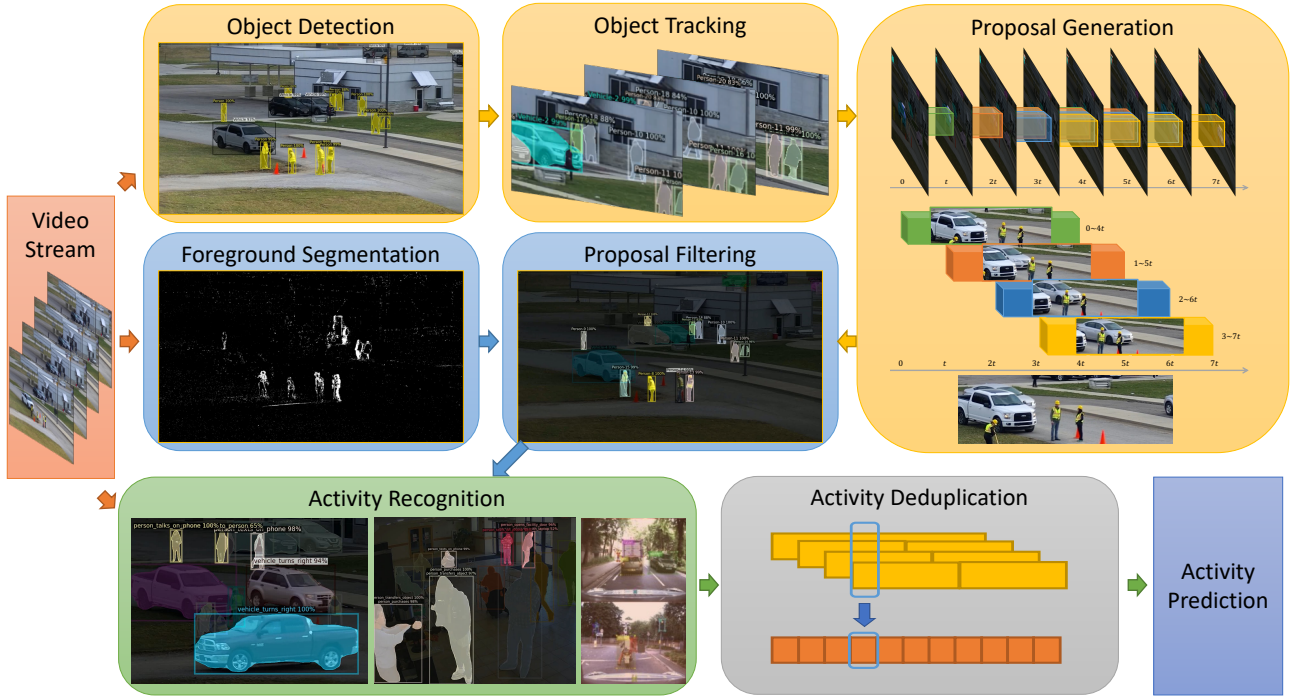
Argus++ in SRL

ActEV SRL Challenge - Metrics

- Time-based false alarm (TFA) -> rate of false alarm (RFA)
  - Match one ground truth with only one prediction
  - Cannot cut into short cubes, but need merging
- Temporal localization -> Spatio-temporal localization
  - Matching require spatial alignment

Argus++ for SRL

- Modifications limited within activity deduplication part
- Trained models from SDL system
- Still runs in real-time



Activity Deduplication

- Filter cubes based on classification confidence
  - Thresholds taken from scorer at 0.02 TFA
- For remaining cubes, merge adjacent ones into one instance
- Use bounding box from central cube
  - Since cube stride is 16, bounding boxes are unions of each 16-frame window
- Applied activity type filter, scene type filter, activity count filter

Results – under different constraints

Correctness Contraints: Loose

Submission Filter: Top per Team

Power Query: submissions=#[,#]*

Rank	Team Name	Submission ID	Submission Date	System Name	Correctness Constraint	AOD Protocol	AOD mean PMiss @0.1rfa	AOD mean nMODE @0.1rfa	AOD mean nAUDC @0.2rfa	AD Protocol	AD mean PMiss @0.1rfa	AD mean nAUDC @0.2rfa	SDL AD Protocol
1	CMU-DIVA	26965	2021-12-14	Argus	Loose	SRL_AOD_V2	0.6415	0.0107	0.6758	SRL_AD_V2	0.6016	0.6391	
2	BUPT-MCPRL	26945	2021-12-13	MCPRL_S0	Loose	SRL_AOD_V2	0.6810	0.0300	0.7065	SRL_AD_V2	0.6483	0.6724	
3	UCF	26912	2021-12-01	UCF-P	Loose	SRL_AOD_V2	0.7025	0.0347	0.7281	SRL_AD_V2	0.6570	0.6864	
4	UMD	26915	2021-12-02	UMD-JHU	Loose	SRL_AOD_V2	0.7552	0.0823	0.7736	SRL_AD_V2	0.7268	0.7468	
5	autohome	26876	2021-12-01	test	Loose	SRL_AOD_V2	0.7872	0.0236	0.8042	SRL_AD_V2	0.7425	0.7638	
6	dev-niu	26852	2021-11-29	Dev	Loose					SRL_AD_V2	0.7909	0.8090	

Results – under different constraints

Correctness Contraints: Medium

Submission Filter: Top per Team

Power Query: submissions=#[,#]*

Rank	Team Name	Submission ID	Submission Date	System Name	Correctness Constraint	AOD Protocol	AOD mean PMiss @0.1rfa	AOD mean nMODE @0.1rfa	AOD mean nAUDC @0.2rfa	AD Protocol	AD mean PMiss @0.1rfa	AD mean nAUDC @0.2rfa	SDL AD Pro
1	UCF	27015	2021-12-29	UCF-P	Medium	SRL_AOD_V1	0.6831	0.0579	0.7128	SRL_AD_V1	0.6288	0.6662	ActEV_SDL
2	CMU-DIVA	26965	2021-12-14	Argus	Medium	SRL_AOD_V1	0.6912	0.0290	0.7230	SRL_AD_V1	0.6462	0.6816	ActEV_SDL
3	BUPT-MCPRL	26945	2021-12-13	MCPRL_S0	Medium	SRL_AOD_V1	0.6997	0.0313	0.7257	SRL_AD_V1	0.6686	0.6925	ActEV_SDL
4	UMD	26915	2021-12-02	UMD-JHU	Medium	SRL_AOD_V1	0.7946	0.1086	0.8107	SRL_AD_V1	0.7479	0.7664	ActEV_SDL
5	autohome	26876	2021-12-01	test	Medium	SRL_AOD_V1	0.8268	0.0168	0.8414	SRL_AD_V1	0.7726	0.7923	ActEV_SDL
6	dev-niu	26852	2021-11-29	Dev	Medium					SRL_AD_V1	0.8172	0.8323	ActEV_SDL

Results – under different constraints

Correctness Contraints: Tight

Submission Filter: Top per Team

Power Query: submissions=#[,#]*

Rank	Team Name	Submission ID	Submission Date	System Name	Correctness Constraint	AOD Protocol	AOD mean PMiss @0.1rfa	AOD mean nMODE @0.1rfa	AOD mean nAUDC @0.2rfa	AD Protocol	AD mean PMiss @0.1rfa	AD mean nAUDC @0.2rfa	SDL AD Protocol
1	BUPT-MCPRL	26945	2021-12-13	MCPRL_S0	Tight	SRL_AOD_V3	0.7949	0.0272	0.8106	SRL_AD_V3	0.7547	0.7734	
2	CMU-DIVA	26965	2021-12-14	Argus	Tight	SRL_AOD_V3	0.8610	0.0424	0.8686	SRL_AD_V3	0.7891	0.8038	
3	UCF	26912	2021-12-01	UCF-P	Tight	SRL_AOD_V3	0.8878	0.0625	0.8980	SRL_AD_V3	0.8047	0.8212	
4	autohome	26876	2021-12-01	test	Tight	SRL_AOD_V3	0.9071	0.0463	0.9143	SRL_AD_V3	0.8557	0.8653	
5	UMD	26915	2021-12-02	UMD-JHU	Tight	SRL_AOD_V3	0.9279	0.1010	0.9349	SRL_AD_V3	0.8378	0.8534	
6	dev-niu	26852	2021-11-29	Dev	Tight					SRL_AD_V3	0.8737	0.8829	

Future Work

- Optimization for AOD
  - Use frame-level object detection bounding boxes, at additional computation costs, depending on the efficiency requirements
- Optimization for RFA
  - Refine deduplication algorithm, with joint optimization with recognition, e.g., classification of activity start, middle, and end

Acknowledgements

This research is supported in part by the Intelligence Advanced Research Projects Activity (IARPA) via Department of Interior/Interior Business Center (DOI/IBC) contract number D17PC00340. This research is supported in part through the financial assistance award 60NANB17D156 from U.S. Department of Commerce, National Institute of Standards and Technology. This project is funded in part by Carnegie Mellon University's Mobility21 National University Transportation Center, which is sponsored by the US Department of Transportation.


Thank You!