UCF | Center for Research in Computer Vision
UNIVERSITY OF CENTRAL FLORIDA

ICPR 2020

Gabriella: An Online System for Real-Time Activity Detection in Untrimmed Security Videos

Mamshad N Rizve, Ugur Demir, Praveen Tirupattur,
Aayush J Rana, Kevin Duarte, Ishan Dave, Yogesh S
Rawat, Mubarak Shah

Introduction

Detect activities in untrimmed security videos

- Human and Vehicles
- Activity types
    - Single actors
    - Interaction between actors
    - Actor-object interactions

Challenges

- Untrimmed nature
- Multiple activities
- Varying length of activities
- Multiple actors

Challenges

- Untrimmed nature
- Multiple activities
- Varying length of activities
- Multiple actors
- Multiple scales

Motivations

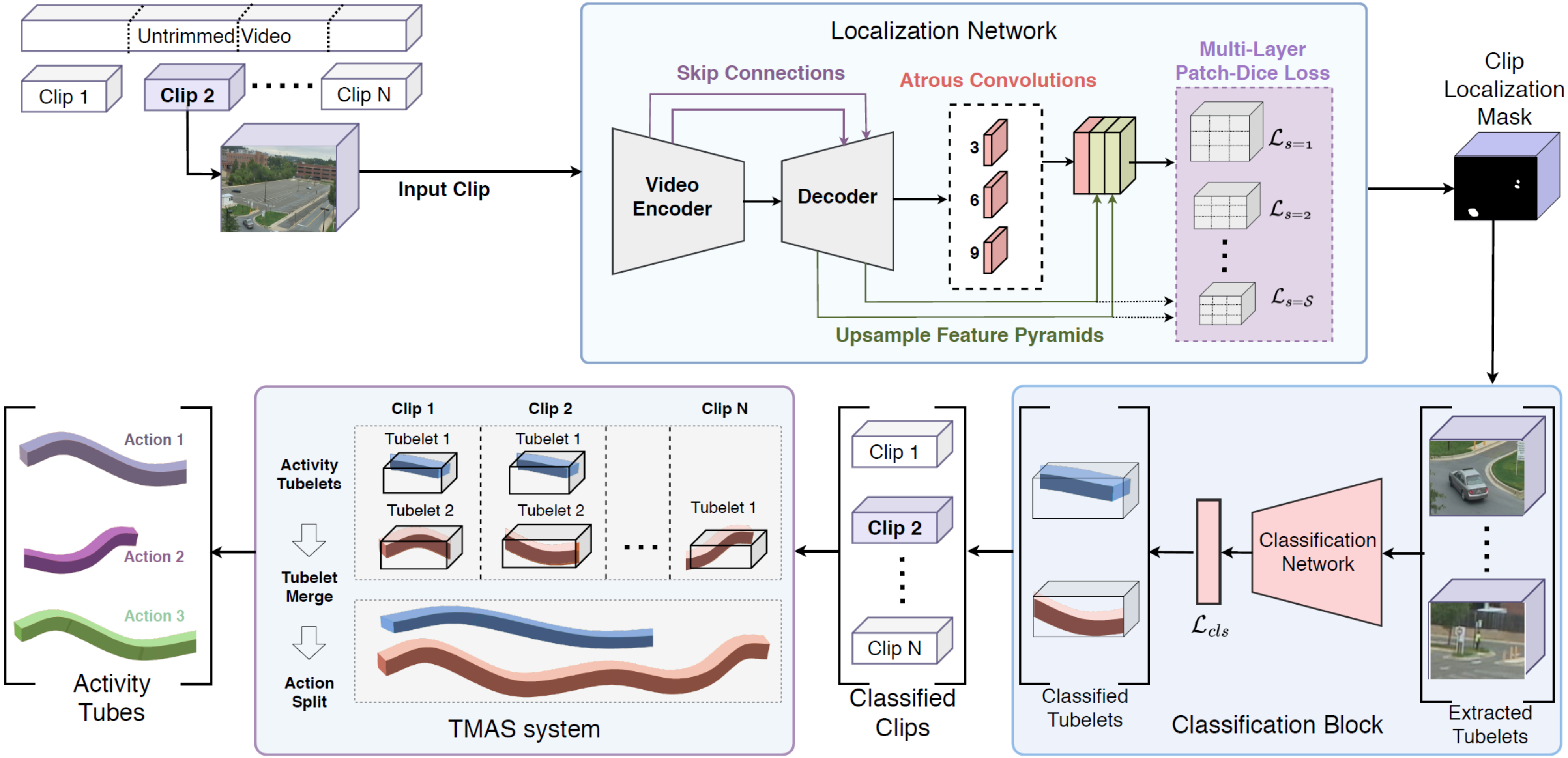
- Region proposal based approach [1, 2]
  - Scaling issue with videos
  - Multiple actors
    - How to pair?
- Object detection [3]
  - Time consuming
  - Multiple actors
    - How to pair?

[1] Hui et al. "Tube convolutional neural network (T-CNN) for action detection in videos." In IEEE international conference on computer vision. 2017.
[2] He et al. "Mask r-cnn." In Computer Vision (ICCV), 2017 IEEE International Conference on, pp. 2980-2988. IEEE, 2017.
[3] Gleason, Joshua, et al. "A proposal-based solution to spatio-temporal action detection in untrimmed videos." 2019 IEEE WACV. IEEE, 2019.

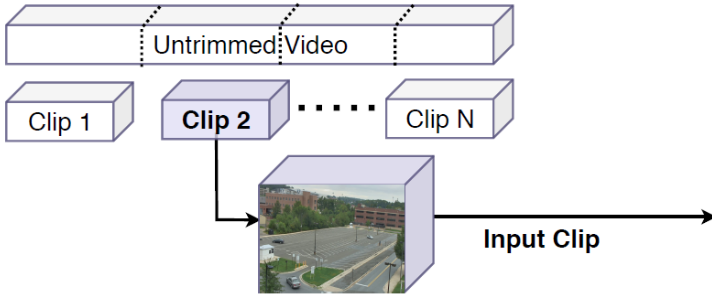
Approach

- A two-stage process
  - Detect activity tubelets from long untrimmed videos
  - Recognize activities in the detected tubelets
- Encoder-decoder architecture
  - No region proposal
- Video level detection
  - No object detection

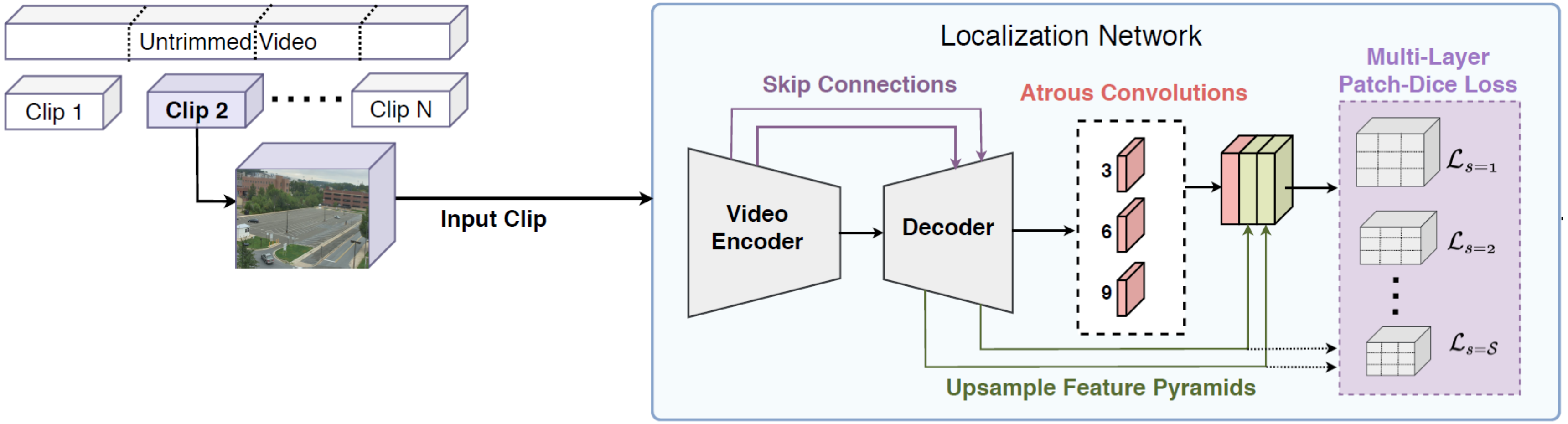
Our Approach



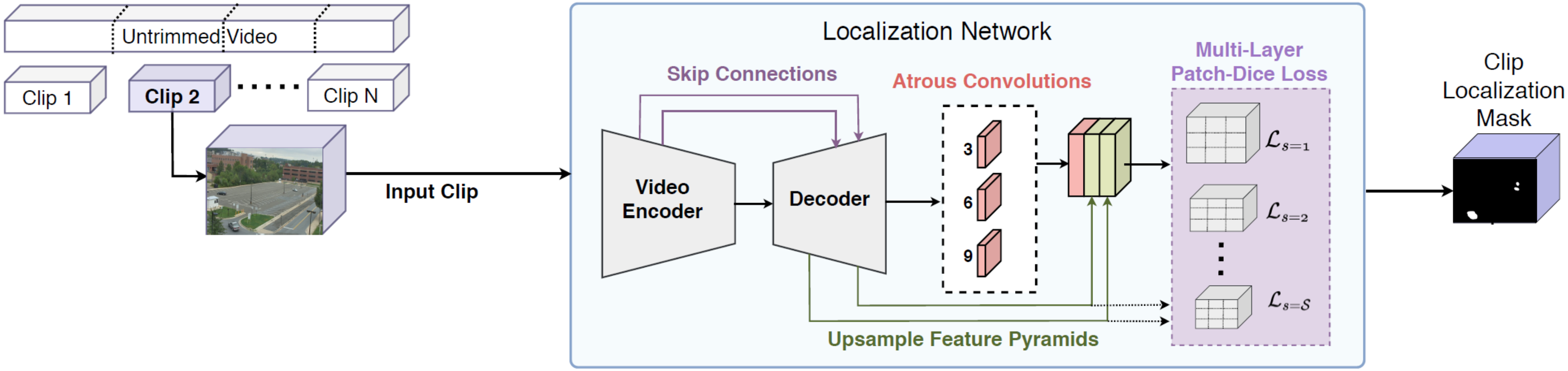
Our Approach



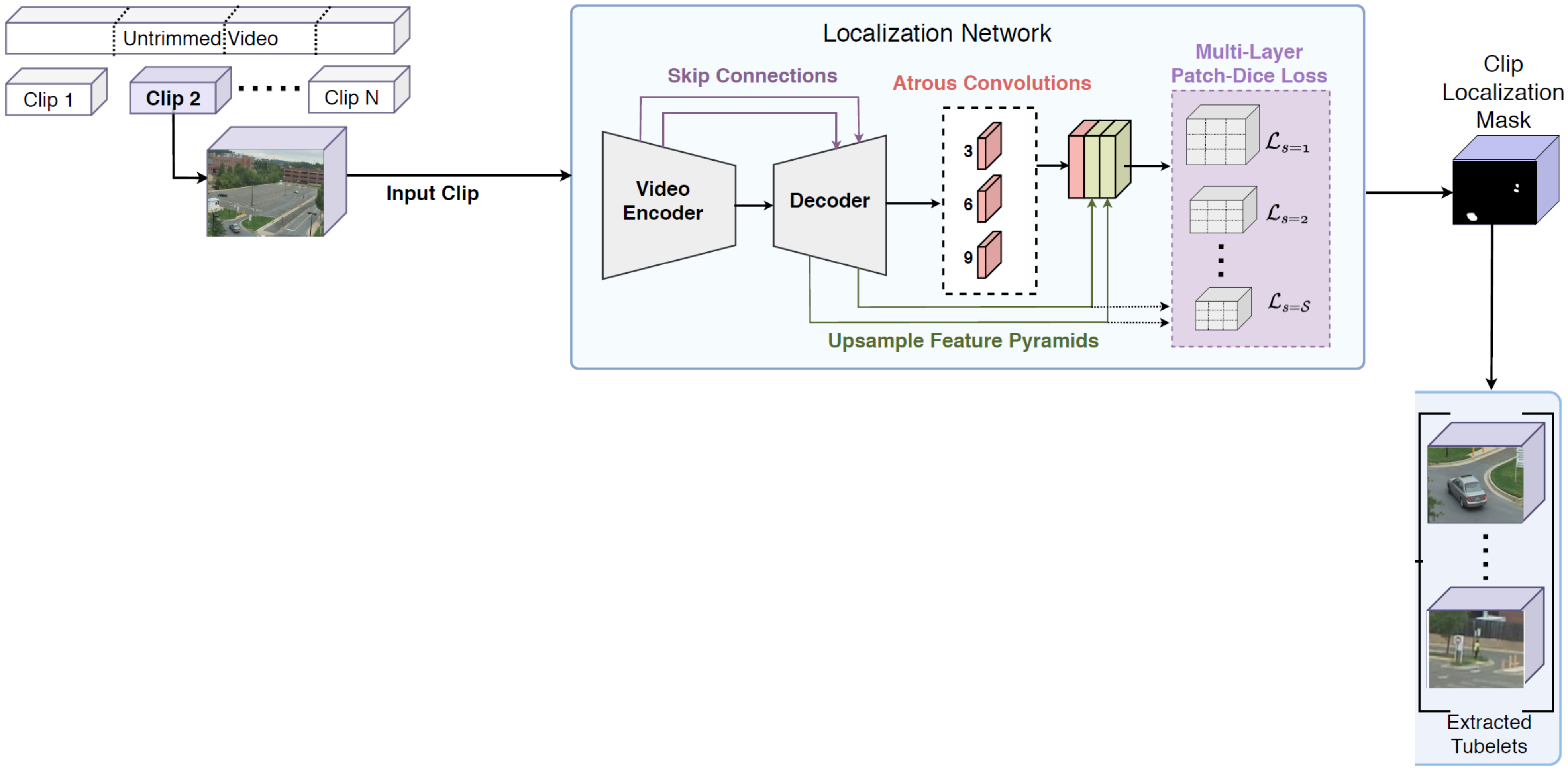
Our Approach



Our Approach



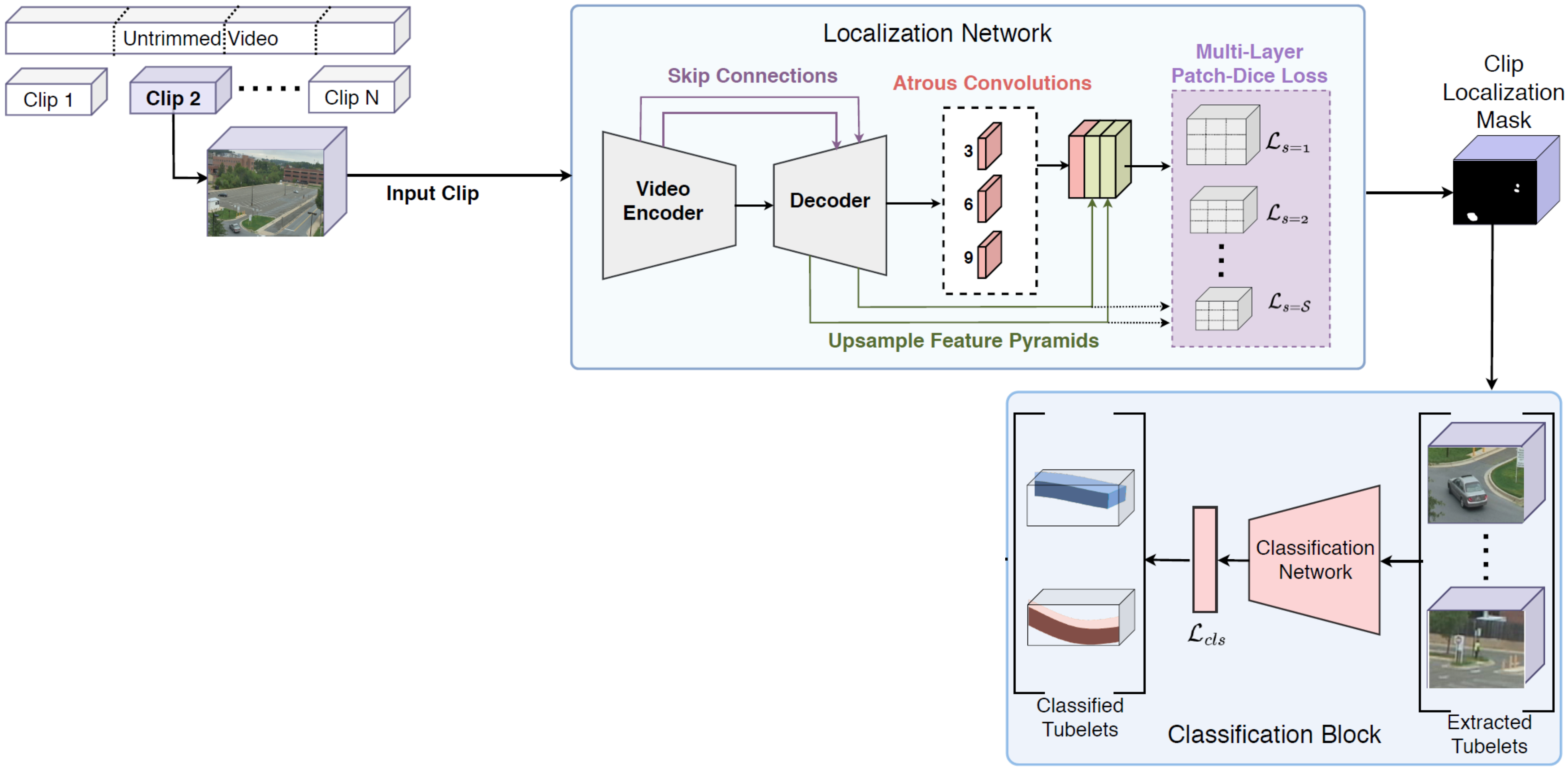
Our Approach



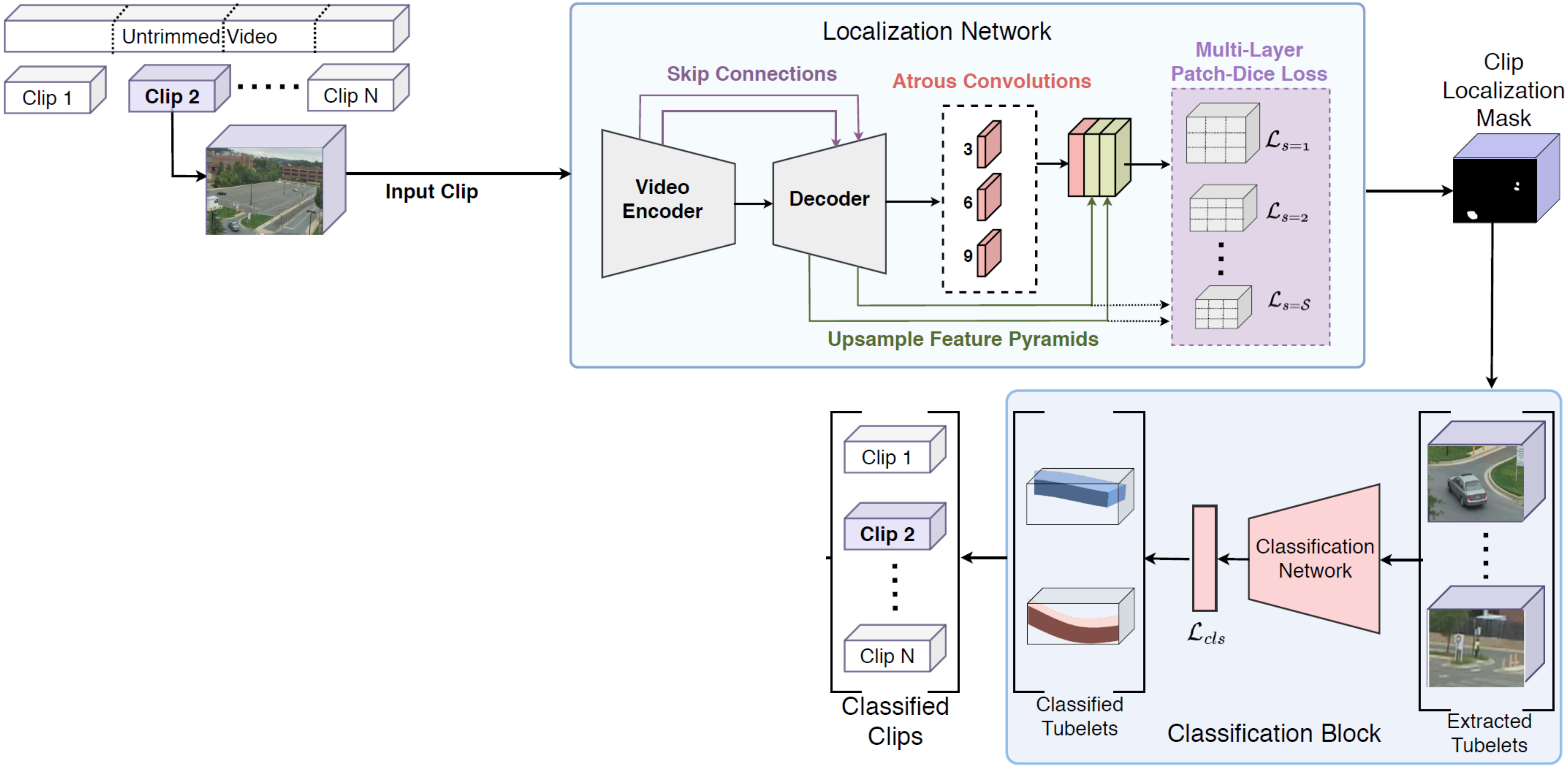
Center for Research in Computer Vision

UNIVERSITY OF CENTRAL FLORIDA

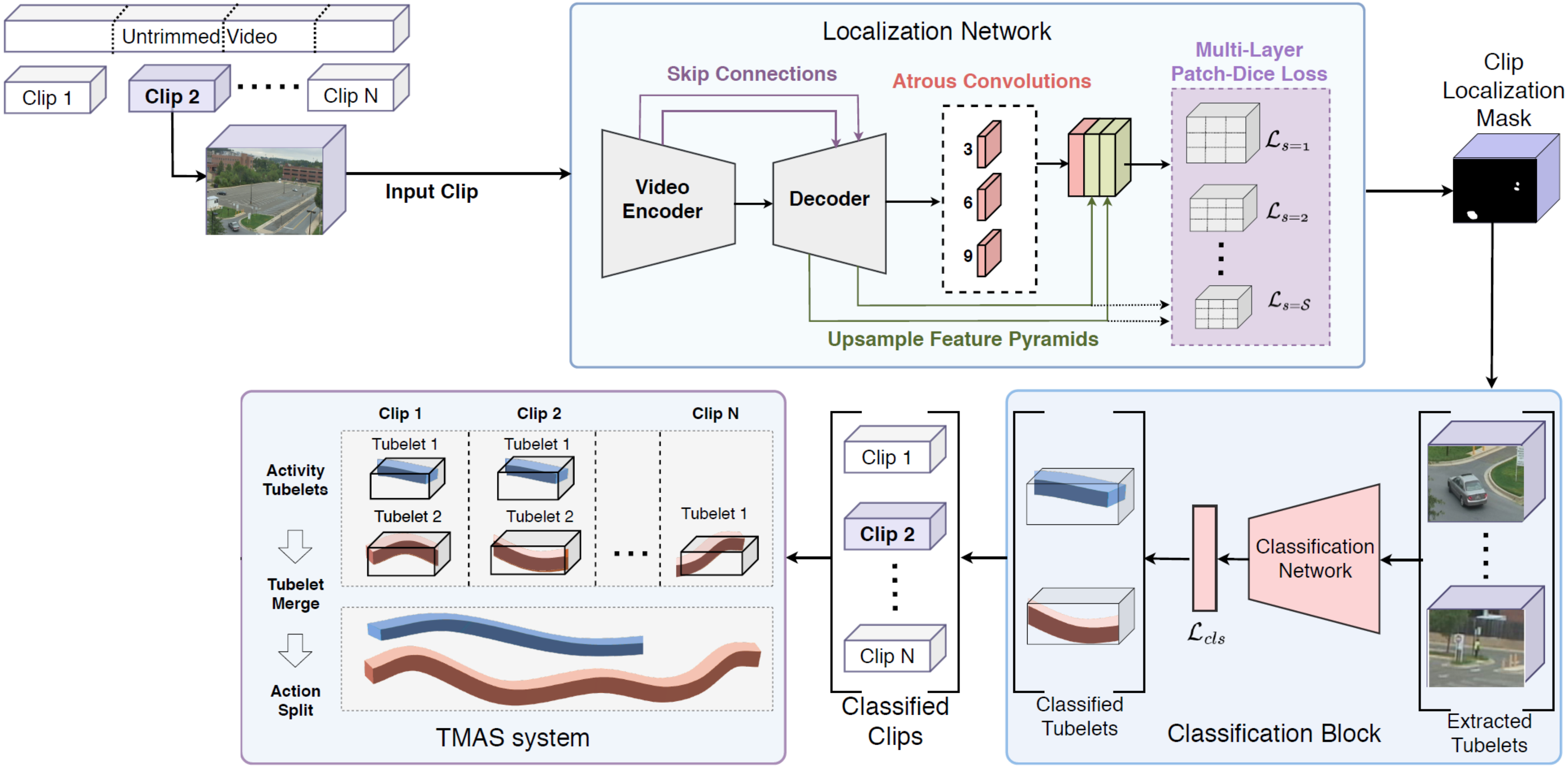
Our Approach



Our Approach



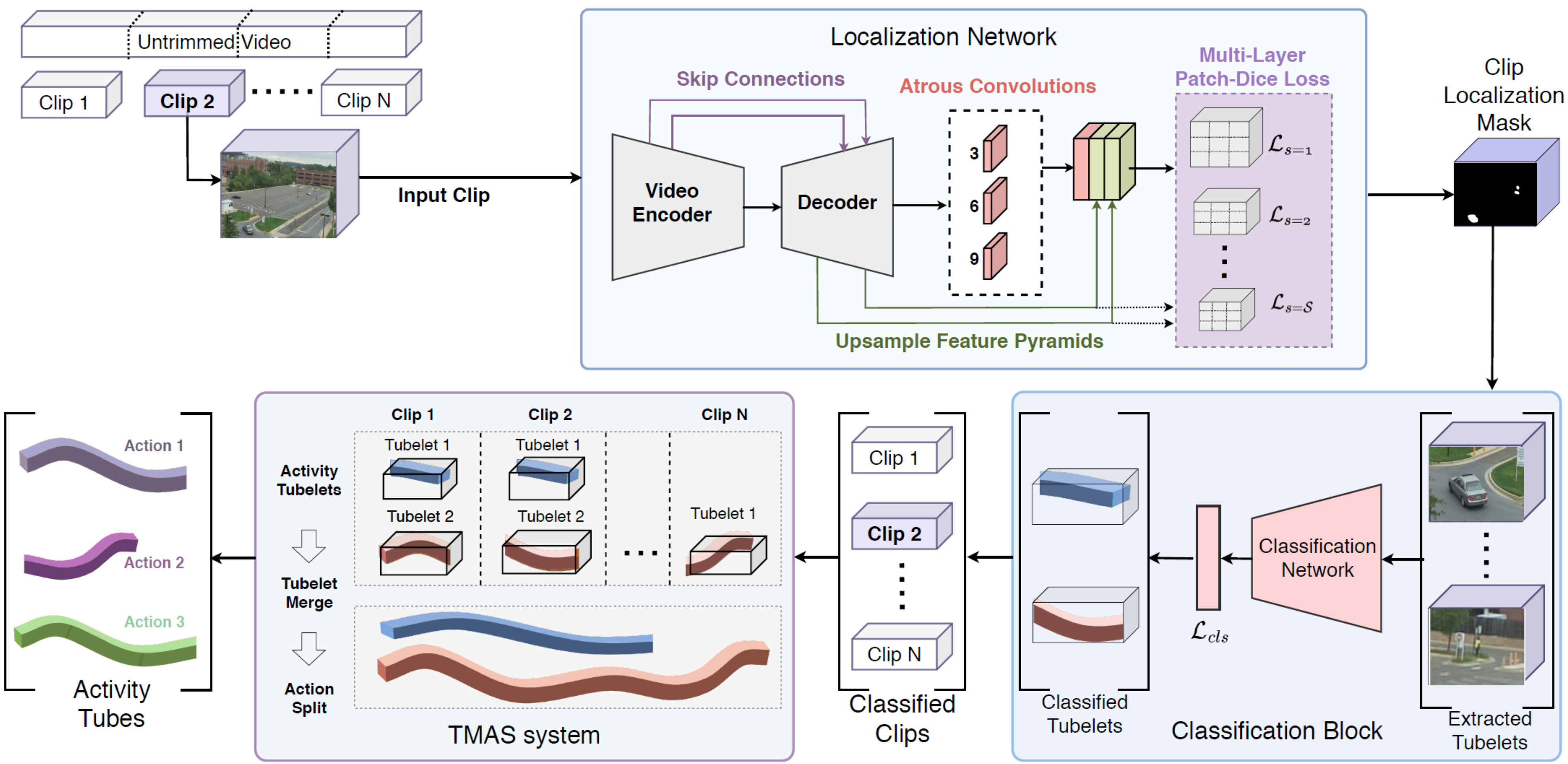
Our Approach



Center for Research in Computer Vision
UCF UNIVERSITY OF CENTRAL FLORIDA

ICPR 2020

Our Approach

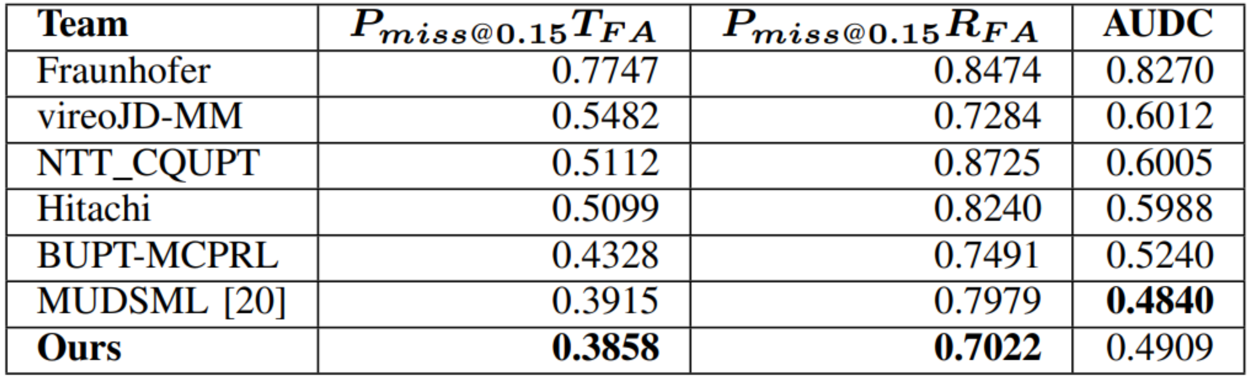


Datasets

- VIRAT [1]
  - 64 (2.47 hours) videos for training
  - 54 videos (1.93 hours) for validation
  - 40 activities
- MEVA [2]
  - 1056 videos (88 hours)
  - 37 activities

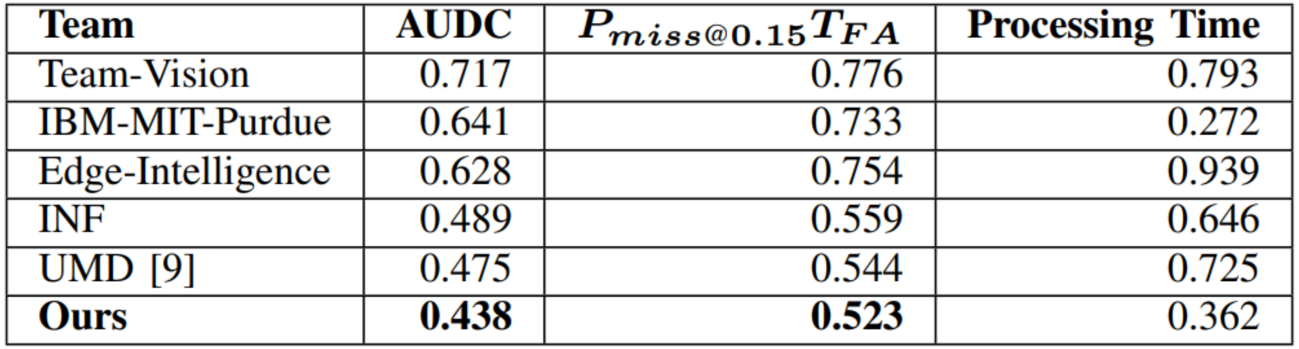
[1] Oh et al. "A Large-scale Benchmark Dataset for Event Recognition in Surveillance Video." In IEEE international conference on computer vision. 2011.
[2] Kitware inc, the multiview extended video with activities (meva) dataset.

Quantitative Results (VIRAT Dataset)

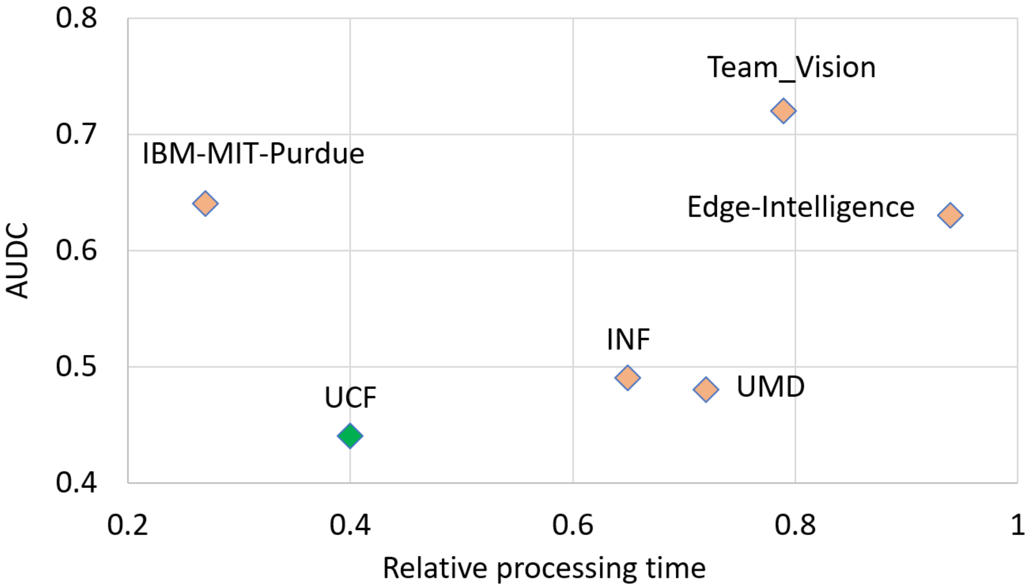


Team	$P_{miss@0.15}T_{FA}$	$P_{miss@0.15}R_{FA}$	**AUDC**
Fraunhofer	0.7747	0.8474	0.8270
vireoJD-MM	0.5482	0.7284	0.6012
NTT_CQUPT	0.5112	0.8725	0.6005
Hitachi	0.5099	0.8240	0.5988
BUPT-MCPRL	0.4328	0.7491	0.5240
MUDSML [20]	0.3915	0.7979	**0.4840**
Ours	**0.3858**	**0.7022**	0.4909

Quantitative Results (MEVA Dataset)

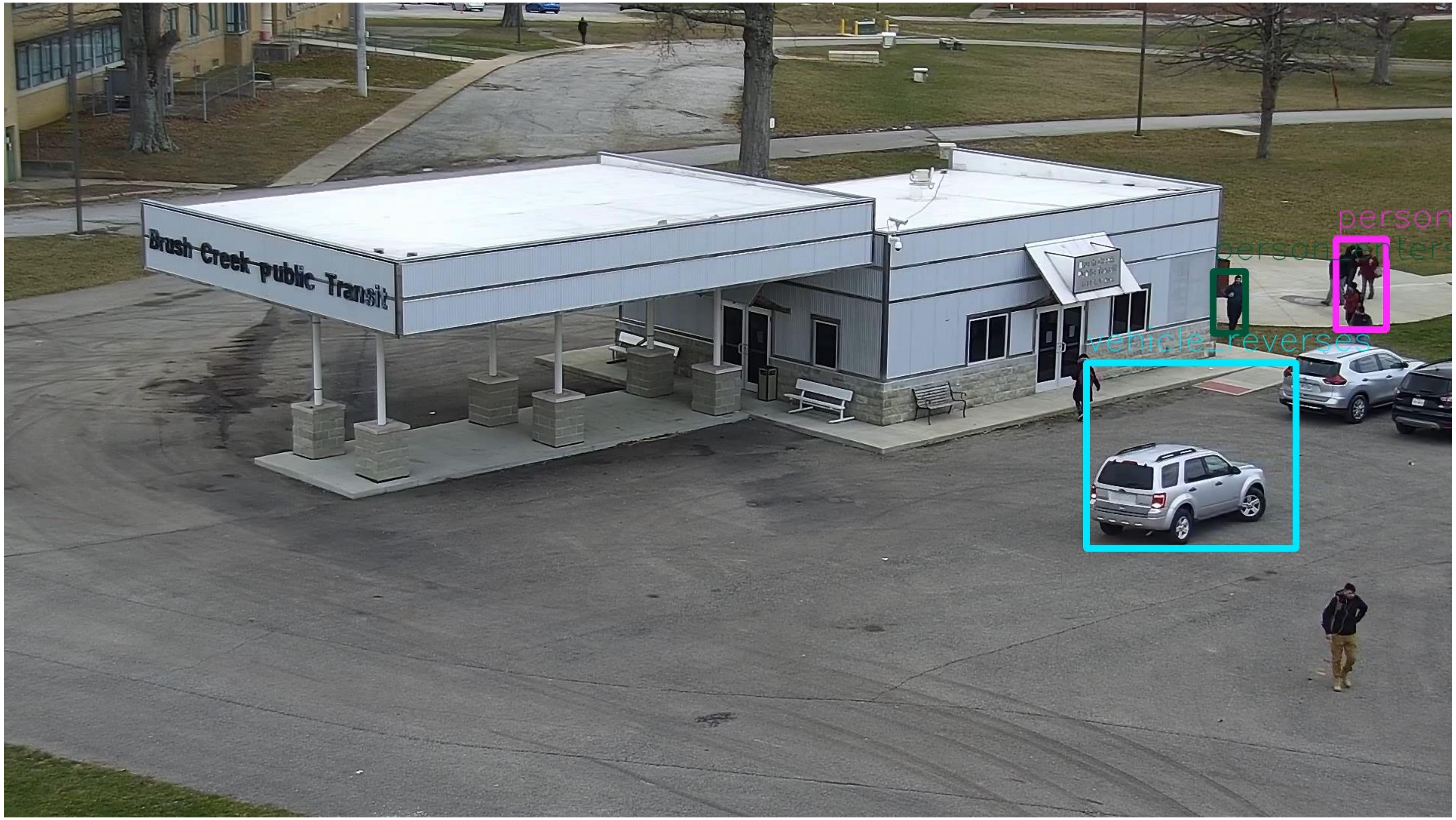


Team	AUDC	$P_{miss@0.15}T_{FA}$	Processing Time
Team-Vision	0.717	0.776	0.793
IBM-MIT-Purdue	0.641	0.733	0.272
Edge-Intelligence	0.628	0.754	0.939
INF	0.489	0.559	0.646
UMD [9]	0.475	0.544	0.725
Ours	**0.438**	**0.523**	0.362



Qualitative Results (Localization)

Qualitative Results



Thank You

Project Page:
https://tinyurl.com/y6gv8dpl