UCF

Yogesh S Rawat, Aayush Rana, Praveen Tirupattur, and Mubarak Shah

Center for Research in Computer Vision
University of Central Florida

Contents

- Activity Detection in Untrimmed Videos
  - AD Task
- Activity Object Detection in Untrimmed Videos
  - AOD Task

Contents

- **Activity Detection in Untrimmed Videos**
  - **AD Task**
- Activity Object Detection in Untrimmed Videos
  - AOD Task

Activity Detection (AD) in Untrimmed Videos

Action Analysis in Video

Given Untrimmed videos

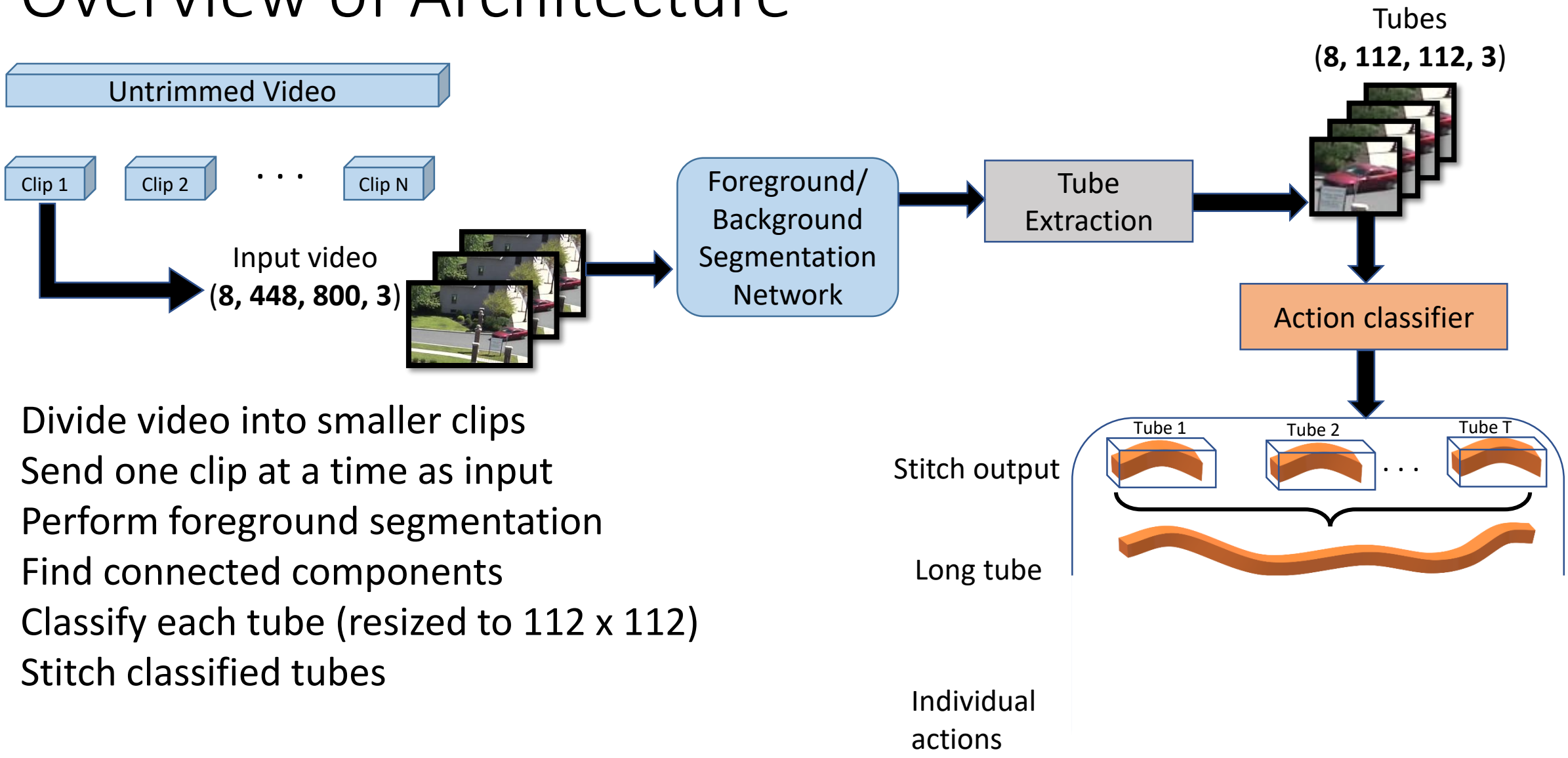
- Containing multiple
  - actors
  - actions
  - action labels per actor
- Varying length of action
- Unbalanced dataset (low samples)

- **We want to**
  - Localize all actions
  - Classify each action

Key Points

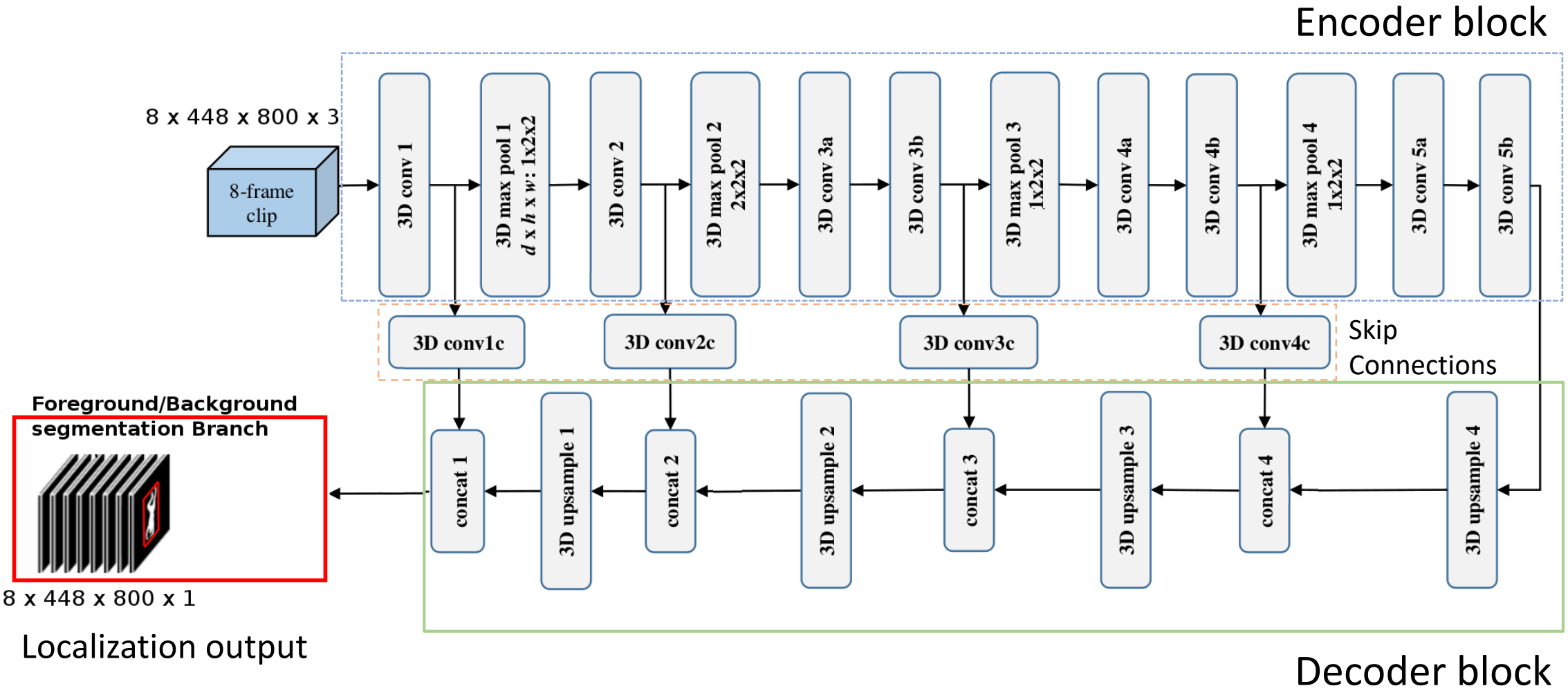
- Bottom up foreground background segmentation
- Detect actions tubes from long untrimmed videos
- Classify each instance individually
- Activity tube generation

Overview of Architecture

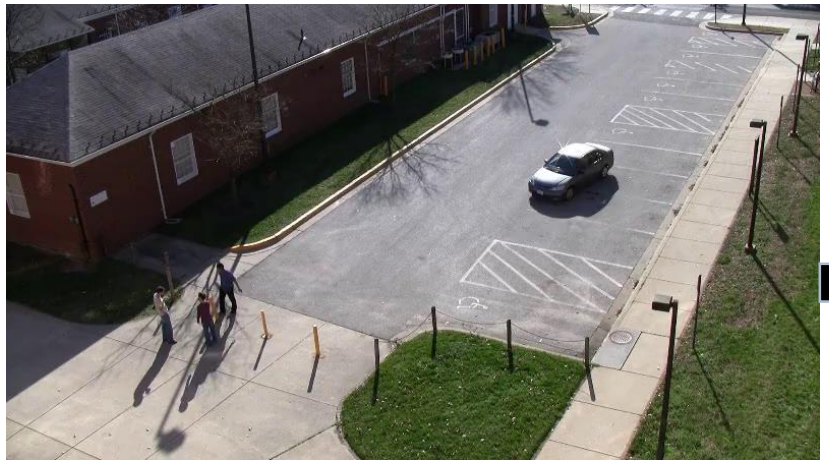


- Divide video into smaller clips
- Send one clip at a time as input
- Perform foreground segmentation
- Find connected components
- Classify each tube (resized to 112 x 112)
- Stitch classified tubes

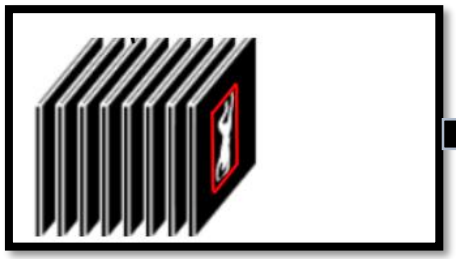
Foreground/ Background Segmentation Network



Tube Extraction



Input Video
(448 x 800 x 3)



Localization Mask
(448 x 800 x 1)

Multiply and segment foreground

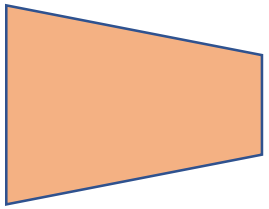
Connected components

Output Tubes
(112 x 112 x 3)

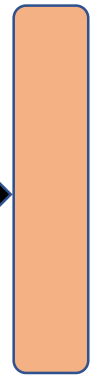
Action Classification

Output Tubes
(8 x 112 x 112 x 3)

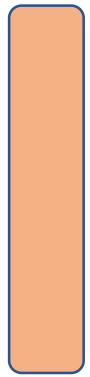
Classification Block



Network
(ResNet 3D)



Extracted
Features



Output

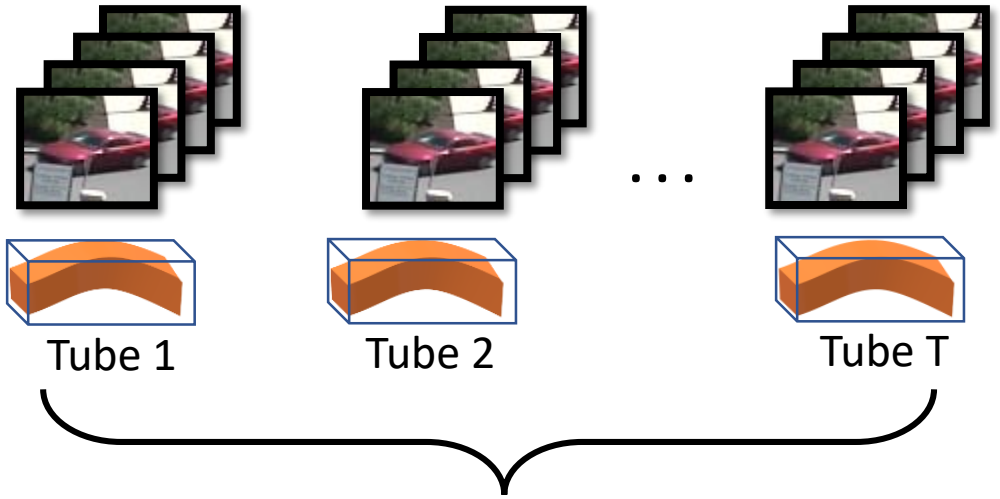
Classification output

Transport
Heavycarry: 0.69
Walking: 0.81

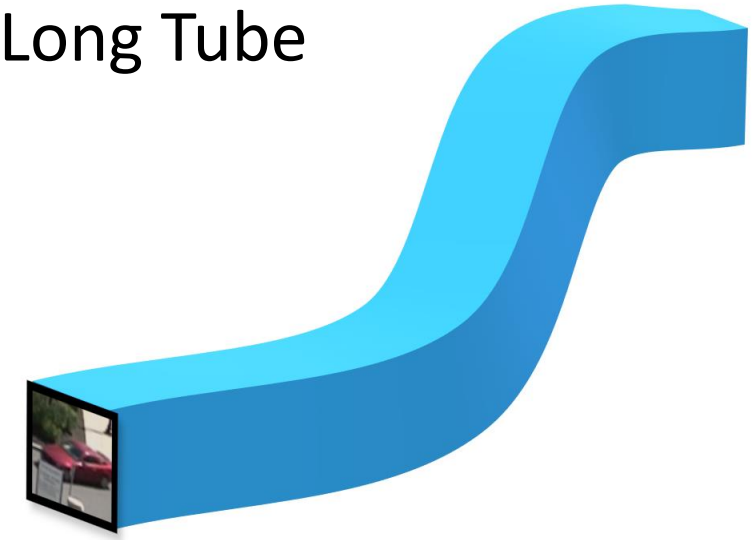
Vehicle
moving: 0.86

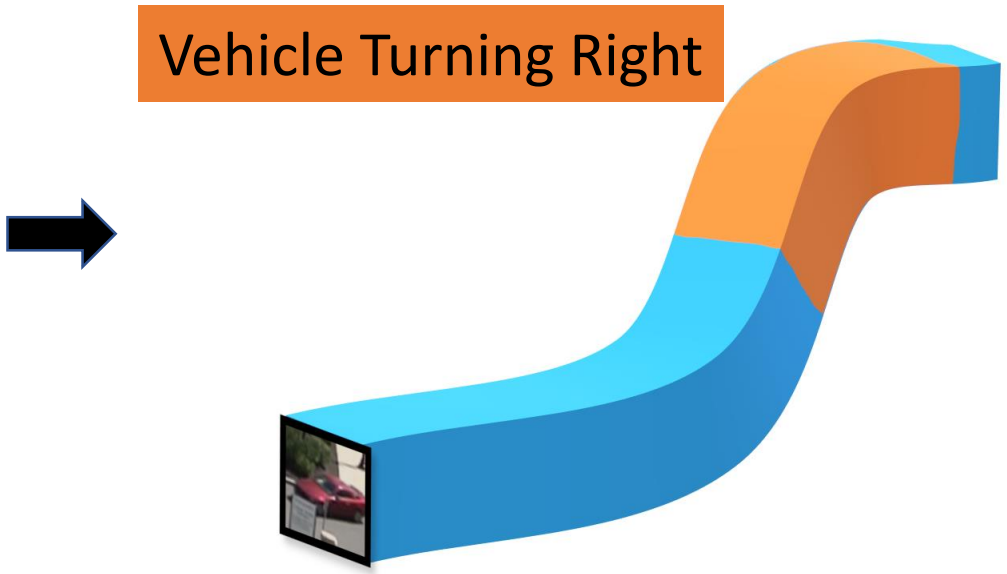
Standing: 0.73
Talking: 0.65
Interacts: 0.77

Tube Stitching



Long Tube





Final Output (Example-1)

Final Output (Example-2)

Final Output (Example-3)

NIST Evaluation on Validation Set

Metric name	Metric Value
Mean-p_miss @ 0.01 rfa	0.9066
Mean-p_miss @ 0.03 rfa	0.8478
Mean-p_miss @ 0.1 rfa	0.6973
Mean-p_miss @ 0.15 rfa	**0.6608**
Mean-p_miss @ 0.2 rfa	0.6279
Mean-p_miss @ 1 rfa	0.4633
N-mide	0.2045

Issues

- Imbalanced Dataset
  - Extremely low samples for some classes
- Similar activities being confused by classifier
- Activities far from camera
  - Very small activities, hard to locate

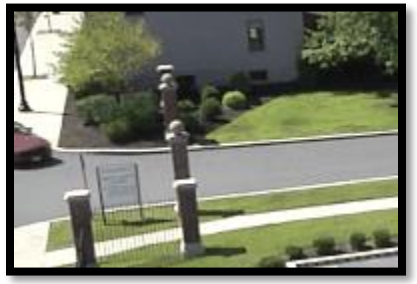
Contents

- Activity Detection in Untrimmed Videos
  - AD Task
- **Activity Object Detection in Untrimmed Videos**
  - **AOD Task**

Activity Detection based on Actor-Object Interaction

Actor Object Interaction in Videos

- Given an untrimmed video, localize
  - all actors present
  - all objects interacted with
- Classify Activities based on the actor-object interaction



Challenges

- Multiple actor-object instances in single clip
  - Multiple actors and objects
- Same actor-object combination in multiple classes
  - Opening door, closing door
- Same actor-object instance with multiple labels
  - Exiting, closing door

Approaches

- Region Proposals
  - Based on bounding box proposals T-CNN [1], Mask-CNN [2]
  - Bottom-up approach
  - Regression over full space
- Encoder-Decoder
  - Unified semantic segmentation ST-CNN [3], SegNet [4]
  - Issue with multiple activity instances
  - Need of connected components and post processing

[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] Rui et al. "An End-to-end 3D Convolutional Neural Network for Action Detection and Segmentation in Videos." arXiv preprint arXiv:1712.01111 (2017).
[4] Badrinarayananet al. "Segnet: A deep convolutional encoder-decoder architecture for image segmentation." arXiv preprint arXiv:1511.00561 (2015).

Motivation

- End-to-end training framework
  - Completely remove region proposal and ToI/RoI pooling
  - Use actor-object attention instead
- Multiple tasks
  - Foreground/background
  - Objects
  - Actions
- Model convergence using multiple losses
- Joint actor-object action classification

Action Classification in Videos

Object:
- Vehicle

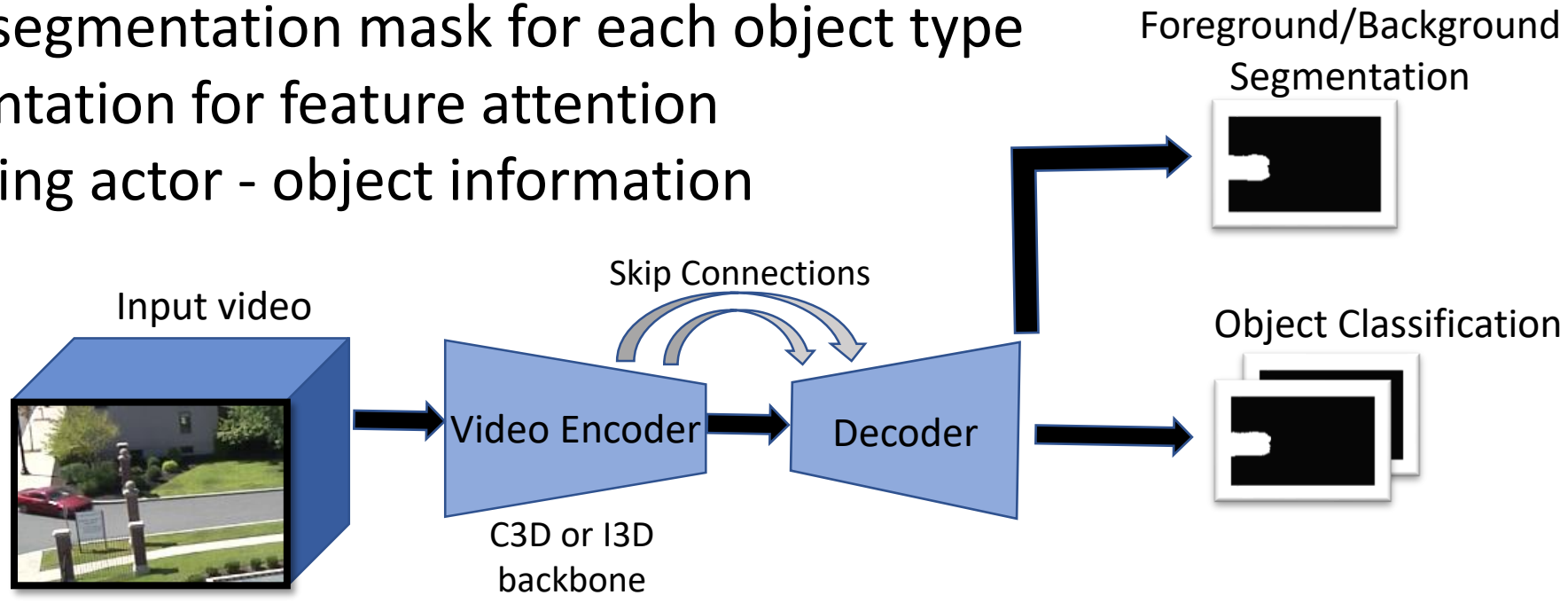
Action: Vehicle turning left

Objects:
- Person

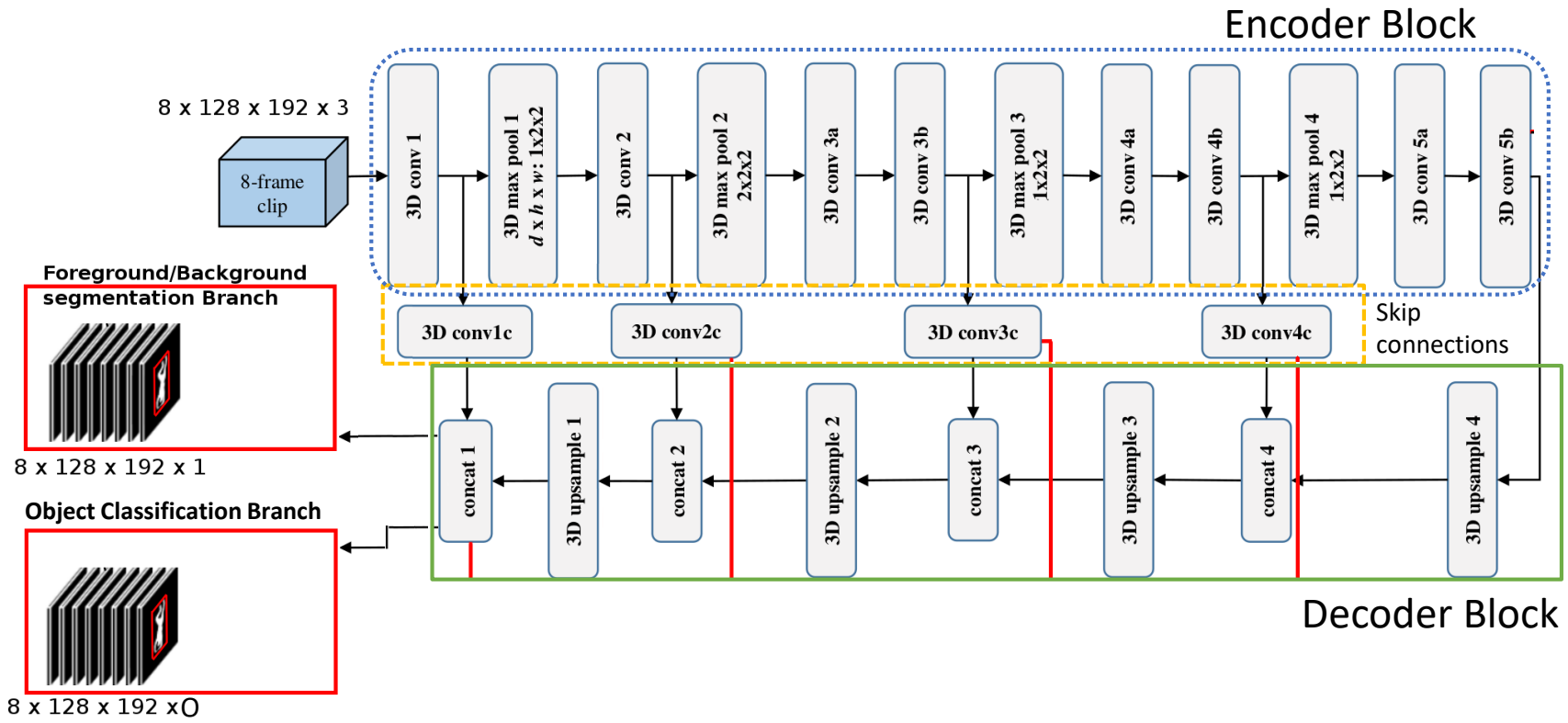
Action: Activity Talking (red),
Activity Carrying (green)

Overview of Proposed Architecture

- Get 8 frame video clip
- Generate foreground / background segmentation mask
- Generate object segmentation mask for each object type
- Use fg/bg segmentation for feature attention
- Classify action using actor - object information



End-to-end network for Video Action Segmentation



- Encode video features (Conv 3D)
- Decode features (Deconv 3D) with skip connection
- Segment foreground/background
- Segment each object class

Quantitative Results

- DIVA data subset
  - Smaller clips focusing on activity used (128 x 192 resolution)
  - 64 training videos, 55 validation videos
  - 19 action classes (DIVA 1B set)
  - 2 object classes (person and vehicle)

- Action object localization IoU: 0.64
- Classification F1 Score (19 classes): 0.46

Qualitative Results

Input

Foreground/Background segmentation
(Only moving objects)

Object segmentation
(3 people)

Action classification
Talking (Red)
Carrying (Green)

Qualitative Results

Input

Foreground/Background segmentation
(Only moving objects)

Object segmentation
Vehicle (Green)
Person (Red)

Action classification
Vehicle turning left

NIST Evaluation on Validation Set

Activity Detection

Metric	Value
mean-p_miss@0.01rfa	0.954337382386
mean-p_miss@0.03rfa	0.925133046316
mean-p_miss@0.15rfa	**0.757087143515**
mean-p_miss@0.1rfa	0.784522064048
mean-p_miss@0.2rfa	0.739966420528
mean-p_miss@1rfa	0.605960537865

NIST Evaluation on Validation Set

Object detection

Metric	Value
mean-mean-object-p_miss@0.033rfa	0.7397920634
mean-mean-object-p_miss@0.1rfa	0.673425676293
mean-mean-object-p_miss@0.2rfa	0.624957826044
mean-mean-object-p_miss@0.5rfa	0.538296977439

Thank you!