香港中文大學
The Chinese University of Hong Kong

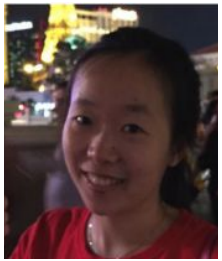
商汤
sensetime

Learning to Reason Relations for Spatio-Temporal Action Localization 1st Place Winning Solution

AVA-Kinetics Crossover Challenge 2020

MMLab-ACAR Team





Siyu Chen* **Junting Pan*** Jing Shao Hongsheng Li Yu Liu

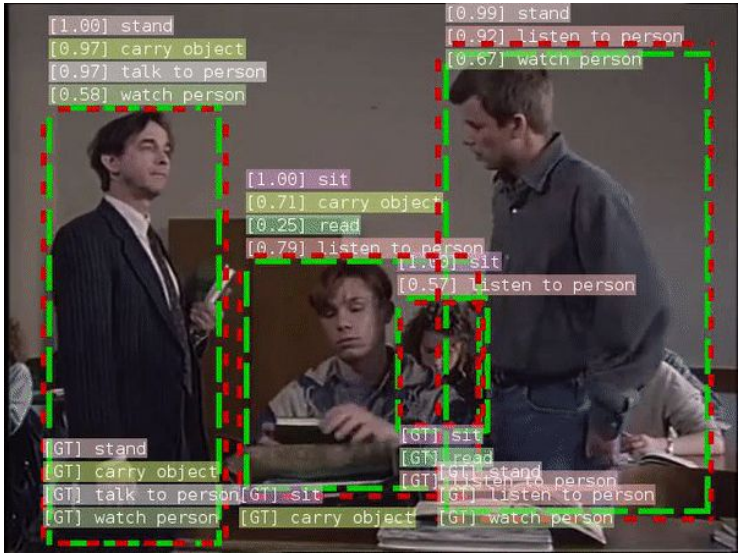
* Equal contribution

- Overview of the task and challenge
- Details of our solution
- Analysis and Rethinking

- **Overview of the task and challenge**
- Details of our solution
- Analysis and Rethinking

Task: Spatio-temporal action localization

- Localize atomic actions in both space and time
- Frames are labeled at 1FPS over 80 action classes
- Evaluation: Frame mAP at keyframes



(Picture from PySlowFast[1])

1. https://github.com/facebookresearch/SlowFast/blob/master/slowfast/datasets/DATASET.md

What's New in AVA-Kinetics Challenge 2020?

- [**Additional Data**] Kinetics-700 videos with AVA-style annotations
  - +238k unique videos (x500 of AVA v2.2)
  - Youtube videos vs Movies
  - AVA challenge >> AVA-Kinetics Crossover challenge

- Overview of the task and challenge
- **Details of our solution**
- Analysis and Rethinking

Overview of our solution

- SlowFast[1] backbone + Relation Reasoning
  - Actor-Context-Actor Relation (ACAR)[2]
  - Long-term Feature banks

1. Feichtenhofer et. al. Slowfast Networks for Video Recognition. ICCV, 2019.
2. Pan et. al. Actor-Context-Actor Relation Network for Spatio-Temporal Action Localization. In submission 2020.

Overview of our solution

- SlowFast[1] backbone + Relation Reasoning
  - Actor-Context-Actor Relation (ACAR)[2]
  - Long-term Feature banks

- Training on both AVA and Kinetics datasets
  - Use whole Kinetics-700 for pretraining (classification task)
  - Use AVA-Kinetics to train Reasoning Module (localization task)

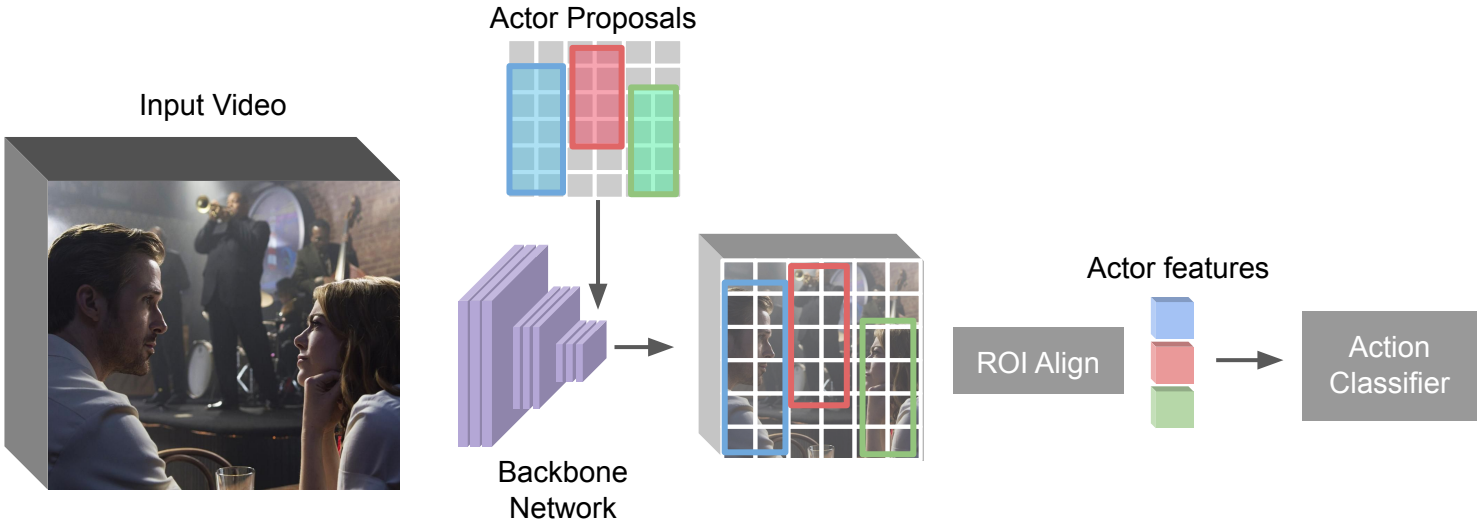
1. Feichtenhofer et. al. Slowfast Networks for Video Recognition. ICCV, 2019.
2. Pan et. al. Actor-Context-Actor Relation Network for Spatio-Temporal Action Localization. In submission 2020.

Overview of our solution

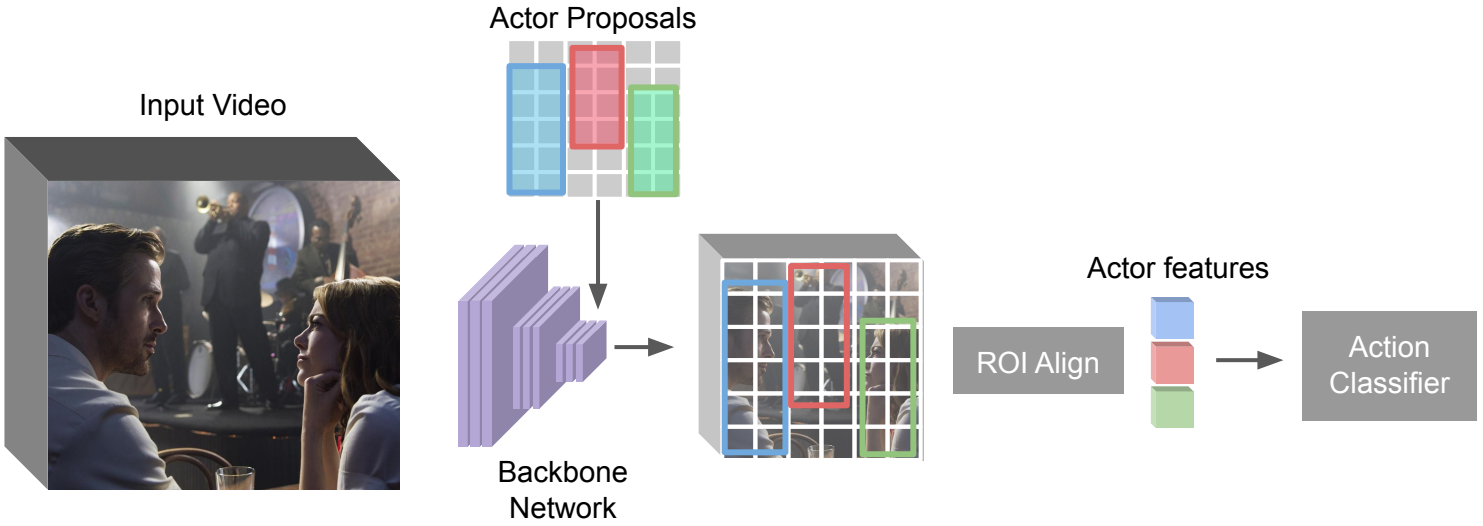
- Code and model will be released at:
  - https://github.com/Siyu-C/ACAR-Net

- Full preprint of ACAR (winning solution):
  - https://arxiv.org/pdf/2006.07976.pdf

Detection Pipeline - Simple Approach

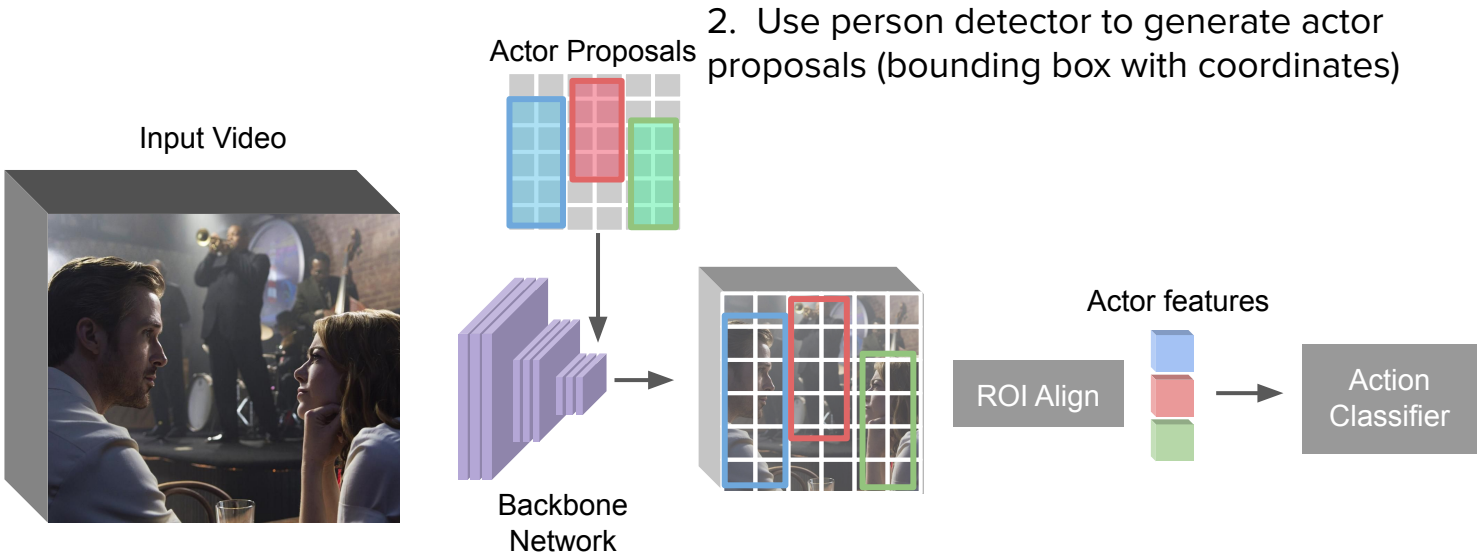


Detection Pipeline - Simple Approach

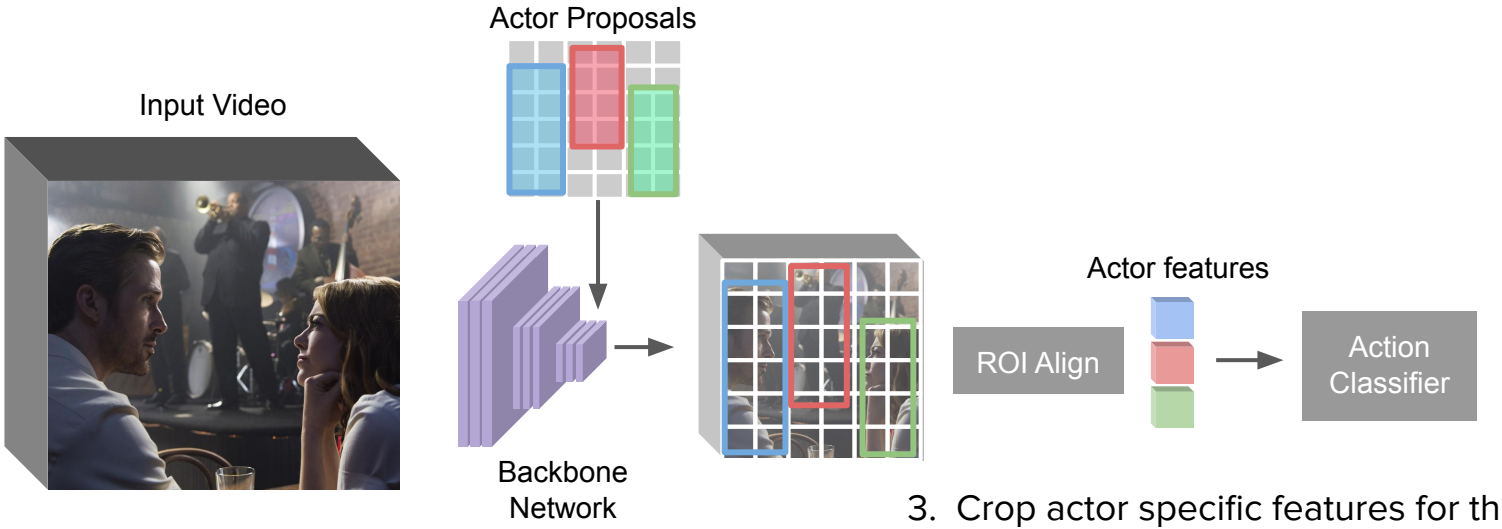


1. We extract spatio-temporal feature with 3D-Convnets.

Detection Pipeline - Simple Approach

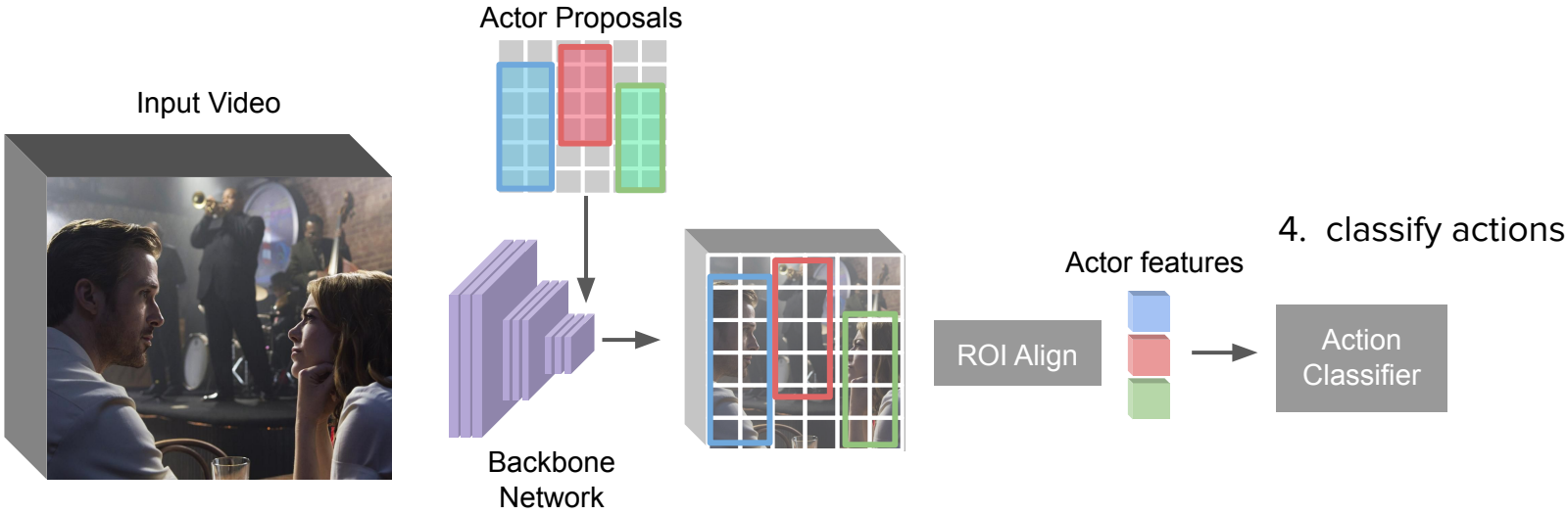


Detection Pipeline - Simple Approach

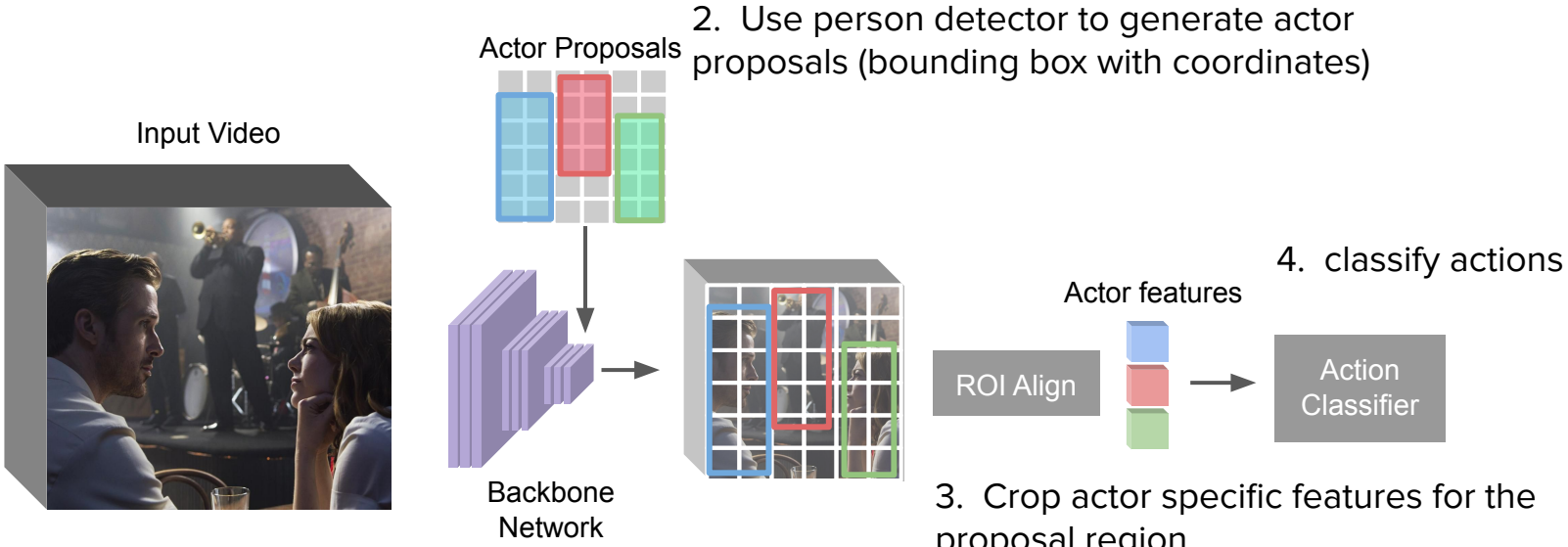


3. Crop actor specific features for the proposal region.

Detection Pipeline - Simple Approach



Detection Pipeline - Simple Approach



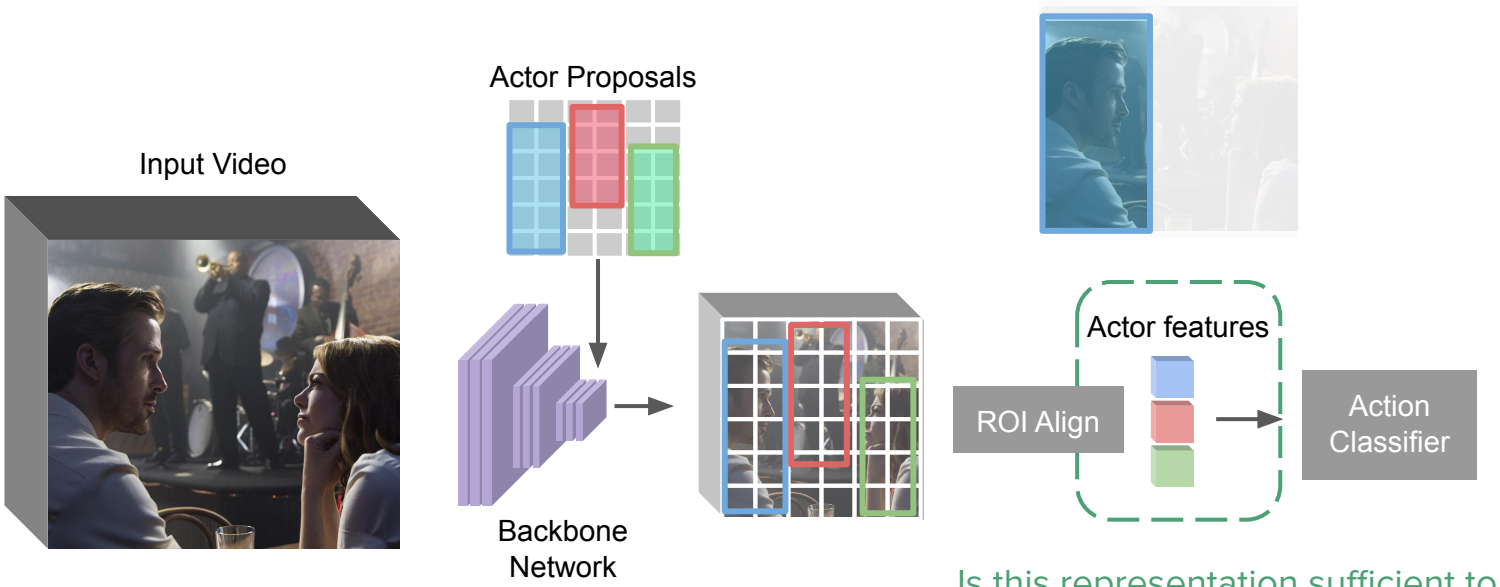
2. Use person detector to generate actor proposals (bounding box with coordinates)

4. classify actions

3. Crop actor specific features for the proposal region.

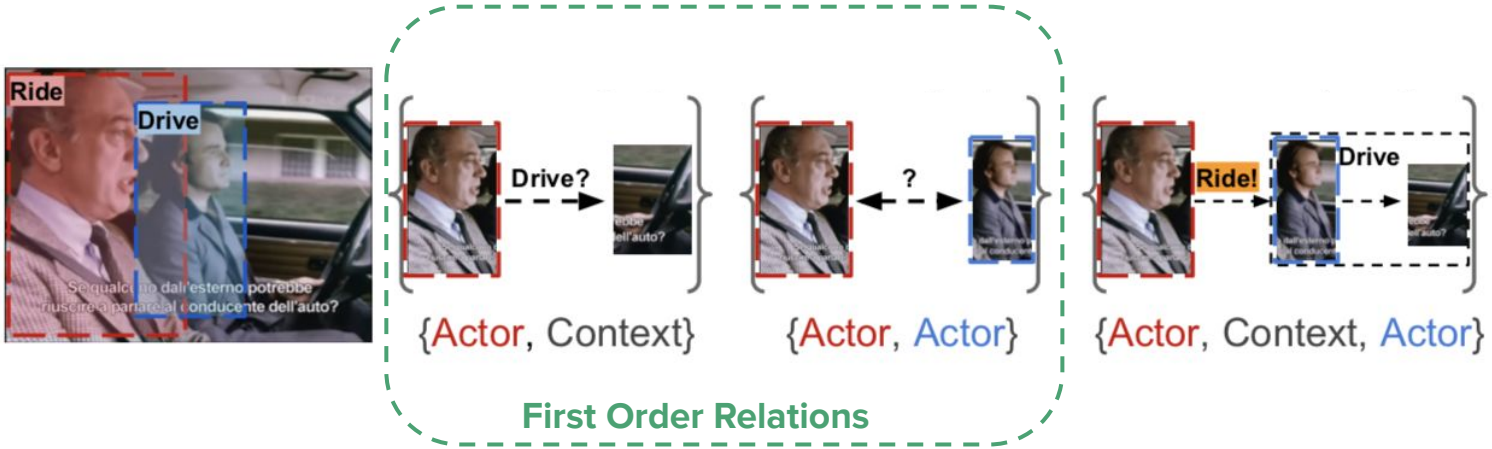
1. We extract spatio-temporal feature with 3D-Convnets.

Detection Pipeline - Simple Approach



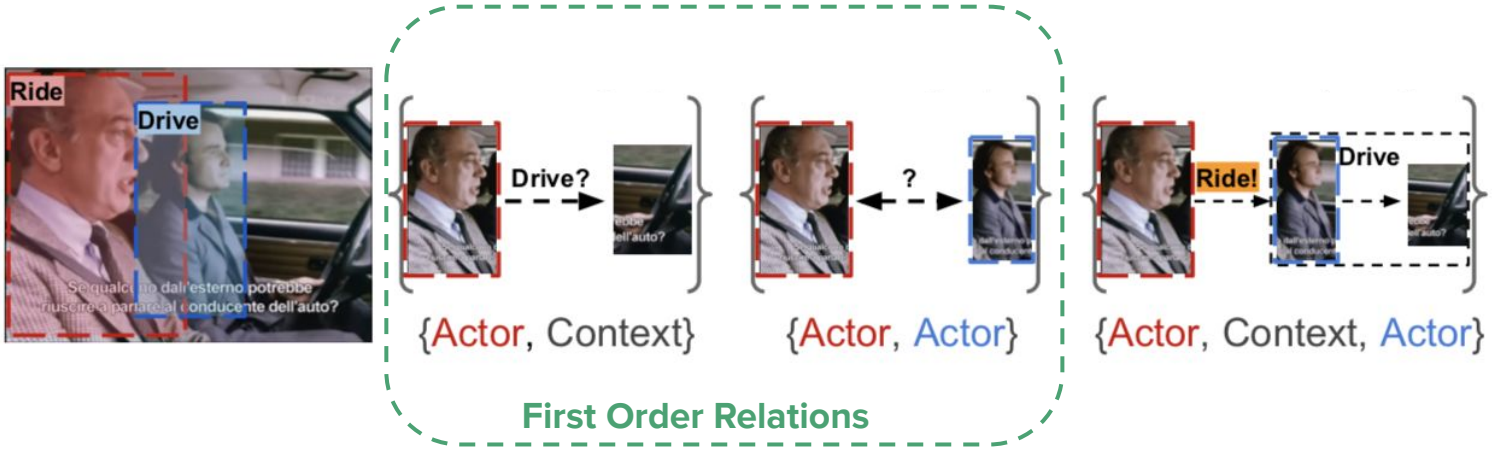
Is this representation sufficient to recognize actions?

Insights



Pan et. al. Actor-Context-Actor Relation Network for Spatio-Temporal Action Localization. In submission 2020.

Insights

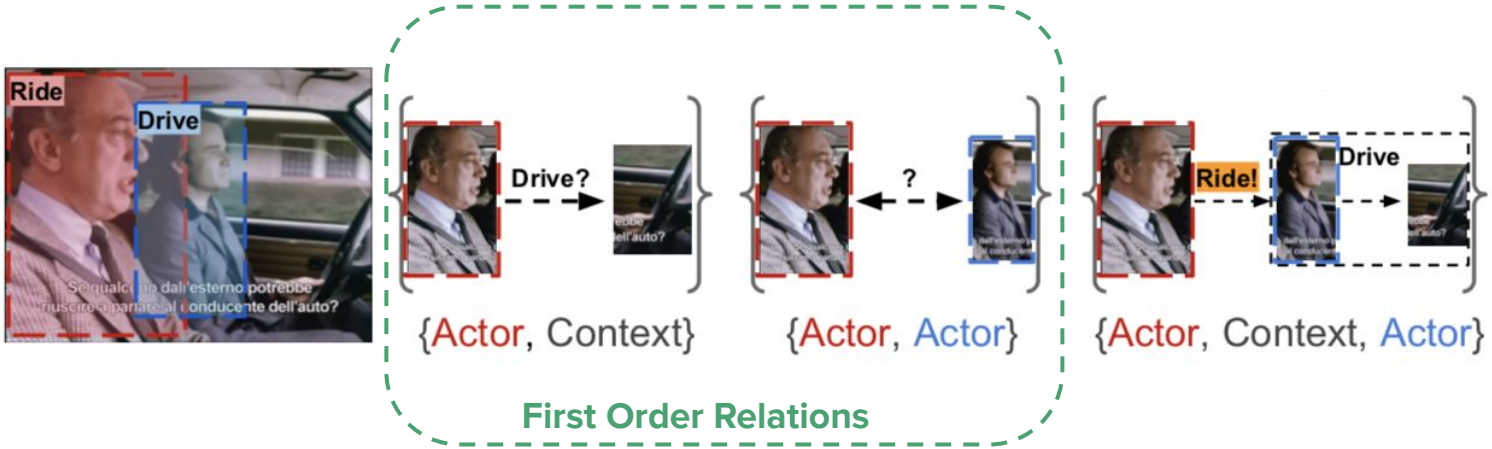


- Relation between **the blue actor** and the steering wheel (drive)
  >> clue for recognizing the action of **the red actor**

- Connections between different actor-context relations.

Pan et. al. Actor-Context-Actor Relation Network for Spatio-Temporal Action Localization. In submission 2020.

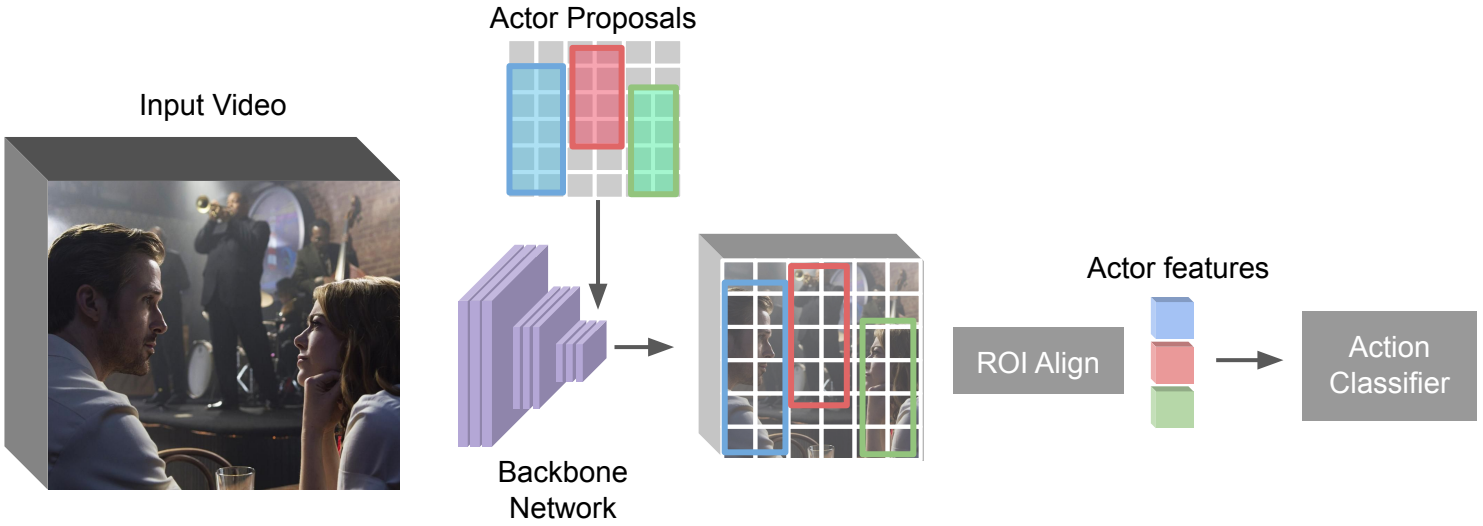
Insights



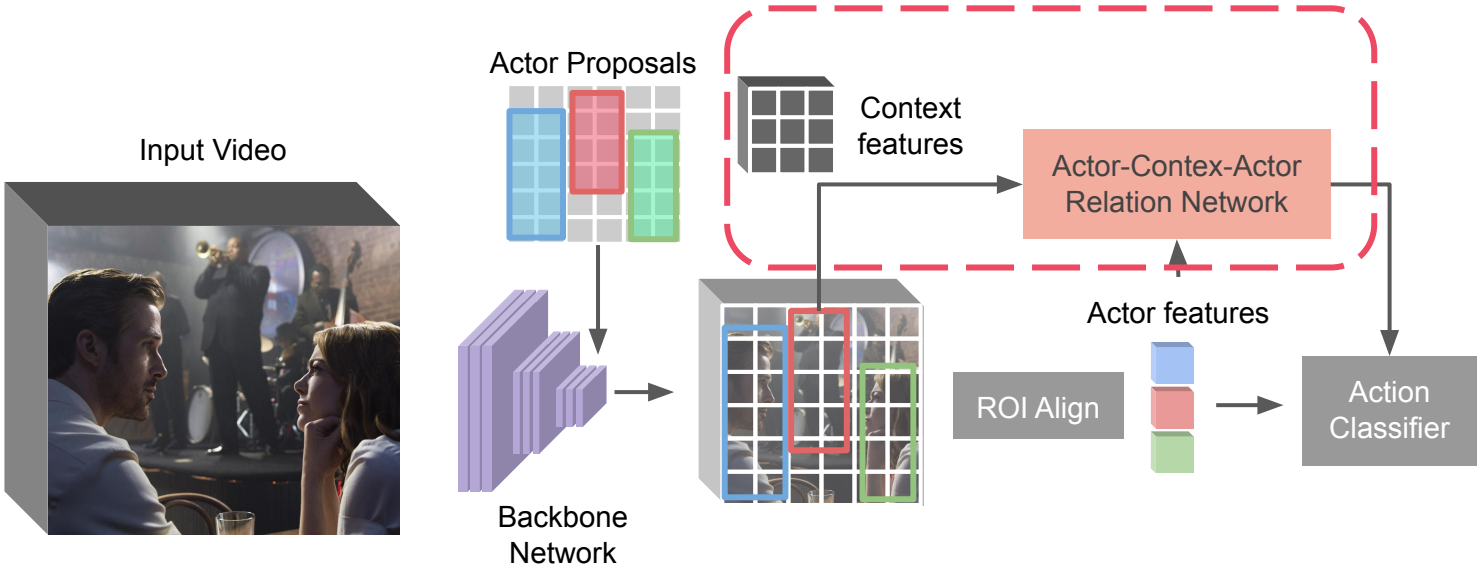
- **<u>Actor-context-actor relations need to be modeled to achieve accurate localization</u>**

Pan et. al. Actor-Context-Actor Relation Network for Spatio-Temporal Action Localization. In submission 2020.

Detection Pipeline - Simple Approach

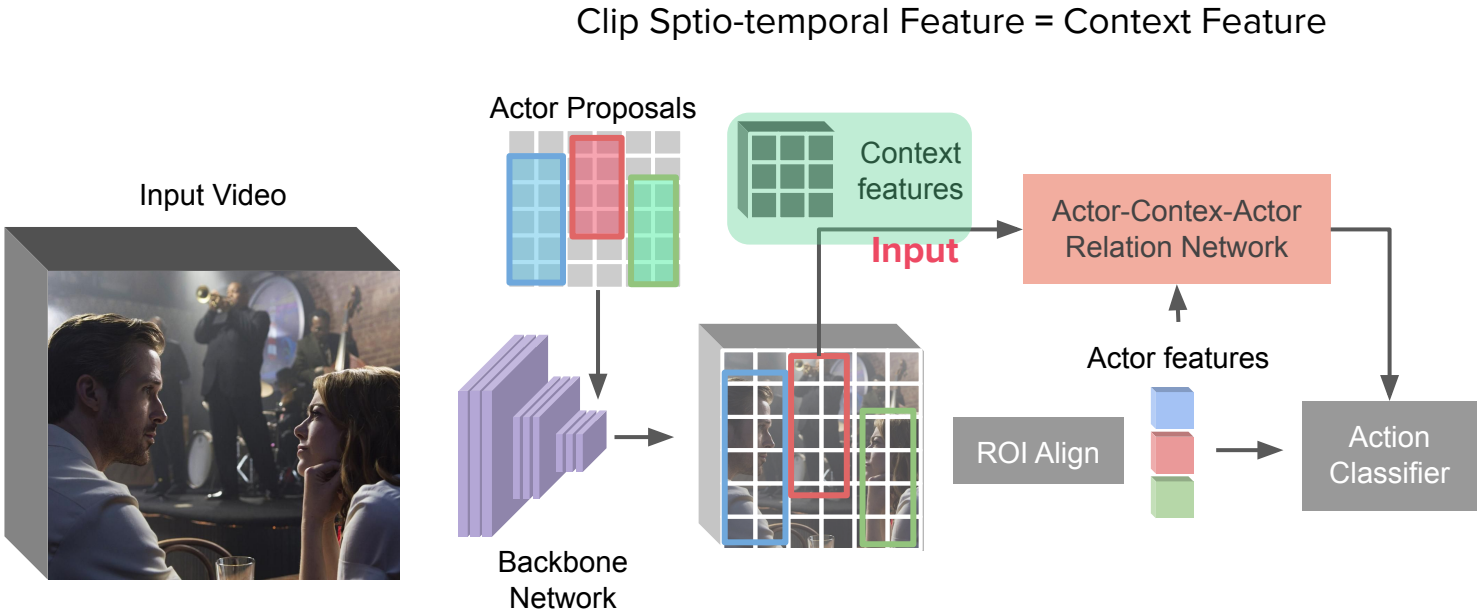


Detection Pipeline - Actor-Context-Actor Relation



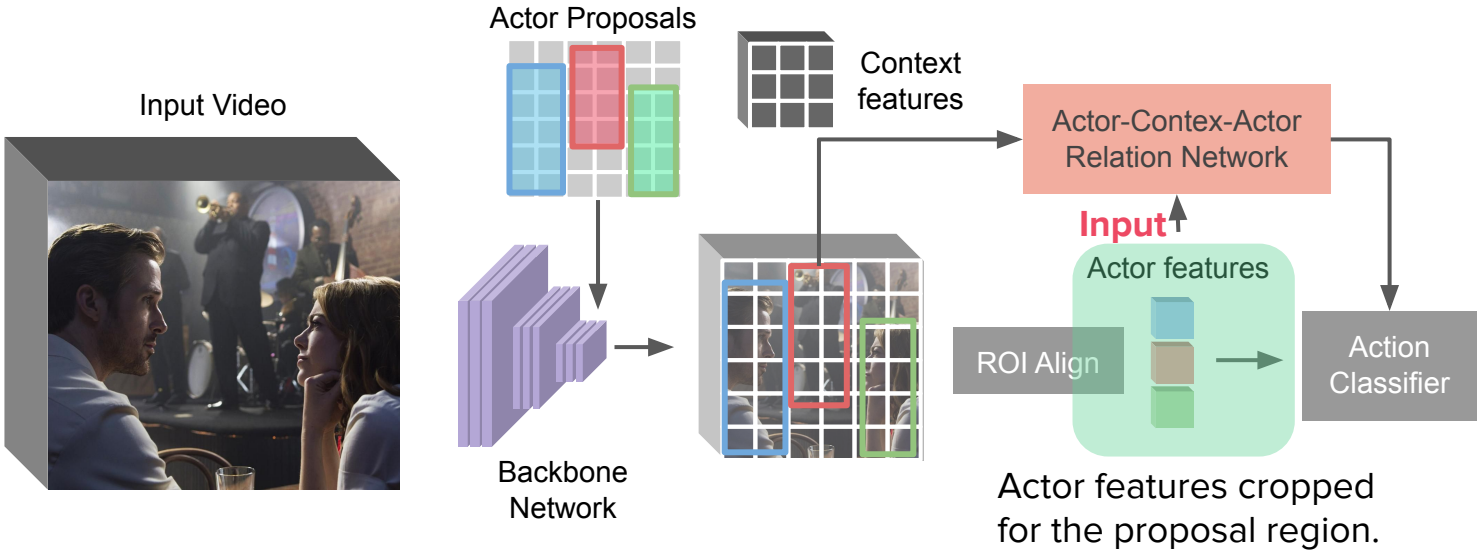
Pan et. al. Actor-Context-Actor Relation Network for Spatio-Temporal Action Localization. In submission 2020.

Detection Pipeline - Actor-Context-Actor Relation



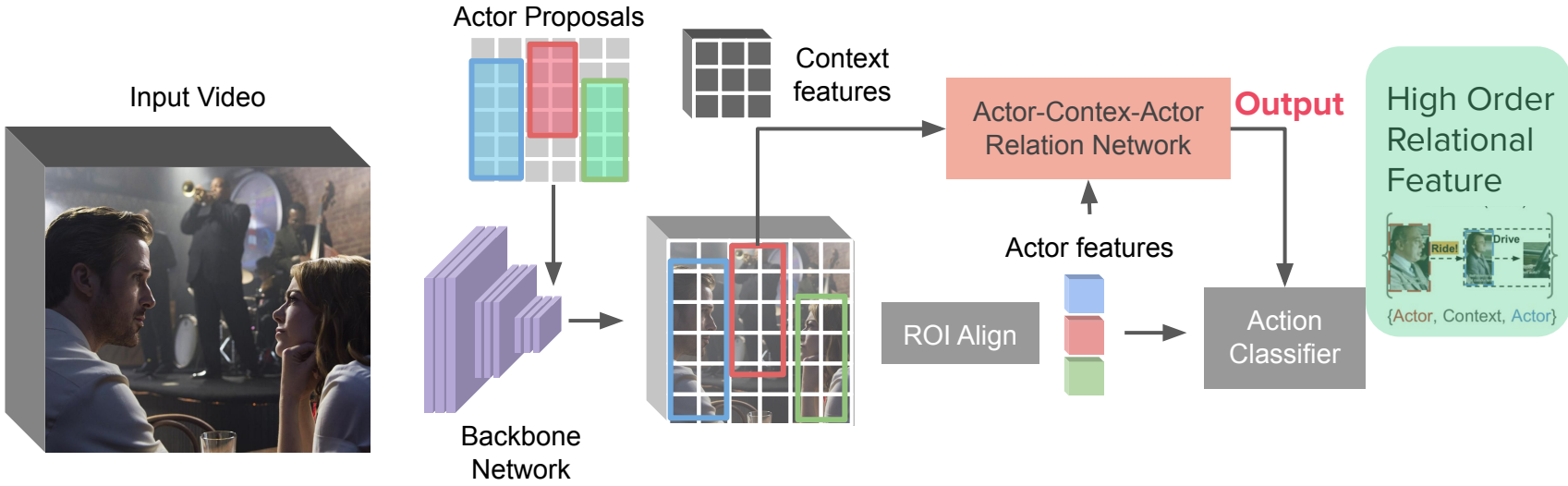
Pan et. al. Actor-Context-Actor Relation Network for Spatio-Temporal Action Localization. In submission 2020.

Detection Pipeline - Actor-Context-Actor Relation



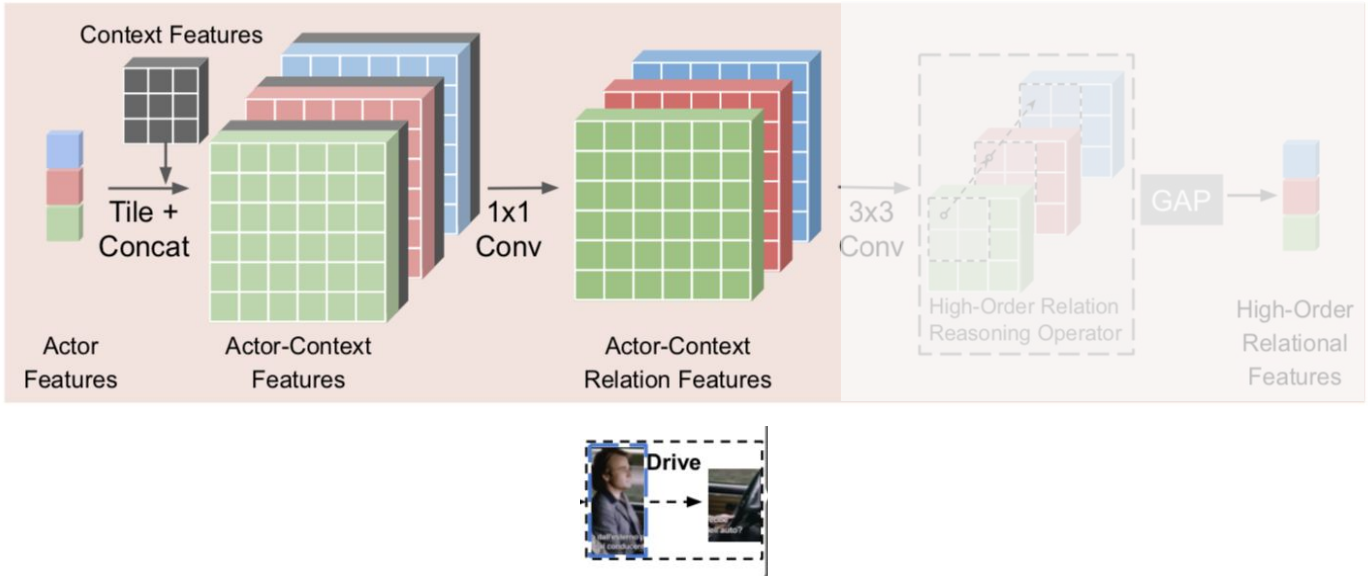
Pan et. al. Actor-Context-Actor Relation Network for Spatio-Temporal Action Localization. In submission 2020.

Detection Pipeline - Actor-Context-Actor Relation



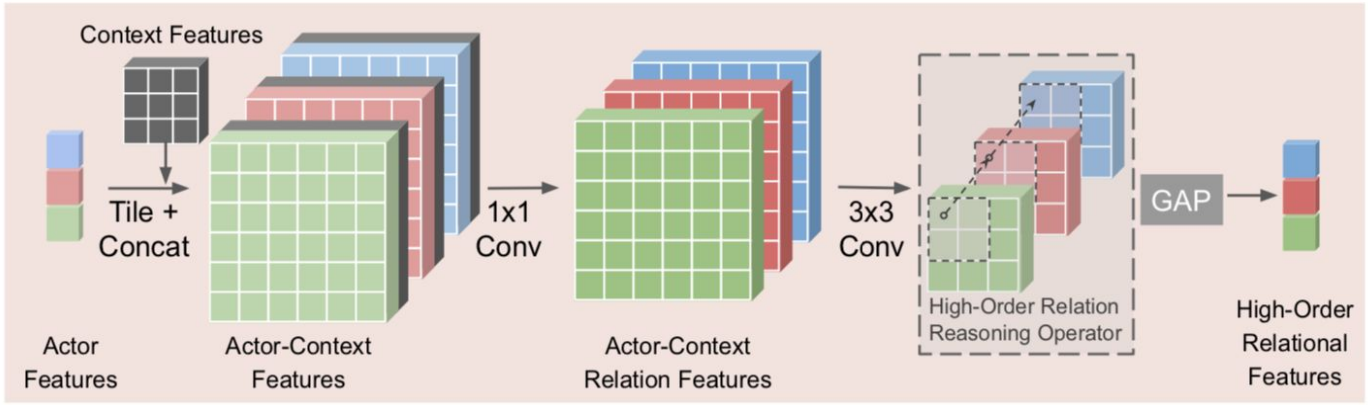
Pan et. al. Actor-Context-Actor Relation Network for Spatio-Temporal Action Localization. In submission 2020.

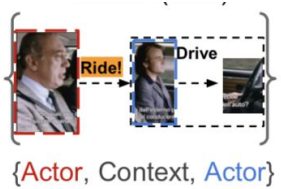
Actor-Context-Actor Relation



Pan et. al. Actor-Context-Actor Relation Network for Spatio-Temporal Action Localization. In submission 2020.

Actor-Context-Actor Relation

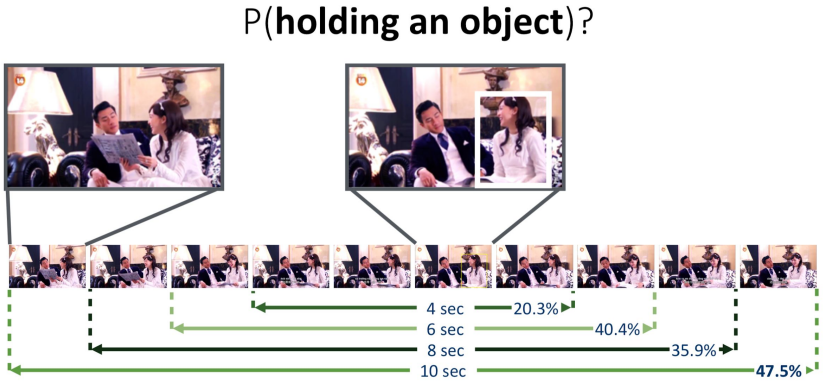




Pan et. al. Actor-Context-Actor Relation Network for Spatio-Temporal Action Localization. In submission 2020.

Long-term Temporal Reasoning

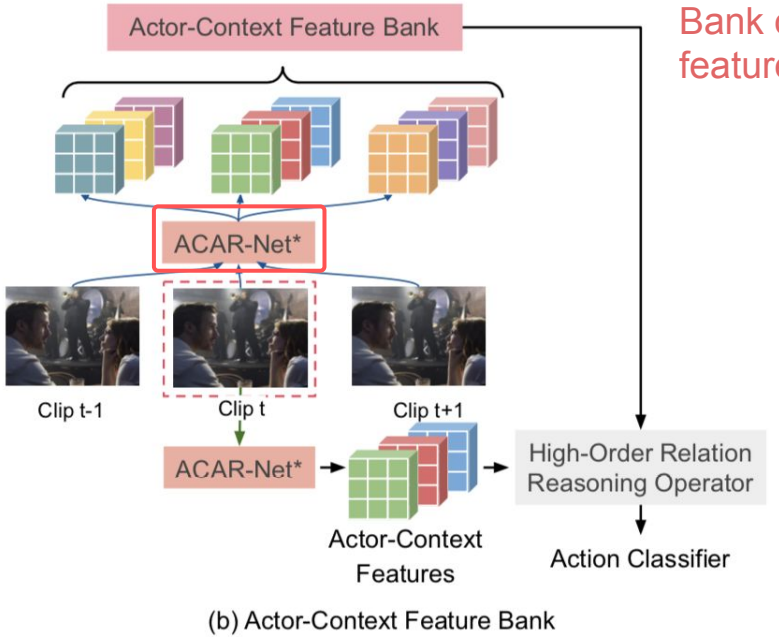
- For Complex scenes
  - Watch longer to understand better.
  - But current model is limited to 2-4s
- Long term Feature Banks (Wu et. al.)
  - Precomputed action features
- Actor-Context Feature Banks (ACFB)
  - Precomputed actor-context relation features.



(Picture from ICCV'19 Recognition in Video tutorial)

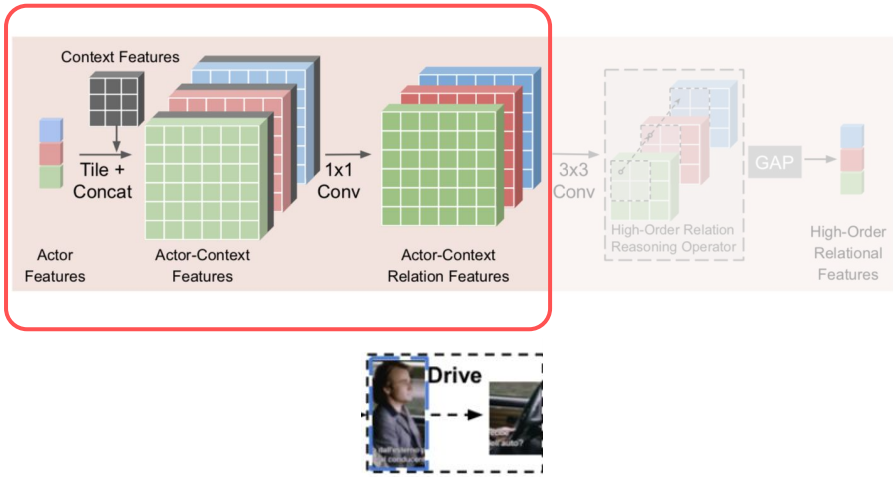
Wu et. al. Long-Term Feature Banks for Detailed Video Understanding. CVPR, 2019.

Actor-Context Feature Bank



(b) Actor-Context Feature Bank

Bank of first-order relation features.



Pan et. al. Actor-Context-Actor Relation Network for Spatio-Temporal Action Localization. In submission 2020.

Summary of Architectures

- Backbone
  - SlowFast (R101, R152) with input sampling 8x8 and 16x8

1. Song et. al. Revisiting the Sibling Head in Object Detector. CVPR, 2020.

Summary of Architectures

- Backbone
  - SlowFast (R101, R152) with input sampling 8x8 and 16x8
- Actor Proposal
  - Faster-RCNN with SENet-154-FPN-TSD[1]

1. Song et. al. Revisiting the Sibling Head in Object Detector. CVPR, 2020.

Summary of Architectures

- Backbone
  - SlowFast (R101, R152) with input sampling 8x8 and 16x8
- Actor Proposal
  - Faster-RCNN with SENet-154-FPN-TSD[1]
- Heads
  - Linear Classifier
  - Actor-Context-Actor Relation (ACAR)

1. Song et. al. Revisiting the Sibling Head in Object Detector. CVPR, 2020.

Summary of Architectures

- Backbone
  - SlowFast (R101, R152) with input sampling 8x8 and 16x8
- Actor Proposal
  - Faster-RCNN with SENet-154-FPN-TSD[1]
- Heads
  - Linear Classifier
  - Actor-Context-Actor Relation (ACAR)
- Long-term Support
  - Actor Feature Bank (Wu et. al)
  - Actor-Context Feature Bank (ACFB)

1. Song et. al. Revisiting the Sibling Head in Object Detector. CVPR, 2020.

Implementation Details

- Pre-train
  - All backbones pre-trained on **Kinetics-700** classification task

Implementation Details

- Pre-train
  - All backbones pre-trained on **Kinetics-700** classification task

- Training schedule
  - Short schedule: **Only 6 epochs**
  - Linear warm-up, stepwise decay

Implementation Details

- Pre-train
  - All backbones pre-trained on **Kinetics-700** classification task

- Training schedule
  - Short schedule: **Only 6 epochs**
  - Linear warm-up, stepwise decay

- Training augmentations
  - **NO** scale jittering / random crop
  - Scale shorter side to 256
  - BBox jittering

Implementation Details

- Default inference
  - Scale shorter side to 256

Implementation Details

- Default inference
  - Scale shorter side to 256

- Ensemble & test
  - For test, train on both training and validation data
  - 3 scales [256, 288, 320] & horizontal flips
  - 20 models 40.49mAP (val) / **39.62mAP** (test)
    (For each action class, assign ensemble weights to models according to their APs on this class[1])

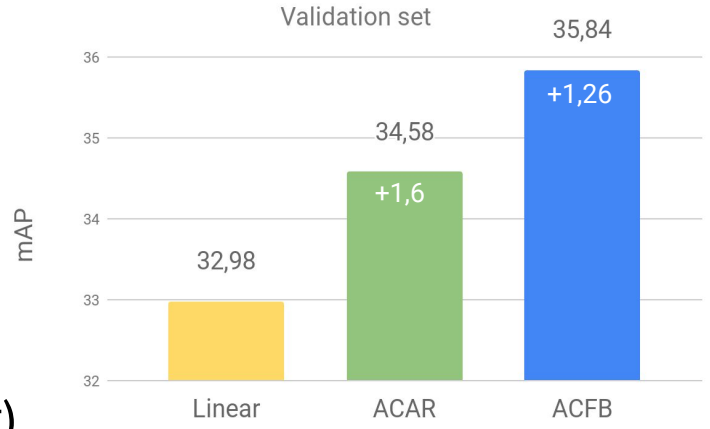
1. Akiba et. al. PFDet: 2nd Place Solution to Open Images Challenge 2018 Object Detection Track. preprint

Single Model Performance on **AVA-Kinetics**

- Default backbone: SlowFast R101 8x8
  - Linear 32.98mAP
  - ACAR 34.58mAP
  - ACFB **35.84mAP**

+2.86

- SlowFast R152 8x8
  - ACAR 35.12mAP(val) / **34.99mAP** (test)

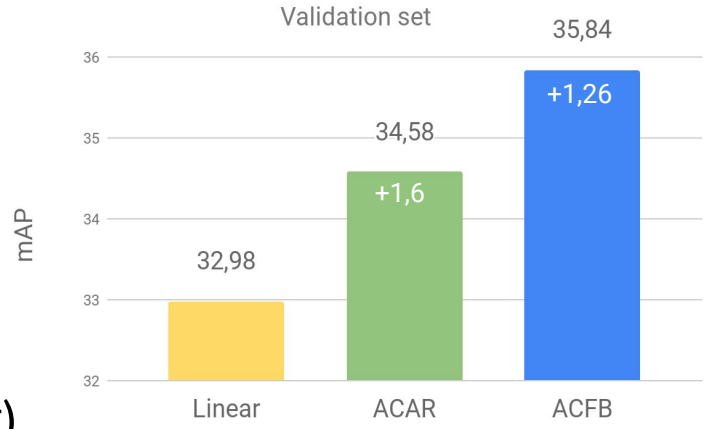


Single Model Performance on **AVA-Kinetics**

- Default backbone: SlowFast R101 8x8
  - Linear 32.98mAP
  - ACAR 34.58mAP
  - ACFB **35.84mAP**

+0.54

- SlowFast R152 8x8
  - ACAR 35.12mAP(val) / **34.99mAP** (test)



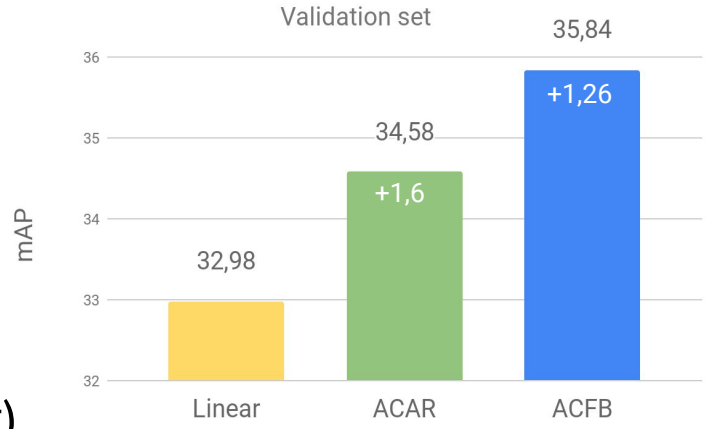
Single Model Performance on **AVA-Kinetics**

- Default backbone: SlowFast R101 8x8
  - Linear 32.98mAP
  - ACAR 34.58mAP
  - ACFB **35.84mAP**

+0.54

- SlowFast R152 8x8
  - ACAR 35.12mAP(val) / **34.99mAP** (test)

-0.13

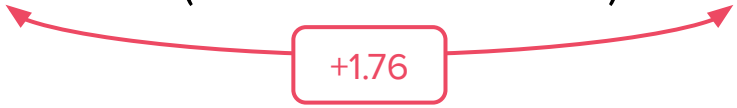


Single Model Performance on **AVA v2.2**

- Gain on AVA by adding Kinetics (SlowFast R101 8x8 + ACAR)
  - **34.15mAP** (AVA-Kinetics train) / 32.29mAP (AVA only train)

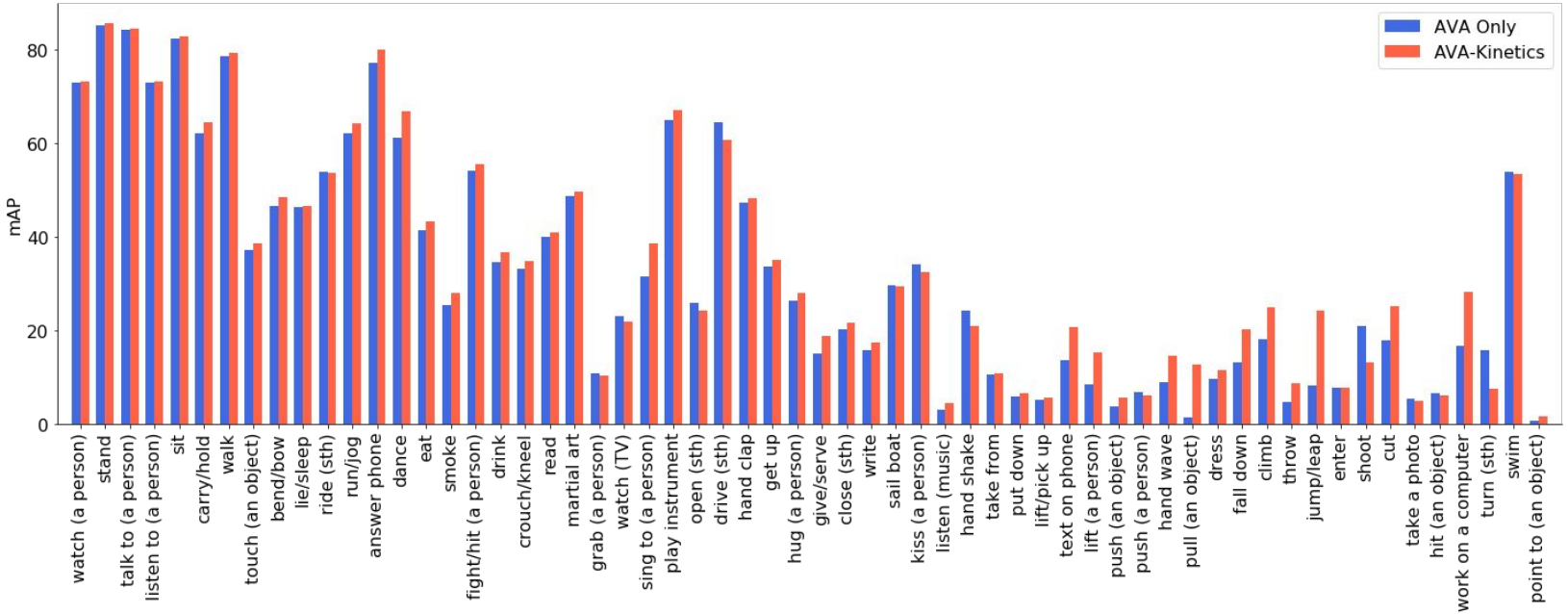
Single Model Performance on **AVA v2.2**

- Gain on AVA by adding Kinetics (SlowFast R101 8x8 + ACAR)
  - **34.15mAP** (AVA-Kinetics train) / 32.29mAP (AVA only train)



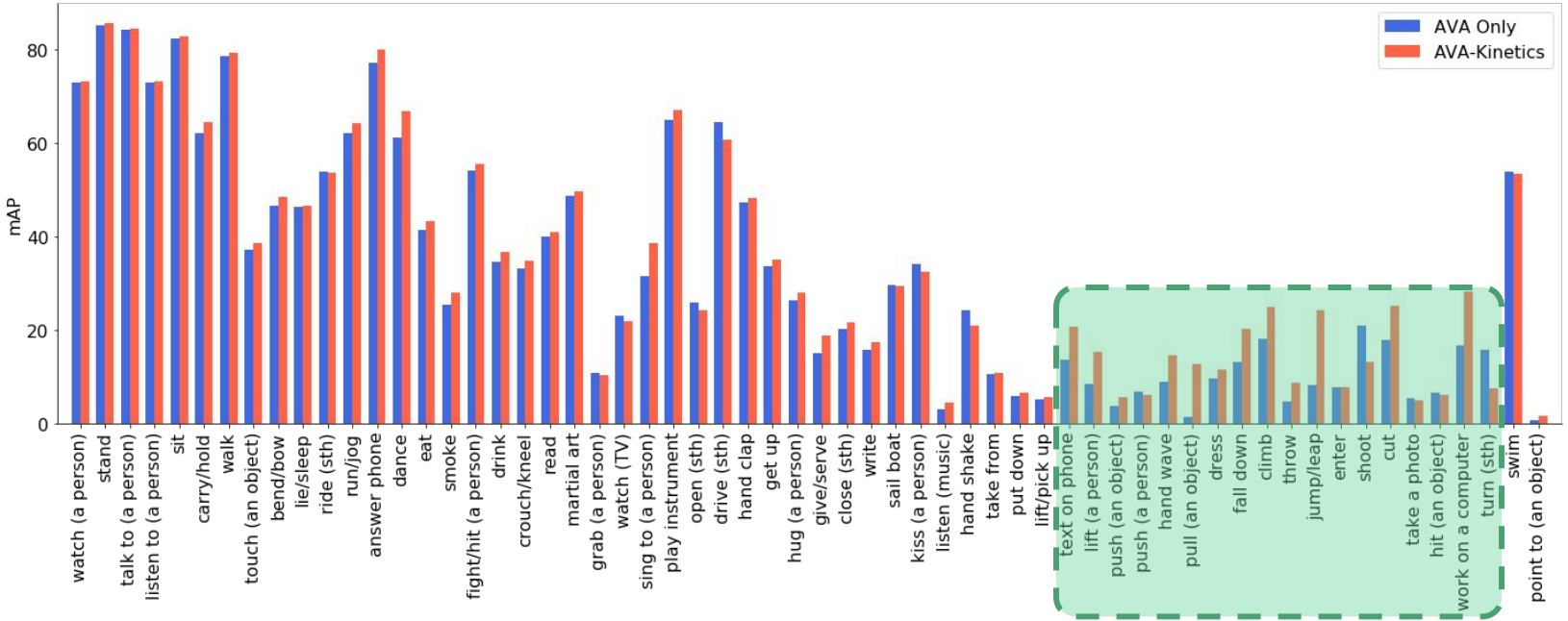
Single Model Performance on **AVA v2.2**

- Gain on AVA by adding Kinetics (SlowFast R101 8x8 + ACAR)
  - **34.15mAP** (AVA-Kinetics train) / 32.29mAP (AVA only train)



Single Model Performance on **AVA v2.2**

- Gain on AVA by adding Kinetics (SlowFast R101 8x8 + ACAR)
  - **34.15mAP** (AVA-Kinetics train) / 32.29mAP (AVA only train)



Single Model Performance on **AVA v2.2**

- Finetune two detectors for AVA and Kinetics respectively
- 95.8AP@50 on AVA vs 84.4AP@50 on Kinetics
  (77.2AP@50 if we use the same AVA detector)

Single Model Performance on **AVA v2.2**

- Finetune two detectors for AVA and Kinetics respectively
- 95.8AP@50 on AVA vs 84.4AP@50 on Kinetics
  (77.2AP@50 if we use the same AVA detector)

- Effect of Different Person Detectors on **AVA**
  (SlowFast R101 8x8 + ACAR trained on AVA-Kinetics)
  - 33.73mAP (Detection from PySlowFast[1], 93.9AP@50)
  - 34.15mAP (TSD[2], 95.8AP@50)
  - **42.25mAP** (Ground Truth BBox)

+8.10

1.https://github.com/facebookresearch/SlowFast/blob/master/slowfast/datasets/DATASET.md
2.Song et. al. Revisiting the Sibling Head in Object Detector. CVPR, 2020.

Single Model Performance on **Kinetics v1.0**

- Finetune two detectors for AVA and Kinetics respectively
- 95.8AP@50 on AVA vs 84.4AP@50 on Kinetics
  (77.2AP@50 if we use the same AVA detector)

- Effect of Different Person Detectors on **Kinetics**
  (SlowFast R101 8x8 + ACAR trained on AVA-Kinetics)
  - 28.41mAP (AVA detector, 77.2AP@50)
  - 30.88mAP (Kinetics detector, 84.4AP@50)
  - **43.60mAP** (Ground Truth BBox)

+12.72

- Overview of the task and challenge
- Details of our solution
- **Analysis and Rethinking**

Analysis and Rethinking

- Failure Case

[GT]: get up
Prediction: [0.51] bend

[GT]: watch, stand, carry, turn
Prediction: [0.99] stand, [0.96] carry

Analysis and Rethinking

- Missing Ground Truth (frames that were not annotated)

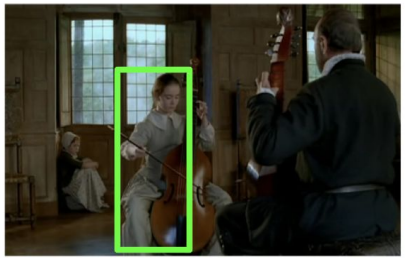


Analysis and Rethinking

- Missing Ground Truth (frames that were not annotated)



- Inconsistent annotations:
  - hold, carry object / cut / read / play instrument / write / smoke ...


[GT] play instrument, sit


[GT] stand, cut

[GT] Talk to , Read, Stand

Future Directions

- Current solution struggles at detect fine-grained details
  - Fine-grained Action Recognition

- Current performance is poor on tail classes
  - Deal with Long-tailed Distribution

- Given Kinetics with both clip level & spatio-temporal annotations
  - Multitask learning (classification & localization)

Acknowledgement

- Guanglu Song (for providing the person detectors)
- Manyuan Zhang and Hao Shao (for providing the pre-trained models)
- Ziyi Lin (for helping optimizing training speed)
- Zheng Shou (for insightful discussions)
- PySlowFast codebase

Thanks for watching!

Q&A

- Please feel free to contact us if you have any question! ;)
  - junting.pa@gmail.com
  - microrunnerup@gmail.com

- Code and model (coming soon):
  - https://github.com/Siyu-C/ACAR-Net
- Full preprint of ACAR:
  - https://arxiv.org/pdf/2006.07976.pdf