Domain Adaptation in Action Detection using Video Clip Order Prediction

Presenter: Rishabh Singh
Mentors: Dr. Yuhua Chen, Dr. Suman Saha
Supervisor: Prof. Dr. Luc Van Gool

ETH zürich

Overview

- Motivation
- Proposed Method
- Experiments
- Results
- Conclusions

Overview

Motivation

Proposed Method

Experiments

Results

Conclusions

Motivation

- Why Domain Adaptation for Action Detection?
- Why Self-Supervision for Domain Adaptation?
- Why Clip-order prediction as Self-Supervision?

Motivation

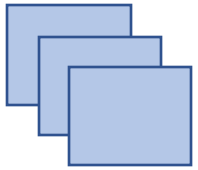
- Why Domain Adaptation for Action Detection?
- Why Self-Supervision for Domain Adaptation?
- Why Clip-order prediction as Self-Supervision?

Domain Shift in Action Detection

Video Input

Model

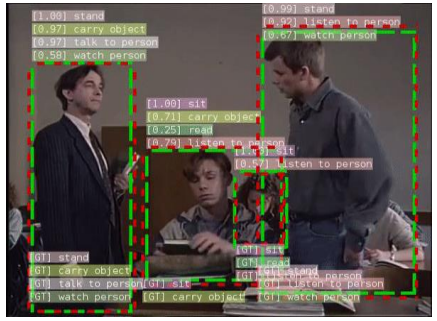
Source:



DNN

Label (Classification)

Bounding Box Coordinates
(Localisation)

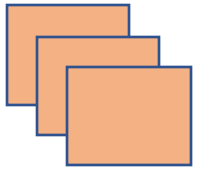


New Domain

Video Input

Model

Target:

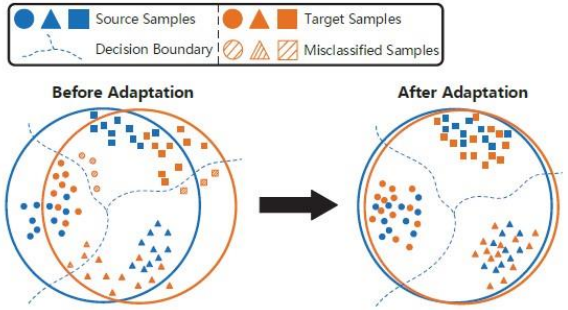


DNN

Accuracy

Drop
in performance
during Test

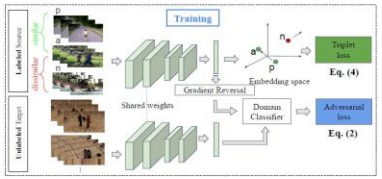
What's missing?



[1] Li, Mengxue, et al.

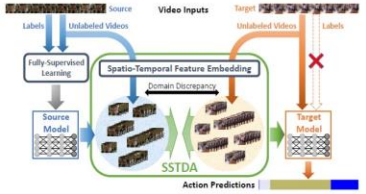
Domain Adaptation

Classification



[2] Choi, Jinwoo, et al.

Localisation (temporal)



[3] Chen, Min-Hung, et al.

No work on Domain Adaptation in Action Detection (Classification + Localisation)!

Motivation

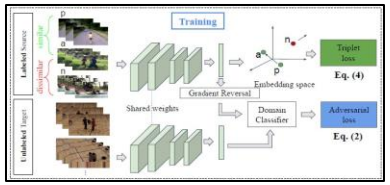
- Why Domain Adaptation for Action Detection?
- Why Self-Supervision for Domain Adaptation?
- Why Clip-order prediction as Self-Supervision?

Ongoing Research

Video Unsupervised Domain Adaptation

(still under-explored)

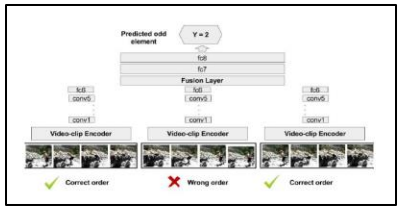
Adversarial Learning



[2] Choi, Jinwoo, et al.

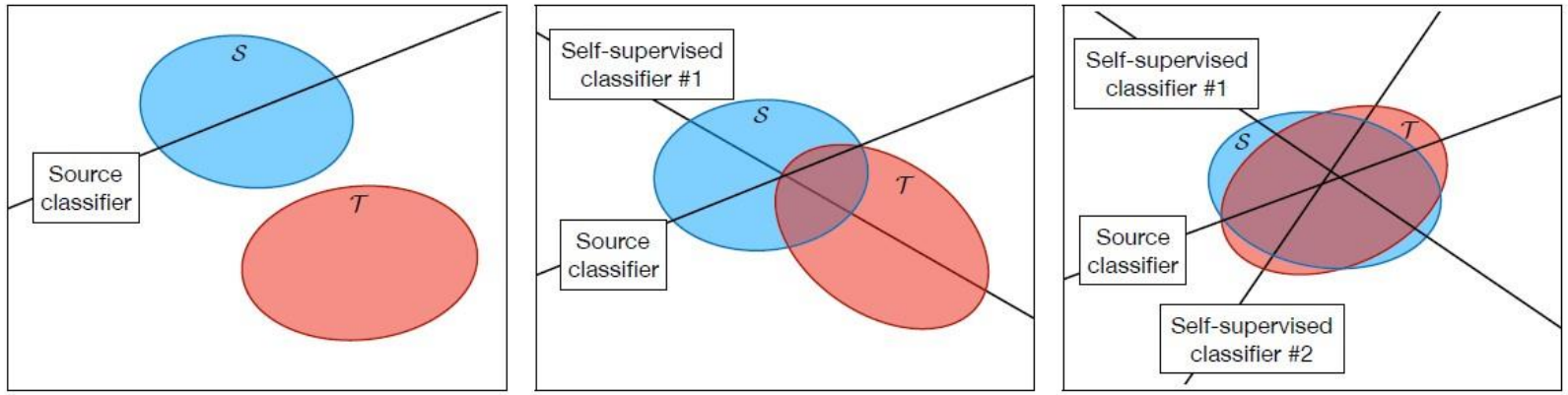
- No additional data
- Circumvent annotation of bounding boxes (expensive, cumbersome)
- Easier training

Self-Supervision



[4] Fernando, Basura, et al.

Alignment of Representations



(a) Source classifier only (b) adding a self-supervised task (c) adding more tasks

Self-Supervision aligns the learned representations of two domains in a shared feature space, thus enabling the source classifier to generalize better to the target domain [5]

Motivation

- Why Domain Adaptation for Action Detection?
- Why Self-Supervision for Domain Adaptation?
- Why Clip-order prediction as Self-Supervision?

Self-Supervision in Videos

Spatial

(2D-CNN)

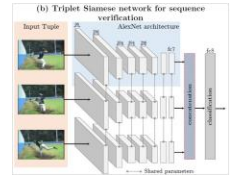
[7] Wang, Xiaolong, et al.

Spatio-temporal

(3D-CNN)

Temporal

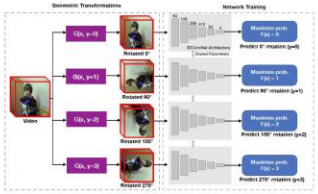
(2D-CNN)



[6] Misra, Ishan et al.

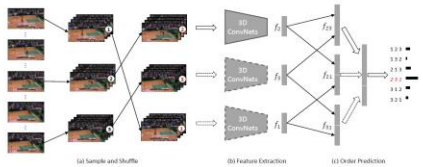
Spatial Pretext

Eg. Rotation



[8] Jing, Longlong, et al.

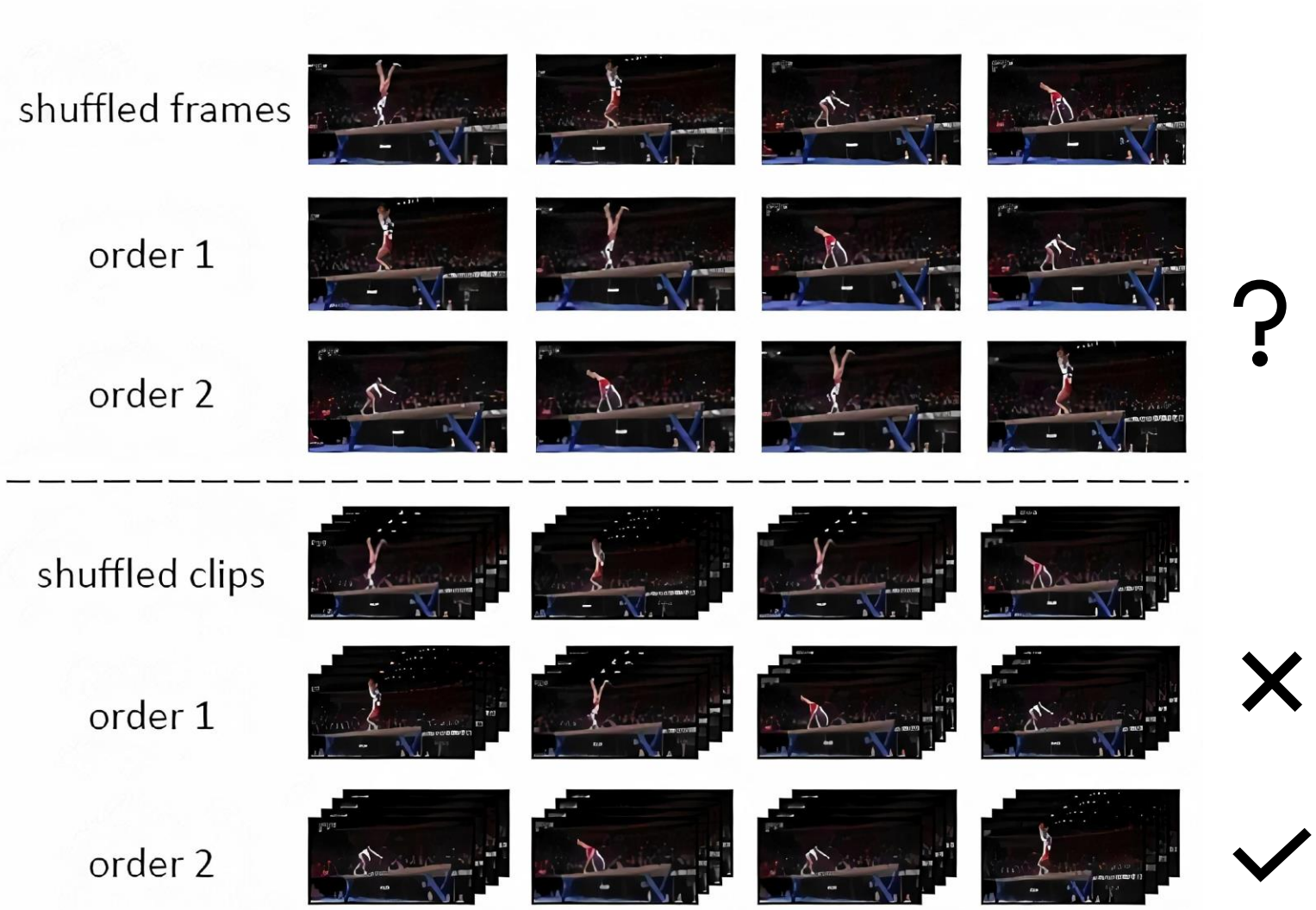
Temporal Pretext



Eg. Clip Order Prediction

[9] Xu, Dejing, et al.

Why Clip Order?



- Learns Temporal Coherence of Action Sequence
- Spatio-temporal features more aligned with 3D-CNN backbone of our Action Detector
- Temporal Pretext task resonates with temporal pathways of the backbone

Overview

Motivation

Proposed Method

Experiments

Results

Conclusions

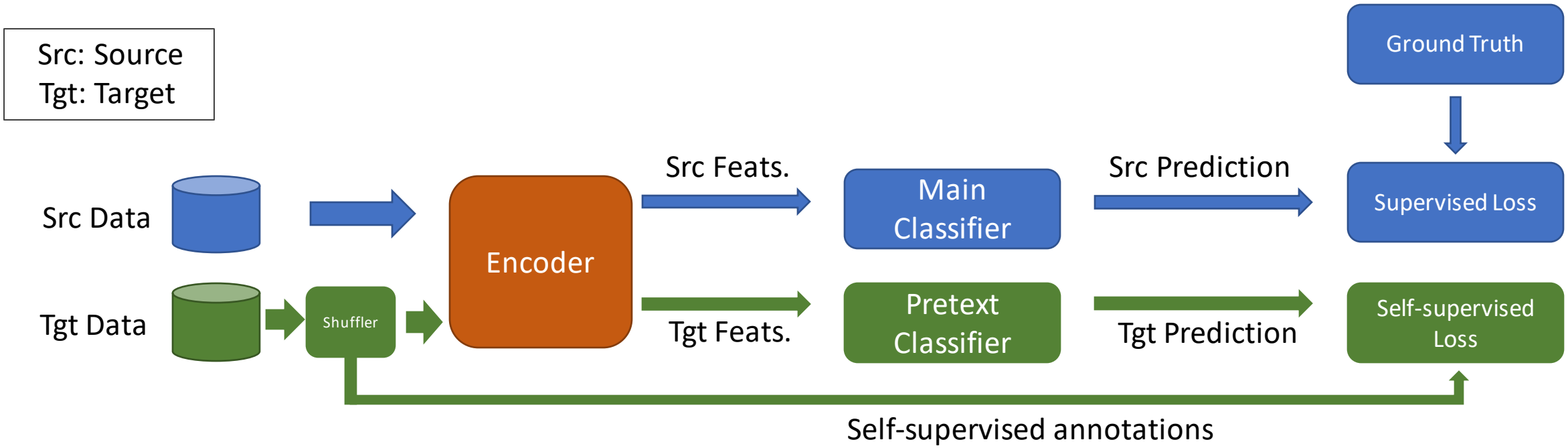
Proposed Method

- Overview
- Knowledge Transfer
- Encoder
- Action Detection
- Video clip order prediction

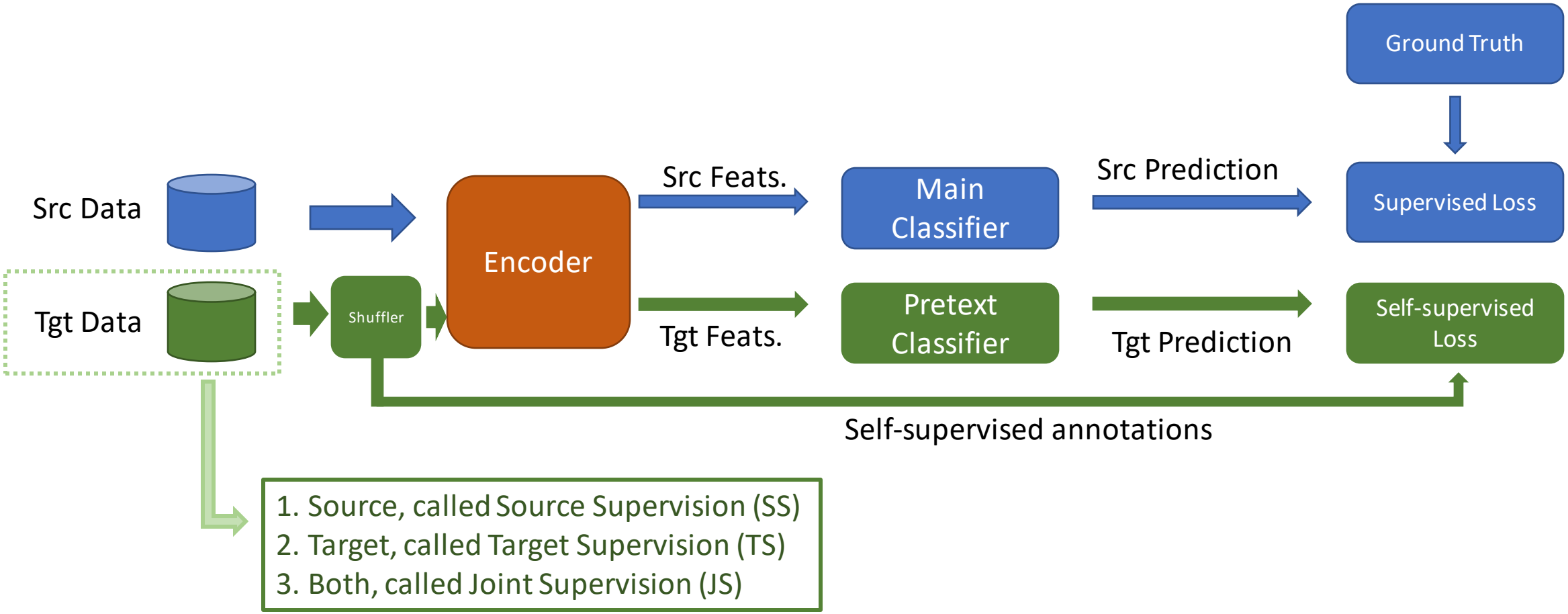
Proposed Method

- Overview
- Knowledge Transfer
- Encoder
- Action Detection
- Video clip order prediction

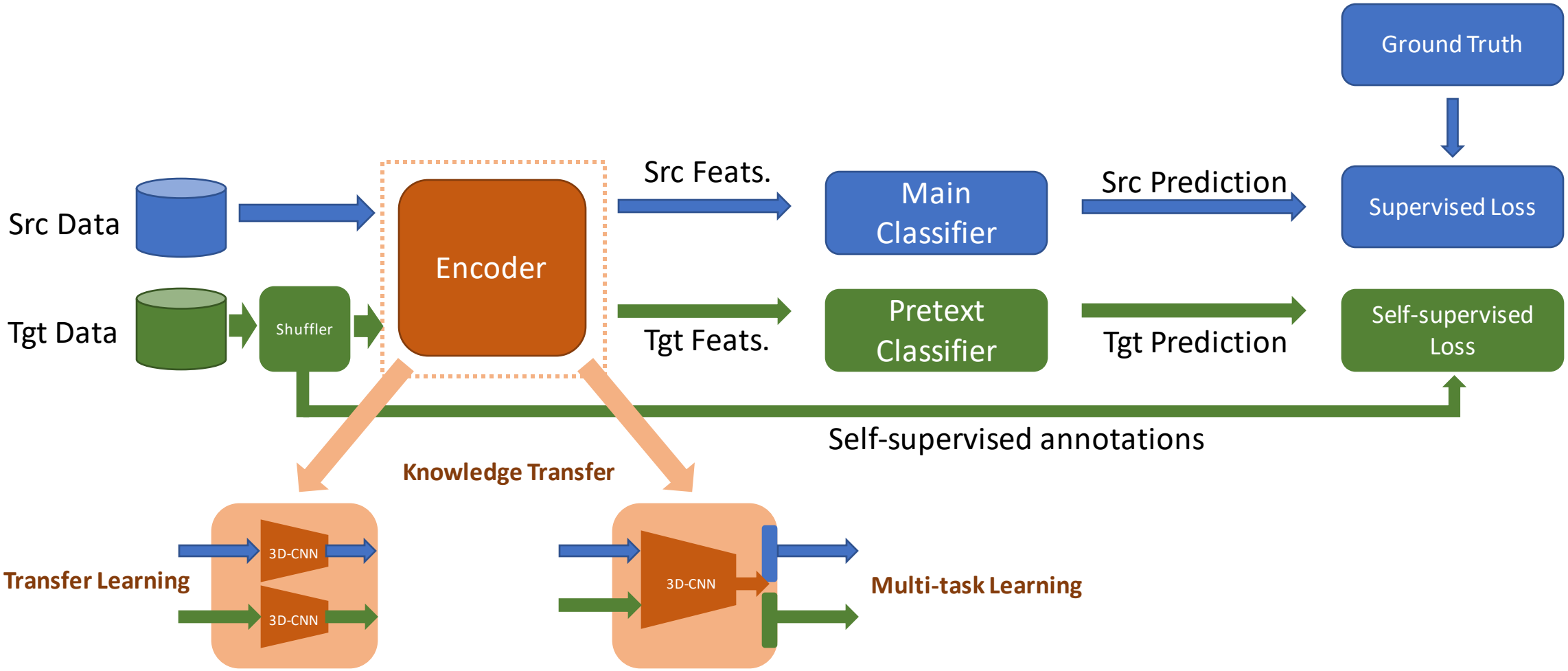
Proposed method



Proposed method



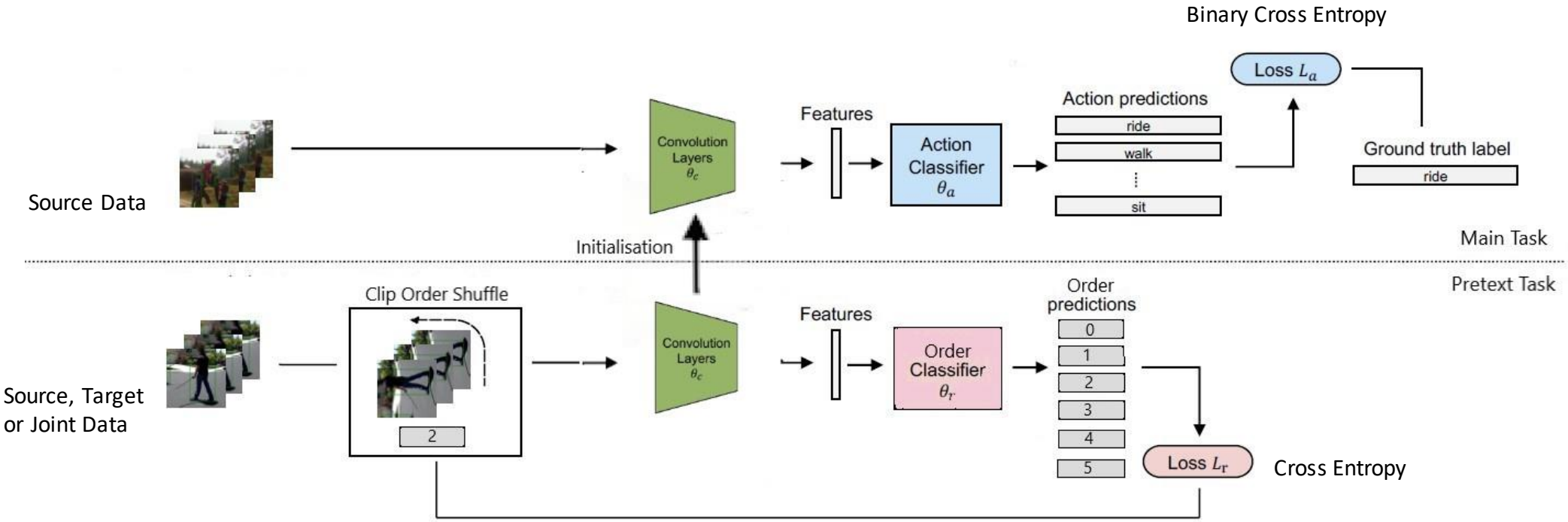
Proposed method



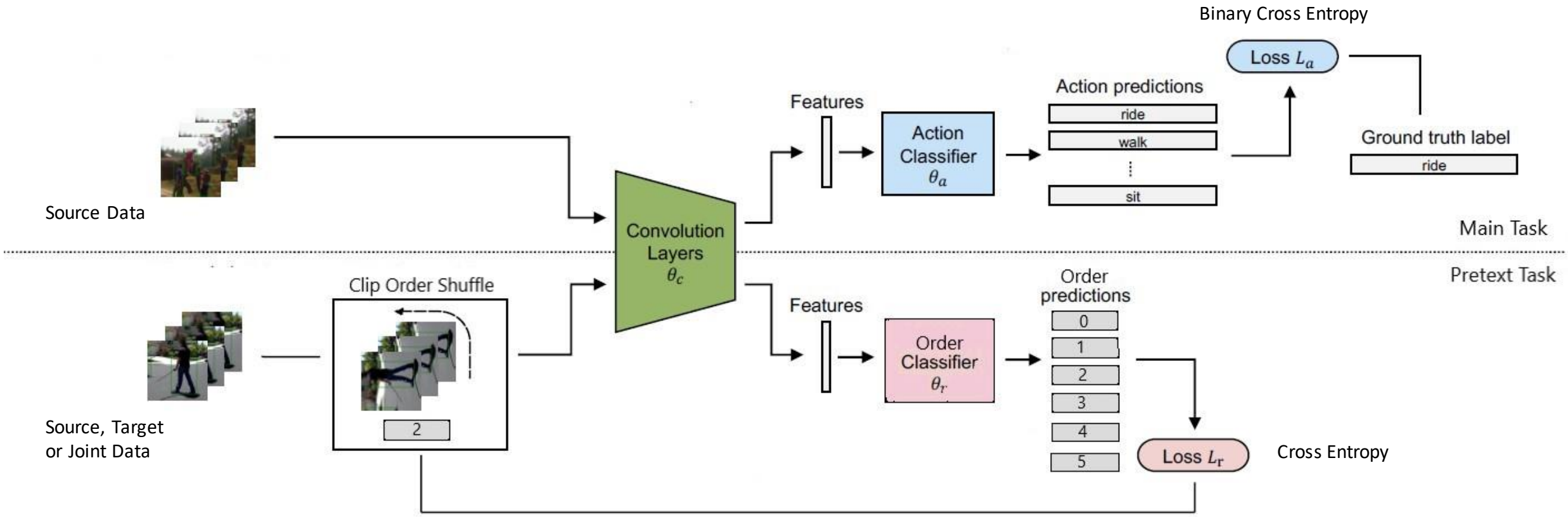
Proposed Method

- Overview
- Knowledge Transfer
- Encoder
- Action Detection
- Video clip order prediction

Transfer Learning



Multi-task Learning

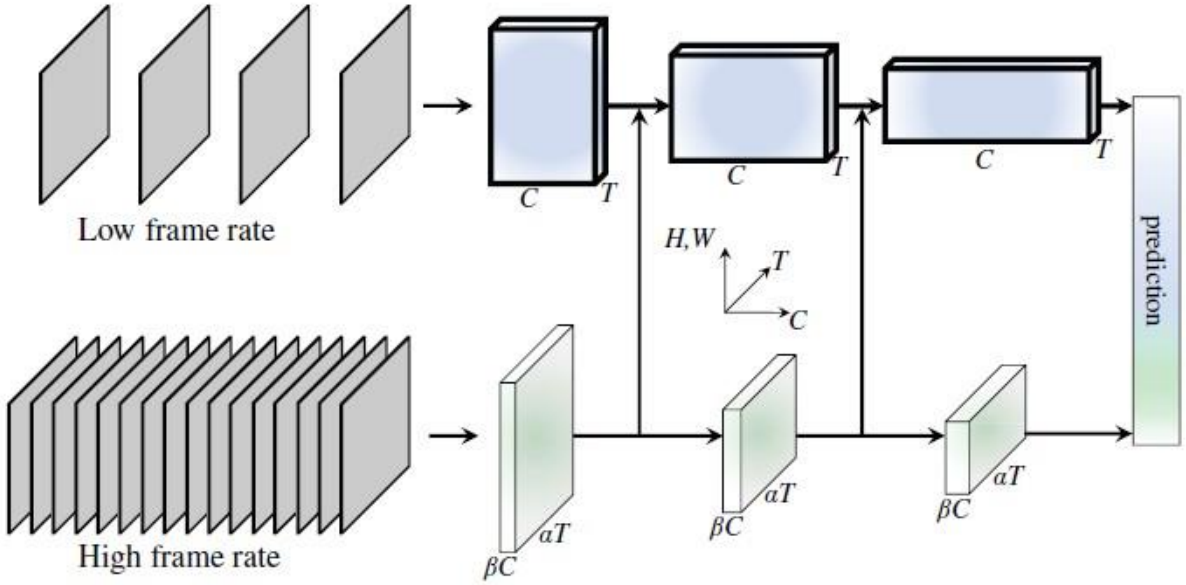


Proposed Method

- Overview
- Knowledge Transfer
- Encoder
- Action Detection
- Video clip order prediction

SlowFast Backbone

- Slow pathway:
  - Low frame rate
  - Low temporal resolution
  - Focus on spatial domain and semantics
- Fast pathway:
  - High frame rate
  - Alpha x higher temporal resolution
  - Beta x number of channels
  - 20% of total computation
  - Capture fast changing motion
- Pathways fused by lateral connections

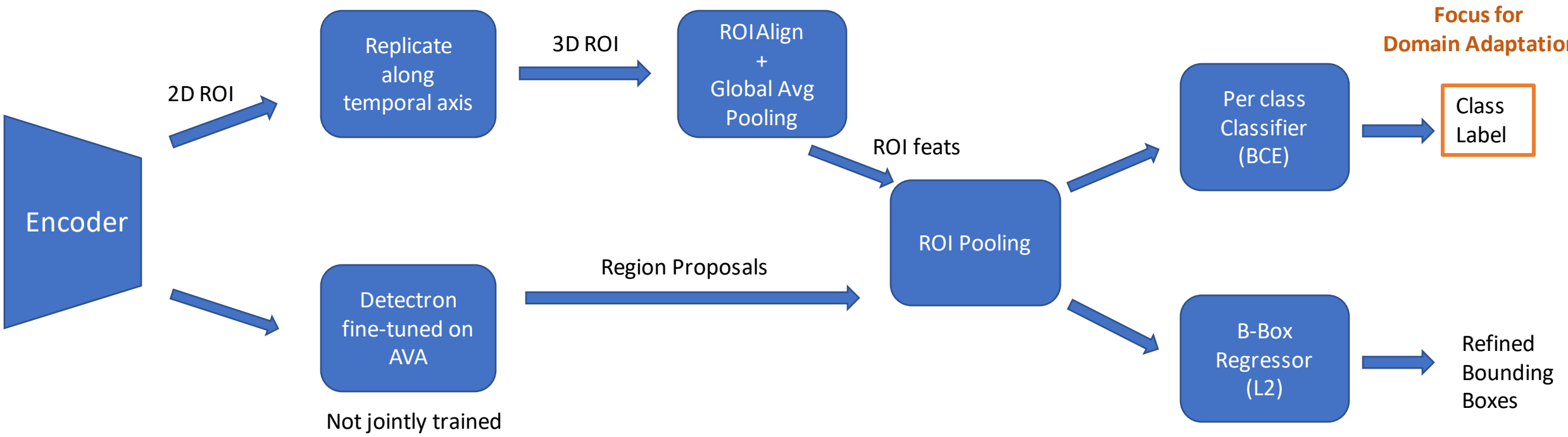


Src: SlowFast Networks for Video Recognition, FAIR

Proposed Method

- Overview
- Knowledge Transfer
- Encoder
- Action Detection
- Video clip order prediction

Action Detection

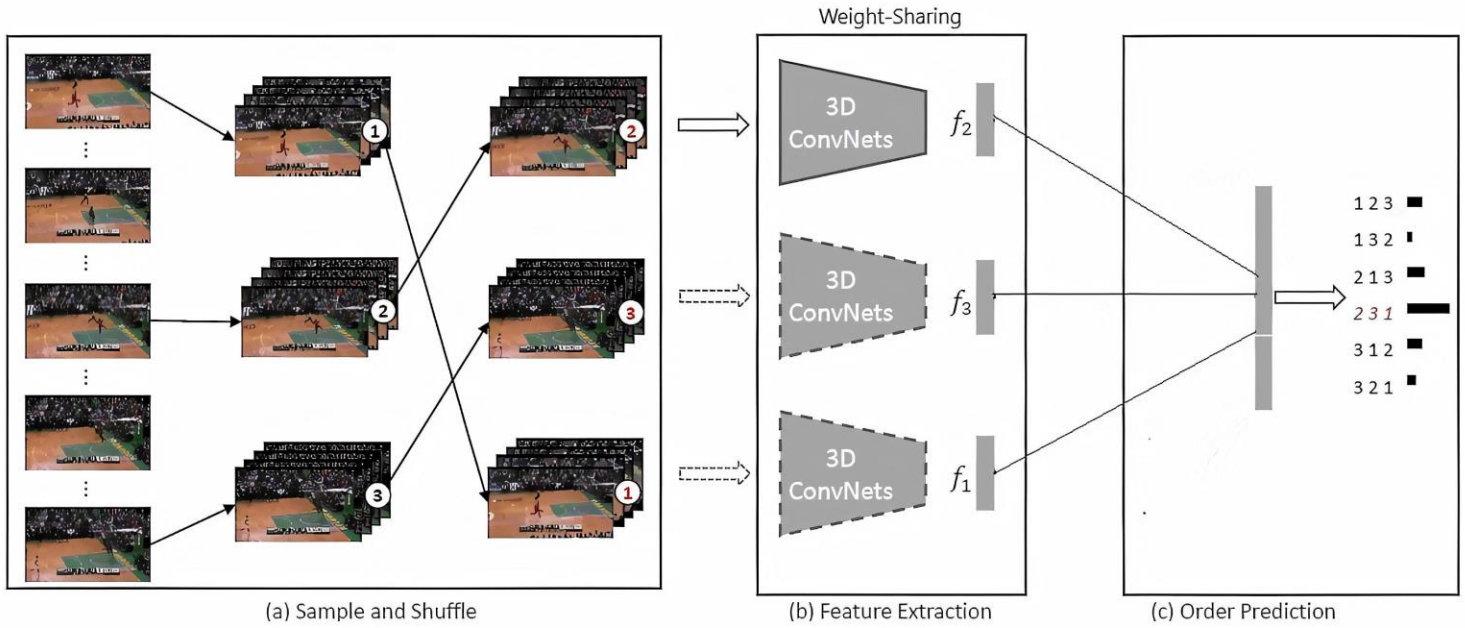


Proposed Method

- Overview
- Knowledge Transfer
- Encoder
- Action Detection
- Video clip order prediction

Clip Order Prediction

- *N* Clips sampled uniformly from video, interval of *m* frames
- Random Shuffling
- 3D CNNs (eg. Slowfast): extract features per clip
- *N* features concatenated to single vector
- Softmax with *N!* classes
- Cross Entropy Loss



Overview

Motivation

Proposed Method

Experiments

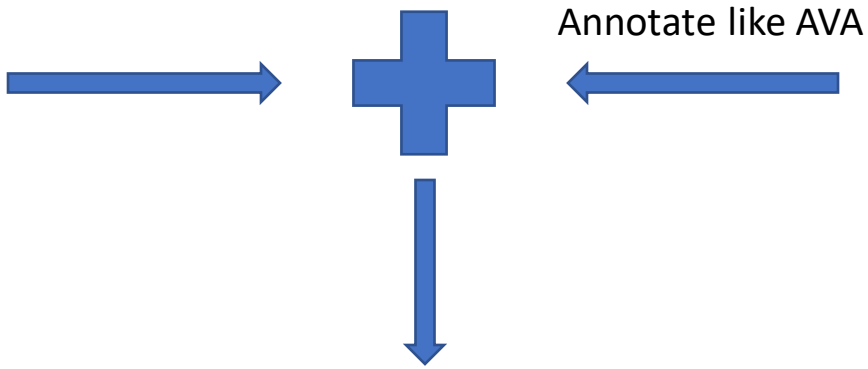
Results

Conclusions

Dataset

AVA [10]

- 80 classes
- 15 min clips
- Bounding box + action label
- Multiple Labels/person
- 430 video samples



Kinetics-700 [11]

- 700 human action classes
- 10s clips
- Action label per frame
- 238,476 video samples

Huge Increase !

AVA-Kinetics [12]

- 80 classes
- Bounding box + action label
- Multiple Labels/person
- 238,906 video samples
- Useful for Domain Adaptation

Reduced Dataset

- Class Reduction
  - 10/80 classes
  - watch (a person)’, 'talk to (e.g., self, a person, a group)’, 'listen to (a person)’, 'sit’, 'carry/hold (an object)’, 'walk’, ‘touch (an object)’, ‘bend/bow (at the waist)’, ‘lie/sleep’, ‘ride (e.g., a bike, a car, a horse)’
- Sample Reduction
  - 5000 training, 1000 validation

	AVA-5k		**Kinetics-5k**	
	Train	Val	Train	Val
Annotations	5′000	1′000	5′000	1′000
Unique boxes	3′847	832	4′121	790
Key-frames	2′972	700	3′140	639
Videos	27	6	3′140	639
Frames	729′814	162′182	565′200	115′020

Baseline

Tags

Src: Source
Tgt: Target
WD: Within Domain
CD: Cross Domain

mAP

- Mean over all classes of the per class Average Precisions (AP)
- AP: area under the Precision Recall curve

Loss

- Main Task: Binary Cross Entropy
- Pretext Task: Cross Entropy

Tag	Src	Tgt	mAP
WD	AVA	AVA	67.03
CD	KTS	AVA	55.8

-16.8%

Tag	Src	Tgt	mAP
WD	KTS	KTS	58.76
CD	AVA	KTS	48.1

-18.1%

Overview

Motivation

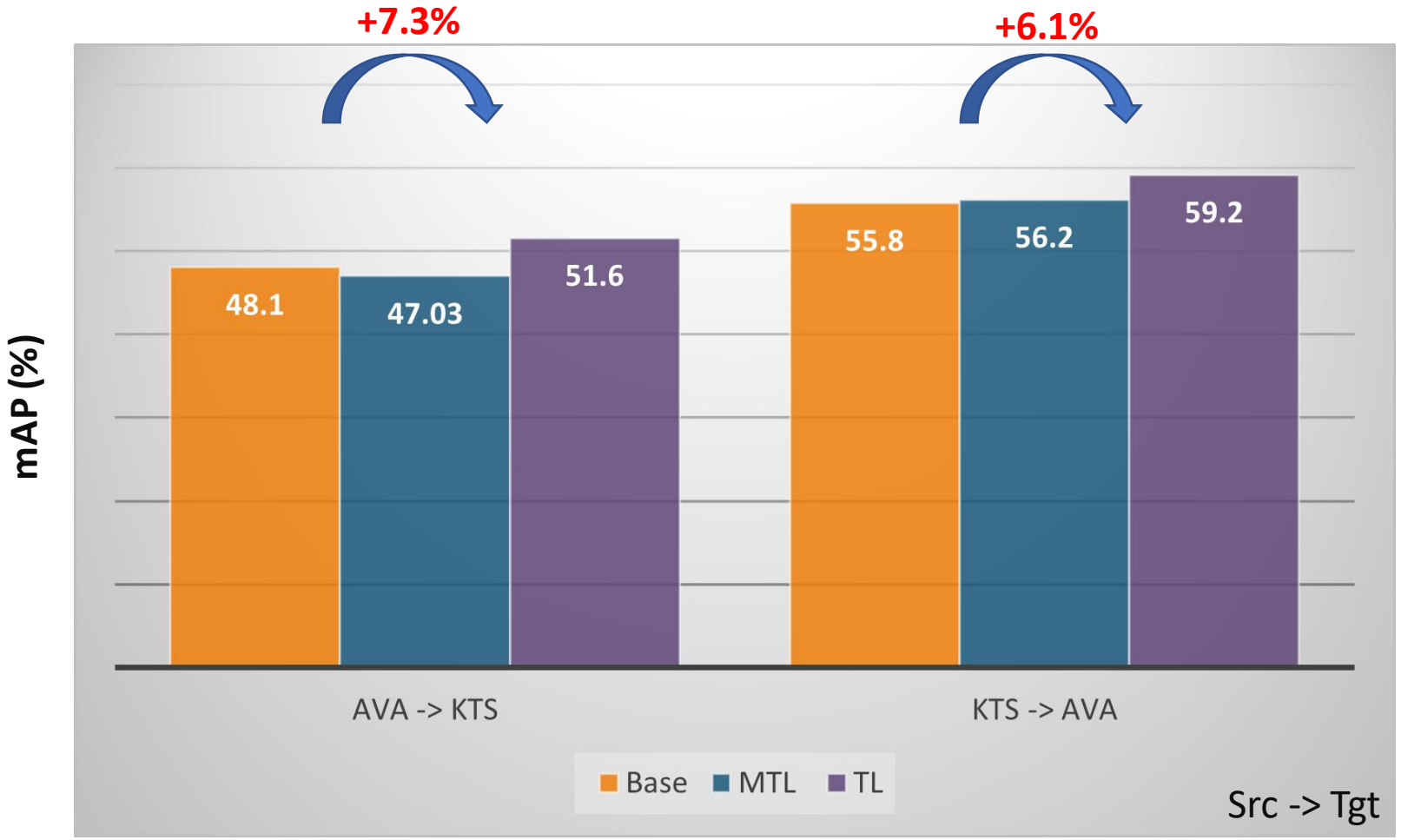
Proposed Method

Experiments

Results

Conclusions

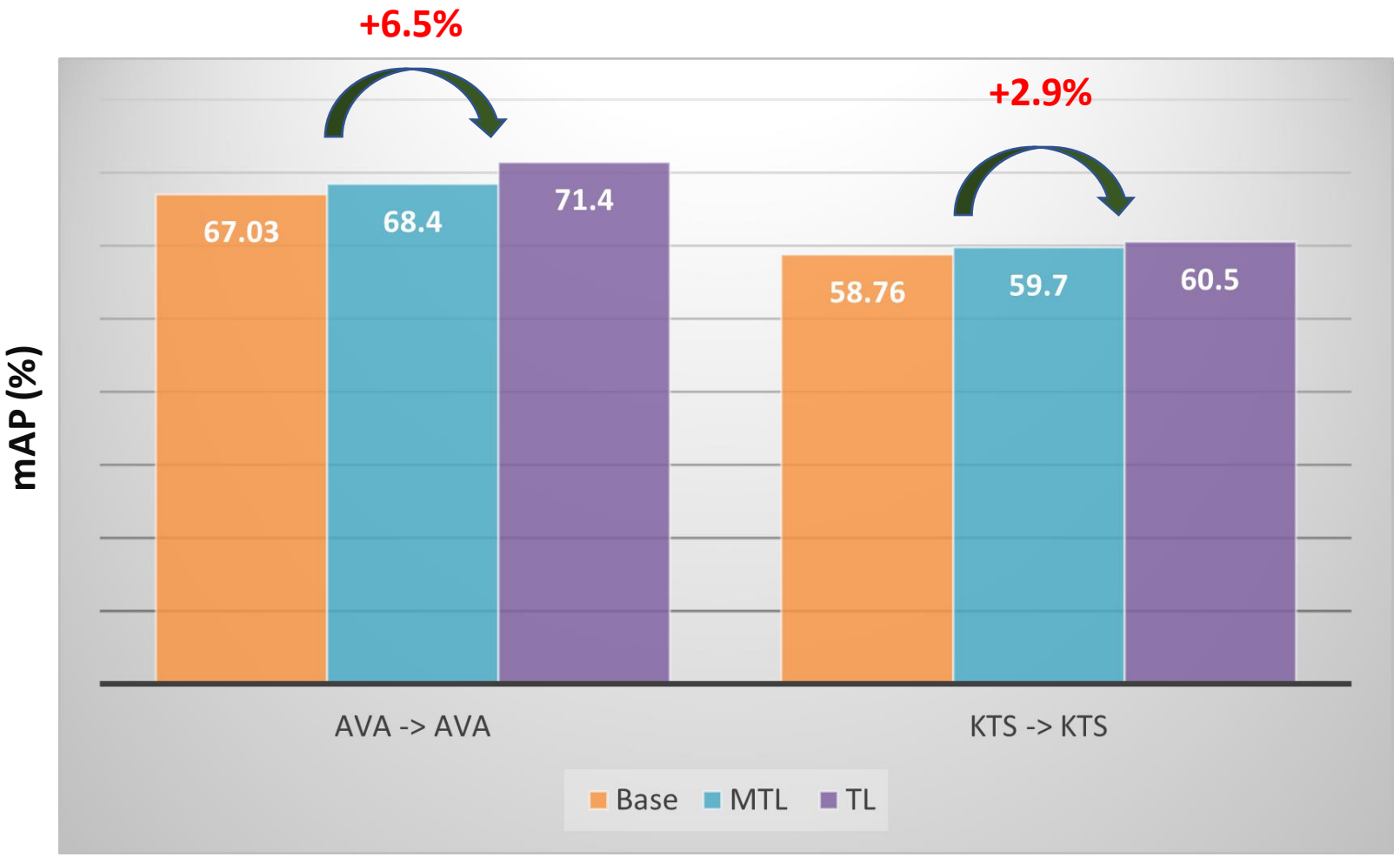
Cross-Domain Performance



MTL: Multi-task Learning
TL: Transfer Learning

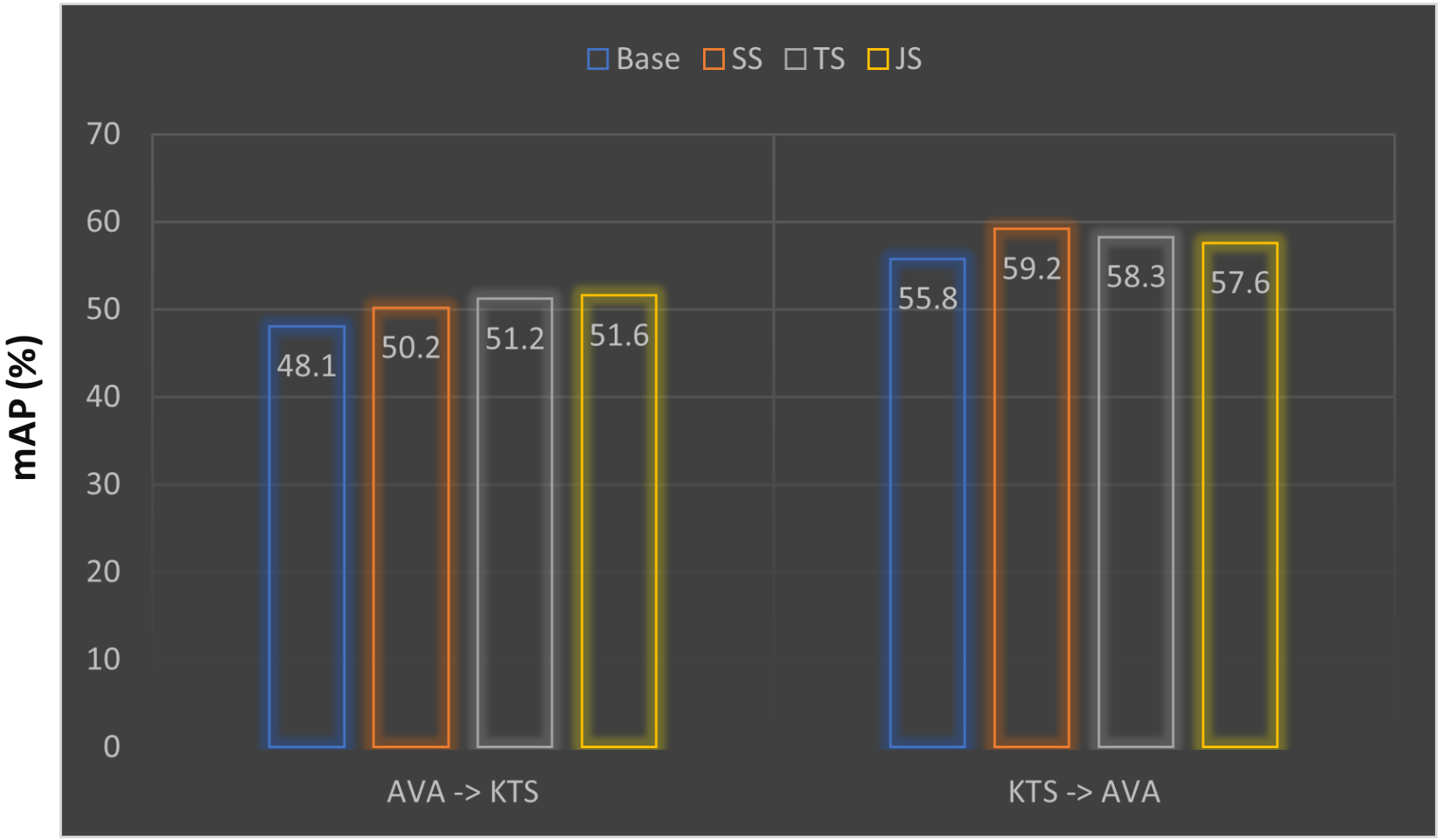
Within Domain Performance

Self-supervision boosts the performance of action detection on the same domain



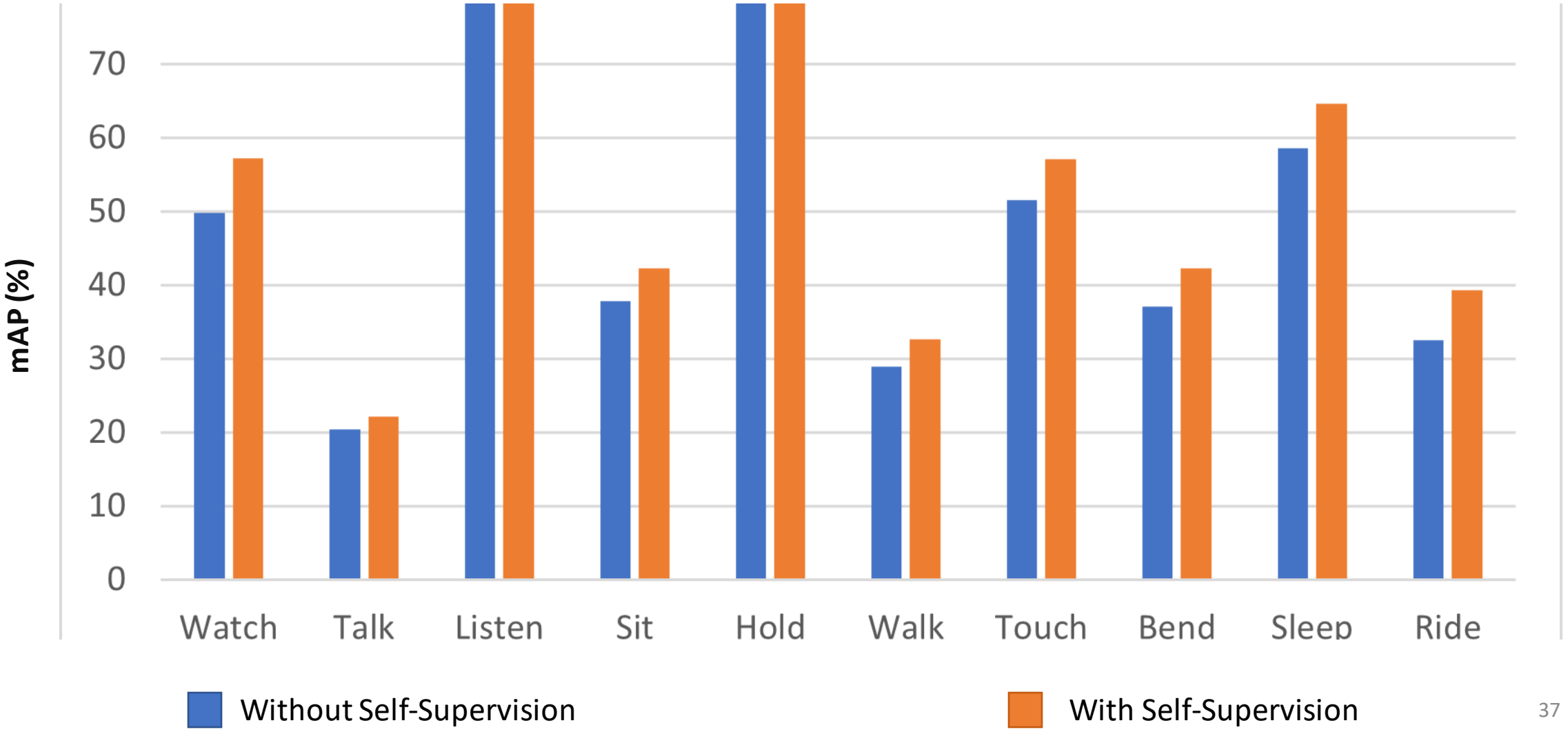
MTL: Multi-task Learning
TL: Transfer Learning

Which supervision works best?



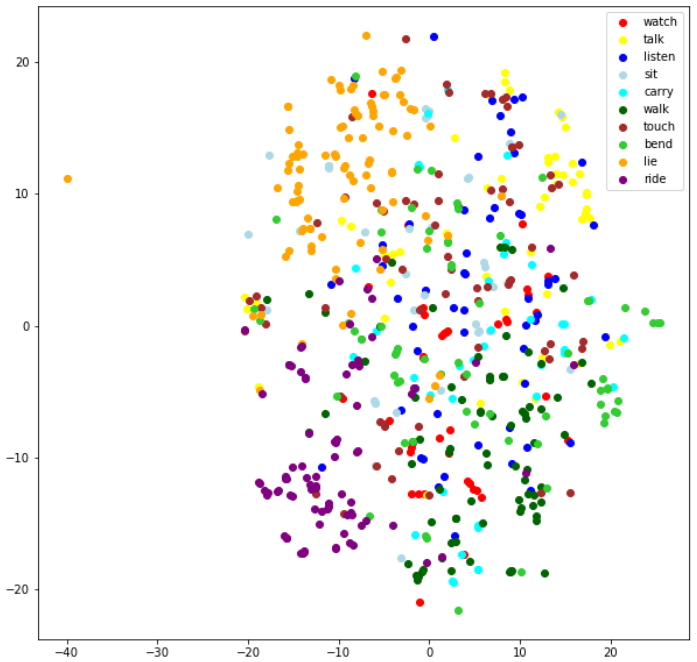
Base: No supervision
SS: Source Supervision
TS: Target Supervision
JS: Joint Supervision
Model: TL

Class-wise Analysis (K->A)

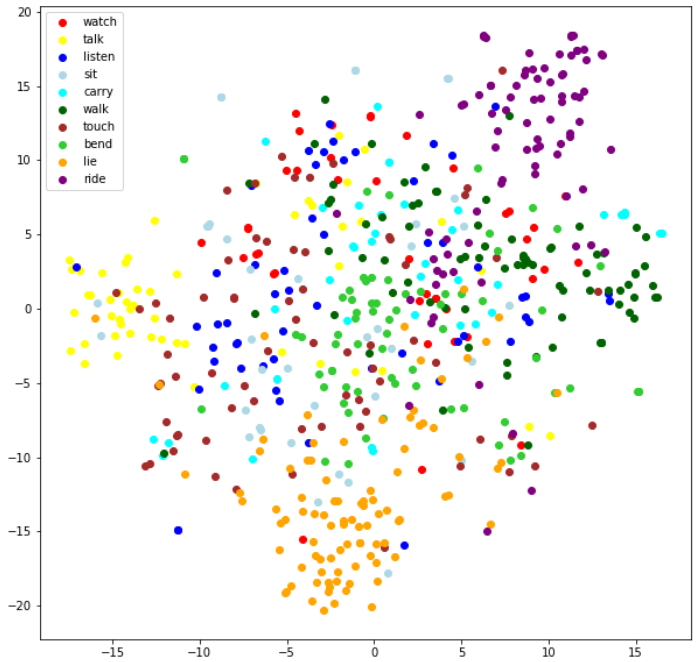


T-SNE: Cross Domain

Self-supervision improves inter-class separability



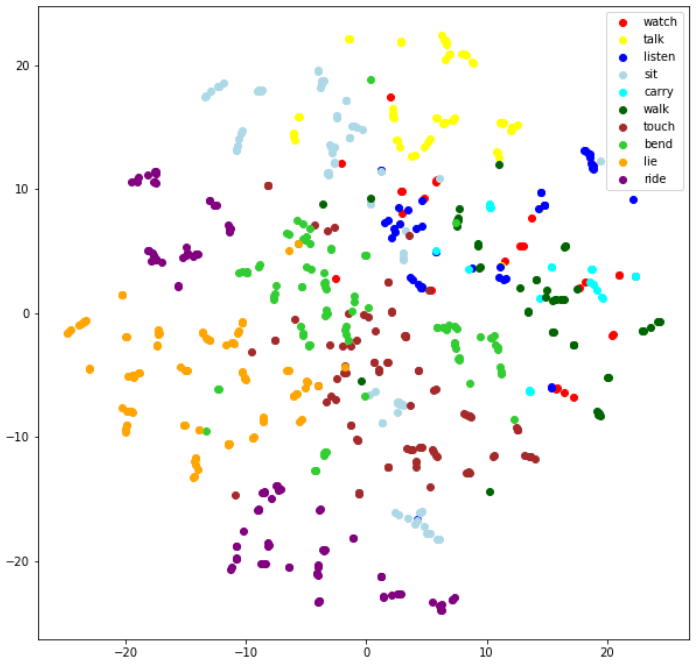
Baseline: without self-supervision (AVA -> KTS)



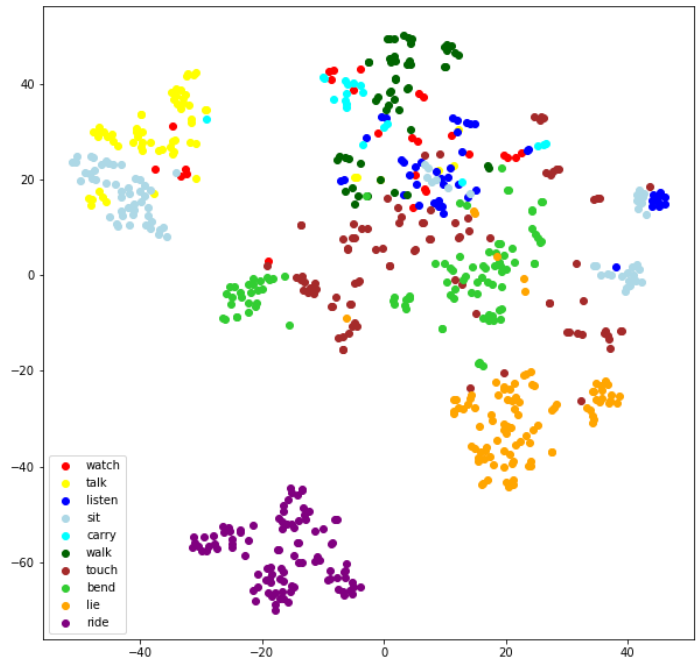
Proposed: Transfer Learning with joint-supervision (AVA -> KTS)

T-SNE: Within Domain

Self-supervision improves intra-class variability

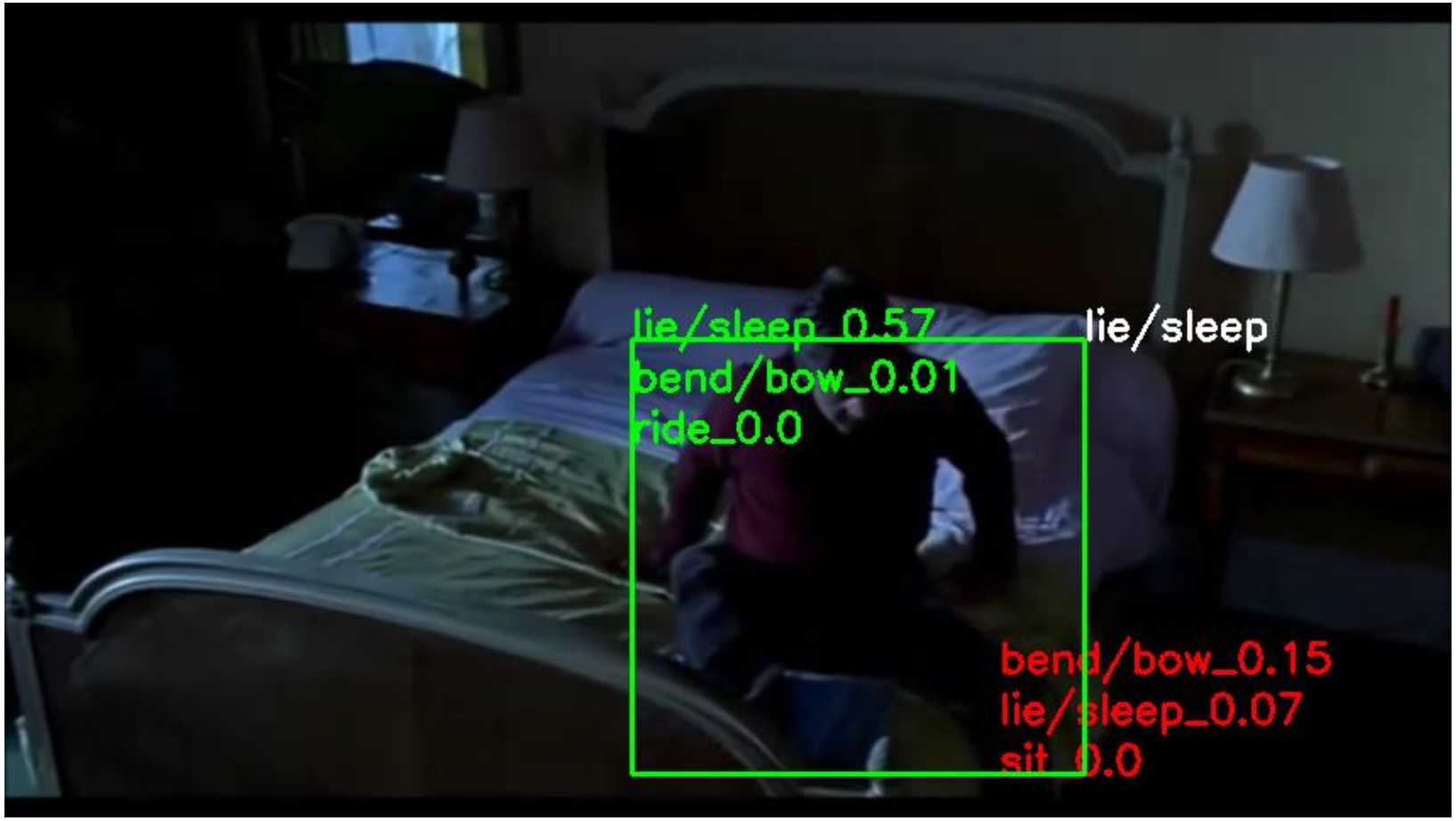


Baseline: without self-supervision (AVA -> AVA)



Proposed: Transfer Learning with joint-supervision (AVA -> AVA)

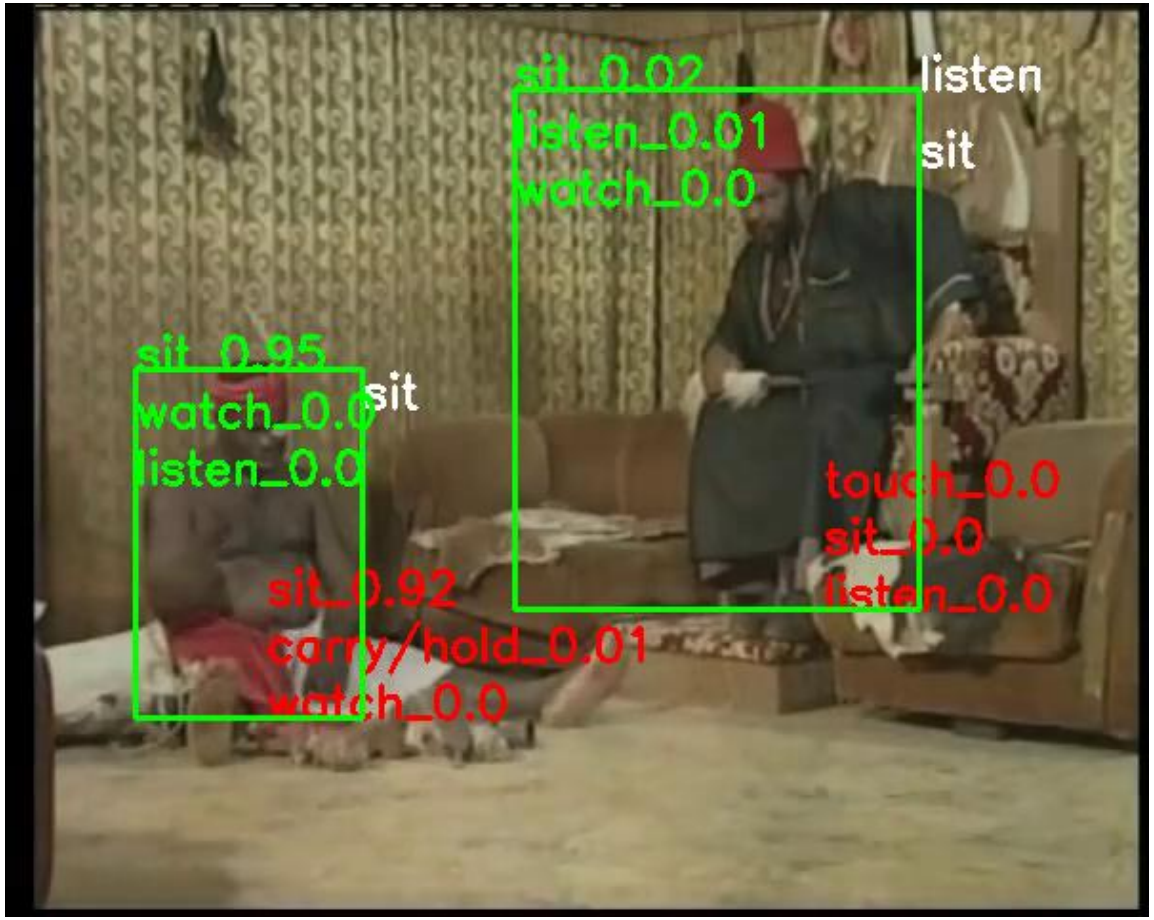
Qualitative Results



We adapt better to the target domain

Qualitative Results

When correct, we are more confident on target domain



We handle multiple action scenarios better

Qualitative Results

When incorrect, we consider related scenarios rationally

Overview

Motivation

Proposed Method

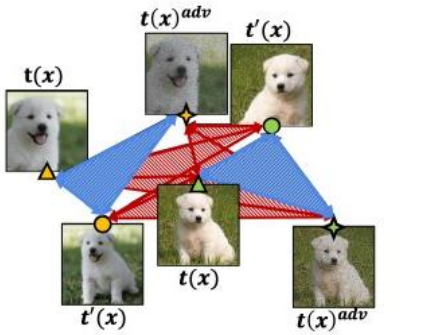
Experiments

Results

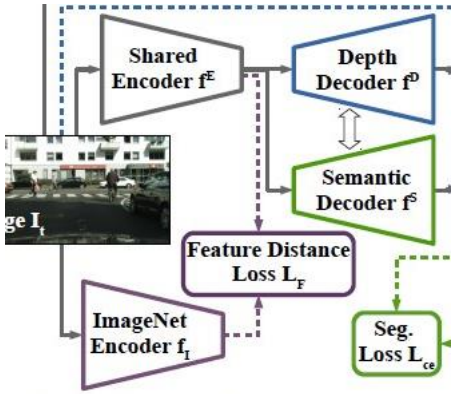
Conclusions

Conclusion

- Self-Supervision improves Domain Adaptation
  - 7.3% and 6.1% improvement in AVA-> KTS and vice-versa, resp.
- Self-Supervision boosts performance within domain
  - 6.5% and 2.9% improvement in AVA and KTS resp.
- Consistent performance improvement in individual classes
  - Max: 20.9% Min: 2.9%
- Different supervision types help in different scenarios
- Improved inter-class and intra-class variations
- More confident, accurate and semantically rational predictions



(b) RoCL: Robust contrastive learning



Future Works

- Go Adversarial
  - Adversarial Self-Supervised Contrastive Learning, NIPS'20
- Regularize MTL features
  - Feature Distance Loss in '*Three Ways to Improve Semantic Segmentation with Self-Supervised Depth Estimation*'
- Pandora of Pretext
  - Spatial and Temporal 'Creativity'

References

1. Li, Mengxue, et al. "Enhanced transport distance for unsupervised domain adaptation." *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2020.
2. Choi, Jinwoo, et al. "Unsupervised and semi-supervised domain adaptation for action recognition from drones." *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*. 2020.
3. Chen, Min-Hung, et al. "Action segmentation with joint self-supervised temporal domain adaptation." *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2020.
4. Fernando, Basura, et al. "Self-supervised video representation learning with odd-one-out networks." *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2017.
5. Sun, Yu, et al. "Unsupervised domain adaptation through self-supervision." *arXiv preprint arXiv:1909.11825* (2019).
6. Misra, Ishan, C. Lawrence Zitnick, and Martial Hebert. "Shuffle and learn: unsupervised learning using temporal order verification." *European Conference on Computer Vision*. Springer, Cham, 2016.
7. Wang, Xiaolong, and Abhinav Gupta. "Unsupervised learning of visual representations using videos." *Proceedings of the IEEE international conference on computer vision*. 2015.
8. Jing, Longlong, et al. "Self-supervised spatiotemporal feature learning via video rotation prediction." *arXiv preprint arXiv:1811.11387* (2018).
9. Xu, Dejing, et al. "Self-supervised spatiotemporal learning via video clip order prediction." *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2019.
10. Gu, Chunhui, et al. "Ava: A video dataset of spatio-temporally localized atomic visual actions." *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2018.
11. Carreira, Joao, et al. "A short note on the kinetics-700 human action dataset." *arXiv preprint arXiv:1907.06987* (2019).
12. Li, Ang, et al. "The ava-kinetics localized human actions video dataset." *arXiv preprint arXiv:2005.00214* (2020).

Thank You

Questions?

Appendix

Model Configuration

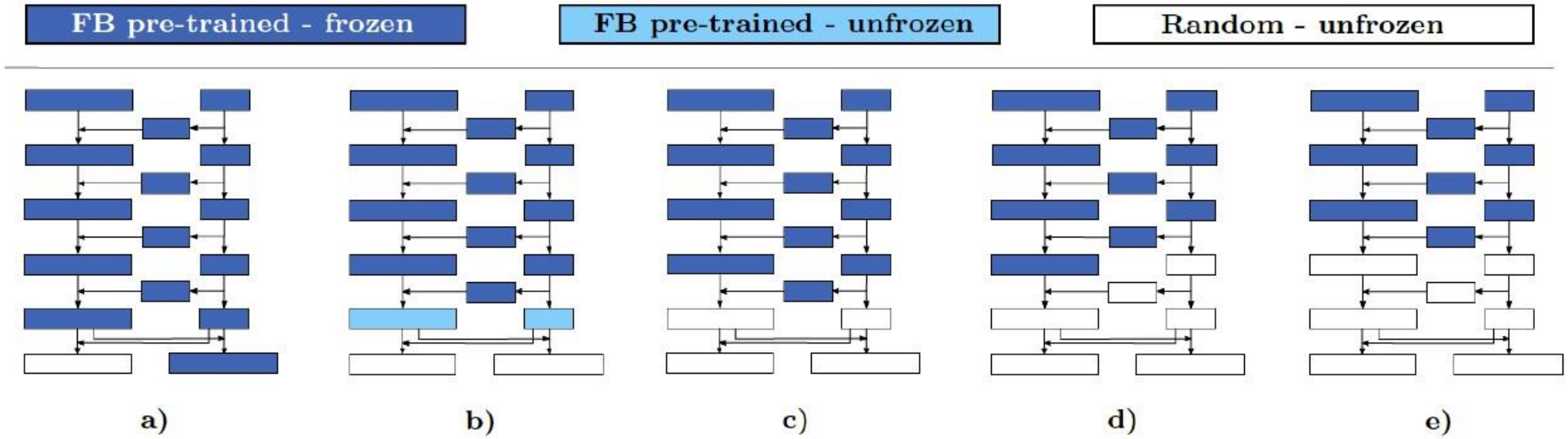
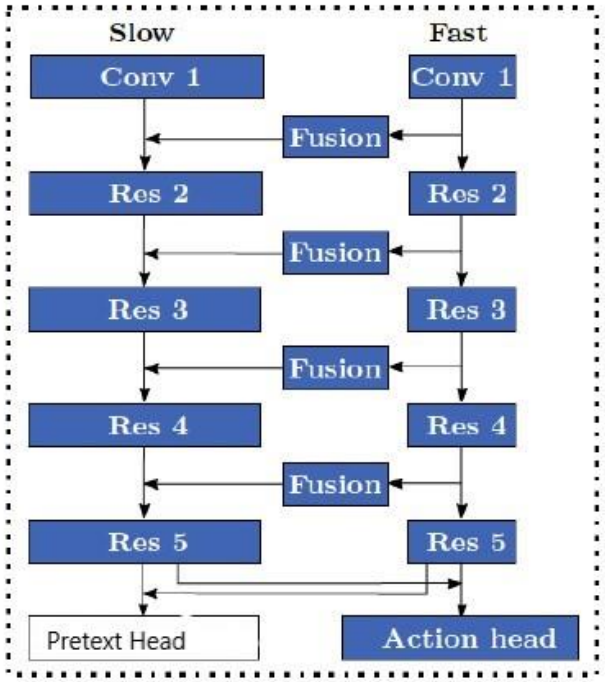


Figure 3: Model initialisation methods and unfrozen weights

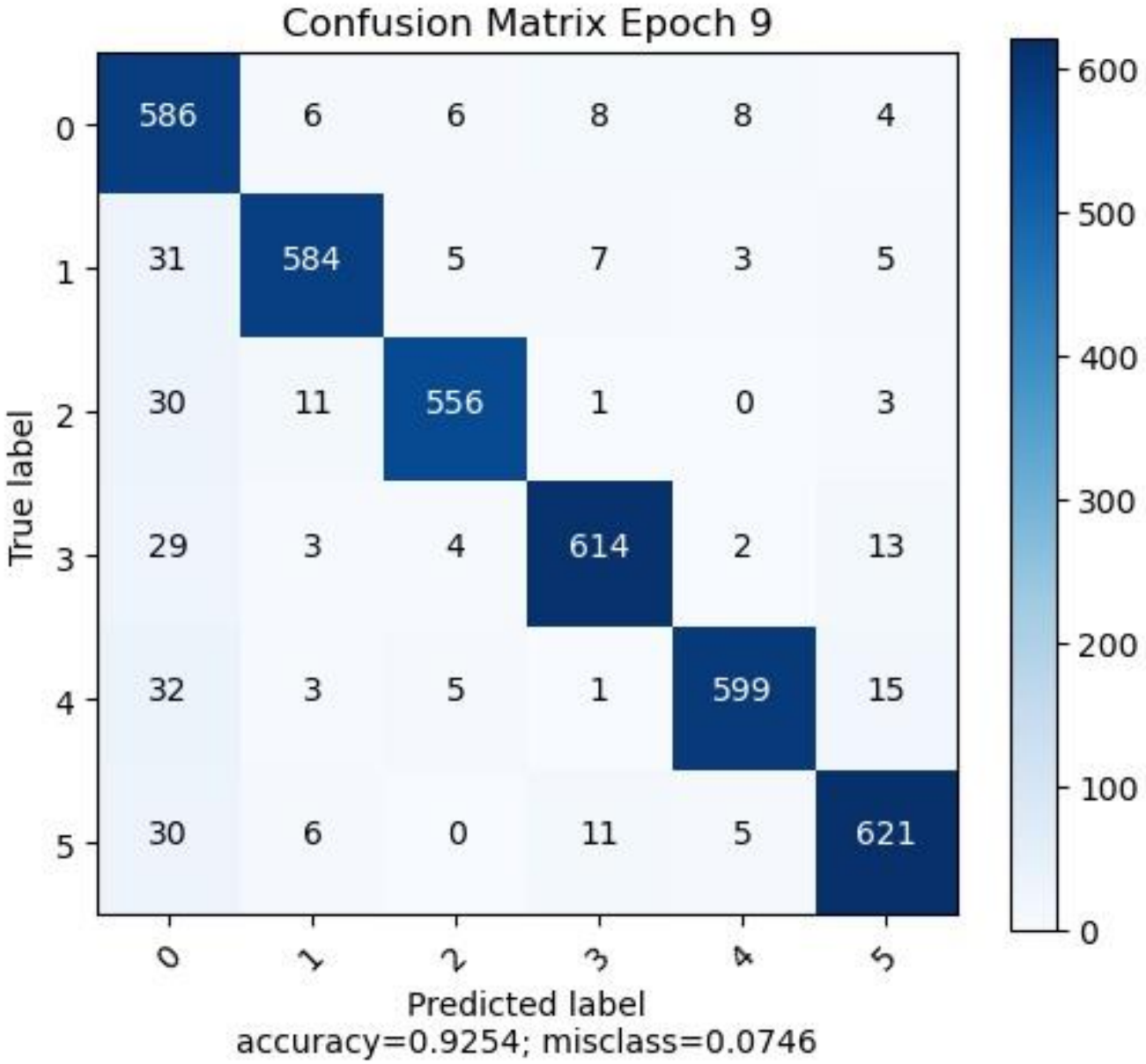
Model Overview



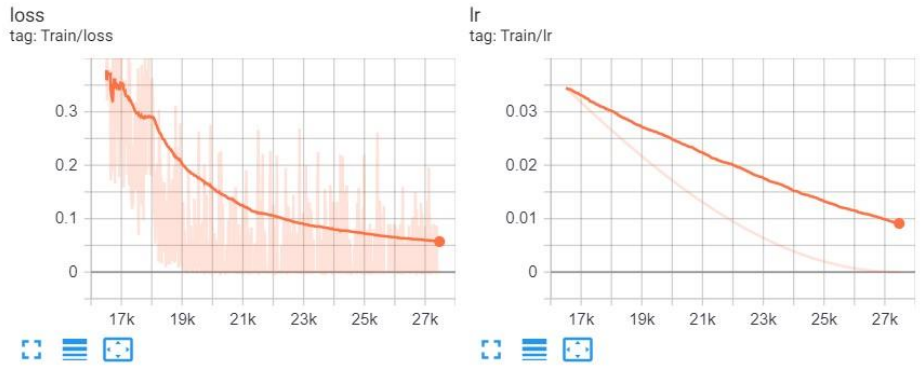
Layer	Parameters			Output dimensions ($C \times T \times S^2$)		
	Slow	Fast	Fusion	Slow	Fast	Fusion
Input				$3 \times 8 \times 224^2$	$3 \times 32 \times 224^2$	
Conv 1	10	6	1	$64 \times 8 \times 56^2$	$8 \times 32 \times 56^2$	$16 \times 8 \times 56^2$
Res 2	221	5	10	$256 \times 8 \times 56^2$	$32 \times 32 \times 56^2$	$64 \times 8 \times 56^2$
Res 3	1′261	27	41	$512 \times 8 \times 28^2$	$64 \times 32 \times 28^2$	$128 \times 8 \times 28^2$
Res 4	35′509	461	164	$1024 \times 8 \times 14^2$	$128 \times 32 \times 14^2$	$256 \times 8 \times 14^2$
Res 5	21′125	318		$2048 \times 8 \times 14^2$	$256 \times 32 \times 14^2$	
Per Pathway	58′126	817	216	2′048	256	
Total		**59′159**			**2′304**	
	Action		Pretext	Action		Pretext
Fully connected	23		9	10 classes		6 classes

Table 2: SlowFast: number of parameters (in 1k, rounded to nearest) and feature map dimensions after each layer

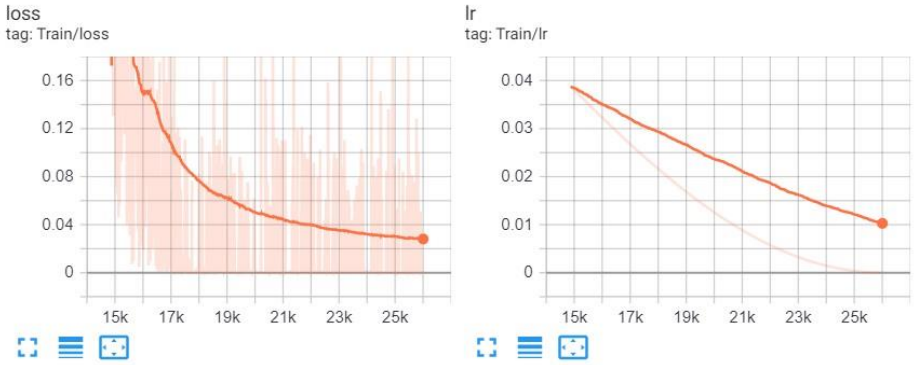
Pretext task performance

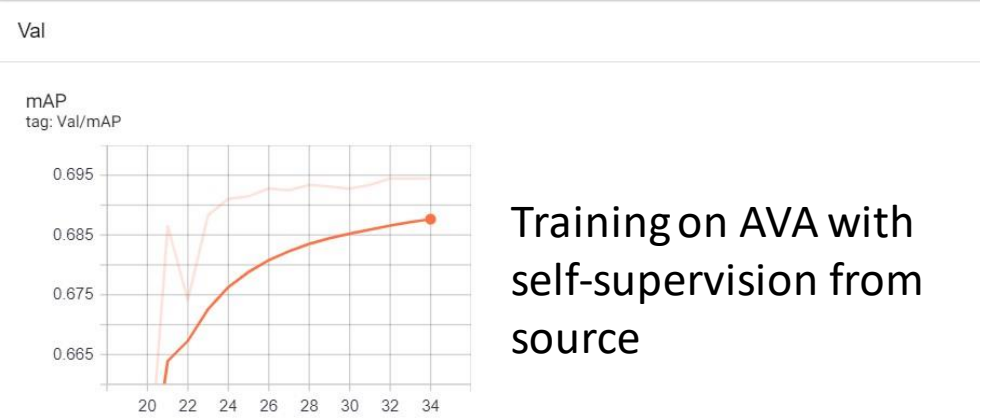


Training Curves: TL



Training on Kinetics700 with self-supervision from source

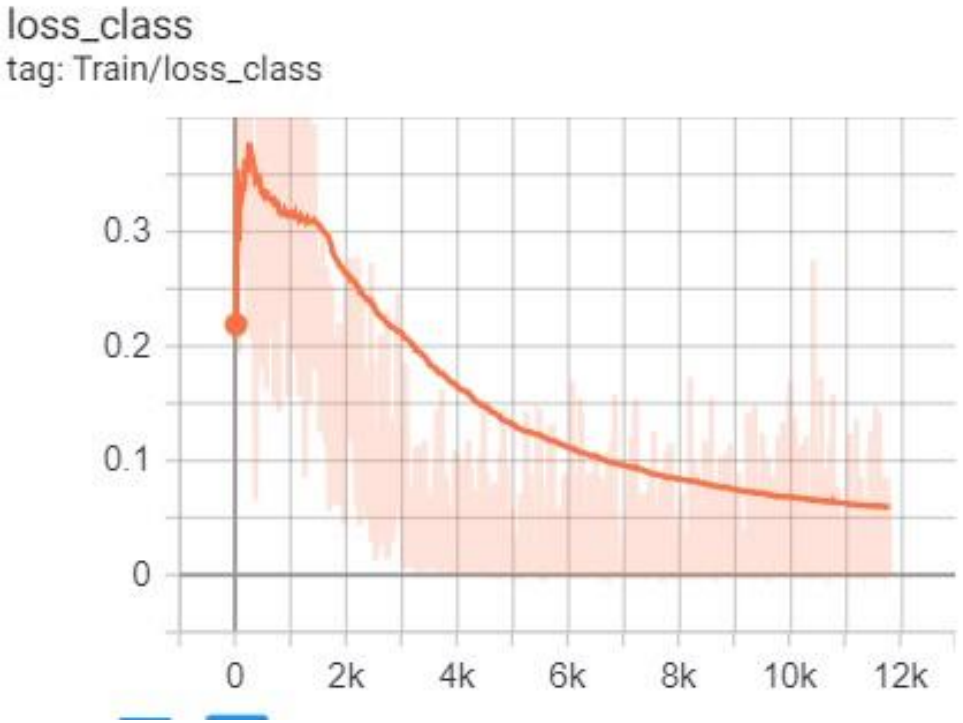


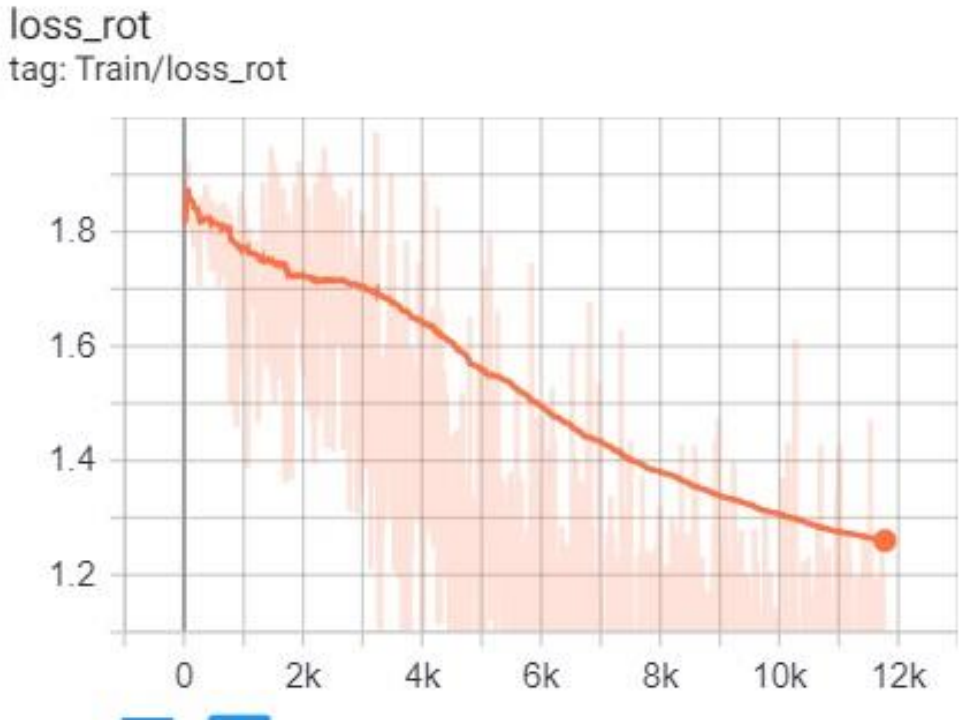


Training on AVA with self-supervision from source

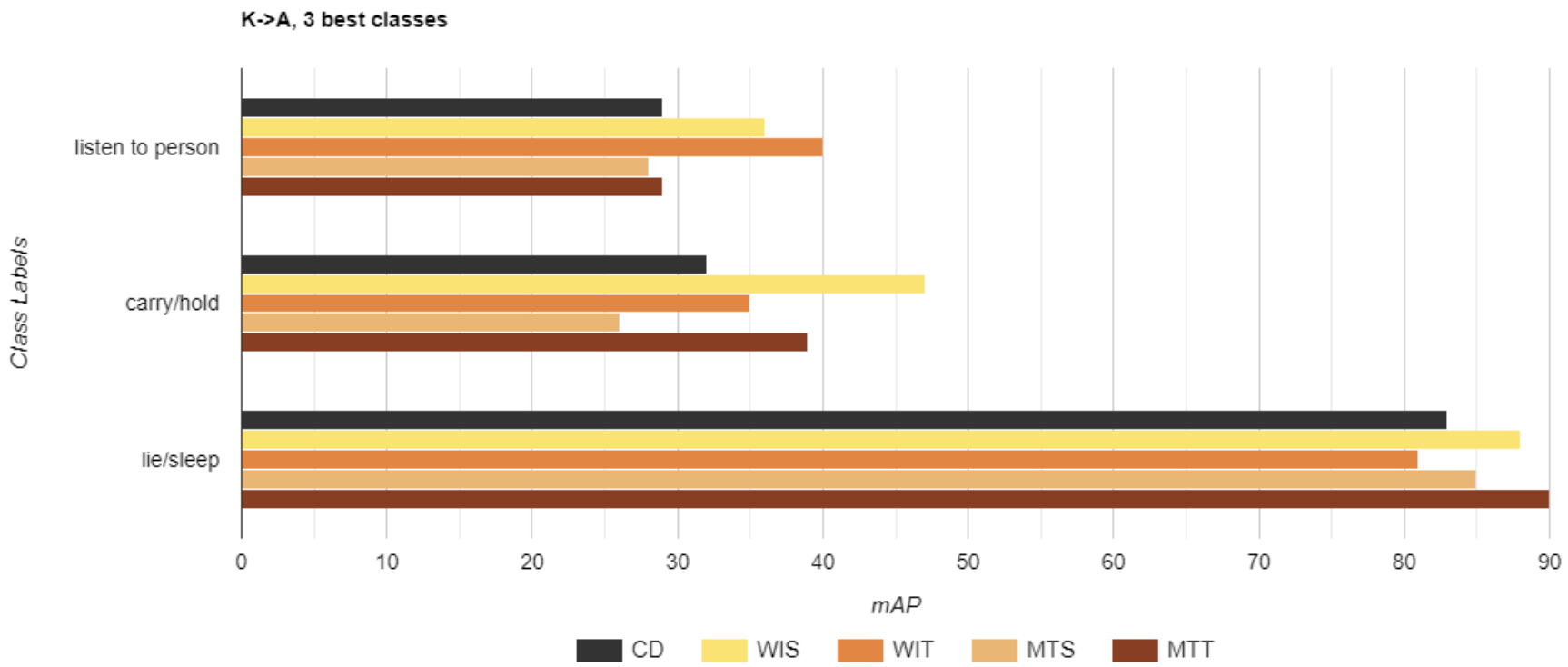
Training curves: MTL

- CD: Kinetics -> AVA
- Target supervision

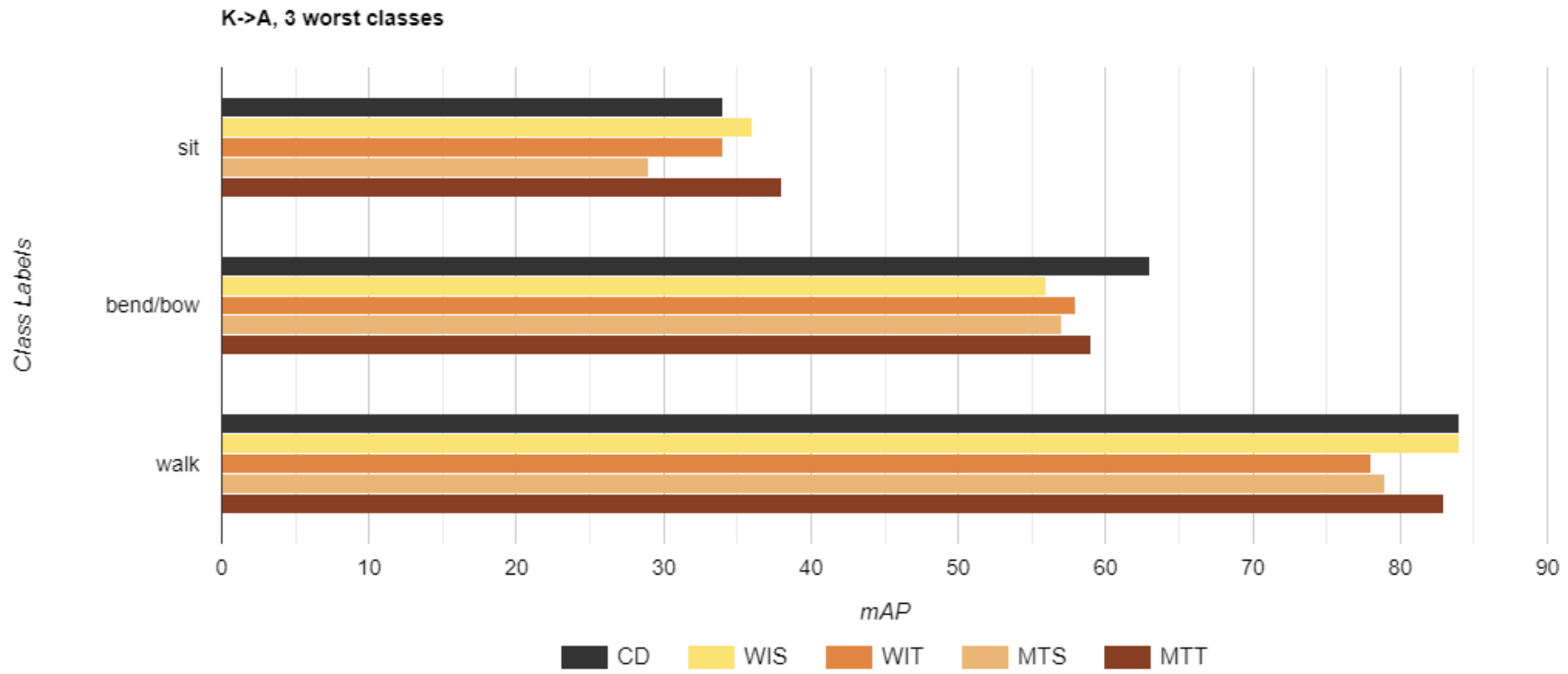


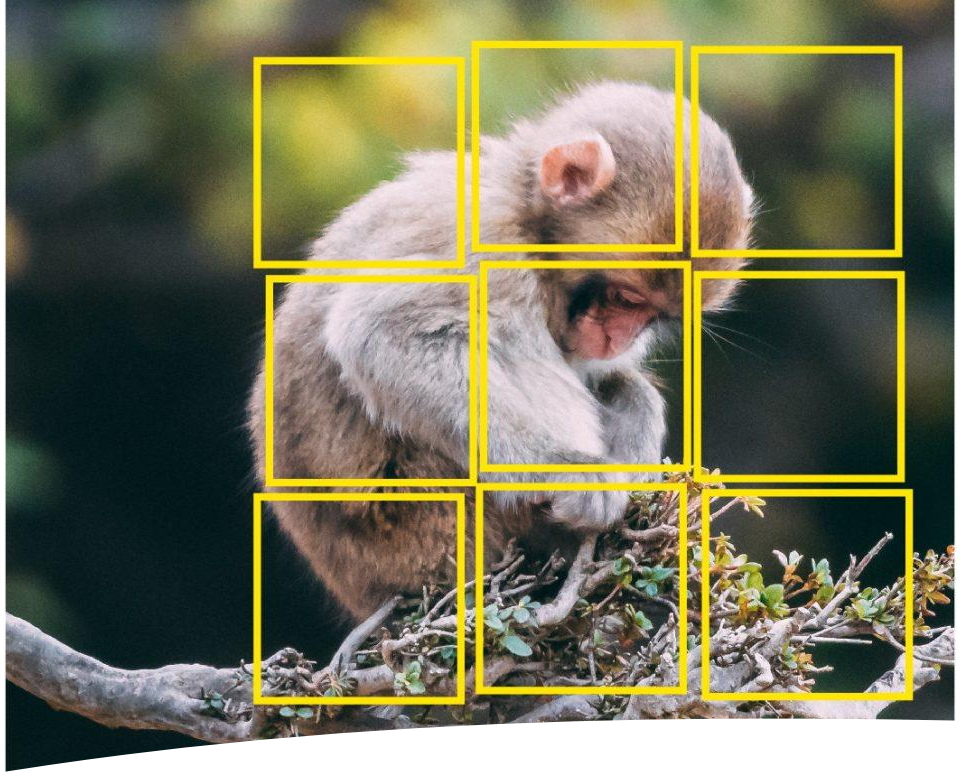


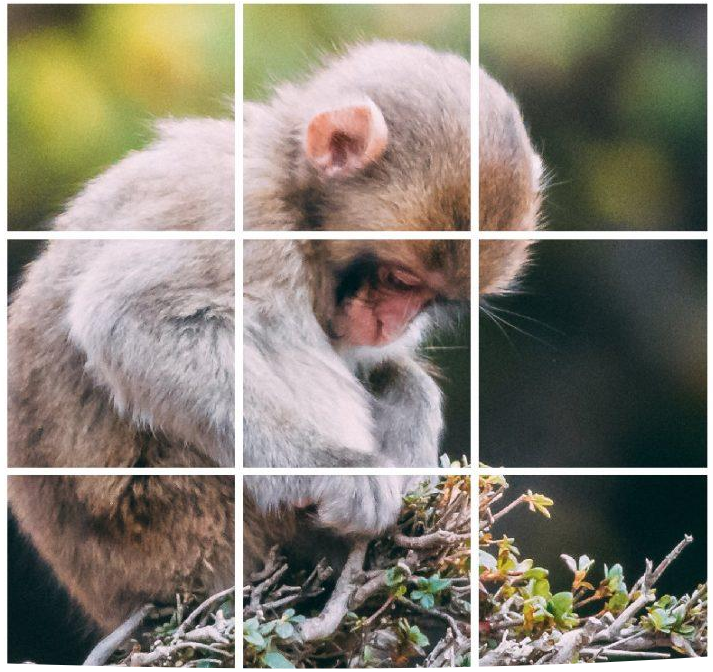
TL vs MTL? Class-wise analysis



TL vs MTL? Class-wise analysis







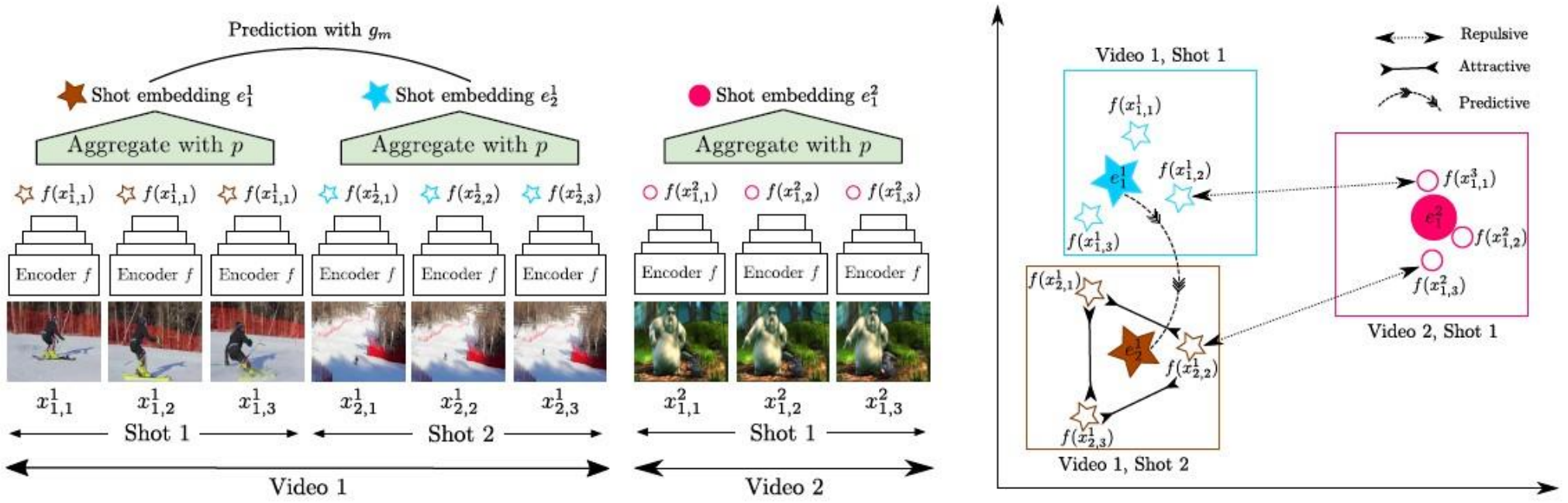
Self-supervision

- Auxiliary (pretext) tasks: data used as supervision
- Learn deep features useful for downstream (main) tasks
- Spatial, temporal or spatio-temporal supervision in vidoes
- Robustness, domain generalization, few-shot learning etc.

Self-Supervised Learning of Video-Induced Visual Invariances

Google Research, Brain Team

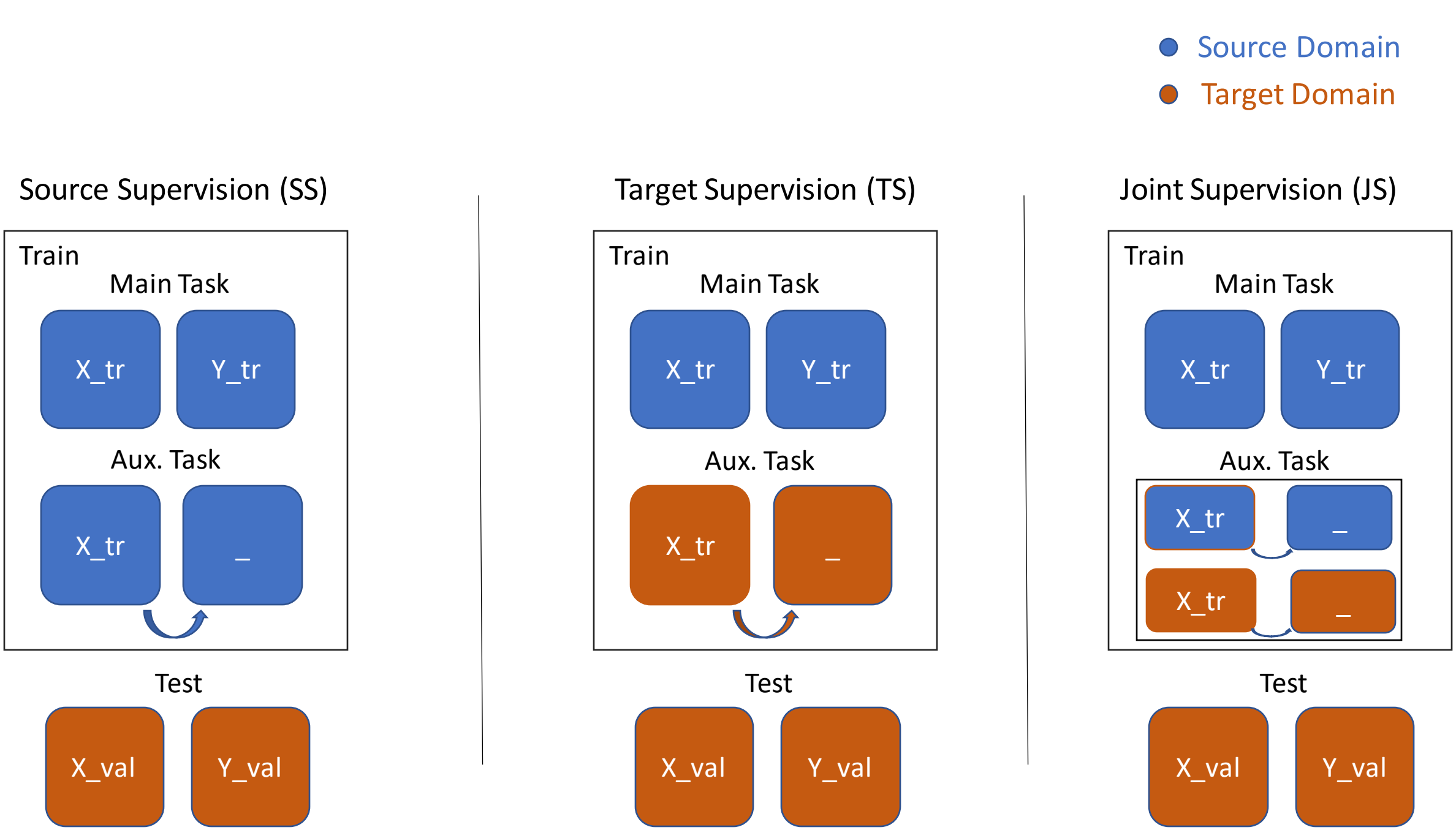
Framework



Src: Self-Supervised Learning of Video-Induced Visual Invariances

Types of Supervision

Which domain to use?


Source Domain
Target Domain
Source Supervision (SS)
Train
Main Task
X_tr
Y_tr
Aux. Task
X_tr
_
Test
X_val
Y_val
Target Supervision (TS)
Train
Main Task
X_tr
Y_tr
Aux. Task
X_tr
_
Test
X_val
Y_val
Joint Supervision (JS)
Train
Main Task
X_tr
Y_tr
Aux. Task
X_tr
_
X_tr
_
Test
X_val
Y_val

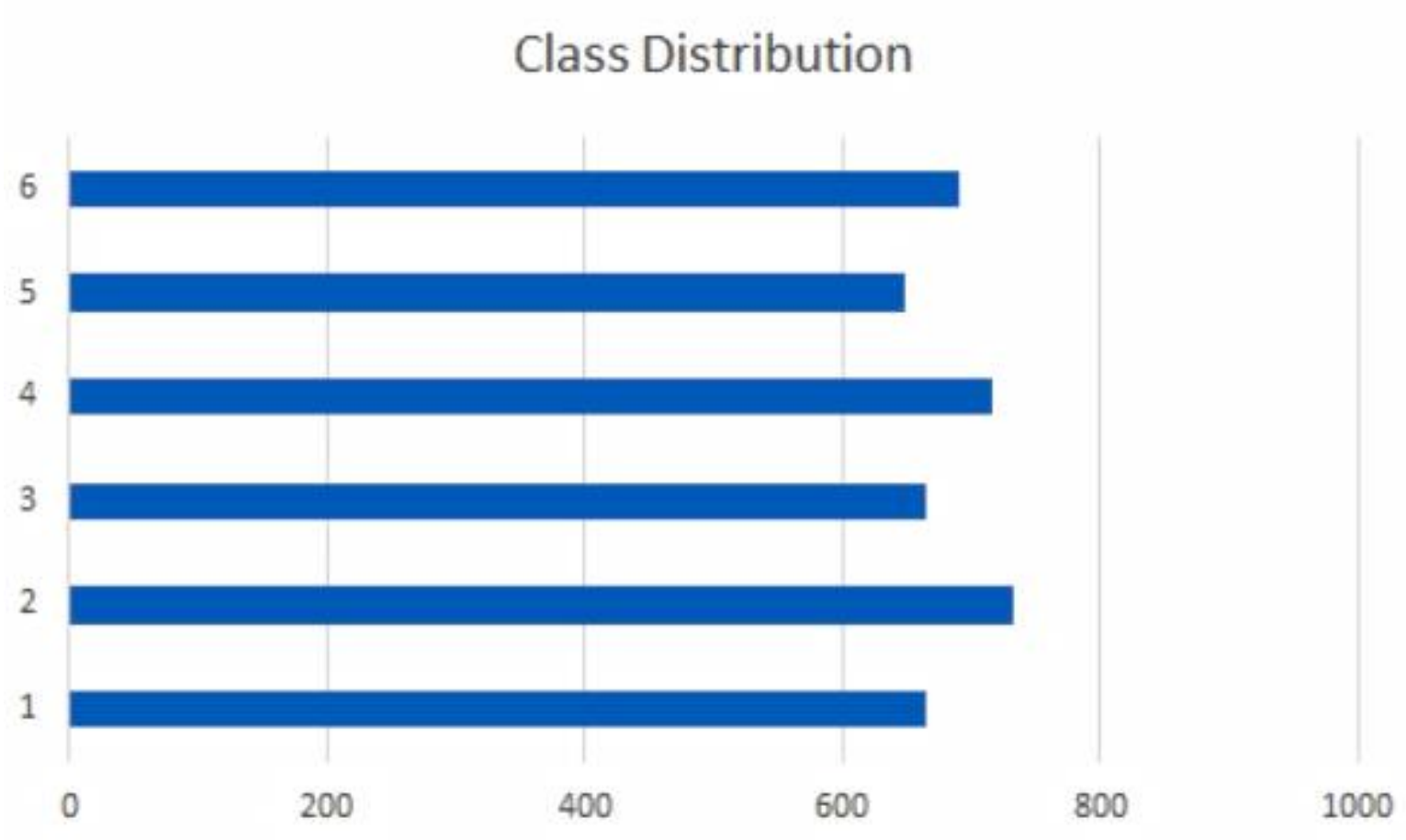
Dataset

- AVA: 80 classes, 15 min clips, bounding box + action label / person
- Kinetics: 700 classes, 10s clips, action label per frame
- AVA-Kinetics: Kinetics clips annotated with 80 AVA action classes for each human in key-frames, added with AVA dataset

	# unique frames			# unique videos		
	AVA	Kinetics	AVA-Kinetics	AVA	Kinetics	AVA-Kinetics
Train	210,634	141,457	352,091	235	141,240	141,475
Val	57,371	32,511	89,882	64	32,465	32,529
Test	117,441	65,016	182,457	131	64,771	64,902
Total	385,446	238,984	624,430	430	238,476	238,906

Src: The AVA-Kinetics Localized Human Actions Video Dataset

Clip order: Class Distribution



Class Label

Number of ordered tuple of clips per class