Domain Adaptation in Videos

Final Presentation

Chen Zhou, Jayant Jain, Je-Hoon Michael Oh, Anirudh Choudhary

Problem Statement

Problem: Domain adaptation (DA) for action recognition across video datasets.

Motivation:

- Large number of un-annotated human action videos; Tedious video annotation process
- Domain Adaptation is relatively unexplored in videos

Challenge

Videos suffer from domain discrepancy along spatial and temporal dimensions

Fencing - HMDB(upper row), UCF(bottom row)

Spatial and temporal discrepancy

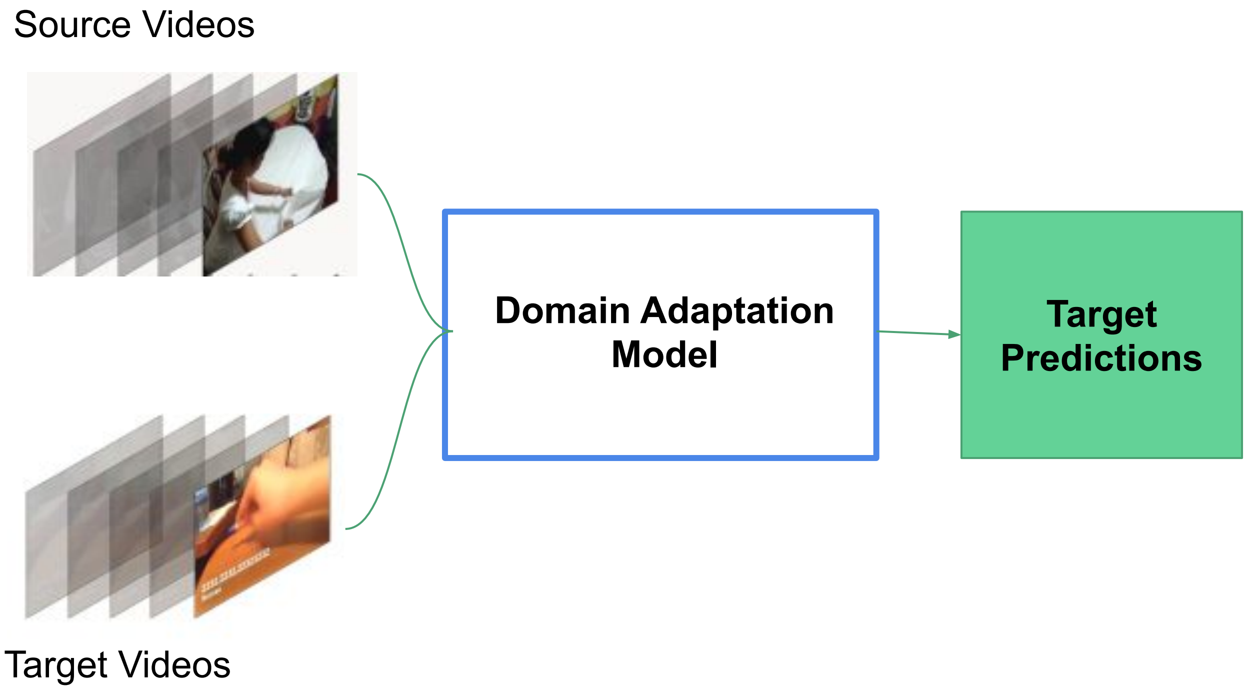
Image credit: HACS Dataset

Problem Statement

Technical problem: Unsupervised DA for action recognition

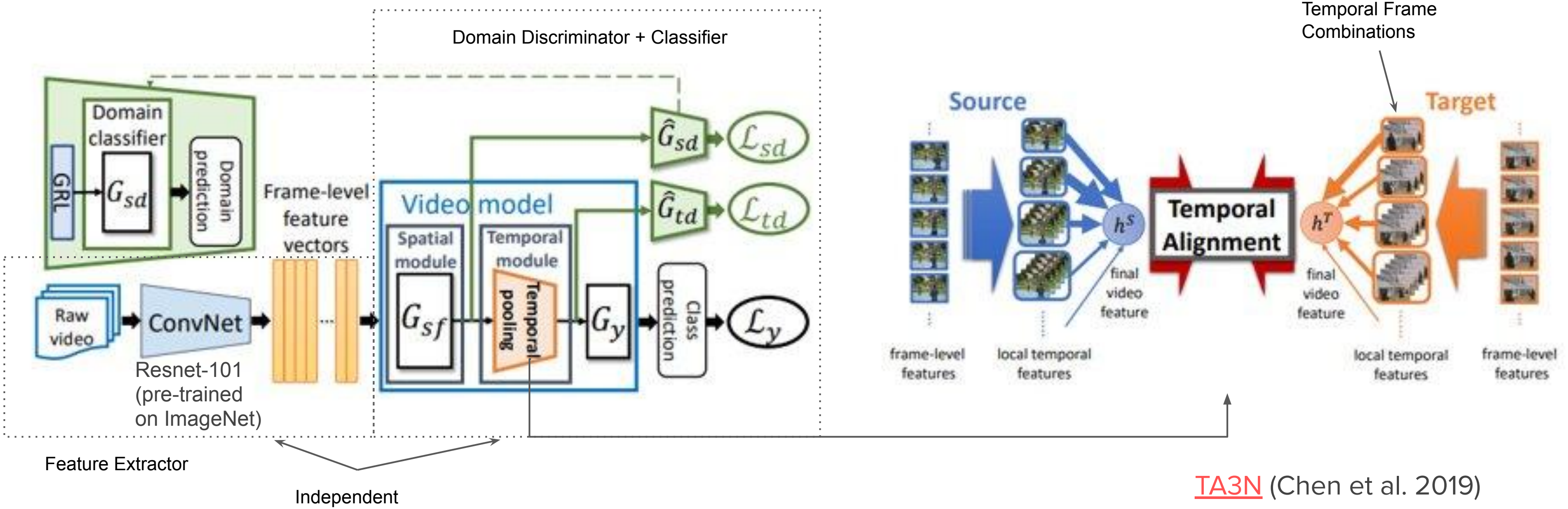
Input: Labeled videos from source and unlabeled videos from target domain

Output: Prediction results on unlabeled video dataset



Related Work: Temporal Attentive Alignment Network

- Frame Attention-based DA
- Temporal Relation network to perform temporal pooling
- Pre-extracted spatial features
- DANN on individual spatial features and pooled temporal features



TA3N (Chen et al. 2019)

Approach: Overview

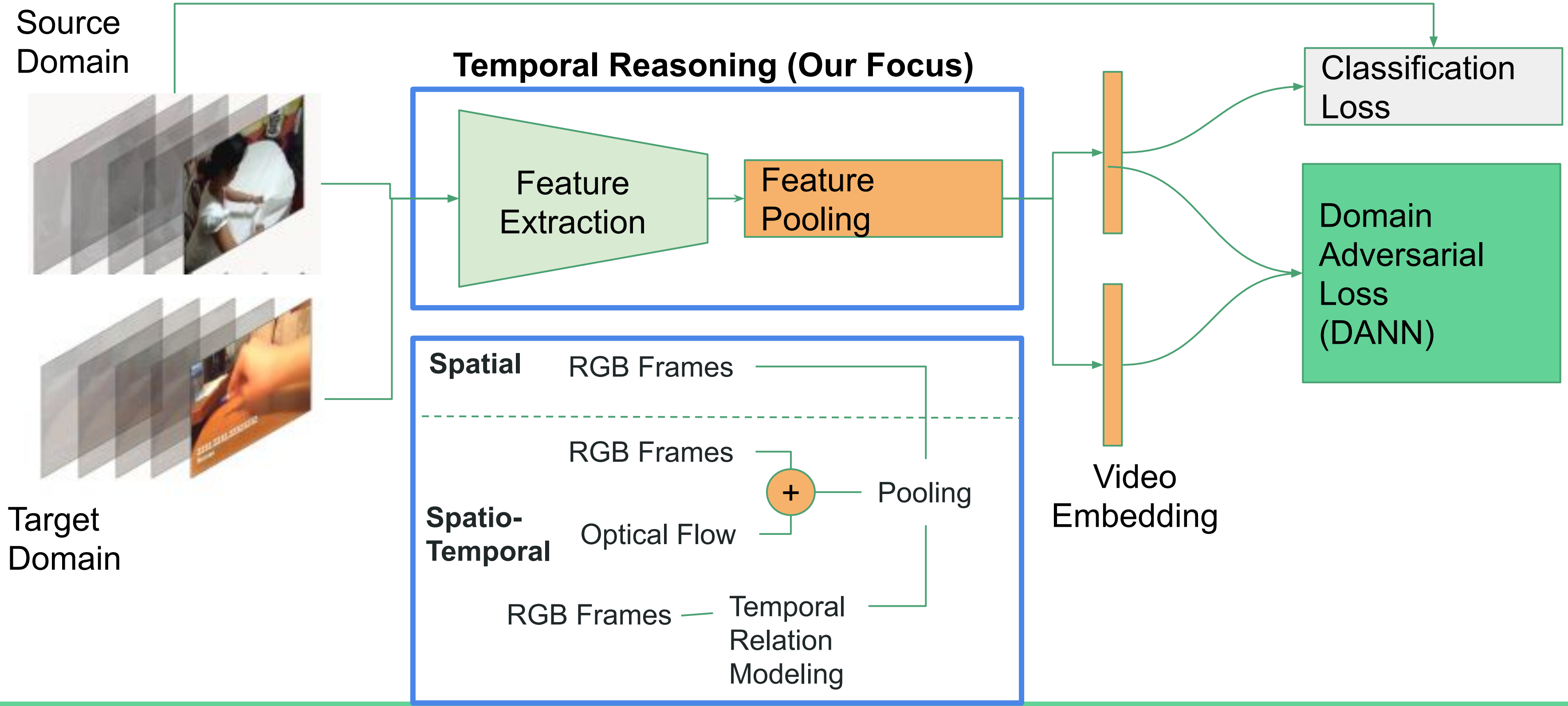
Goal: Leverage rich temporal information in videos to improve alignment and recognition performance

Our Contributions:

- Simultaneous learning & alignment of temporal relations benefit video DA
- Explore alternative frame sampling
- Explore temporal pooling mechanisms

Frame selection

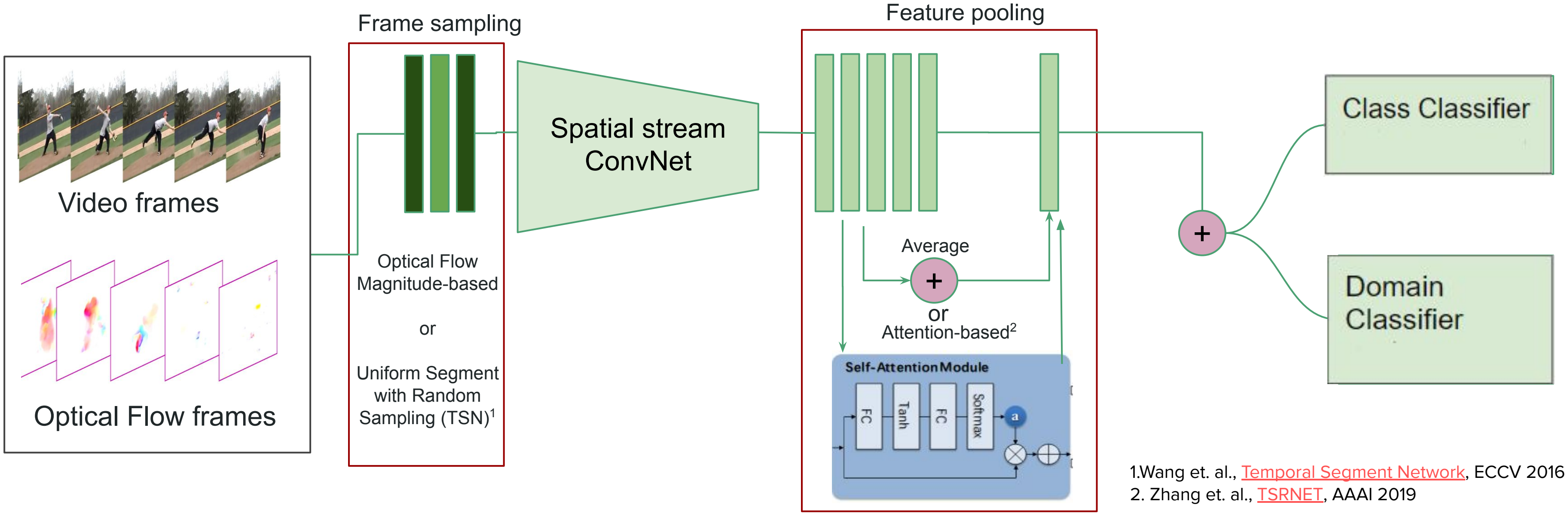
Approach: Methodology



Spatial DA

Hypothesis: Improving spatial feature selection and pooling should improve spatial DA

Approach: Optical flow based spatial frame-sampling led to slightly better performance

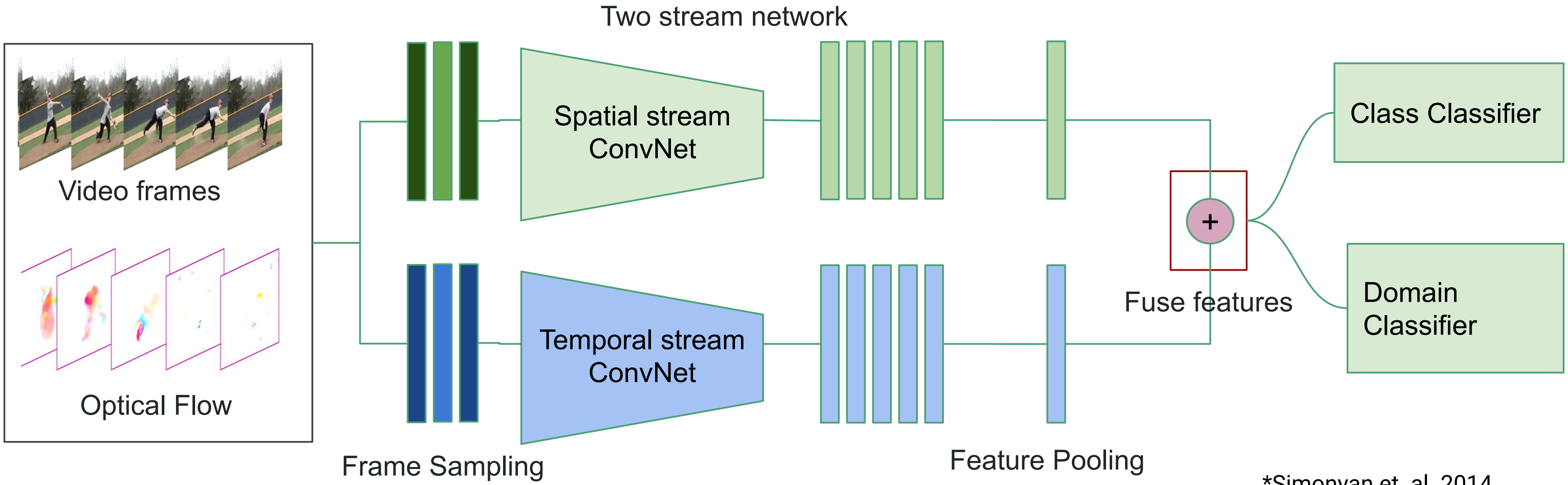


1.Wang et. al., Temporal Segment Network, ECCV 2016
2. Zhang et. al., TSRNET, AAAI 2019

Spatio-Temporal DA

Hypothesis: Incorporating motion-based feature improve performance over spatial DA

Approach: DANN on fused spatial and optical-flow features in two-stream network*

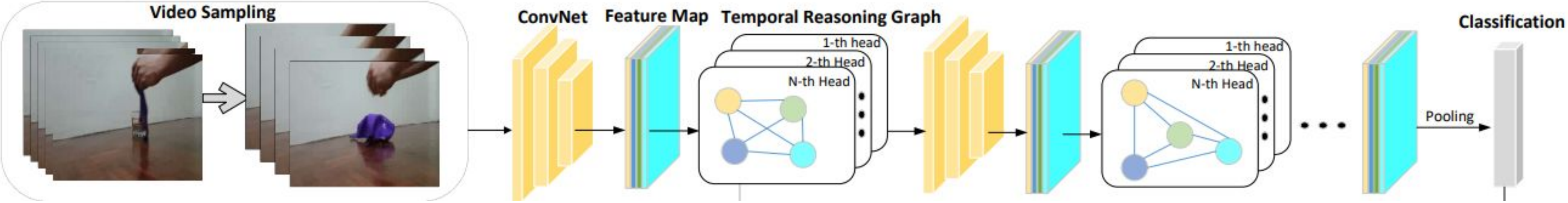


*Simonyan et. al. 2014

Spatio-Temporal DA: Integrated temporal modeling

Hypothesis: Improved feature maps with temporal relation modeling should beat spatial

Approach: Learn short and long term relationship between convolutional feature maps of RGB frames using a Attention-based Graph Convolutional Network

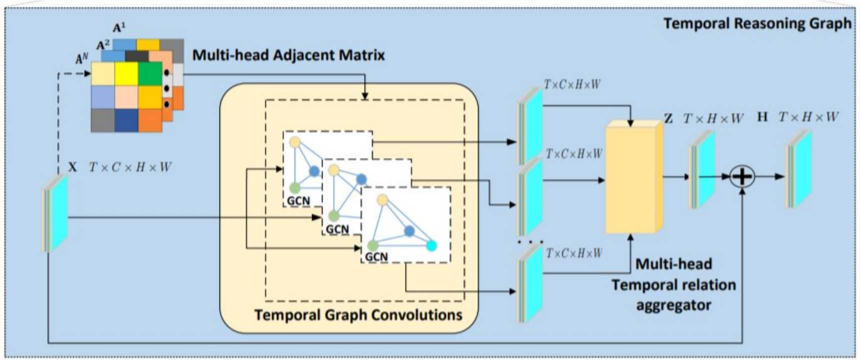


Overall architecture of Temporal Graph Convolutional Net

Zhang, Jingran, et al. "Temporal Reasoning Graph for Activity Recognition." arXiv preprint arXiv:1908.09995 (2019).

Spatio-Temporal DA: Integrated temporal modeling

- Stacked graph convolution layers
- Multiple **learnable adjacency matrices** at each layer to learn different relations
- Node: Frame feature vector at that layer
- Edge: Temporal "relation" between frames



Single graph convolutional layer with multi-head adjacency matrix

Zhang, Jingran, et al. "Temporal Reasoning Graph for Activity Recognition." arXiv preprint arXiv:1908.09995 (2019).

Experiments

Setup: Labeled Source Dataset + Unlabeled Target Dataset

- Non - DA : Source only, Target only
- DA : Spatial Module (Baseline)
- DA : Spatial-Temporal Module

Dataset: UCF101 - HMDB51

- 12 overlapping classes *
- UCF: 2009 videos; HMDB: 1200 videos (Train/Test - 70/30)

Metrics:

Gain (prec@1) : Model with DA compared to model trained only on Source

Network Architecture : Resnet - 34

UCF101

HMDB51

* *Climb, fencing, golf, kick_ball, pullup, punch, walk, pushup, ride_bike, ride_horse, shoot_ball, shoot_bow*

Dataset Discrepancy*

accuracy metric: precision@1

Spatial Model **Source dataset**	Target dataset	
	UCF	HMDB
UCF	90.54	61.01 (-22.32)
HMDB	64.45 (-26.09)	83.33

Motion Model **Source dataset**	Target dataset	
	UCF	HMDB
UCF	90.89	56.94 (-13.34)
HMDB	68.65 (-22.24)	70.28

*measured using standard 2-stream network configuration for Spatial and Temporal CNN

Domain Adaptation Results - UCF > HMDB

Temporal Reasoning Module	4 spatial frames		8 spatial frames	
	Prec@1	Gain vs source only	Prec@1	Gain vs source only
Target only	87.22	-	85.28	-
Source only	67.02	-	68.61	-
Spatial	68.06	1.04	71.17	2.56
Spatial + Optical Flow (concatenate)	69.34	2.32	**72.92**	**4.31**
Spatial + Optical Flow (conv)	**69.64**	**2.62**	71.73	3.12
Spatial + Optical Flow (Separate DA)	69.04	2.02	**72.92**	**4.31**
Spatial + Temporal Graph	67.50	0.48	68.89	0.28
TemRelation*	75.28	3.61*	-	-
TA3N (TemRelation + Domain Attention)*	78.33	6.66*	-	-

Domain Adaptation

*Chen et. al., “Temporal Attentive Alignment for Large-Scale Video Domain Adaptation”, ICCV 2019

Domain Adaptation Results - HMDB > UCF

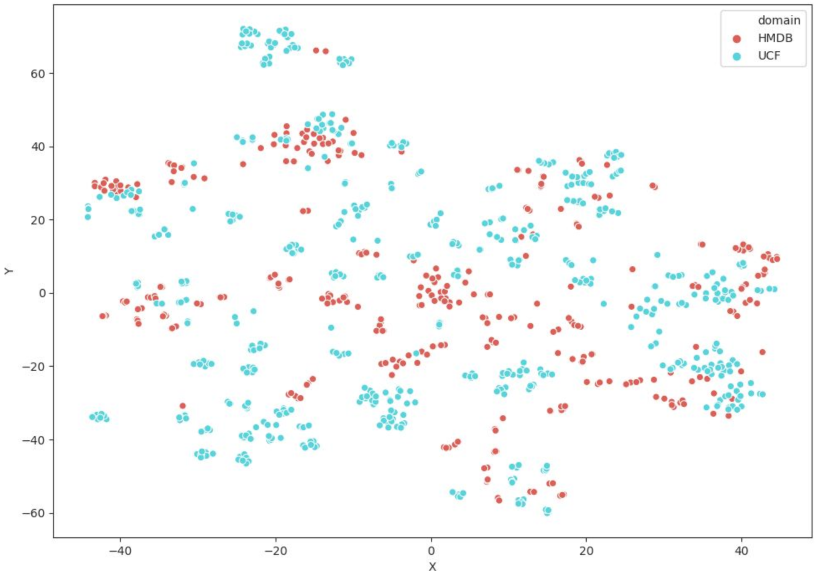
	Temporal Reasoning Module	4 spatial frames		8 spatial frames	
		Prec@1	**Gain (w.r.t. source only)**	**Prec@1**	**Gain (w.r.t. source only)**
	Target only	94.31	-	94.95	-
	Source only	71.59	-	72.63	-
Domain Adaptation	Spatial	74.21	2.63	76.32	3.69
	Spatial + Optical Flow (concatenate)	75.31	3.72	78.46	5.83
	Spatial + Optical Flow (conv)	76.18	4.59	**79.51**	**6.88**
	Spatial + Optical Flow (Separate DA)	**76.36**	**4.77**	77.06	4.43
	Spatial + Temporal Graph	71.80	0.21	73.68	1.05
	TemRelation*	76.36	4.77*	-	-
	TA3N (TemRelation + Domain Attention)*	81.79	10.20*	-	-

*Chen et. al., “Temporal Attentive Alignment for Large-Scale Video Domain Adaptation”, ICCV 2019

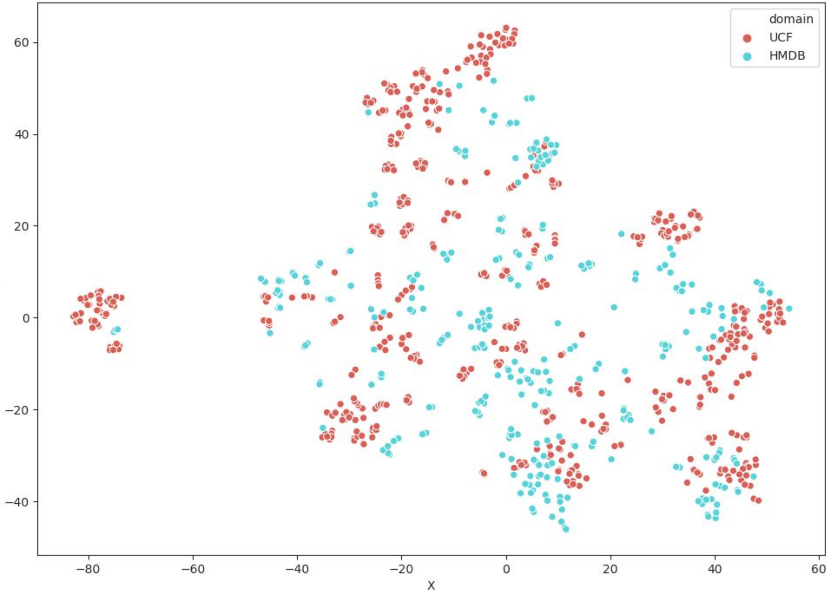
Analysis/Takeaways

1. Optical Flow features as complementary temporal information help alignment and improve the performance on target data
2. UCF→HMDB is a harder adaptation task than HMDB→UCF
3. Different pooling strategies do not show a significant difference in performance
4. Temporal relation graph does not do much better than the spatial DANN
    a. It overfits on the non-DA activity recognition task
    b. Has more parameters, and may require a larger dataset (like in the original paper)

tSNE Visualization (Spatial + Optical Flow)

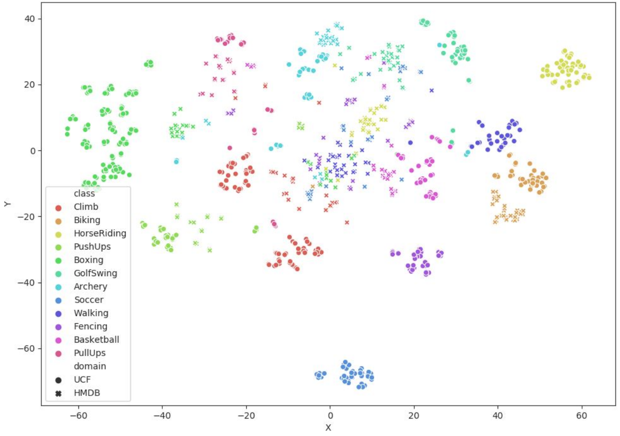


HMDB > UCF (8 frames)

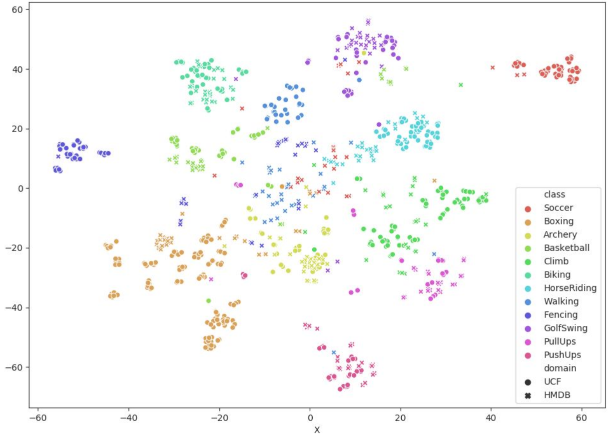


UCF > HMDB (8 frames)

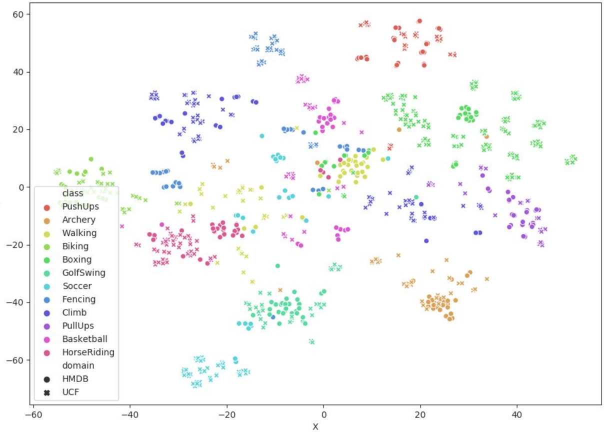
tSNE Visualization (Class-wise Alignment)



Source Only (No DA)



Spatial DA (UCF > HMDB)



Spatial DA (HMDB > UCF)

classes difficult to align : soccer, fencing, walking

*spatial DA using 4 frames

Sampling and Pooling Strategies

Approaches		4 spatial frames	
		UCF > HMDB	**HMDB > UCF**
Spatial Feature Sampling	Uniform Segments + Random	68.06	**74.21**
	Probabilistic (optical flow)	**71.67**	70.70
Feature Pooling	Average	68.06	**74.21**
	Attention-based	**70.00**	72.81

Optical flow-based frame sampling leads to better performance in UCF>HMDB

Train/Test: 70/30 split; Network Architecture: Resnet-34; DANN for adaptation

Discussion: Conclusion and Challenges

Conclusion

Investigated the domain shift problem on cross videos action recognition

Learning & alignment of temporal relations achieves better domain alignment

Fusing optical flow features as complementary to RGB lead to better alignment

Challenges

Global alignment of temporal features could confuse the model for prediction

Smaller scale dataset constraints on network architecture

Discussion: Future Work

- Better spatial-temporal learning and alignment for cross video DA, especially using only RGB frames
- Auxiliary pre-text tasks on target dataset to provide self-supervision
- DA on larger scale cross-domain video datasets
- Other cross-domain video tasks: segmentation and detection

Thank You