Temporal Attentive Alignment for Large-Scale Video Domain Adaptation

Min-Hung Chen[1] Zsolt Kira[1] Ghassan AlRegib[1] Jaekwon Yoo[2] Ruxin Chen[2] Jian Zheng[3]

[1]Georgia Institute of Technology [2]Sony Interactive Entertainment LLC [3]Binghamton University

ghassanalregib.info/

ICCV 2019 Seoul, Korea

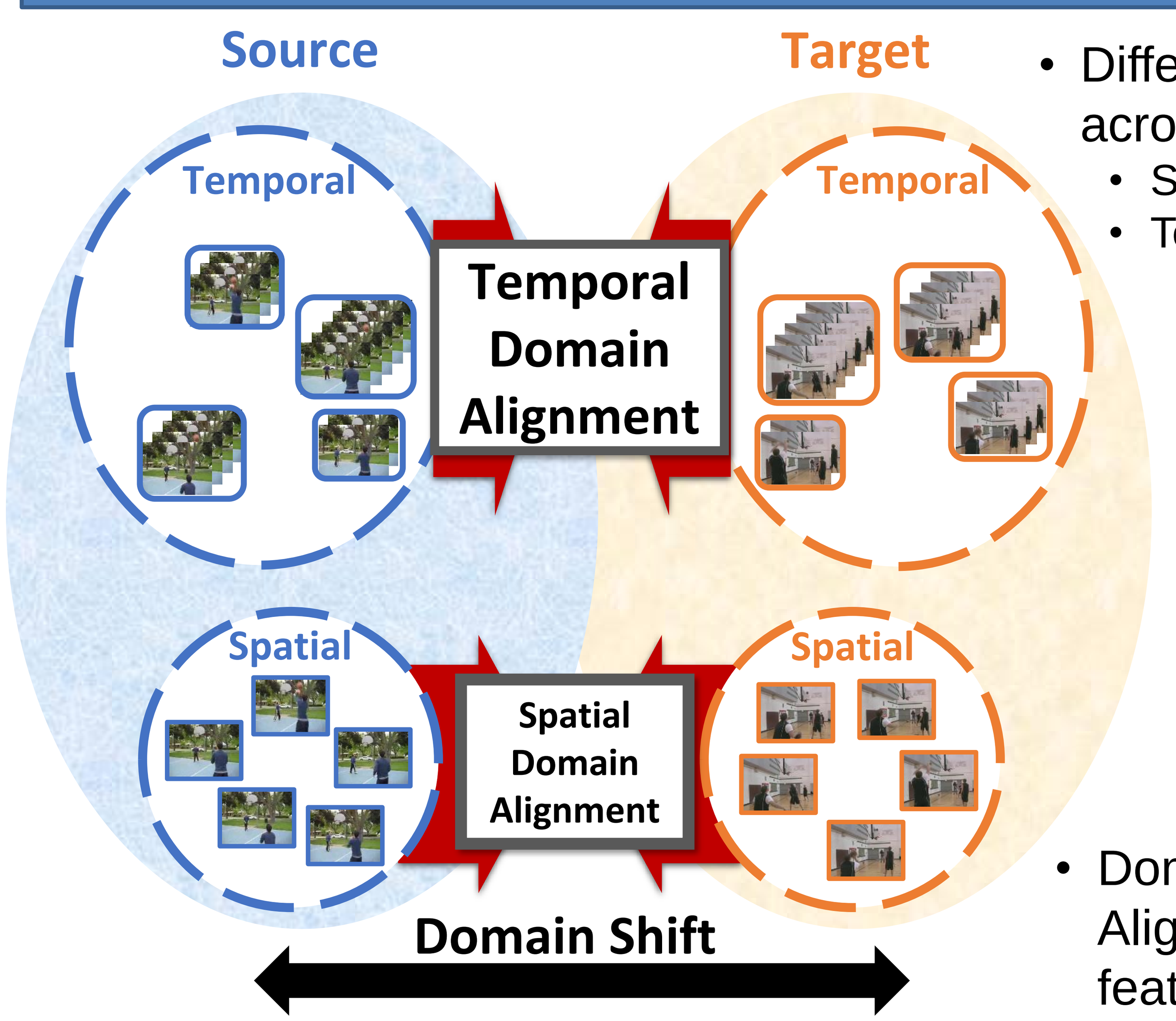
Code: https://github.com/cmhungsteve/TA3N

Paper: https://arxiv.org/abs/1907.12743

Summary

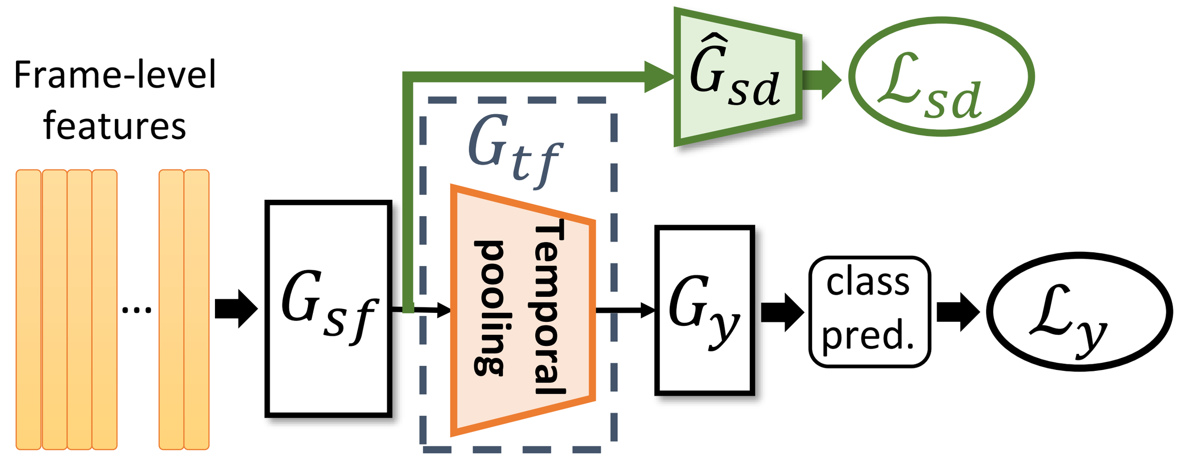
- Domain adaptation for videos is an under-explored real-world problem.
- Three contributions:
  - Large-scale Video DA Dataset Collection: UCF-HMDB$_{full}$ & Kinetics-Gameplay
  - Exploration of Feature Alignment for Video DA: Exploration of how to effectively align spatio-temporal features
  - Temporal Attentive Adversarial Adaptation Network (TA^{3}N): Simultaneously attend, align and learn temporal dynamics across domains
- State-of-the-art performance on four cross-domain video datasets

Visual Domain Shift



- Different data distributions across domains
  - Spatial: frames
  - Temporal: video clips
- Domain adaptation (DA): Align the spatio-temporal feature representations

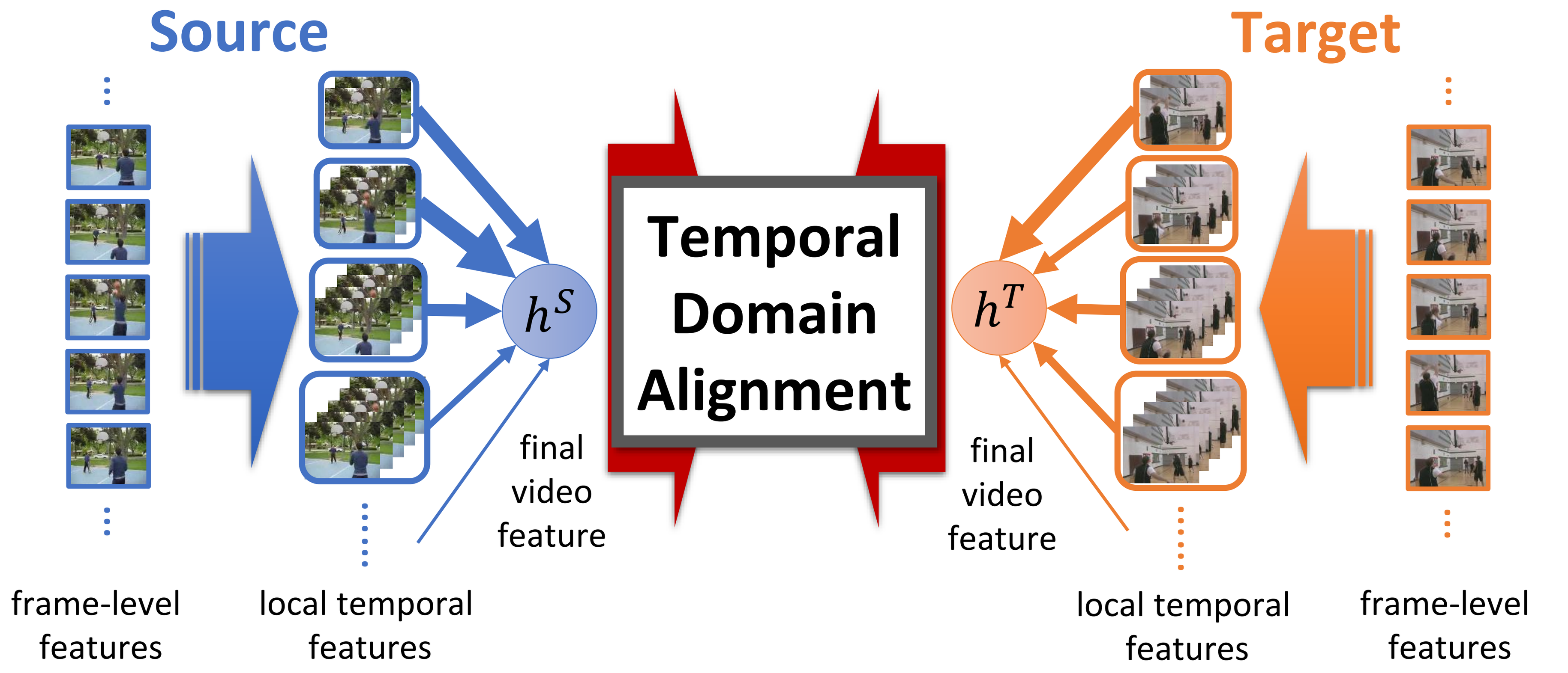
Baseline: DANN [1] + TemPooling



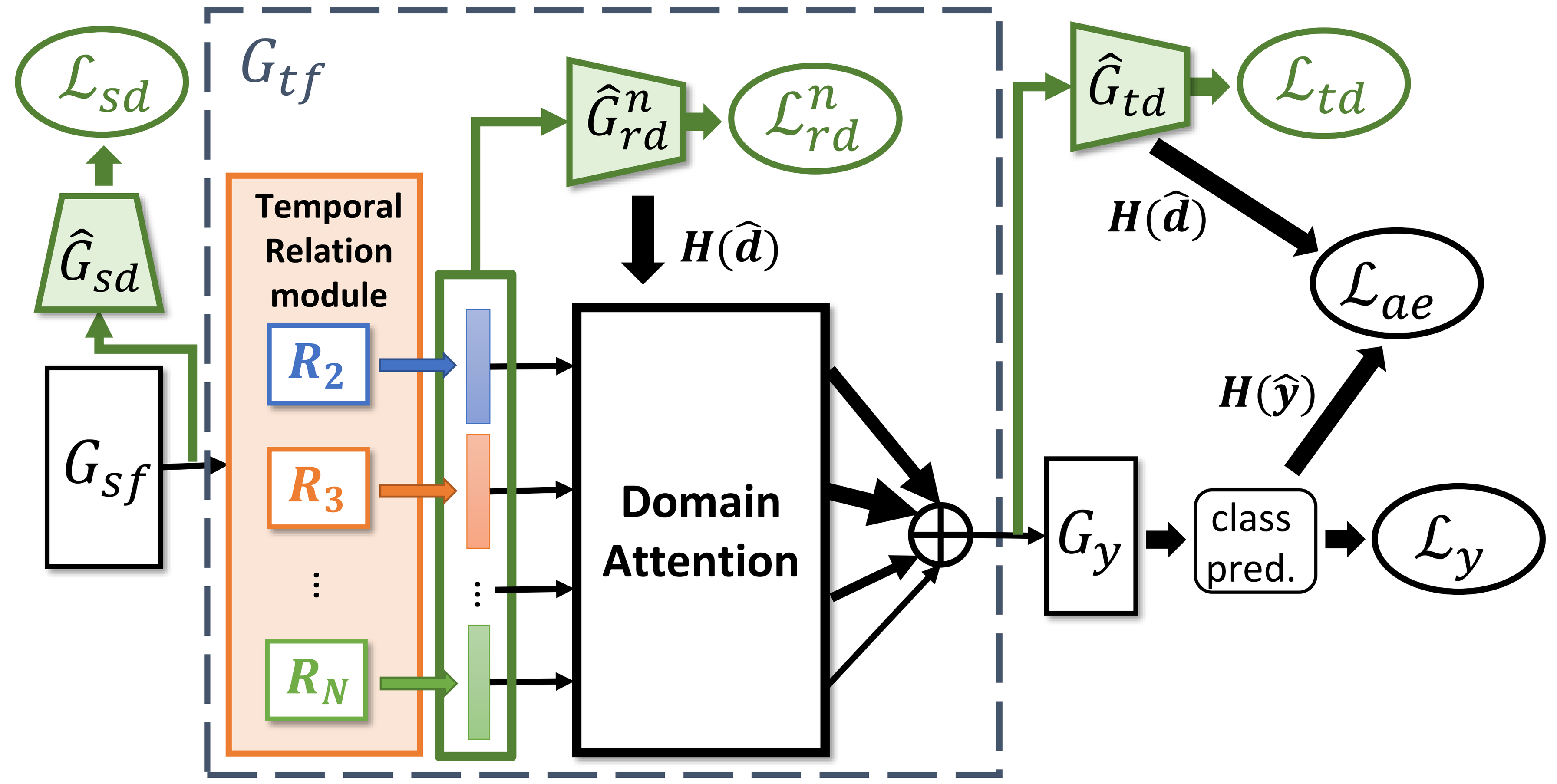
- Spatial module G_{sf}
- Temporal module G_{tf}
  - TemPooling
- Adversarial discriminator $\hat{G}_{sd}$
  - Domain loss $\mathcal{L}_{sd}$

Main Idea

- Focus more on aligning the video clips with larger domain discrepancy ➔ higher contribution to overall domain shift
  - $video\ feature\ \boldsymbol{h} = \sum attention\ weight \cdot local\ temporal\ features$
  - Use domain discrepancy to calculate the attention weights
- Simultaneously align and learn temporal dynamics



TA^{3}N: Temporal Attentive Adversarial Adaptation Network



- TemRelation module: n-frame relation features as local temporal features
- Domain attention: get attention weights using domain entropy $H(\hat{d})$
- Attentive entropy loss $\mathcal{L}_{ae}$: aim to minimize entropy within each domain

Datasets

U: UCF; H: HMDB; O: Olympic; K: Kinetics; G: Gameplay

Our datasets: U-H$_{full}$, K-G

	U-H$_{small}$	U-O	U-H$_{full}$	K-G
Class #	5	6	12	30
Video #	1171	1145	3209	49998

Detroit: Become Human™ ©Sony Interactive Entertainment Europe, developed by Quantic Dream

Experimental Results

UCF-Olympic / UCF-HMDB$_{small}$

Source → Target (S → T)	U → O	O → U	U → H	H → U
W. Sultani et al. [2]	33.3	47.9	68.7	68.7
T. Xu et al. [3]	87.0	75.0	82.0	82.0
AMLS [4]	84.6	86.4	89.5	95.4
DAAA [4]	91.6	90.0	-	-
TemPooling (Source-only)	96.3	87.1	98.7	97.4
TemPooling + DANN [1]	**98.2**	90.0	**99.3**	98.4
TA^{3}N	**98.2**	**92.9**	**99.3**	**99.5**

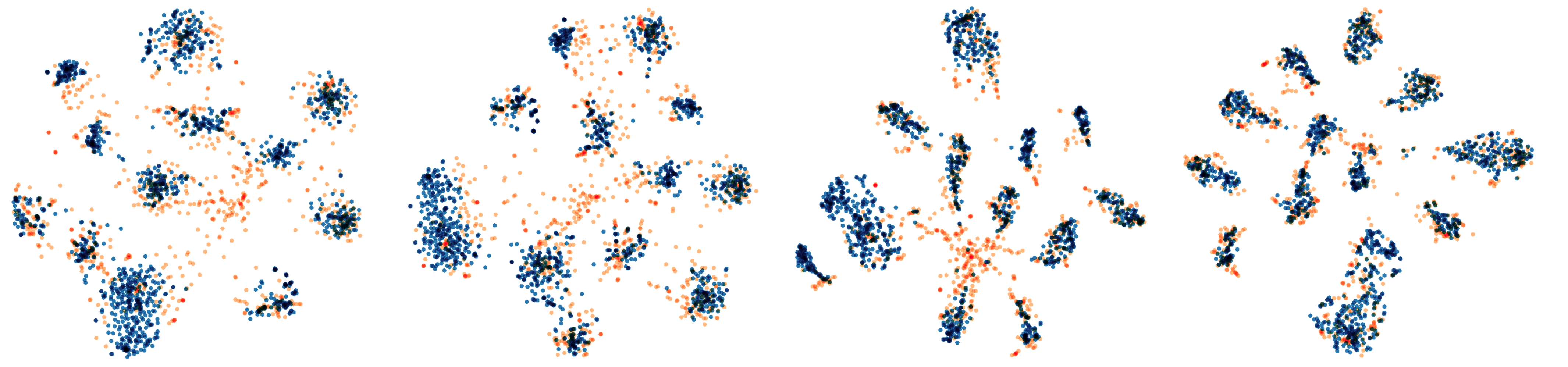
Kinetics-Gameplay / UCF-HMDB$_{full}$

S → T	U → H		H → U		K → G	
Target-only	82.8	-	94.9	-	64.5	-
Source-only	71.7	Gain	73.9	Gain	17.2	Gain
DANN [1]	74.4	2.7	74.4	0.5	20.6	3.4
JAN [5]	74.7	3.0	79.7	5.8	18.2	1.0
AdaBN [6]	72.2	0.5	77.4	3.5	20.3	3.1
MCD [7]	73.9	2.2	79.3	5.4	19.8	2.6
TA^{3}N	**78.3**	**6.6**	**81.8**	**7.9**	**27.5**	**10.3**

Source-only: trained with source data only

Gain: accuracy improvement over Source-only

What to align >>> *What DA methods* to use



TemPooling   TemPooling + DANN [1]   TemRelation   TA^{3}N

[1] (JMLR 16), [2] (CVPR 14), [3] (IVC 16), [4] (BMVC 18), [5] (ICML 17), [6] (ICLRW 17), [7] (CVPR 18)