Spatio-temporal Contrastive Domain Adaptation for Action Recognition (Supplementary Material)

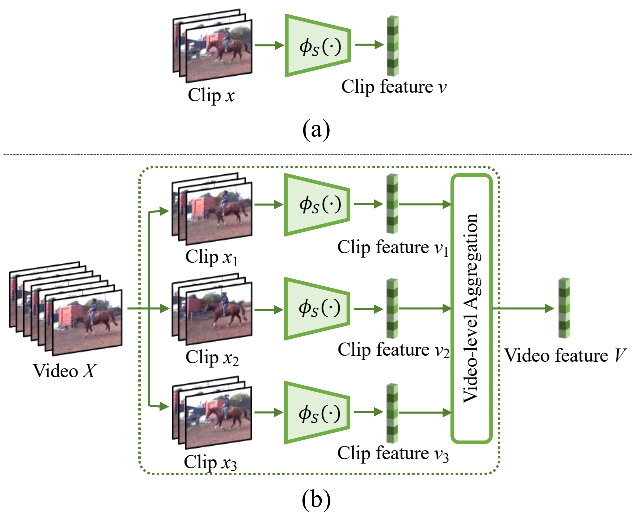

Figure 1. The feature representation on (a) the clip level and (b) the video level. Use the RGB modality as the example.

1. More Method Details

Feature representation on different level. We conduct our spatio-temporal contrastive domain adaptation (STCDA) framework with clip-level and video-level features, which are utilized in both components of spatio-temporal contrastive learning (STCL) and video-based contrastive alignment (VDA).The feature representation on each level is extracted as Figure 1. In particular, video-level feature is aggregated with sampled clip features at different times.

Memory bank. The memory bank mechanism for STCL conducts a non-parametric network branch without back-propagation, which aims at storing the representations computed from clip-level/video-level feature extraction. The representation of a sample in the memory bank is updated when the feature appears with same index [10].

2. Datasets

Olympic Sports. Olympic Sports dataset [3] contains videos from YouTube of athletes practicing different sports with 16 categories.

Table 1. The category list of UCF–Olympic.

UCF50	Olympic
Basketball	basketball_layup
CleanAndJerk	clean_and_jerk
ThrowDiscus	discus_throw
Diving	diving_springboard_3m
PoleVault	pole_vault
TennisSwing	tennis_serve

Table 2. The category list of UCF–HMDB$_{small}$.

UCF101	HMDB51
GolfSwing	golf
PullUps	pullup
Biking	ride_bike
HorseRiding	ride_horse
Basketball	shoot_ball

HMDB51. HMDB51 dataset [4] has 51 action categories, which totally contain 6,766 manually annotated videos, which are extracted from a variety of sources, with face actions, body movements, and human-object interaction.

UCF50 and UCF101. These two datasets consist of realistic action recognition videos collected from Youtube. UCF50 dataset [6] has 50 action categories with a total of 6,676 videos. UCF101 dataset [8] is a extension of UCF50 with extra action categories, which consists of 101 action classes with 13,320 videos from YouTube, with realistic user-uploaded videos with large variations in motion, pose and scales, containing camera motion and cluttered background.

UCF–Olympic. UCF–Olympic dataset [2] consists of 6 common categories from UCF50 and Olympic Sports datasets, and the shared classes are listed in Table 1.

UCF–HMDB$_{small}$. UCF–HMDB$_{small}$ dataset [9] consists of 5 common categories from UCF101 and HMDB51 datasets, and the shared classes are listed in Table 2.

UCF–HMDB$_{full}$. UCF–HMDB$_{full}$ dataset [1] consists of 12 common categories from UCF101 and HMDB51 datasets, and the shared classes are listed in Table 3.

EPIC Kitchens. EPIC Kitchens [5] is a fine-grained cross-domain action recognition dataset, with eight action categories (‘put’, ‘take’, ‘open’, ‘close’, ‘wash’, ‘cut’, ‘mix’,

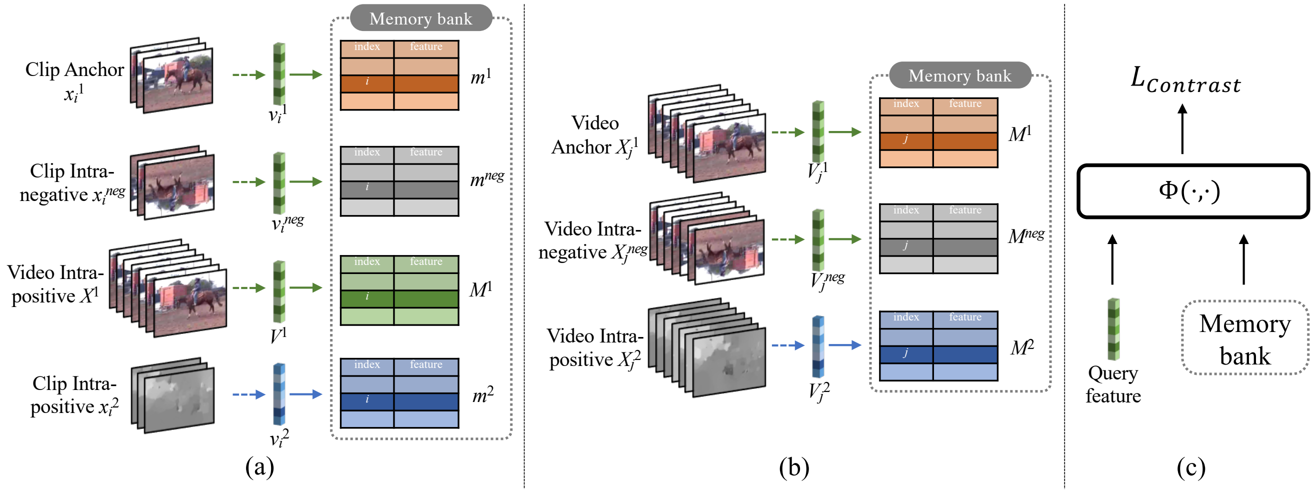

Figure 2. Clip-level and video-level contrastive learning in STCL. Use the RGB modality as the example. (a) Clip-level memory bank. (b) Video-level memory bank. (c) Contrastive learning with memory bank.

Table 3. The category list of UCF–HMDB$_{full}$.

UCF101	HMDB51
RockClimbingIndoor, RopeClimbing	climb
Fencing	fencing
GolfSwing	golf
SoccerPenalty	kick_ball
PullUps	pullup
Punch, BoxingPunchingBag, BoxingSpeedBag	punch
PushUps	pushup
Biking	ride_bike
HorseRiding	ride_horse
Basketball	shoot_ball
Archery	shoot_bow
WalkingWithDog	walk

and 'pour'). The dataset is imbalanced with different numbers of training data in each category. It contains 3 domains ('D1', 'D2', and 'D3'), and the evaluation is involved on pairs for each other with 6 different settings ('D1→D2', 'D1→D3', 'D2→D1', 'D2→D3', 'D3→D1', and 'D3→D2').

3. More Results

Experimental results on RGB and optical flow. We have compared our STCDA with different modalities of RGB and optical flow on each benchmark. In Table 4, Table 5 and Table 6, the framework obtain the results on UCF–HMDB$_{small}$, UCF–Olympic, UCF–HMDB$_{full}$, and EPIC Kitchens, respectively.

Visualization. We indicate more samples of target videos and predictions in Figure 3 and Figure 4 on different datasets, to present the heat map of activation region for corresponding prediction. Besides, we show the confidence score of each predicted results. The visualization results show that the network focuses on relevant action position with a higher confidence score using the proposed STCDA framework.

References

[1] Min-Hung Chen, Zsolt Kira, Ghassan AlRegib, Jaekwon Yoo, Ruxin Chen, and Jian Zheng. Temporal attentive alignment for large-scale video domain adaptation. In *ICCV*, 2019. 1

[2] Arshad Jamal, Vinay P Namboodiri, Dipti Deodhare, and KS Venkatesh. Deep domain adaptation in action space. In *BMVC*, 2018. 1

[3] Chih-Wei Chen Juan Carlos Niebles and Li Fei-Fei. Modeling temporal structure of decomposable motion segments for activity classification. In *ECCV*, 2010. 1

[4] H. Kuehne, H. Jhuang, E. Garrote, T. Poggio, and T. Serre. Hmdb: A large video database for human motion recognition. In *ICCV*, 2011. 1

[5] Jonathan Munro and Dima Damen. Multi-modal domain adaptation for fine-grained action recognition. In *CVPR*, 2020. 1

[6] Kishore K. Reddy and Mubarak Shah. Recognizing 50 human action categories of web videos. *Machine Vision and Applications*, 24(5):971–981, 2013. 1

[7] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. *IJCV*, 128(2):336–359, 2020. 4

[8] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. Ucf101: A dataset of 101 human actions classes from videos in the wild. *Computer ence*, 2012. 1

[9] J. Tang, H. Jin, S. Tan, and D. Liang. Cross-domain action recognition via collective matrix factorization with graph laplacian regularization. volume 55, pages 119–126, 2016. 1

Table 4. Comparison of accuracy (%) on UCF–HMDB$_{small}$ and UCF–Olympic.

Method	Modality	Backbone	UCF→HMDB	HMDB→UCF	UCF→Olympic	Olympic→UCF
Source only	RGB	BN-Inception	94.7	97.7	91.6	90.4
STCDA	RGB	BN-Inception	97.3	99.3	94.4	93.3
Target only	RGB	BN-Inception	98.7	99.5	96.3	98.3
Source only	Flow	BN-Inception	92.0	94.2	87.0	85.4
STCDA	Flow	BN-Inception	95.3	95.2	92.6	92.1
Target only	Flow	BN-Inception	96.7	98.9	96.3	96.3
Source only	RGB + Flow	BN-Inception	96.7	99.3	94.4	92.9
STCDA	RGB + Flow	BN-Inception	**98.7**	**100**	**98.1**	**96.3**
Target only	RGB + Flow	BN-Inception	100	100	98.1	100

Table 5. Comparison of accuracy (%) on UCF–HMDB$_{full}$.

Method	Modality	Backbone	UCF→HMDB	HMDB→UCF
Source only	RGB	BN-Inception	74.1	82.5
STCDA	RGB	BN-Inception	76.9	85.1
Target only	RGB	BN-Inception	91.7	94.7
Source only	Flow	BN-Inception	71.1	75.1
STCDA	Flow	BN-Inception	75.3	83.4
Target only	Flow	BN-Inception	83.9	96.3
Source only	RGB + Flow	BN-Inception	76.1	85.8
STCDA	RGB + Flow	BN-Inception	80.0	87.7
Target only	RGB + Flow	BN-Inception	94.2	96.8
Source only	RGB	I3D	80.8	88.4
STCDA	RGB	I3D	81.9	91.9
Target only	RGB	I3D	94.4	96.3
Source only	Flow	I3D	77.8	85.8
STCDA	Flow	I3D	80.0	88.1
Target only	Flow	I3D	91.9	94.6
Source only	RGB + Flow	I3D	82.8	89.8
STCDA	RGB + Flow	I3D	**83.1**	**92.1**
Target only	RGB + Flow	I3D	95.8	97.7

[10] Stella X. Yu Zhirong Wu, Yuanjun Xiong and Dahua Lin. Unsupervised feature learning via non-parametric instance discrimination. In *CVPR*, 2018. 1

Table 6. Comparison of accuracy (%) on EPIC Kitchens.

Method	Modality	D2→D1	D3→D1	D1→D2	D3→D2	D1→D3	D2→D3
Source only	RGB	37.9	37.4	41.1	37.9	36.0	35.1
STCDA	RGB	44.4	41.1	47.7	45.5	41.2	47.6
Target only	RGB	54.7	54.7	63.3	63.3	64.7	64.7
Source only	Flow	40.2	39.8	42.4	50.5	38.7	45.2
STCDA	Flow	45.3	52.2	45.1	**59.5**	44.0	51.2
Target only	Flow	59.1	59.1	72.7	72.7	63.9	63.9
Source only	RGB + Flow	44.4	48.5	46.5	52.8	40.6	45.3
STCDA	RGB + Flow	**49.0**	**52.6**	**52.0**	55.6	**45.5**	**52.5**
Target only	RGB + Flow	63.9	63.9	74.9	74.9	72.0	72.0

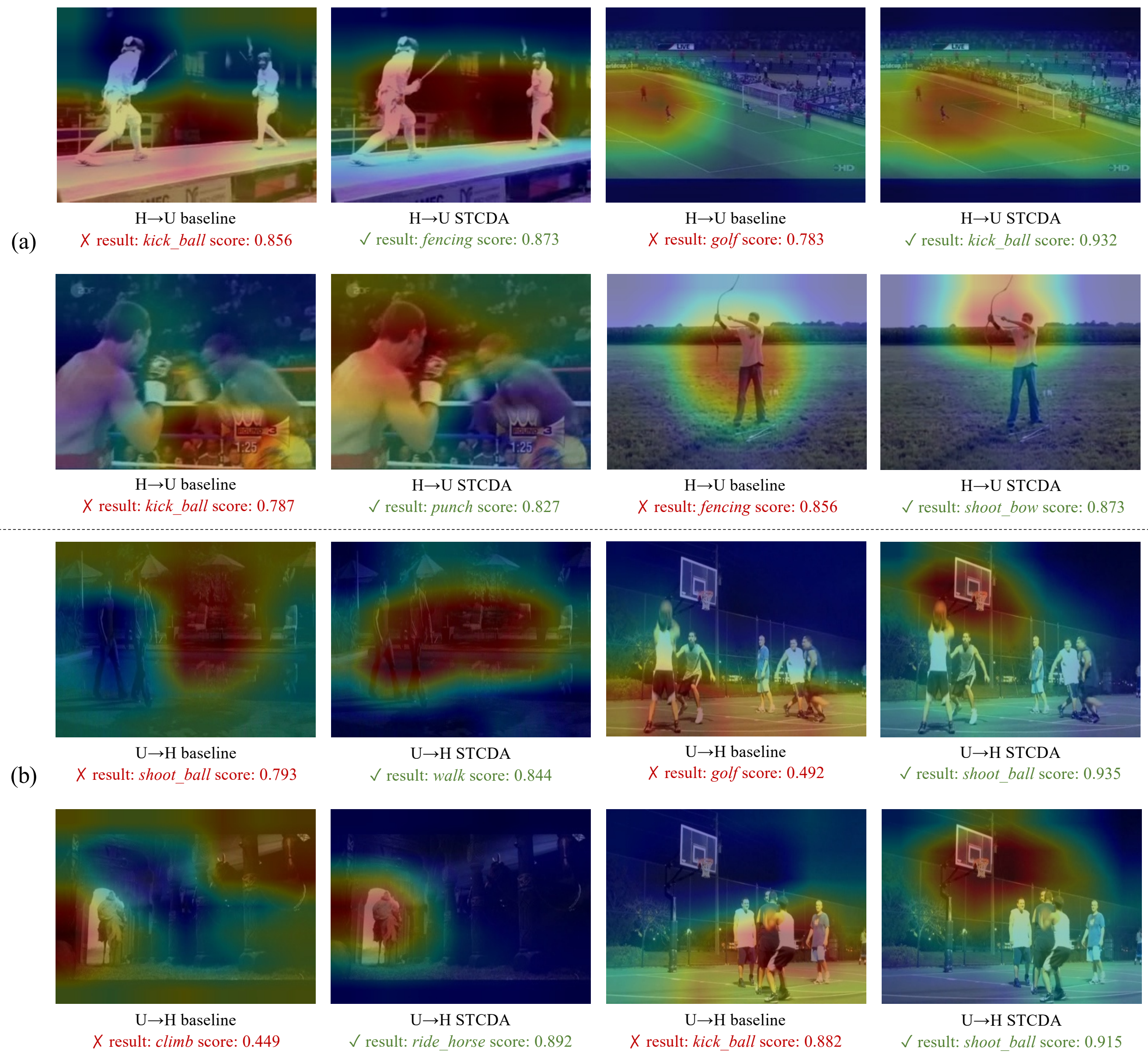

Figure 3. Visualization of Grad-CAM [7] on UCF–HMDB$_{full}$ dataset. Examples are sampled from (a) UCF dataset and (b) HMDB dataset. "score" means the confidence score of the current prediction.

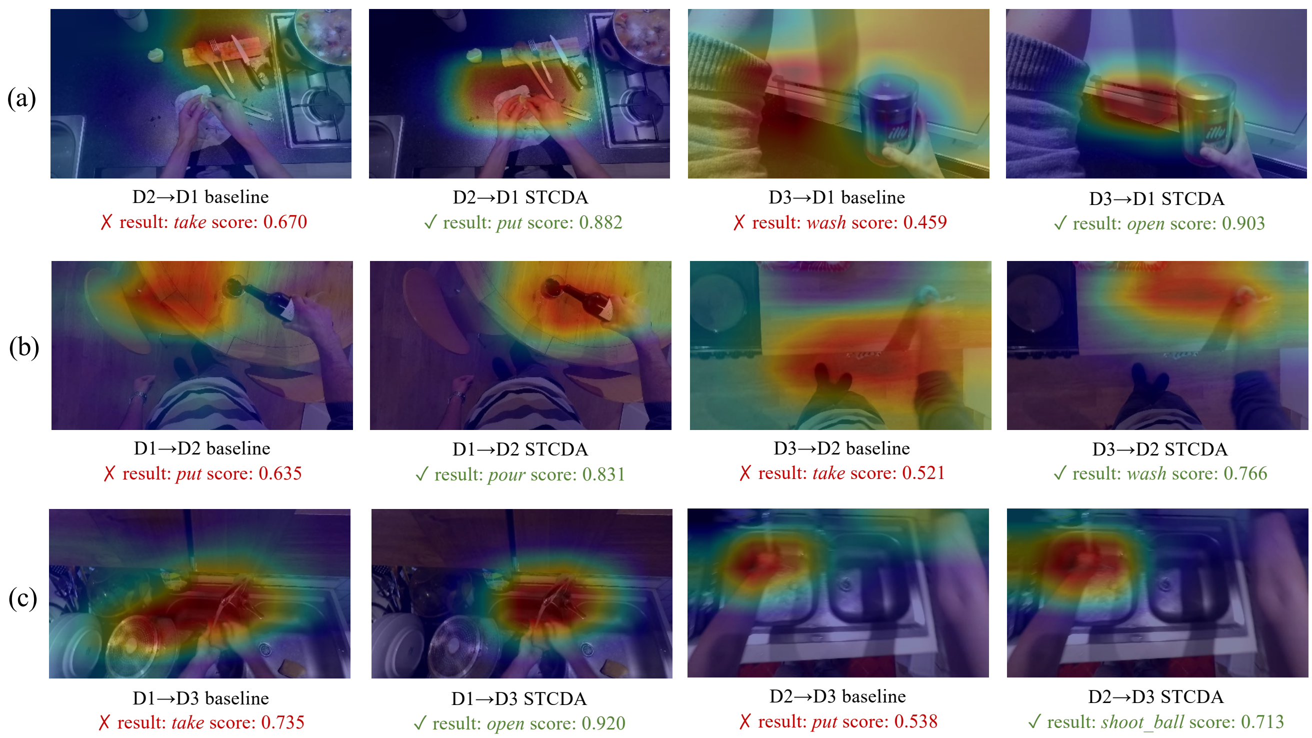

Figure 4. Visualization of Grad-CAM on EPIC Kitchens dataset. Examples are sampled from EPIC Kitchens of (a) D1 subset, (b) D2 subset and (c) D3 subset. “score” means the confidence score of the current prediction.