Source-free Video Domain Adaptation by Learning Temporal Consistency for Action Recognition Supplementary Material

Yuecong Xu[1⋆], Jianfei Yang[2⋆], Haozhi Cao[2], Keyu Wu[1], Min Wu[1], and Zhenghua Chen[1](✉)

[1] Institute for Infocomm Research, A*STAR, Singapore
xuyu0014@e.ntu.edu.sg, {wu_keyu, wumin}@i2r.a-star.edu.sg, chen0832@e.ntu.edu.sg
[2] School of Electrical and Electronic Engineering, Nanyang Technological University, Singapore
{yang0478,haozhi001}@e.ntu.edu.sg

This appendix presents more details of the proposed Attentive Temporal Consistent Network (ATCoN) and is organized as follows: **first**, we introduce the detailed implementation of ATCoN with specific hyperparameter settings supported by hyperparameter sensitivity analysis. **Subsequently**, we visualize local temporal features learned by various models to justify the proposed *cross-temporal hypothesis*. **Lastly**, we present details of the cross-domain action recognition benchmarks for evaluating ATCoN.

1 Detailed Implementation of Attentive Temporal Consistent Network (ATCoN)

In this work, we propose the Attentive Temporal Consistent Network (ATCoN) to address SFVDA by learning temporal consistency, guaranteed by two novel consistency objectives: feature consistency and source prediction consistency, performed across local temporal features. The proposed ATCoN further constructs effective overall temporal features by attending to local temporal features based on prediction confidence. In this section, we further present the implementation of ATCoN in detail, whose structure is displayed in Fig. 1. To obtain video features, we instantiate Temporal Relation Network (TRN) [18] with ResNet-50 [4] as the model backbone. TRN has been widely adopted in previous video domain adaptation tasks, such as VUDA [2,15], MSVDA [17], and PVDA [14], bringing state-of-the-art results, thanks to its ability in extracting accurate temporal features by reasoning over correlations between spatial representations which coincides with the process of humans recognizing actions. Following [9], a Batch Normalization [5] and an additional fully connected layer are inserted while weight normalization [12] is applied to the last fully connected layer. *Code is provided at https://github.com/xuyu0010/ATCoN.*

⋆ Equal Contributions

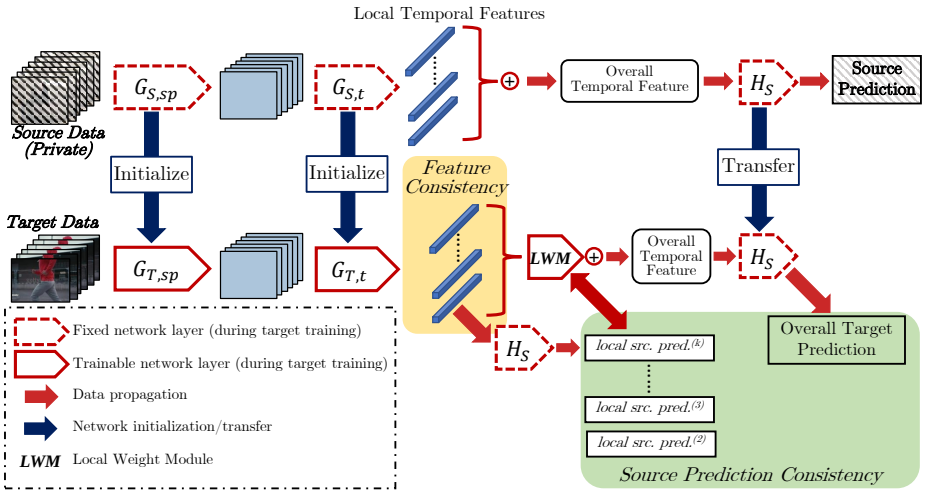


Fig. 1. Structure of the proposed ATCoN. Dashed shapes indicate fixed network layers during adaptation. *Best viewed in color.*

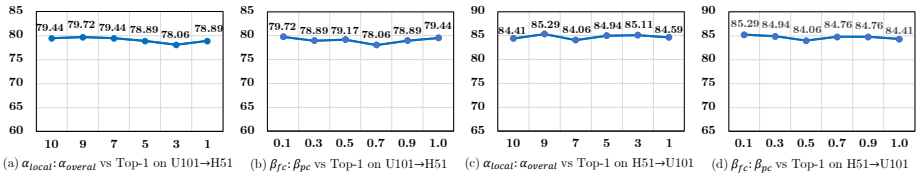
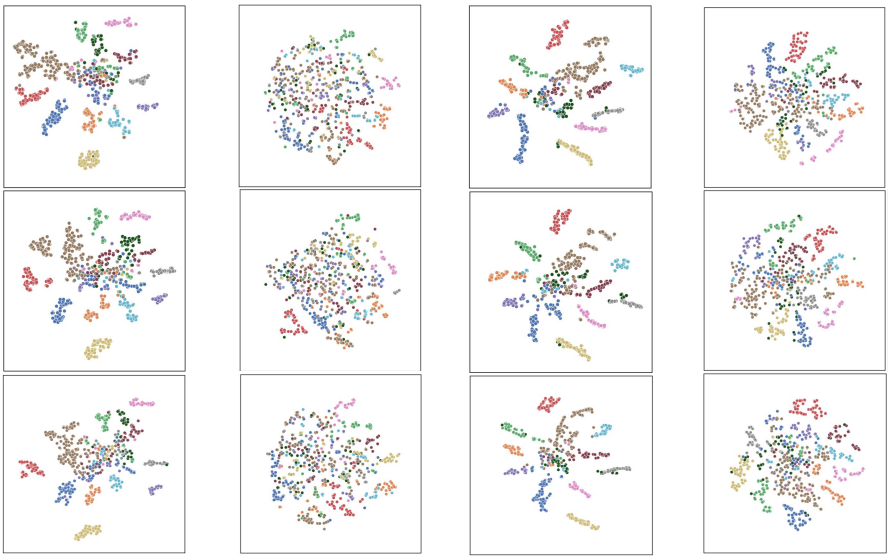


Fig. 2. Top-1 on **UCF-HMDB**$_{full}$ with differed hyperparameters.

To train the source model, the TRN is pre-trained on ImageNet [3] and trained for 100 epochs. All new layers are trained from scratch, with their learning rates set to be 10 times that of the pretrained-loaded layers. Subsequently, the spatial and temporal feature extractors of the target model are initialized by the source model, while the classifier is transferred directly from the source classifier. The source classifier remains fixed throughout the training of the target model. The final target model is obtained by a training process with 20 epochs for the UCF-HMDB$_{full}$ dataset, 30 epochs for the Daily-DA dataset, and 50 epochs for the Sports-DA dataset. The stochastic gradient descent (SGD) algorithm [1] is used for optimization, with the weight decay set to 0.0001 and the momentum set to 0.9. The batch size is set to 32 input videos per GPU. With reference to [9], we set the tradeoff constants of the information maximization and the cross-entropy loss as $\beta_{IM} = 0.5$, $\beta_{ce} = 0.2$. Hyperparameters $\alpha_{overall}$, β_{pc}, β_{tc} are set to 1.0 directly, while $\lambda = 0.004$, $\alpha_{local} = 9.0$, $\beta_{fc} = 0.1$ are set empirically.



(a) Source local temporal features with source-only model

(b) Target local temporal features with source-only model

(c) Target local temporal features with ATCoN-*TC*

(d) Target local temporal features with CPGA

Fig. 3. t-SNE embeddings of local temporal features with class information. Different colors represent different classes. *Best viewed in color.*

To understand the effects of hyperparameter selection, we perform hyperparameter sensitivity analysis over α_{local} and β_{fc} with the ratios $\alpha_{local} : \alpha_{overall}$ and $\beta_{fc} : \beta_{pc}$ being tuned on the **UCF-HMDB**$_{full}$ dataset. Fig. 2 shows that the performances of ATCoN are robust to $\alpha_{local}, \beta_{fc}$. Despite slight variations of around 1% on the two tasks of **UCF-HMDB**$_{full}$, ATCoN still achieves the best results with all the hyperparameter settings. Fig. 2 also confirms that the selected hyperparameter settings results in the best performances.

2 Visualization of Local Temporal Features

In this work, we introduce the *cross-temporal hypothesis*, which states that the local temporal features in effective source-like video representations should not only be discriminative but also consistent across each other and possess similar feature distribution patterns. In this section, we justify the *cross-temporal hypothesis* by visualizing local temporal features learned by the various models via plotting their t-SNE embeddings [10]. Specifically, we visualize local temporal features learned by the source-only model on the source data and the target data, and local temporal features learned by ATCoN-*TC* and CPGA on the target data for the H51→U101 task, as presented in Fig. 3. Each column denotes a method for a specific domain and each row indicates the corresponding three local temporal features. Note that the ATCoN-*TC* is leveraged for visualiza-

tion to rule out any possible effects brought by applying either the Information Maximization loss or pseudo-labeling.

The results further validate the *cross-temporal hypothesis*, where the local temporal features of the source data can be observed to share similar distribution patterns, whereas the data distribution patterns of target data with the source model are inconsistent. Meanwhile, even though the local temporal features of target data extracted with CPGA seem more clustered than the source model, their distributions are also inconsistent. This conforms to the poor performance of CPGA in the H51→U101 task and proves that discriminative yet inconsistent local temporal features would not result in discriminative overall temporal features. On the contrary, by learning temporal consistency, ATCoN-*TC* is able to extract discriminative and relatively consistent local temporal features, satisfying the *cross-temporal hypothesis*. Such observation justifies that learning temporal consistency enables the learning of source-like representations for target data. Therefore, ATCoN is effective in aligning target data to source data distribution.

Table 1. Summary of cross-domain action recognition benchmarks statistics.

Statistics	UCF-HMDB$_{full}$	Daily-DA	Sports-DA
Video Classes #	12	8	23
Training Video #	U101:1,438/H51:840	A11:2,776/H51:560/MIT:4,000/K600:8,959	U101:2,145/S1M:14,754/K600:19,104
Testing Video #	U101:571/H51:360	A11:1,289/H51:240/MIT:400/K600:725	U101:851/S1M:1,900/K600:1,961

3 Cross-domain Action Recognition Benchmarks

In this paper, to evaluate our proposed ATCoN, we utilized three cross-domain action recognition benchmarks. In this section, we provide more details on each benchmark.

UCF-HMDB$_{full}$ UCF-HMDB$_{full}$ [2] contains a total of 3,209 videos from two public datasets: UCF101 (U101) [13] and HMDB51 (H51) [8], with a total of 2,278 training videos and 931 testing videos as shown in Table 1. This benchmark covers videos from 12 categories, while each category may correspond to multiple categories in the UCF101 or HMDB51 dataset [2], as listed in Table 2. While UCF-HMDB$_{full}$ is not the first cross-domain action recognition benchmark, it improves on prior benchmarks in terms of both the number of video classes and the total number of videos. It is currently one of the most widely used cross-domain action recognition benchmark.

Daily-DA Though UCF-HMDB$_{full}$ improves on prior benchmarks, it is still limited to include only two domains, with videos in both domains collected under normal illumination. More recently, the Daily-DA benchmark is introduced

Table 2. List of action classes for UCF-HMDB$_{full}$.

Class ID	UCF101 Class	HMDB51 Class
0	RockClimbingIndoor, RopeClimbing	climb
1	Fencing	fencing
2	GolfSwing	golf
3	SoccerPenalty	kick ball
4	PullUps	pullup
5	Punch, BoxingPunchingBag, BoxingSpeedBag	punch
6	PushUps	pushup
7	Biking	ride bike
8	HorseRiding	ride horse
9	Basketball	shoot ball
10	Archery	shoot bow
11	WalkingWithDog	walk

[17] as a more challenging benchmark that includes videos from four domains and incorporates both normal videos and low-illumination videos. Specifically, Daily-DA is built from four datasets : the dark dataset ARID (A11) [16], as well as HMDB51 (H51), Moments-in-Time (MIT) [11], and Kinetics (K600) [7]. Compared with other action recognition datasets such as Moments-in-Time and Kinetics, ARID is comprised of videos shot under adverse illumination conditions, characterized by low brightness and low contrast. Statistically, videos in ARID possess much lower RGB mean value and standard deviation (std) [15], which strongly suggests a larger domain shift between ARID and the other action recognition datasets. In total, Daily-DA contains 16,295 training videos and 2,654 testing videos from 8 categories, with each category corresponding to one or more categories in the original datasets, as listed in Table 1 and Table 3.

Table 3. List of action classes for Daily-DA.

Class ID	ARID Class	HMDB51 Class	Moments-in-Time Class	Kinetics Class
0	Drink	drink	drinking	drinking shots
1	Jump	jump	jumping	jumping bicycle, jumping into pool, jumping jacks
2	Pick	pick	picking	picking fruit
3	Pour	pour	pouring	pouring beer
4	Push	push	pushing	pushing car, pushing cart, pushing wheelbarrow, pushing wheelchair
5	Run	run	running	running on treadmill
6	Walk	walk	walking	walking the dog, walking through snow
7	Wave	wave	waving	waving hand

Sports-DA To further verify the efficacy of SFVDA approaches on large-scale datasets, we adopt the Sports-DA benchmark. The Sports-DA benchmark includes a total of 36,003 training videos and 4,721 testing videos, collected from three large-scale datasets: UCF101 (U101) [13], Sports-1M (S1M) [6], and Kinetics (K600) [7], as shown in Table 1. These videos are categorized into 23 action classes, with each action class corresponding to one or more categories in the

Table 4. List of action classes for Sports-DA.

Class ID	UCF101 Class	Sports-1M Class	Kinetics Class
0	Archery	archery	archery
1	Baseball Pitch	baseball	catching or throwing baseball, hitting baseball
2	Basketball Shooting	basketball	playing basketball, shooting basketball
3	Biking	bicycle	riding a bike
4	Bowling	bowling	bowling
5	Breaststroke	breaststroke	swimming breast stroke
6	Diving	diving	springboard diving
7	Fencing	fencing	fencing (sport)
8	Field Hockey Penalty	field hockey	playing field hockey
9	Floor Gymnastics	floor (gymnastics)	gymnastics tumbling
10	Golf Swing	golf	golf chipping, golf driving, golf putting
11	Horse Race	horse racing	riding or walking with horse
12	Kayaking	kayaking	canoeing or kayaking
13	Rock Climbing Indoor	rock climbing	rock climbing
14	Rope Climbing	rope climbing	climbing a rope
15	Skate Boarding	skateboarding	skateboarding
16	Skiing	skiing	skiing crosscountry, skiing mono
17	Sumo Wrestling	sumo	wrestling
18	Surfing	surfing	surfing water
19	Tai Chi	t'ai chi ch'uan	tai chi
20	Tennis Swing	tennis	playing tennis
21	Trampoline Jumping	trampolining	bouncing on trampoline
22	Volleyball Spiking	volleyball	playing volleyball

original datasets as presented in Table 4. The Sports-DA benchmark is one of the largest cross-domain action recognition benchmarks introduced.

References

1. Bottou, L.: Large-scale machine learning with stochastic gradient descent. In: Proceedings of COMPSTAT'2010, pp. 177–186. Springer (2010) 2
2. Chen, M.H., Kira, Z., AlRegib, G., Yoo, J., Chen, R., Zheng, J.: Temporal attentive alignment for large-scale video domain adaptation. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 6321–6330 (2019) 1, 4
3. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: 2009 IEEE conference on computer vision and pattern recognition. pp. 248–255. Ieee (2009) 2
4. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016) 1
5. Ioffe, S., Szegedy, C.: Batch normalization: Accelerating deep network training by reducing internal covariate shift. In: International conference on machine learning. pp. 448–456. PMLR (2015) 1
6. Karpathy, A., Toderici, G., Shetty, S., Leung, T., Sukthankar, R., Fei-Fei, L.: Large-scale video classification with convolutional neural networks. In: Proceedings of the IEEE conference on Computer Vision and Pattern Recognition. pp. 1725–1732 (2014) 5
7. Kay, W., Carreira, J., Simonyan, K., Zhang, B., Hillier, C., Vijayanarasimhan, S., Viola, F., Green, T., Back, T., Natsev, P., Suleyman, M., Zisserman, A.: The kinetics human action video dataset (2017) 5

8. Kuehne, H., Jhuang, H., Garrote, E., Poggio, T., Serre, T.: Hmdb: a large video database for human motion recognition. In: 2011 International Conference on Computer Vision. pp. 2556–2563. IEEE (2011) 4
9. Liang, J., Hu, D., Wang, Y., He, R., Feng, J.: Source data-absent unsupervised domain adaptation through hypothesis transfer and labeling transfer. IEEE Transactions on Pattern Analysis and Machine Intelligence (2021) 1, 2
10. Van der Maaten, L., Hinton, G.: Visualizing data using t-sne. Journal of machine learning research **9**(11) (2008) 3
11. Monfort, M., Andonian, A., Zhou, B., Ramakrishnan, K., Bargal, S.A., Yan, T., Brown, L., Fan, Q., Gutfreund, D., Vondrick, C., et al.: Moments in time dataset: one million videos for event understanding. IEEE transactions on pattern analysis and machine intelligence **42**(2), 502–508 (2019) 5
12. Salimans, T., Kingma, D.P.: Weight normalization: A simple reparameterization to accelerate training of deep neural networks. Advances in neural information processing systems **29** (2016) 1
13. Soomro, K., Zamir, A.R., Shah, M.: Ucf101: A dataset of 101 human actions classes from videos in the wild. arXiv preprint arXiv:1212.0402 (2012) 4, 5
14. Xu, Y., Yang, J., Cao, H., Chen, Z., Li, Q., Mao, K.: Partial video domain adaptation with partial adversarial temporal attentive network. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 9332–9341 (2021) 1
15. Xu, Y., Yang, J., Cao, H., Mao, K., Yin, J., See, S.: Aligning correlation information for domain adaptation in action recognition (2021) 1, 5
16. Xu, Y., Yang, J., Cao, H., Mao, K., Yin, J., See, S.: Arid: A new dataset for recognizing action in the dark. In: International Workshop on Deep Learning for Human Activity Recognition. pp. 70–84. Springer (2021) 5
17. Xu, Y., Yang, J., Cao, H., Wu, K., Wu, M., Zhao, R., Chen, Z.: Multi-source video domain adaptation with temporal attentive moment alignment. arXiv preprint arXiv:2109.09964 (2021) 1, 5
18. Zhou, B., Andonian, A., Oliva, A., Torralba, A.: Temporal relational reasoning in videos. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 803–818 (2018) 1