Supplementary Materials for Source-Free Video Domain Adaptation with Spatial-Temporal-Historical Consistency Learning

Kai Li, Deep Patel, Erik Kruus, Martin Renqaing Min
NEC Labs, America
{kaili,dpatel,kruus,renqiang}@nec-labs.com

1. More Experimental Details

Here we introduce more details about the experiments.

1.1. Standard SFVDA

We follow the protocols of ATCoN [3] to learn the source models. Specifically, we adopt Temporal Relation Network (TRN) [4] as our action recognition model which uses ResNet-50 [1] as the frame feature extractor and a one-layer MLP as the temporal feature extractor. The temporal features are processed by a Batch Normalization (BN) layer and an additional fully connected layer before they are sent to the last fully connected layer for predictions. We initialize the frame feature extractor ResNet-50 with weights pretrained on ImageNet. We train the model with 100 epochs at an initial learning rate of 0.005 and set the learning rate of the frame feature extractor to be 1/10 of the other randomly initialized layers. After learning the source models, we adapt them with unlabeled target videos following the procedures described in the main text.

1.2. Partial Domain Adaptation (PDA)

For the PDA setting, everything is the same as the standard setting except that we remove the class-balancing term in the IM loss. We still learn to adapt the same source model using exactly the same strategy though there are less classes in the target domain. For the evaluation benchmark, as introduced in the main text, we utilize the *UCF101* and *HMDB51* datasets. We collect $2,780$ videos from the 14 common classes[1]. The number training/test samples from the two datasets are shown in Table 1.

1.3. Open-Set Domain Adaptation (OSDA)

We use the same datasets as the PDA setting for the OSDA experiments. The difference is that we choose the first (according to alphabetically sorted class names) 7 categories from *UCF101* to train the source model and adapt it using all 14 classes from the *HDMB51*. Table 1 shows the statistical numbers. For both source model training, we use the same training protocols as for the standard SFVDA setting. For the adaptation process, we adopt the strategy proposed in [2] to exclude samples from unknown classes for calculating the loss. Specifically, for every 15 iterations, we perform inference on the training data with the model H, producing the prediction $\mathbf{P} = \{\mathbf{p}_i\}_i^M$ where $\mathbf{p}_i = H(\mathbf{U}_i)$. Then, we apply K-Means clustering on the entropy of the predictions, i.e.,

$$\{C_k\}_{k=1}^2, \{L_i\}_{i=1}^M = \mathrm{K-Mean}(\mathcal{E}), \tag{1}$$

where $\mathcal{E} = \{e_i\}_{i=1}^M$ with $e_i = -\sum_j \mathbf{p}_i^j \log \mathbf{p}_i^j$, $C_k (k = 0, 1)$ is the cluster centers for known classes and unknown classes, and L_i is an indicator specifying which cluster each sample belongs to. We regard $\mathbf{C}_k$ as a known class center if $C_k < \frac{1}{M}\sum_{e_i \sim E} e_i$ because known class samples should be more confidently predicted and hence have smaller entropy. Then, samples assigning to C_k are regarded as known class samples and used for loss calculation; samples assigned to the other cluster are excluded.

	#Class (Src/Trgt)	#Training/test
PDA	14/7	$(1,323/529)$ - $(489/209)$
OSDA	7/14	$(657/265)$ - $(979/979)^\dagger$

Table 1. Statistics of the partial domain adaptation (PDA) and open-set domain adaptation (OSDA) benchmarks. †Training videos are used for test. Note this does not violate the basic evaluation rules as the label of training data is not used for training.

1.4. Black-Box Domain Adaptation (BBDA)

We adopt a simple two-step approach to extend our method to the BBDA setting. We first train a student model from scratch using the black-box model as the teacher via

[1] Although the same datasets as the standard SFVDA setting are used, the granularity of action categories is different, which results in more categories.

	#Class	#Training/test
UCF-HMDB	12	$(1,438/571)$ - $(840/360)$
UCF-Kinetics	23	$(2,145/851)$ - $(19,104/1,961)$
Jester	7	$(45,899/5,599)$ - $(45,827/5,588)$
DailyDA	8	**A**: $(2,776/1,289)$ - **H**: $(560/240)$ - **M**: $(4,000/400)$ - **K**: $(8,959/725)$

Table 2. Statistics of the benchmarks.

	C1	C2	C3	C4	C5	C6	C7	C8	C9	C10	C11	C12	Avg.
						UCF→HMDB							
TRN (source)	100	80.0	96.7	36.7	100	73.3	100	96.7	96.7	93.3	93.3	20.0	82.2
SHOT	90.0	90.0	96.7	53.3	100	56.7	100	96.7	96.7	93.3	93.3	16.7	82.2
ATCoN	83.3	86.7	100	50.0	100	76.7	100	96.7	93.3	93.3	93.3	46.7	85.6
STHC (ours)	100	73.3	97.2	82.7	100	76.3	100	97.2	100	93.3	93.3	78.8	90.9
						HMDB→UCF							
TRN (source)	92.0	97.1	92.3	100	100	77.6	100	100	100	97.1	85.4	36.1	88.1
SHOT	84.0	100	94.9	100	100	46.4	100	92.1	100	100	100	36.1	81.2
ATCoN	89.3	100	100	100	100	68.0	100	94.7	100	100	100	83.3	90.2
STHC (ours)	93.3	100	100	100	100	69.6	100	100	100	97.1	100	97.2	92.1

Table 3. Detailed results on the *UCF − HMDB* benchmark. C1∼C12 represent the 12 classes from the datasets.

knowledge distillation on unlabeled target samples. To simply this process, we use the same procedures as we train the source models, but replace the cross-entropy loss with the KL-divergence loss, as

$$L_{\text{div}} = \mathbb{E}_{\mathbf{U}\sim\mathcal{U}}\Big[\text{KLD}(p^s(\mathbf{U}), p^t(\mathbf{U}))\Big], \tag{2}$$

where $p^s(\mathbf{U})$ and $p^t(\mathbf{U})$ are the predictions by the student model and teacher model, respectively. With the intermediate source model, we perform adaptation using the same practices as adapting a standard source model.

2. Benchmark Details

Table 2 shows the statistics of the four benchmarks employed for experiments.

3. Complete Results with Per-class Accuracy

In the main text, we show the average accuracy for most experiments to save space. For reference, Tables 3-11 show the complete results with per-class accuracy provided.

References

[1] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 1

[2] Jian Liang, Dapeng Hu, and Jiashi Feng. Do we really need to access the source data? source hypothesis transfer for unsupervised domain adaptation. In *ICML*, 2020. 1

[3] Yuecong Xu, Jianfei Yang, Haozhi Cao, Keyu Wu, Min Wu, and Zhenghua Chen. Source-free video domain adaptation by learning temporal consistency for action recognition. In *ECCV*, 2022. 1

[4] Bolei Zhou, Alex Andonian, Aude Oliva, and Antonio Torralba. Temporal relational reasoning in videos. In *ECCV*, 2018. 1

	Kinetics→UCF				*UCF→Kinetics*			
	TRN (source)	SHOT	ATCoN	STHC (ours)	TRN (source)	SHOT	ATCoN	STHC (ours)
C1	87.8	95.1	95.1	95.1	80.0	80.0	80.0	80.0
C2	93.0	88.3	86.0	83.7	89.7	97.4	100	94.9
C3	100	97.1	94.3	100	100	90.0	95.0	95.0
C4	97.4	97.4	100	100	100	100	100	100
C5	100	100	97.7	100	100	100	100	100
C6	100	100	100	100	76.2	95.2	95.2	100
C7	100	100	100	100	100	100	94.1	100
C8	100	100	100	100	31.1	84.4	80.0	68.9
C9	97.5	100	100	100	95.3	14.0	100	100
C10	88.9	100	100	100	50.0	50.0	50.0	50.0
C11	92.3	87.2	92.3	92.3	76.2	38.1	60.3	82.5
C12	34.3	68.6	94.3	100	0	0	0	0
C13	91.7	91.7	94.4	91.7	100	100	50.0	100
C14	95.1	100	97.6	97.6	0	0	100	100
C15	100	100	100	100	92.9	92.9	92.9	100
C16	100	100	96.9	96.9	100	100	100	100
C17	82.5	90.0	85.0	95.0	90.3	88.7	88.7	90.3
C18	100	100	100	100	40.0	60.0	70.0	60.0
C19	100	100	100	100	100	100	100	100
C20	100	100	96.4	100	100	100	100	100
C21	83.7	69.4	71.4	71.4	83.3	83.3	83.3	83.3
C22	100	100	100	100	81.8	90.9	90.9	72.7
C23	94.3	100	100	100	100	100	100	100
Avg.	92.7	94.1	95.3	96.1	82.5	75.3	87.3	89.8

Table 4. Detailed results on *UCF-Kinetics* benchmark. C1∼C23 represent the 23 classes from the datasets.

		C1	C2	C3	C4	C5	C6	C7	C8	Avg.
K→A	TRN (source)	47.2	28.0	0	8.5	24.5	5.1	52.8	13.6	24.4
	SHOT	37.1	12.0	0	0.2	40.6	0	19.3	19.2	20.1
	ATCoN	46.5	32.0	0	0	8.3	0	16.0	13.0	14.6
	STHC (ours)	1.0	1.0	0	0	94.4	1.1	6.9	0	15.5
K→H	TRN (source)	80.0	70.0	0.0	80.0	90.0	0	46.7	33.3	50.0
	SHOT	86.7	40.0	13.3	86.7	76.7	6.7	30.0	46.7	49.1
	ATCoN	83.3	70.0	0.0	86.7	93.3	10.0	23.3	26.7	49.1
	STHC (ours)	73.3	67.0	0.0	87.3	100	10.0	23.3	33.3	48.7
K→M	TRN (source)	26.0	42.0	40.0	36.0	50.0	4.0	42.0	20.0	32.5
	SHOT	26.0	18.0	22.0	12.0	14.0	16.0	18.0	26.0	36.8
	ATCoN	34.0	46.0	48.0	46.0	34.0	10.0	22.0	42.0	35.8
	STHC (ours)	36.0	28.0	44.0	40.0	44.0	18.0	20.0	48.0	34.8
M→A	TRN (source)	73.0	40.8	1.1	45.8	17.2	0	55.7	7.3	31.2
	SHOT	51.6	0	53.8	0	0	28.7	9.0	2.8	16.1
	ATCoN	57.2	0	21.5	0	10.4	0	21.7	0.6	13.6
	STHC (ours)	69.1	0	0	0	9.3	16.9	36.3	1.0	18.4
M→H	TRN (source)	100	6.7	3.3	90.0	86.7	26.7	73.3	20.0	50.8
	SHOT	93.3	20.0	36.7	83.3	86.7	36.7	40.0	30.0	53.3
	ATCoN	96.7	13.3	13.3	90.0	90.0	53.3	66.7	33.3	58.3
	STHC (ours)	97.2	17.3	13.1	83.3	100	57.1	47.4	37.3	56.3
M→K	TRN (source)	58.3	77.3	95.5	83.3	84.3	28.4	84.4	50.0	75.9
	SHOT	0	100	100	100	23.7	0	76.6	0	42.8
	ATCoN	94.3	90.3	100	25.0	43.7	97.2	85.1	0	71.7
	STHC (ours)	97.1	90.3	100	100	62.3	68.9	85.1	0	76.6
H→A	TRN (source)	59.7	0	2.2	1.3	3.1	25.3	34.9	0	17.4
	SHOT	0	0	0	56.9	0	29.8	19.3	2.2	14.3
	ATCoN	0	0	0	9.2	0.5	29.8	26.4	2.8	10.2
	STHC (ours)	0	0	2.2	49.1	0	0	42.3	6.1	13.8
H→M	TRN (source)	46.0	46.0	44.0	20.0	26.0	42.0	8.0	26.0	32.3
	SHOT	16.0	42.0	68.0	44.0	38.0	20.0	12.0	38.0	35.0
	ATCoN	34.0	56.0	60.0	46.0	36.0	26.0	26.0	34.0	38.8
	STHC (ours)	30.0	48.0	60.0	44.0	42.0	20.0	24.0	50.0	39.8
H→K	TRN (source)	31.9	24.4	37.9	91.7	79.9	7.5	2.8	58.3	43.7
	SHOT	5.7	93.5	75.0	100	40.7	0	4.3	1	36.9
	ATCoN	62.9	93.5	97.2	100	42.2	0	6.4	1	45.8
	STHC (ours)	41.7	80.7	93.9	91.7	56.7	0	3.7	66.7	50.1
A→H	TRN (source)	0	0	43.3	3.3	56.7	30.0	10.0	0	17.9
	SHOT	70.0	13.3	30.0	90.0	50.0	0	6.7	13.3	34.2
	ATCoN	90.0	6.7	16.7	90.0	56.7	13.3	3.3	43.3	40.0
	STHC (ours)	70.0	27.1	13.3	87.2	87.4	26.7	27.2	20.0	44.6
A→M	TRN (source)	8.0	4.0	10.0	4.0	32.0	48.0	28.0	12.0	18.3
	SHOT	32.0	12.0	60.0	24.0	34.0	30.0	12.0	18.0	27.3
	ATCoN	40.0	26.0	36.0	36.0	22.0	28.0	12.0	22.0	27.3
	STHC (ours)	48.0	20.0	34.0	30.0	32.0	22.0	16.0	14.0	27.3
A→K	TRN (source)	0	4.2	3.0	0	39.9	40.3	18.3	8.3	22.3
	SHOT	57.1	6.5	97.2	100	33.3	0	66.0	0	41.8
	ATCoN	71.4	0	0	75.0	39.3	0	76.6	100	36.8
	STHC (ours)	86.1	0	0	100	54.3	0	78.4	100	44.7

Table 5. Detailed results on the *DailyDA* benchmark. "**K**", "**A**", "**H**", and "**M**" are short for the *Kinetics*, *HMDB51*, *ARID*, and *Moments-in-Time*, respectively. C1∼C8 represent the 8 classes from the datasets.

	C1	C2	C3	C4	C5	C6	C7	Avg.
TRN (source)	73.3	16.7	80.0	100	100	46.7	100	73.8
SHOT	46.7	20.2	60.0	96.7	90.3	43.3	100	65.2
ATCoN	60.0	40.3	76.7	100	80.1	46.7	100	71.9
STHC (ours)	80.0	20.2	82.9	100	89.8	50.1	100	75.2

Table 6. Detailed results for *UCF* $\rightarrow$ *HMDB* in the partial domain adaptation (PDA) setting.

	C1	C2	C3	C4	C5	C6	C7	Unknown	OS	OS*
TRN (source)	81.4	50.0	84.3	2.9	42.9	97.1	68.6	66.5	61.7	61.0
SHOT	88.6	85.7	95.7	38.6	5.7	92.9	92.9	3.9	63.0	71.4
ATCoN	98.6	87.1	98.6	0	47.1	98.6	91.4	5.3	65.8	74.5
STHC (ours)	85.7	84.3	80.0	5.7	74.3	100	87.1	54.7	69.5	73.9

Table 7. Detailed results for *UCF* $\rightarrow$ *HMDB* in the open-set domain adaptation (OSDA) setting. "Unknown" represents classes absent in the source domain. OS and OS* denote mean accuracy over all classes and mean accuracy over known classes, respectively.

	C1	C2	C3	C4	C5	C6	C7	C8	C9	C10	C11	C12	Avg.
TRN (source)	100	80.0	96.7	36.7	100	73.3	100	96.7	96.7	93.3	93.3	20.0	82.2
TRN (source)†	93.3	86.7	96.7	40.0	100	60.1	100	96.7	90.2	93.3	96.7	20.1	81.1
SHOT	80.0	93.3	100	53.3	100	80.2	100	96.7	96.7	93.3	93.3	50.2	86.4
ATCoN	63.3	90.0	100	23.3	100	66.7	80.2	96.7	93.3	90.0	96.7	13.3	76.7
STHC (ours)	90.1	83.3	97.4	70.2	100	65.9	100	97.3	100	93.2	93.3	65.7	87.8

Table 8. Detailed results for *UCF* $\rightarrow$ *HMDB* in the black-box domain adaptation (BBDA) setting.

	C1	C2	C3	C4	C5	C6	C7	C8	C9	C10	C11	C12	Avg.
w/o spatial consistency	100	83.1	93.3	90.1	100	72.4	100	97.1	100	83.3	93.3	41.1	87.8
w/o temporal consistency	93.1	83.1	97.2	67.2	100	72.4	100	97.1	93.3	97.1	93.3	76.4	88.9
w/o historical consistency	100	90.0	93.3	80.0	100	66.3	100	97.1	100	86.3	93.3	71.9	89.8
w/o training the classifier	100	80.0	93.3	76.9	100	68.9	100	97.1	100	92.6	96.6	51.8	88.1
Full Model	100	89.8	97.2	82.5	100	76.1	100	97.1	97.2	92.6	93.3	65.5	90.9

Table 9. Detailed results for the ablation study with *UCF* $\rightarrow$ *HMDB*.

	C1	C2	C3	C4	C5	C6	C7	Avg.
w/o spatial consistency	55.2	100	82.1	51.1	98.1	95.8	80.0	75.0
w/o temporal consistency	56.1	99.2	84.3	24.2	96.0	98.1	78.1	70.1
w/o historical consistency	44.2	99.6	76.5	63.6	99.1	96.6	92.0	76.6
w/o training the classifier	30.0	99.2	85.3	38.9	98.3	80.1	95.2	70.4
Full Model	66.1	99.4	76.5	60.3	98.3	96.6	83.5	78.4

Table 10. Detailed results for the ablation study with *Jester*.

	C1	C2	C3	C4	C5	C6	C7	C8	C9	C10	C11	C12	Avg.
$\alpha = 0.001$	0	0	0	0	0	100	0	0	0	0	0	0	8.2
$\alpha = 0.01$	100	53.2	100	60.0	100	86.3	100	97.1	97.2	93.1	93.3	17.3	83.0
$\alpha = 0.1$	100	73.3	97.2	82.7	100	76.3	100	97.2	100	93.3	93.3	78.8	90.9
$\alpha = 1$	97.1	80.1	97.2	77.3	100	72.4	100	97.1	100	90.1	93.3	69.4	89.2
$\alpha = 10$	100	77.2	92.8	70.2	100	72.4	100	97.1	100	93.2	93.3	69.4	88.6

Table 11. Detailed results for the sensitivity analysis of α with *UCF* $\rightarrow$ *HMDB*.