CycDA: Unsupervised Cycle Domain Adaptation to Learn from Image to Video Supplementary Material

Wei Lin[1], Anna Kukleva[2], Kunyang Sun[1,3], Horst Possegger[1], Hilde Kuehne[4], and Horst Bischof[1]

[1] Institute of Computer Graphics and Vision, Graz University of Technology, Austria
{wei.lin,possegger,bischof}@icg.tugraz.at
[2] Max-Planck-Institute for Informatics, Germany akukleva@mpi-inf.mpg.de
[3] Southeast University, China sunky@seu.edu.cn
[4] Goethe University Frankfurt, Germany kuehne@uni-frankfurt.de

1 Overview

For additional insights into our unsupervised cycle domain adaptation approach (CycDA), we first analyze the frame thresholding in Sec. 2 and provide additional baselines of image-to-video, frame-to-video and frame&video-to-video adaptation in Sec. 3. We illustrate the performance of multiple iterations of CycDA on two adaptation datasets in Sec. 4. Then, we demonstrate that the domain alignment improves the learned feature representations by searching nearest neighbors to target query frames in the source domain in Sec. 5. In Sec. 6, we inspect the confusion matrices of our generated pseudo labels. Finally, we analyze potential failure cases of CycDA in Sec. 7.

2 Frame Thresholding

In the spatio-temporal learning stage (Sec. 3.2 in the main paper), we train a video model with pseudo labeled data in the target domain. To remove pseudo labels with low confidence, we set the confidence threshold δ_p such that $p{\times}100\%$ of videos with highest confidence remain after the thresholding.

We determine the maximum of the frame-level confidence scores within each video, and sort the maximum frame-level confidence scores of all videos in a descending order. The ordered score sequence is denoted as $[s_1, s_2, ..., s_N]$. For $\delta_p \in [s_{\lfloor p \cdot N \rfloor}, s_{\lfloor p \cdot N \rfloor + 1}]$, there are $\lfloor p \cdot N \rfloor$ videos left after thresholding. We set $\delta_p = (s_{\lfloor p \cdot N \rfloor} + s_{\lfloor p \cdot N \rfloor + 1})/2$.

We illustrate the accuracy of video pseudo labels and the spatio-temporal learning (Stage 2) performance in Fig. 1. Intuitively, with p increasing, the accuracy of video pseudo labels drops, from 85.7% ($p = 10\%$) to 42.7% ($p = 100\%$). For p smaller than 60%, the spatio-temporal learning performance suffers due to the insufficient amount of training data, in spite of the higher accuracy of pseudo labels. For p between 70% and 100%, the performance remains fairly

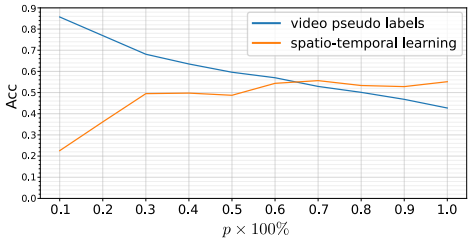


Figure 1. Accuracy of video pseudo labels and the spatio-temporal learning (Stage 2) performance w.r.t. different thresholding percentages $p\times100\%$ on EADs $\rightarrow$ HMDB51.

stable at an accuracy of approximately 55%. We set p to 70% for Stage 2 in all the experiments in the main paper. As Stage 3 further improves the pseudo labels transferred to Stage 4, we increase p by 10% to 80% for the spatio-temporal learning in Stage 4 heuristically.

3 Additional baselines

3.1 Image-to-video adaptation from BU101 to UCF-HMDB

In Sec. 4.4 of the main paper, we sample 50 web images per class from the 12 classes in BU101 and adapt to the UCF-HMDB videos. The BU101 dataset has around 230 images per class on average. Here, we evaluate the image-to-video adaptation with a varying amount of web images. For this, we sample different numbers of web images per class from BU101 and also compare our CycDA with video-to-video adaptation approaches (on UCF-HMDB) in Table 1. Apparently, increasing the number of web images on BU101 leads to a significant performance boost.The image-to-video adaptation with 230 images per class (82.2% for BU$\rightarrow$H and 97.9% for BU$\rightarrow$U) outperforms most state-of-the-art video-to-video adaptation approaches. The performance of BU$\rightarrow$U even exceeds the supervised target model. This demonstrates that CycDA can exploit the large informativity in the web images for enhanced performance on the target video classifier.

3.2 Frame-to-Video and Frame&Video-to-Video Adaptation

We add the baselines of *frame-to-video* in the case of 1 frame and 5 frames from each source video in Table 1. Furthermore, we add a new setting of *frame&video-to-video* adaptation: using 5 frames from each source video as the source image data (instead of web images), together with the videos as the source video data, to adapt to target videos with domain shift. Both, using more frames and using frame together with videos, lead to further improvements.

	DA setting	method	source data		U→H	H→U
			BU web image (# images per class)	videos (U or H) in %		
A:	video-to-video	AdaBN [5]	-	100%	75.5	77.4
		MCD [10]	-	100%	74.4	79.3
		TA^{3}N [1]	-	100%	78.3	81.8
		ABG [6]	-	100%	79.1	85.1
		TCoN [8]	-	100%	**87.2**	89.1
		DANN [3]	-	100%	80.7	88.0
		TA^{3}N [1]	-	100%	81.4	90.5
		SAVA [2]	-	100%	82.2	91.2
		MM-SADA [7]	-	100%	84.2	91.1
		CrossModal [4]	-	100%	84.7	92.8
		CoMix [9]	-	100%	86.7	**93.9**
B:	image-to-video	CycDA	50	-	77.8	88.6
			100	-	81.7	93.9
			230	-	**82.2**	**97.9**
C:	frame-to-video	CycDA	-	1 frame	83.3	80.4
			-	5 frames	84.4	84.6
D:	frame&video-to-video	CycDA	-	5 frames+videos	85.6	87.9
	supervised target		-	-	94.4	97.0

Table 1. Results of image-to-video adaptation (from images in BU101 to videos on UCF-HMDB) in comparison to video-to-video adaptation approaches.

4 Multiple Iterations

We observed that the fluctuation effect of performance among multiple iterations is more pronounced for small datasets with large domain shift (such as E→H), which are more prone to overfitting on either the image or the video side. To verify this, we compare the behaviour on E→H with B→U for 6 iterations in Fig. 2. While in both cases the peak is reached after three iterations, the performance on the large-scale B→U remains more stable.

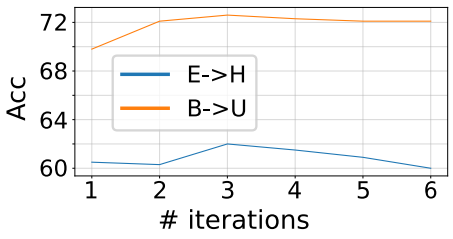


Figure 2. Multiple iterations of CycDA on E→H and B→U.

5 Nearest Neighbor Search in Image Feature Space

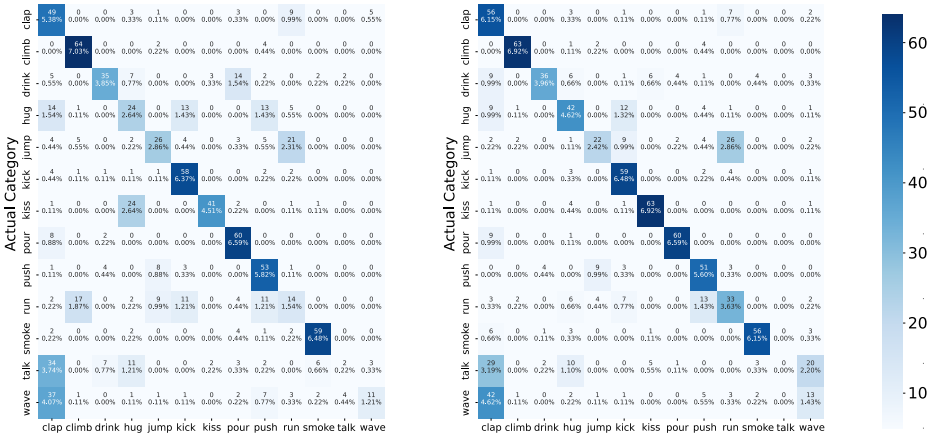
In Sec. 4.5 in the main paper, we perform a stage-wise ablation study to show how each stage contributes to the performance boost. In order to further demonstrate how the domain alignment is improved, we use the sampled target frames as query and search for their nearest neighbor (NN) source web images. For this, we use the following three image feature representations: (i) source-only, (ii) class-agnostic domain alignment (CycDA stage 1) and (iii) class-aware domain alignment (CycDA stage 3). The t-SNE visualizations of these 3 image feature spaces are shown in the main paper in Figure 3(a)–(c). For further investigation, we sample query frames from target videos and show their 5 nearest neighbor source web images in Fig. 4 and 5. In each subfigure, the 3 rows show the nearest neighbor results in the image feature space of the source-only model (1st row), after CycDA stage 1 class-agnostic alignment (2nd row), and after CycDA stage 3 class-aware alignment (3rd row).

We see that the nearest neighbors of the target query frame in the source-only feature space are from different categories, due to the large domain shift between source and target distributions before alignment. After class-agnostic domain alignment, the amount of source nearest neighbors from the same category slightly increases. After class-aware domain alignment in stage 3, most of the nearest neighbors are from the same category. In Fig. 5(b), where the target query frame shows a child *clapping*, both source-only and class-agnostic alignment result in several source nearest neighbors which show children, but none of these belong to the *clap* category. On the contrary, class-aware alignment leads to nearest neighbors with non-baby content in the correct category. This indicates our effective category-level alignment which semantically gathers target frames with source data of the same category, instead of simply aligning images in terms of styles and appearance. Similar examples can be found in Fig. 4(c) and (d).

6 Confusion Matrix

Complementing our pseudo label analysis (Fig. 4(a) in the main paper), we further illustrate the confusion matrices of the video pseudo labels on the target videos after CycDA stage 2 (Fig. 3(a)) and CycDA stage 4 (Fig. 3(b)). Clearly, Fig. 3(b) improves along the diagonal on the difficult classes (*e.g. run, hug, kiss*). Furthermore, Fig. 3(b) shows less confusion in comparison to Fig. 3(a), *e.g.* between *kiss* and *hug*, *drink* and *pour*, or *run* and *climb*. Our complete CycDA (with stages 3 and 4) contributes to a significant performance boost with less confusion among categories.

The remaining confusion in Fig. 3(b) is due to low inter-class variation (*c.f.* Sec. 7), *e.g.* between *wave* and *clap*, or *jump* and *run*. The most difficult category is *talk* (with zero accuracy), which is defined on EADs as the interaction between two or more subjects. This interaction is captured by our model and provides a strong bias for classification. On HMDB51, however, most *talk* videos contain only a single subject speaking and thus, capture no interaction.



(a) CycDA stage 1 + stage 2 (b) CycDA stage 1 to 4

Figure 3. Confusion matrices of video pseudo labels on target videos after (a) CycDA stage 2 and (b) CycDA stage 4 for 12 classes on EADs → HMDB51. Best viewed on screen.

7 Failure Cases

To analyze the limitations of CycDA, Fig. 6 illustrates failure cases of the nearest neighbor search in the image feature space after the step 3 class-aware domain alignment. The majority of our failures can be attributed to unusual backgrounds (where the background has a high visual similarity to the typical scenario of another action category) and ambiguous actions (low inter-class variation). For example, Fig. 6(a) shows a baby *waving* in a crib, where the crib has a similar pattern to carts that are typically found in the *push* category. Similarly, *running* in front of rocks (Fig. 6(b)) or *running* towards a vehicle (Fig. 6(c)) is grouped to *climb* or *push* respectively. *Jumping* on the slope of a bouncy castle (Fig. 6(d)) looks visually similar to the scene of *climb*.

Fig. 6(e)–(h) demonstrate category confusion due to low inter-class variation. For example, *waving* while *talking* to another subject occurs frequently in the web images annotated as *talk* (Fig. 6(e)). *Waving* with both hands is visually similar to *clap* (Fig. 6(f)). *Kissing* while hugging can be confused with *hugging* only (Fig. 6(g)). *Running* can look like *jumping* when the subject is in the air (Fig. 6(h)). These ambiguous actions can also be seen from the confusion matrices in Fig. 3. Although spatio-temporal learning on a video model improves the inference of motion-based actions on the target domain, these visually similar or ambiguous actions still pose a challenge. Deriving motion information from source web images might be a solution to this issue.

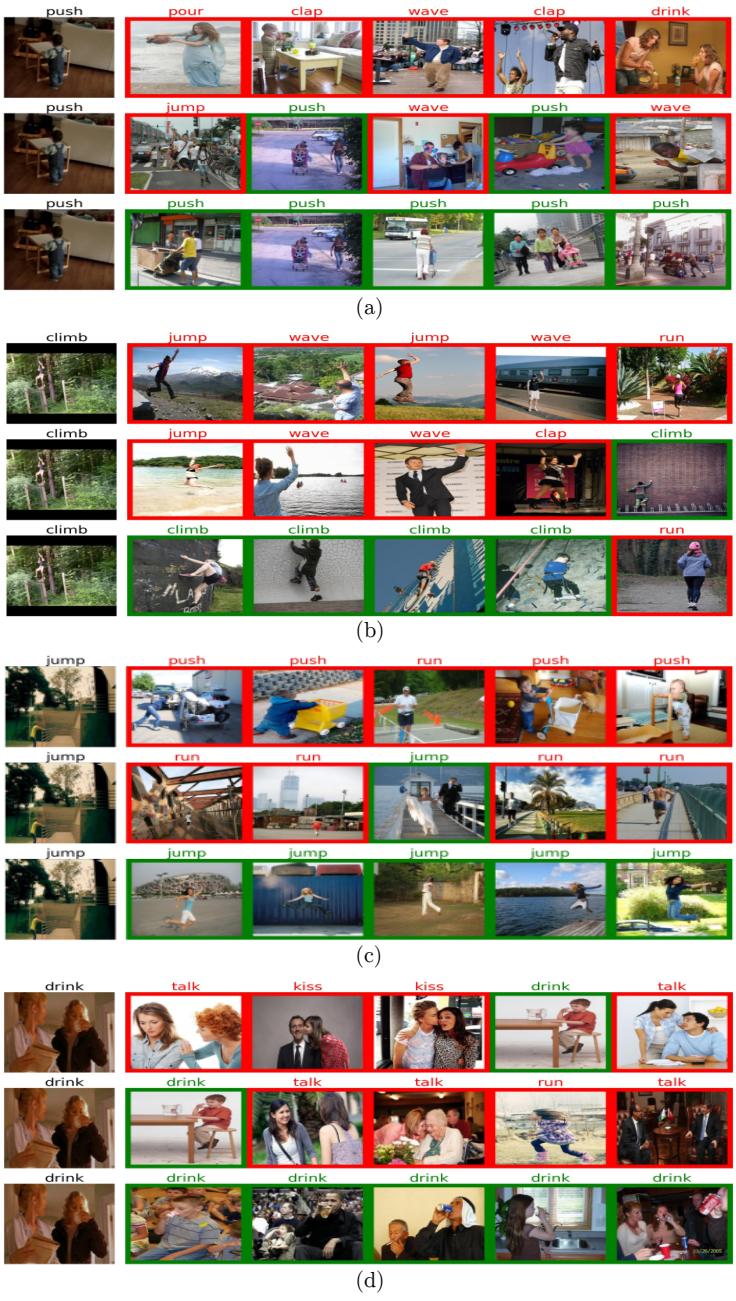


Figure 4. Target video frame and its 5 nearest neighbor (NN) web images in the source domain. Each 3-row group subfigure displays the NN search in image feature space of source only (1st row), CyDA stage 1 class-agnostic alignment (2nd row), CycDA stage 3 class-aware alignment (3rd row). The image border color indicates NN in same (green) or different (red) category.

Figure 5. Target video frame and its 5 nearest neighbor (NN) web images in the source domain. Each 3-row group subfigure displays the NN search in image feature space of source only (1st row), CyDA stage 1 class-agnostic alignment (2nd row), CycDA stage 3 class-aware alignment (3rd row). The image border color indicates NN in same (green) or different (red) category.

Figure 6. Failure cases in search of 5 nearest neighbors (NN) in the image feature space of CycDA stage 3 class-aware domain alignment. The image border color indicates NN in same (green) or different (red) category.

References

1. Chen, M.H., Kira, Z., AlRegib, G., Yoo, J., Chen, R., Zheng, J.: Temporal attentive alignment for large-scale video domain adaptation. In: ICCV. pp. 6321–6330 (2019)
2. Choi, J., Sharma, G., Schulter, S., Huang, J.B.: Shuffle and attend: Video domain adaptation. In: ECCV. pp. 678–695. Springer (2020)
3. Ganin, Y., Ustinova, E., Ajakan, H., Germain, P., Larochelle, H., Laviolette, F., Marchand, M., Lempitsky, V.: Domain-adversarial training of neural networks. JMLR **17**(1), 2096–2030 (2016)
4. Kim, D., Tsai, Y.H., Zhuang, B., Yu, X., Sclaroff, S., Saenko, K., Chandraker, M.: Learning cross-modal contrastive features for video domain adaptation. In: ICCV. pp. 13618–13627 (2021)
5. Li, Y., Wang, N., Shi, J., Hou, X., Liu, J.: Adaptive batch normalization for practical domain adaptation. Pattern Recognition **80**, 109–117 (2018)
6. Luo, Y., Huang, Z., Wang, Z., Zhang, Z., Baktashmotlagh, M.: Adversarial bipartite graph learning for video domain adaptation. In: ACM Multimedia. pp. 19–27 (2020)
7. Munro, J., Damen, D.: Multi-modal domain adaptation for fine-grained action recognition. In: CVPR. pp. 122–132 (2020)
8. Pan, B., Cao, Z., Adeli, E., Niebles, J.C.: Adversarial cross-domain action recognition with co-attention. In: AAAI. vol. 34, pp. 11815–11822 (2020)
9. Sahoo, A., Shah, R., Panda, R., Saenko, K., Das, A.: Contrast and mix: Temporal contrastive video domain adaptation with background mixing. In: NeurIPS (2021)
10. Saito, K., Watanabe, K., Ushiku, Y., Harada, T.: Maximum classifier discrepancy for unsupervised domain adaptation. In: CVPR. pp. 3723–3732 (2018)