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

Wei Lin[1,5], 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
[5] Christian Doppler Laboratory for Semantic 3D Computer Vision

Abstract Although action recognition has achieved impressive results over recent years, both collection and annotation of video training data are still time-consuming and cost intensive. Therefore, image-to-video adaptation has been proposed to exploit labeling-free web image source for adapting on unlabeled target videos. This poses two major challenges: (1) spatial domain shift between web images and video frames; (2) modality gap between image and video data. To address these challenges, we propose Cycle Domain Adaptation (CycDA), a cycle-based approach for unsupervised image-to-video domain adaptation. We leverage the joint spatial information in images and videos on the one hand and, on the other hand, train an independent spatio-temporal model to bridge the modality gap. We alternate between the spatial and spatio-temporal learning with knowledge transfer between the two in each cycle. We evaluate our approach on benchmark datasets for image-to-video as well as for mixed-source domain adaptation achieving state-of-the-art results and demonstrating the benefits of our cyclic adaptation.

Keywords: Image-to-video adaptation, unsupervised domain adaptation, action recognition

1 Introduction

The task of action recognition has seen tremendous success in recent years with top-performing approaches typically requiring large-scale labeled video datasets [10,39,40], which can be impractical in terms of both data collection and annotation effort. In the meanwhile, webly-supervised learning has been explored to leverage the large amount of easily accessible web data as a labeling-free data source for video recognition [12,16,24,38,41,47].

In this work, we address the problem of image-to-video adaptation with webly-labeled images as the source domain and unlabeled videos as the target domain to allow for action classification without video annotation. This setting

provides two major challenges: (1) the spatial domain shift between web images and video frames, based on difference in image styles, camera perspectives and semantic drifts; (2) the modality gap between spatial images and spatio-temporal videos. Specifically, this modality gap restrains that merely spatial knowledge can be transferred from source to target domain. Previous works on action recognition with web supervision either learn from web data directly [11,13] or perform joint training by combining the web source with annotated target data [9,29]. To specifically address the domain shift between web images and target videos, some approaches perform class-agnostic domain-invariant feature learning either within [19] or across modalities [44,43,27], in the absence of ensuring domain-invariance on the category-level.

In this context, we propose Cycle Domain Adaptation (CycDA), *i.e.* alternating knowledge transfer between a spatial model and a spatio-temporal model. Compared to other works, we address the two challenges at hand, domain-alignment and closing the modality gap in separate stages, cycling between both of them. An overview of the CycDA is given in Fig. 1. With the category knowledge from the spatio-temporal model, we achieve enhanced category-level domain invariance on the spatial model. With updated knowledge transferred from the spatial model, we attain better spatio-temporal learning. In this manner, we can better tackle each challenge for the corresponding model, with the updated knowledge transferred from the other.

More specifically, we propose a four stage framework to address the domain shift between images and videos on different levels. In stage 1, we enforce *class-agnostic* domain alignment on the spatial model between images and video frames. In stage 2, we use supervision from the spatial model to learn a spatio-temporal video model, bridging the gap between the two modalities. Stage 3 then focuses on *class-aware* domain alignment on the spatial model, given pseudo labels computed by the video model trained on stage 2. In stage 4, we update the video model with the improved pseudo labels from the spatial model of stage 3.

We first evaluate our approach on several challenging settings for web image based action recognition, where a single cycle already outperforms baselines and state-of-the-arts. Second, we show how CycDA can be flexibly applied for mixed-source image&video-to-video DA settings, leading to a performance competitive to the state-of-the-art requiring only 5% of the provided source videos.

We summarize our contributions as follows: (1) We propose to address web image-to-video domain adaptation by decoupling the domain-alignment and spatio-temporal learning to bridge the modality gap. (2) We propose cyclic alternation between spatial and spatio-temporal learning to improve spatial and spatio-temporal models respectively. (3) We provide an extensive evaluation with different benchmark tasks that shows state-of-the-art results on unsupervised image-to-video domain adaptation and a competitive performance for the mixed-source image&video-to-video setting.
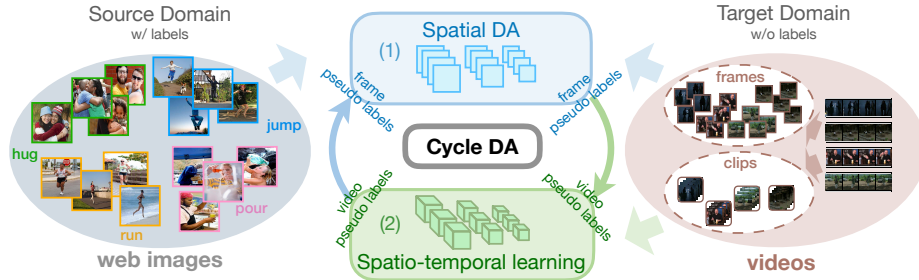
Figure 1. Cycle Domain Adaptation (CycDA) pipeline: we address image-to-video adaption by training a spatial model and a spatio-temporal model alternately, passing pseudo labels to supervise each other in a cycle. The two alternating steps are: (1) domain alignment on the spatial model with pseudo labels from the spatio-temporal model, and (2) training the spatio-temporal model with updated pseudo labels from the spatial model.

2   Related Work

Webly-supervised action recognition. Various works have shown how web images and videos can be used as labeling-free data source to improve the performance of action recognition [9,11,13,29]. Gan *et al.* [11,13] train with web data only, without adaptation to the target domain. Ma *et al.* [29] combine web images and video frames to train a spatial model and achieve comparable performance by replacing parts of annotated videos with web images. Duan *et al.* [9] transform different types of web modalities to trimmed video clips and combine them with labeled target videos for joint training. In our work, we specifically address the domain shift on the spatial level and transfer the knowledge to the spatio-temporal level, without using any annotations on target data.

Image-to-video DA. Compared to webly-supervised learning, image-to-video DA approaches actively address the domain shift between web images and target videos either by spatial alignment between web images and video frames [24,36,45], or through class-agnostic domain-invariant feature learning for images and videos [27,43,44]. Li *et al.* [24] use a spatial attention map for cross-domain knowledge transfer from web images to videos. In this case, the DA is addressed on the spatial level without transfer to the temporal level. Liu *et al.* [27] perform domain-invariant representation learning for images, video keyframes and videos, and fuse features of different modalities. Furthermore, hierarchical GAN [44], symmetric GAN [43] and spatio-temporal causal graph [2] are proposed to learn the mapping between image features and video features. Closest to our work is probably the work of Kae *et al.* [19] which also employs a spatial and a spatio-temporal model for two stages of class-agnostic domain alignment, proposing to copy the weights from the spatial to the spatio-temporal model.

In contrast to these, we propose to transfer knowledge in the form of pseudo labels, without enforcing spatial information from the DA stage onto the spatio-temporal model. Moreover, we alternately conduct spatial alignment and spatio-
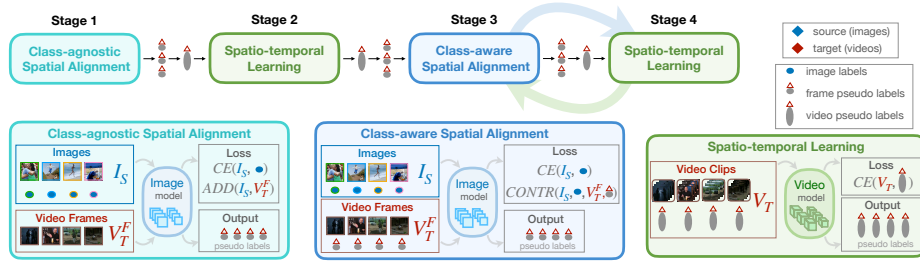
Figure 2. Our CycDA framework alternates between spatial alignment (stage 1 and 3) and spatio-temporal learning (stage 2 and 4). See text for details.

temporal learning with knowledge transfer to each other. With category knowledge transferred to the spatial model, we perform class-aware domain alignment that induces domain-invariance within category-level.

Video-to-video DA. Compared to image-to-video adaption, video-to-video adaptation methods adapt annotated source videos to unlabeled target videos [3,33,22,7,6,31,28,32,18], focusing mainly on the problem of domain alignment. Chen *et al.* [3] align the features spatially on the frame-level and temporally on the scale-level. Others propose feature alignment via self-attention [7], cross-domain co-attention [32], or across two-stream modalities of RGB and optical flow [22,31]. Sahoo *et al.* [33] propose temporal contrastive learning and background mixing for domain-invariance. In this work, we focus on image-to-video adaptation and use only single stream of RGB. We further show that we are able to extend the pipeline to the mixed-source case, where we achieve competitive performance compared to video-to-video adaptation methods while requiring only a small amount of source videos.

3   Cycle Domain Adaptation (CycDA)

We propose four stages to tackle image-to-video adaptation, which we summarize in Sec. 3.1. Afterwards, we detail each stage and motivate the cycling of stages in Sec. 3.2. Our CycDA can be flexibly extended (Sec. 3.3) for mixed-source video adaptation, where a limited amount of annotated source videos are available.

3.1   System Overview

The task of unsupervised image-to-video DA is to learn a video classifier given labeled source images and unlabeled target videos. In order to close the domain gap across these two different modalities, we employ (1) a spatial (image) model to train on source web images and frames sampled from target videos, and (2) a spatio-temporal (video) model to train on target video clips. We propose a training pipeline that alternately adapts the two models by passing pseudo labels to supervise each other in a cycle. This facilitates the knowledge transfer between both models, where pseudo labels efficiently guide the model through the

corresponding task, *i.e.* semantic alignment (image model) or spatio-temporal learning (video model).

As shown in Fig. 2, our CycDA pipeline consists of four training stages. The initial pseudo labels from *class-agnostic* domain alignment (Stage 1) are improved via video spatio-temporal learning (Stage 2). In Stage 3, these pseudo labels on target data, together with ground truth labels on source, enable semantic alignment through the class-aware domain alignment. This again refines the pseudo labels, which further provide enhanced supervision for spatio-temporal learning in stage 4. In this manner, one iteration of CycDA facilitates the alternating knowledge transfer between the image and video models, whereby pseudo labels are improved on each stage and further provide strengthened supervision for better learning on the next stage.

Notations: First, we denote the feature extractor as $E(\cdot; \theta_E)$, the classifier as $C(\cdot; \theta_C)$, and the domain discriminator as $D(\cdot; \theta_D)$. Then, we have the image model $\phi^I = \{E^I(\cdot; \theta_E^I), \ C^I(\cdot; \theta_C^I), \ D^I(\cdot; \theta_D^I)\}$ and the video model $\phi^V = \{E^V(\cdot; \theta_E^V), \ C^V(\cdot; \theta_C^V)\}$. We use the superscripts I, V and F to denote modalities of image, video and video frame, correspondingly. S and T stand for *source* and *target* domains respectively. The labeled source image domain is denoted as $I_S = \{(i_j, l(i_j))|_{j=1}^{N_S^I}\}$, where $l(\cdot)$ is the ground truth label of the corresponding image. The unlabeled target video domain is $V_T = \{v_j|_{j=1}^{N_T^V}\}$ and each video v_j has M_j frames, the set of frames of unlabeled target videos $V_T^F = \{\{v_{j,m}^F|_{m=1}^{M_j}\}|_{j=1}^{N_T^V}\}$.

3.2   Stages

Stage 1 - Class-agnostic Spatial Alignment. In the first stage, we learn the class-agnostic domain alignment between source web images and frames sampled from unlabeled target videos. Thus, we reduce the domain gap between the appearance of the web images and target videos even if the classes could be incorrectly aligned during this stage. We train the image model ϕ^I with a supervised **c**ross **e**ntropy loss $\mathcal{L}_{CE}(I_S)$ and an **a**dversarial **d**omain **d**iscrimination loss $\mathcal{L}_{ADD}(I_S, V_T^F)$ on source images and target frames. With the classification loss on source images given as

$$\mathcal{L}_{CE}(I_S) = \sum_{(i_j, l(i_j)) \in I_S} -l(i_j) \cdot \log(C^I(E(i_j; \theta_E^I); \theta_C^I)) \tag{1}$$

and the binary cross entropy loss for domain discrimination given as

$$\mathcal{L}_{ADD}(I_S, V_T^F) = \sum_{i_j, v_{j',m}^F} \log D^I(E^I(i_j; \theta_E^I); \theta_D^I) + \log(1 - D^I(E^I(v_{j',m}^F; \theta_E^I); \theta_D^I)), \tag{2}$$

the overall objective is $\min_{\theta_E^I, \theta_C^I} \mathcal{L}_{CE}(I_S) + \beta \max_{\theta_E^I} \min_{\theta_D^I} \mathcal{L}_{ADD}(I_S, V_T^F)$, where β is the trade-off weight between the two losses. We train the domain discriminator D^I to distinguish between extracted features from different domains, while the feature extractor is trained adversely based on the domain discrimination

task. In this case, the feature extractor learns to yield domain invariant features that the domain discriminator is unable to differentiate. The adversarial training is performed with a gradient reversal layer (GRL) [14,15] which reverses the sign of gradients of the domain discrimination loss on the feature extractor during back-propagation. The domain alignment on this stage is class-agnostic as there is not yet any pseudo label for category knowledge in the target domain. The alignment is performed globally at the domain level.

Stage 2 & Stage 4 - Spatio-Temporal Learning. In this stage, we use the trained image model ϕ^I from the previous stage to generate pseudo labels for the target videos. Then we perform supervised spatio-temporal learning with the pseudo labeled target data.

Specifically, we first use ϕ^I to predict the pseudo label $\hat{l}(\cdot)$ for each frame of the target videos. We employ a spatio-temporal model that trains on target videos capturing both spatial and temporal information in the target domain only. To select new pseudo label candidates, we temporally aggregate frame-level predictions into a video-level prediction. We discard predictions with confidence lower than a threshold δ_p and perform a majority voting among the remaining predictions to define the video label. From all videos, we only keep those that have at least one frame with a minimum confidence. We set the confidence threshold δ_p such that $p \times 100\%$ of videos remain after the thresholding.

We denote the target video set after thresholding as $\tilde{V}_T$. For stage 2, the supervised task on pseudo labeled target videos is to $\min_{\theta_E^V, \theta_C^V} \hat{\mathcal{L}}_{CE}(\tilde{V}_T)$, with

$$\hat{\mathcal{L}}_{CE}(\tilde{V}_T) = \sum_{v_j \in \tilde{V}_T} -\hat{l}(v_j) \cdot \log(C^V(E^V(V_j; \theta_E^V); \theta^V)). \tag{3}$$

In stage 4, we repeat the process as described above and re-train the video model on the target data with the updated pseudo labels from the third stage.

Stage 3 - Class-aware Spatial Alignment. The adversarial learning for domain discrimination in the first stage aligns features from different domains globally, but not within each category (*c.f.* Fig. 3 (b) and (c)). In this case, target samples in a category A can be incorrectly aligned with source samples in a different category B. This would lead to inferior classification performance of the target classifier. To evade this misalignment, we perform class-aware domain alignment in the third stage between the source web images and the target video frames. Since the source data consists exclusively of images, we apply alignment on the spatial model between images and frames. Furthermore, as the target data is unlabeled, in order to align features across both domains within each category, we generate pseudo labels by the model ϕ^V from the second stage to provide category knowledge. Specifically, we use the video model to generate video-level labels that we disseminate into frame-level labels. To align images and video frames we use cross-domain contrastive learning by maximizing the similarity between samples across domains of the same class and minimizing the similarity between samples from different classes. We use $z = E^I(i; \theta_E^I)$ to denote the feature computed by the feature extractor on image i. The set of source

image features is $Z_S^I = \{E^I(i; \theta_E^I) | i \in I_S\}$ and the set of target frame features is $Z_T^F = \{E^I(v^F; \theta_E^I) | v^F \in V_T^F\}$. During training, for each pseudo labeled target sample $z_j^F \in Z_T^F$, we randomly choose two samples from the source domain: a positive sample of the same label and a negative sample of a different label, $i.e.$ $z_{j+}^I, z_{j-}^I \in I_S$. The contrastive loss is formulated as

$$\mathcal{L}_{CONTR}(I_S, V_T^F) = -\sum_{z_j^F \in Z_T^F} \log \frac{h(z_j^F, z_{j+}^I)}{h(z_j^F, z_{j+}^I) + h(z_j^F, z_{j-}^I)}. \tag{4}$$

Following [5], we set $h(u, v) = \exp(\text{sim}(u, v)/\tau)$, where we use the cosine similarity $\text{sim}(u, v) = u^\mathsf{T} v / (\|u\| \|v\|)$ and τ is the temperature parameter. Thus, the objective of stage 3 on the image model is $\min_{\theta_E^I, \theta_C^I} \mathcal{L}_{CE}(I_S) + \mathcal{L}_{CONTR}(I_S, V_T^F)$.

In the third stage, an alternative of exploiting pseudo labels from the video model from the second stage is to self-train the video model on the target data, as self-training is a common practice in DA [26,46,48,49]. However, the category-level domain alignment with supervision from the source domain further regularizes the learning of class distribution in the target domain. This results in a significantly improved target classifier, as we show in Sec. 4.5.

Cycling of the Stages. The pseudo labels from the video model are exploited for class-aware domain alignment on the image model (stage 3) and the updated pseudo labels from the image model can supervise the training of the video model (stage 4). In this manner, stage 3 and stage 4 can be performed iteratively. We show in the evaluation (Table 1 and Fig. 5) the impact of this cyclic learning setup and how several iterations of CycDA can further improve the performance of the target classifier.

3.3   Mixed-source Video Adaptation

Image-to-video DA applies to the case in which the source domain consists only of web images. However, other possible settings presume limited amount of annotated videos with the domain shift to the unlabeled target videos. We refer to this case as mixed-source video adaptation. CycDA can be adjusted for this setting as follows. We denote the labeled source video domain as $V_S = \{(v_j, l(v_j))|_{j=1}^{N_S^V}\}$. For the class-agnostic (stage 1) and class-aware domain alignment (stage 3) stages we replace the source image domain $\{I_S\}$ by the mixed-source domain data $\{I_S, F_S\}$ which consists of web images and frames sampled from source videos. The supervised classification, adversarial domain discrimination and cross-domain contrastive learning are adapted accordingly. For the spatio-temporal learning of the video model ϕ^V (stage 2 and 4) we include additional supervised classification w.r.t. the ground truth labels for the source videos, therefore the overall loss is $\mathcal{L}_{CE}(V_S) + \hat{\mathcal{L}}_{CE}(\tilde{V}_T)$. In this case, the annotated source videos are utilized to regularize domain alignment on the image model, and provide further supervision for learning the classification task on the video model. In Sec. 4.4, we demonstrate that in the context of mixed-source video adaptation, even a limited amount of source videos is sufficient to achieve results competitive to video-to-video adaptation approaches that employ the entire source video dataset.

4   Experiments

4.1   Datasets

To evaluate our CycDA framework for image-to-video adaptation, we conduct experiments on 3 real-world image-video action recognition benchmark settings. Videos are from two large-scale action recognition datasets, UCF101 [35] and HMDB51 [23]. Web images are from the EADs (Extensive Action Dataset) [2], Stanford40 [42] and the BU101 dataset [29].

The three image-to-video adaptation tasks are: (1) Stanford40 → UCF101: the UCF101 action dataset contains 13320 videos collected from YouTube with 101 action classes. The Stanford40 dataset contains 9532 images collected from Google, Bing and Flickr, comprised of 40 action classes. Following [2,43,44], we select the 12 common action classes between the two datasets for image-to-video action recognition. (2) EADs→HMDB51: HMDB51 has 6766 videos with 51 action classes collected from online videos and movie clips. The EADs dataset consists of Stanford40 and the HII dataset [37]. It has 11504 images from 50 action classes. There are 13 shared action classes between the two datasets. (3) BU101→UCF101: BU101 consists of 23.8K web action images from 101 classes that completely correspond to classes on UCF101. We use data of all classes for evaluation on large-scale image-to-video adaptation.

UCF101 and HMDB51 both have three splits of training and test sets. Following [24,44,43], we report the average performance over all three splits.

The UCF-HMDB dataset [3] is a benchmark for video-to-video DA. It consists of the 12 common classes between UCF101 and HMDB51. On this dataset, we perform two types of evaluations: (1) *frame-to-video adaptation*: we use only a single frame from each source video to adapt to target videos; and (2) mixed-source video adaptation: we use source and target videos of UCF-HMDB, and extend the source domain with web images from BU101.

4.2   Implementation Details

For the image model, we use a ResNet-18 [17] pretrained on ImageNet [8]. We freeze the first 6 sequential blocks and train with a learning rate of 0.001 to perform the domain alignment between web images and frames sampled from target videos. To avoid redundancy in the video frames, we uniformly divide a video into 5 segments. In each training epoch, we randomly sample one frame from each segment. As trade-off weight for domain discrimination, we follow the common practices in [14,15,4] to gradually increase β from 0 to 1. The temperature parameter τ is set to 0.05.

For the video model, we employ I3D Inception v1 [1] pretrained on the Kinetics dataset [21], which is common practice, *e.g.* [2,7,22,31,33]. We train the RGB stream only. To validate the efficacy of our CycDA pipeline, we use the I3D backbone with a shallow classifier of 2 FC layers, without any temporal aggregation module (*e.g.* GCN in [33] or self-attention module in [7]). We extract a clip of 64 frames from each video. Following [33], we use a learning rate of

Method	Backbone	E→H	S→U	B→U
source only	ResNet18	37.2	76.8	54.8
DANN [15]*	ResNet18	39.6	80.3	55.3
UnAtt [24]	ResNet101	-	-	66.4
HiGAN [44]	ResNet50, C3D	44.6	95.4	-
SymGAN [43]	ResNet50, C3D	55.0	97.7	-
CycDA (1 iteration)	ResNet50, C3D	56.6	98.0	-
DANN [15]+I3D*	ResNet18, I3D	53.8	97.9	68.3
HPDA [2]*	ResNet50, I3D	38.2	40.0	-
CycDA (1 iteration)	ResNet18, I3D	60.5	99.2	69.8
CycDA (2 iterations)	ResNet18, I3D	60.3	**99.3**	72.1
CycDA (3 iterations)	ResNet18, I3D	**62.0**	99.1	**72.6**
supervised target	ResNet18, I3D	83.2	99.3	93.1

Table 1. Results on E→H (13 classes), S→U (12 classes) and B→U (101 classes), averaged over 3 splits. ResNet, C3D and I3D are pretrained on ImageNet[17], Sports-1M[20] and Kinetics400[21]. * denotes our evaluation.

0.001 for the backbone and 0.01 on other components. We keep $p = 70\%$ and 80% of videos in stage 2 and stage 4.

4.3   Image-to-video DA

We compare the proposed approach to other image-to-video adaptation methods on the three described benchmark settings as shown in Table 1. As CycDA enables the iterative knowledge transfer between the image model and the video model, we can repeat stage 3 and stage 4 multiple times. We therefore report the performance for the first three iterations. We add the lower bound (source only) and the upper bound (ground truth supervised target) for reference.

We compare against several approaches: DANN [15] is classical adversarial domain discrimination on the image-level. UnAtt [24] applies a spatial attention map on video frames. HiGAN [44] and SymGAN [43] employ GANs for feature alignment (on backbone of ResNet50 and C3D) and define the current state-of-the-art on E→H and S→U. We also evaluate CycDA with the same backbones for fair comparison. DANN [15]+I3D is a strong baseline that trains the I3D model with pseudo labels from an adapted image model. HPDA [2] is a recent partial DA approach and for a fair comparison, we re-run its official implementation in a closed-set DA setting. Our CycDA outperforms all other approaches already after the first iteration. Except for the saturation on S→U, running CycDA for more iterations leads to a further performance boost on all evaluation settings.

We further explore the potential of CycDA on UCF-HMDB, which is a benchmark for video-to-video adaptation. For a strict comparison, we select data from the same source video dataset used in the video-to-video adaptation methods, without using any auxiliary web data for training. However, instead of directly using the source videos, we perform *frame-to-video* adaptation where we use only one frame from each source video to adapt to target videos. Here we sample the middle frame from each video and report the results in Table 2 (case B).

| DA setting | Method | Video backbone | Source data | | U→H | H→U |
			web image	videos (U or H) in %		
A: video-to-video	AdaBN [25]	ResNet101	-	100%	75.5	77.4
	MCD [34]	ResNet101	-	100%	74.4	79.3
	TA^{3}N [3]	ResNet101	-	100%	78.3	81.8
	ABG [28]	ResNet101	-	100%	79.1	85.1
	TCoN [32]	ResNet101	-	100%	87.2	89.1
	DANN [15]	I3D	-	100%	80.7	88.0
	TA^{3}N [3]	I3D	-	100%	81.4	90.5
	SAVA [7]	I3D	-	100%	82.2	91.2
	MM-SADA [31]	I3D	-	100%	84.2	91.1
	CrossModal [22]	I3D	-	100%	84.7	92.8
	CoMix [33]	I3D	-	100%	86.7	93.9
B: frame-to-video	CycDA	I3D	-	one frame	83.3	80.4
C: mixed-source to video	CycDA	I3D	BU*	0%	77.8	88.6
			BU*	5%	82.2	93.1
			BU*	10%	82.5	93.5
			BU*	50%	84.2	95.2
			BU*	100%	88.1	98.0
supervised target		I3D	-	-	94.4	97.0

Table 2. Results of Cycle Adaption on UCF-HMDB in comparison to video-to-video adaptation (case A) approaches. For frame-to-video adaptation (case B), we use only one frame from each source video to adapt to target videos. For mixed-source video adaptation (case C), we combine BU101 web images and source videos as the source data. *We sample 50 web images per class from 12 classes in BU101.

We see that even when using only one frame per source video, on U→H, CycDA (83.3%) can already outperform TA^{3}N [3] (81.4%) and SAVA [7] (82.2%) which use all source videos for video-to-video adaptation. This demonstrates the strength of CycDA to exploit the large informativity in single images such that they could potentially replace videos as the source data. On H→U, our source domain contains only 840 frames from the 840 videos in the HMDB training set on UCF-HMDB, which leads to an inferior performance. We show that this can be easily addressed by adding auxiliary web data in Sec. 4.4.

4.4   Mixed-source image&video-to-video DA

For mixed-source adaptation, we assume to have both web images and some amount of source videos in the source domain. To evaluate this case, we use the source and target videos on UCF-HMDB, and extend the source domain with web images of the 12 corresponding action classes in BU101. We notice that using all web data from BU101 leads to performance saturation on the target video set of UCF-HMDB. Therefore, to validate the efficacy of CycDA, we only sample 50 web images per class as auxiliary training data. We vary the amount of source videos in the mixed-source domain and report the results in Table 2 (case C). First, by training with only sampled web images (without any source videos), we achieve baseline results of 77.8% (BU→H) and 88.6% (BU→U). By adding only 5% of videos to the mixed-source domain, we already achieve performance com-

	Experiment	stage 1	stage 2	stage 3	stage 4	Acc
A:	source only	$\mathcal{L}_{CE}(I_S)$	-	-	-	39.0
B:	source only + video model	$\mathcal{L}_{CE}(I_S)$	$\hat{\mathcal{L}}_{CE}(V_T)$	-	-	50.5
C:	class-agnostic DA + video model	$\mathcal{L}_{CE}(I_S),$ $\mathcal{L}_{ADD}(I_S, V_T^F)$	$\hat{\mathcal{L}}_{CE}(V_T)$	-	-	52.3
D:	case C + vid. self-train×1	$\mathcal{L}_{CE}(I_S),$ $\mathcal{L}_{ADD}(I_S, V_T^F)$	$\hat{\mathcal{L}}_{CE}(V_T)$	$\hat{\mathcal{L}}_{CE}(V_T)$	-	55.4
E:	case C + vid. self-train×2	$\mathcal{L}_{CE}(I_S),$ $\mathcal{L}_{ADD}(I_S, V_T^F)$	$\hat{\mathcal{L}}_{CE}(V_T)$	$\hat{\mathcal{L}}_{CE}(V_T)$	$\hat{\mathcal{L}}_{CE}(V_T)$	56.4
F:	CycDA	$\mathcal{L}_{CE}(I_S),$ $\mathcal{L}_{ADD}(I_S, V_T^F)$	$\hat{\mathcal{L}}_{CE}(V_T)$	$\mathcal{L}_{CE}(I_S),$ $\mathcal{L}_{CONTR}(I_S, V_T^F)$	$\hat{\mathcal{L}}_{CE}(V_T)$	**60.8**

Table 3. Stage-wise ablation study of the CycDA training pipeline on EADs $\rightarrow$ HMDB51 split 1. A: source only on image model. B: source only training on image model and video model training. C: class-agnostic DA (stage 1) and video model training (stage 2). D: case C + one stage of self-training the videos model. E: case C + two stages of self-training the video model. F: CycDA with stage 1~4. Category-level pseudo labels of case C and F are compared in Fig. 4.

parable to the video-to-video adaptation methods, *i.e.* 82.0% (BU+U→H) and 93.1% (BU+H→U). Furthermore, increasing the amount of source videos from 5% to 50% leads to another improvement of 2%. As web images are more informative than sampled video frames, using web images as auxiliary training data can thus significantly reduce the amount of videos required. Finally, with sampled web images and all source videos, we outperform all video-to-video adaptation methods, even exceeding the supervised target model for BU+H→U by 1%. Considering that we only use 50 web images per class, this further demonstrates that CycDA can exploit both, the information in web images and knowledge from the source data with domain shift, for a potentially improved learning.

4.5 Ablation study

We perform several ablation studies to validate the proposed CycDA pipeline. We conduct these experiments on the setting of EADs $\rightarrow$ HMDB51 (split 1).

Stage-wise ablation study. We first validate the efficacy of the CycDA pipeline by switching the stages with alternate counterparts. We report the quantitative results of six ablation settings in Table 3. Case A (source only), training the image model on web images only and predicting on target test set, demonstrates the lower bound of 39%. Case B (source only + video model), training with pseudo labeled target videos, is a vanilla baseline. It shows that training the video model with supervision from the image model already significantly improves performance by 11.5%. In case C, with class-agnostic domain alignment on the image model, the performance of B is improved by 1.8%. In case F, after completing the cycle with class-aware domain alignment in stage 3 and training the video model in stage 4 with the updated pseudo labels, we achieve the best performance with 60.8%.
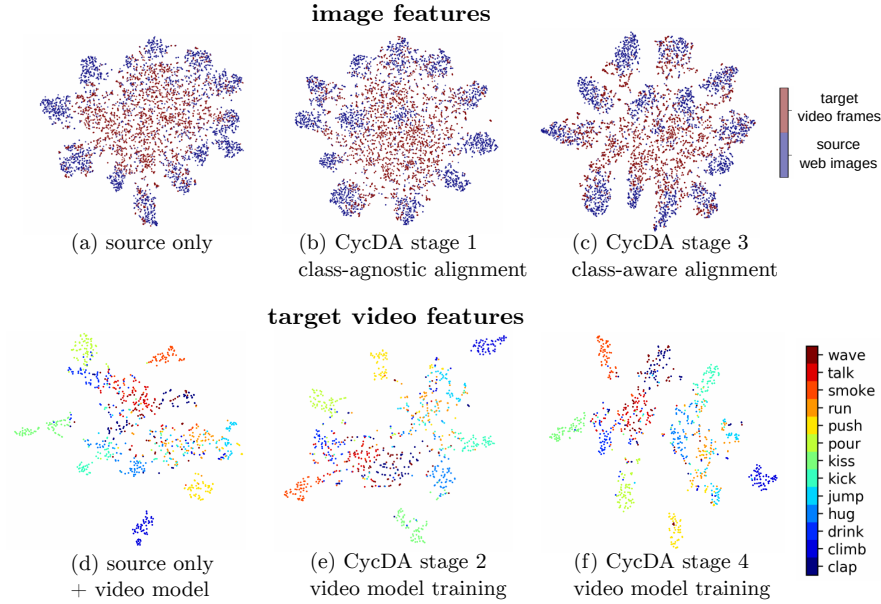
Figure 3. t-SNE visualizations of image features (for both source and target) and target video features (colored w.r.t. ground truth). We plot the results of source only, source only and video model training, and results of each stage in CycDA.

Additionally, we conduct ablation experiments using pseudo labels from case C to self-train the video model. Although self-training the video model for 1 (case D) or 2 (case E) stages exhibits performance improvement compared to case C, it is still clearly outperformed by CycDA. This indicates that the elaborate step of knowledge transfer from the video model to the image model and class-aware domain alignment are critical for a good performance.

We further illustrate the t-SNE [30] feature visualizations of ablation cases in Fig. 3. By observing image features of the source only case (Fig. 3(a)), we see that web images (blue) gather in category clusters after supervised training, while target video frames (red) are far less discriminative and highly misaligned with source due to large domain shift. The class-agnostic domain alignment (Fig. 3(b)) in stage 1 results in a slightly better global alignment of source and target features. With category knowledge from the video model, the class-aware domain alignment (Fig. 3(c)) demonstrates distinctly better category-level association between source and target. By passing pseudo labels from the image model of Fig. 3(a)(b)(c) to supervise the spatio-temporal training, we get the corresponding video models whose features are plotted in Fig. 3(d)(e)(f). The vanilla baseline of source only + video model (Fig. 3(d)) leads to highly undiscriminative features on the difficult classes, which is slightly improved by class-agnostic alignment (Fig. 3(e)). Class-aware domain alignment (Fig. 3(f)) results in more pronounced category clusters with larger inter-class distances.
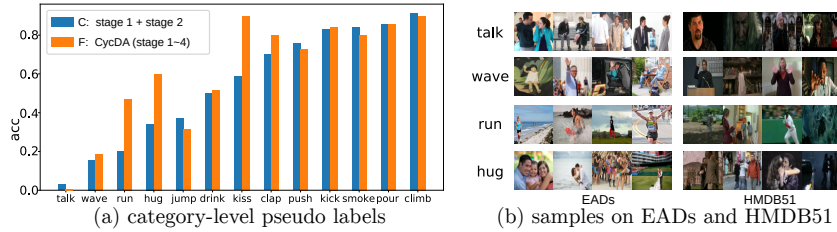
(a) category-level pseudo labels   (b) samples on EADs and HMDB51

Figure 4. (a) Category-level pseudo label analysis on target. We compare the accuracy of category-level pseudo labels on target videos for the ablation case C and F from Table 3. Here we plot the accuracy of pseudo labels of the 13 common classes on EADs → HMDB51. (b) Samples of four categories on EADs and HMDB51.

Category-level pseudo label analysis on target. In Fig. 4(a), we plot the accuracy of category-wise pseudo labels on target videos for the ablation cases C and F from Table 3. On case C (blue) which consists of stage 1 (class-agnostic spatial domain alignment) and stage 2 (spatio-temporal learning), image-to-video action recognition has varying performance on different action classes. More specifically, it yields better results on appearance-based actions with distinct background (*e.g. climb*), or on actions with discriminative pose or gesture (*e.g. smoke, pour*). On the contrary, there is inferior performance on fine-grained actions with subtle movements (*e.g. talk*) or actions that are semantically highly generalized with large inter-class variation (*e.g. run, wave*), *c.f.* in Fig. 4(b).

By comparing the pseudo label accuracy of cases C and F, we see that the complete CycDA with stage 3 and stage 4 contributes to a significant performance boost on the difficult classes (*e.g. run, hug, kiss*) while keeping the performance on the easy classes. This can be attributed to the class-aware domain alignment that improves cross-domain association on the difficult classes while keeping the alignment on easy ones. Note that better class-aware domain alignment also leads to more support samples for difficult classes, which could result in slight performance drop on the easier classes.

Domain adaptation strategies in stage 3. In stage 3, we transfer knowledge from the video model to the image model by passing pseudo labels from the video model to provide category information in domain alignment. We compare different DA strategies that use pseudo labels from the video model in Table 4. Intuitively, performing the supervised task on both source and target (case B) outperforms the case of pseudo labeled target only (case A). Adding the adversarial domain discrimination (case C) leads to a further boost. The cross-domain contrastive learning (case D) has the best performance among the 4 cases. In comparison to the case of only stage 1 and 2, a complete cycle with 4 stages leads to performance improvement of at least 3.3% for all cases. This indicates the benefits of transferring the knowledge from the video model onto the image model. The proposed CycDA generalizes well on different strategies using pseudo labels from the previous stage.

Experiment		stage 3	stage 4	Acc
stage 1~2		-	-	52.3
stage 1~4	A: B: C: D:	$\mathcal{L}_{CE}(V_T^F)$ $\mathcal{L}_{CE}(I_S, V_T^F)$ $\mathcal{L}_{CE}(I_S, V_T^F), \mathcal{L}_{ADD}(I_S, V_T^F)$ $\mathcal{L}_{CE}(I_S), \mathcal{L}_{CONTR}(I_S, V_T^F)$	$\hat{\mathcal{L}}_{CE}(V_T)$	55.6 56.4 58.0 **60.8**

Table 4. Comparison of DA strategies in stage 3 on EADs → HMDB51 split 1. A: supervised classification on pseudo labeled target frames. B: supervised classification on source and pseudo labeled target. C: B + adversarial domain discrimination. D: supervised classification on source + cross-domain contrastive learning.

	Aggregation	Acc
A:	avg. + class-balanced thresh.	50.3
B:	avg. + thresh.	55.9
C:	thresh. + avg.	**60.8**

Table 5. Temporal aggregation of frame-level pseudo labels into video-level pseudo labels on EADs → HMDB51 split 1.

Figure 5. Performance of multiple iterations of CycDA (stage 3 and 4) on EADs → HMDB51. Average on 3 splits.

Temporal aggregation of frame-level pseudo labels. Before video model training, we temporally aggregate frame-level pseudo labels. Here we compare three strategies of temporal aggregation in Table 5. Frame-level thresholding followed by temporal averaging (case C) outperforms the other two cases which perform temporal averaging before video-level thresholding. Frame-level thresholding filters out frames of low confidence scores and effectively increases the accuracy of video pseudo labels.

Number of iterations. Finally, we illustrate the performance after several iterations on EADs→HMDB51 in Fig. 5. It shows that within the first five iterations, performing CycDA iteratively results in a slight increase of performance. Further, within all the 9 iterations, CycDA delivers relatively stable performance, with the result fluctuating between 60.0% and 62.2% without dropping below the ablated alternatives shown in Table 3.

5   Conclusions

We presented CycDA to address the image-to-video adaptation problem. We proposed to alternately perform spatial domain alignment to address the domain shift between images and target frames, and spatio-temporal learning to bridge the modality gap. Our evaluations across several benchmarks and datasets demonstrate that CycDA exploits the large informativity in the source images for enhanced performance on the target video classifier.

References

1. Carreira, J., Zisserman, A.: Quo vadis, action recognition? a new model and the kinetics dataset. In: CVPR. pp. 6299–6308 (2017)
2. Chen, J., Wu, X., Hu, Y., Luo, J.: Spatial-temporal causal inference for partial image-to-video adaptation. In: AAAI. vol. 35, pp. 1027–1035 (2021)
3. 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)
4. Chen, M.H., Li, B., Bao, Y., AlRegib, G., Kira, Z.: Action segmentation with joint self-supervised temporal domain adaptation. In: CVPR. pp. 9454–9463 (2020)
5. Chen, T., Kornblith, S., Norouzi, M., Hinton, G.: A simple framework for contrastive learning of visual representations. In: ICML. pp. 1597–1607. PMLR (2020)
6. Choi, J., Sharma, G., Chandraker, M., Huang, J.B.: Unsupervised and semi-supervised domain adaptation for action recognition from drones. In: WACV. pp. 1717–1726 (2020)
7. Choi, J., Sharma, G., Schulter, S., Huang, J.B.: Shuffle and attend: Video domain adaptation. In: ECCV. pp. 678–695. Springer (2020)
8. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: CVPR. pp. 248–255. Ieee (2009)
9. Duan, H., Zhao, Y., Xiong, Y., Liu, W., Lin, D.: Omni-sourced webly-supervised learning for video recognition. In: ECCV. pp. 670–688. Springer (2020)
10. Feichtenhofer, C.: X3d: Expanding architectures for efficient video recognition. In: CVPR (2020)
11. Gan, C., Sun, C., Duan, L., Gong, B.: Webly-supervised video recognition by mutually voting for relevant web images and web video frames. In: ECCV. pp. 849–866. Springer (2016)
12. Gan, C., Sun, C., Nevatia, R.: Deck: Discovering event composition knowledge from web images for zero-shot event detection and recounting in videos. In: AAAI. vol. 31 (2017)
13. Gan, C., Yao, T., Yang, K., Yang, Y., Mei, T.: You lead, we exceed: Labor-free video concept learning by jointly exploiting web videos and images. In: CVPR. pp. 923–932 (2016)
14. Ganin, Y., Lempitsky, V.: Unsupervised domain adaptation by backpropagation. In: ICML. pp. 1180–1189. PMLR (2015)
15. 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)
16. Guo, S., Huang, W., Zhang, H., Zhuang, C., Dong, D., Scott, M.R., Huang, D.: Curriculumnet: Weakly supervised learning from large-scale web images. In: ECCV. pp. 135–150 (2018)
17. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: CVPR. pp. 770–778 (2016)
18. Jamal, A., Namboodiri, V.P., Deodhare, D., Venkatesh, K.: Deep domain adaptation in action space. In: BMVC. vol. 2, p. 5 (2018)
19. Kae, A., Song, Y.: Image to video domain adaptation using web supervision. In: WACV. pp. 567–575 (2020)
20. Karpathy, A., Toderici, G., Shetty, S., Leung, T., Sukthankar, R., Fei-Fei, L.: Large-scale video classification with convolutional neural networks. In: CVPR. pp. 1725–1732 (2014)

21. Kay, W., Carreira, J., Simonyan, K., Zhang, B., Hillier, C., Vijayanarasimhan, S., Viola, F., Green, T., Back, T., Natsev, P., et al.: The kinetics human action video dataset. arXiv preprint arXiv:1705.06950 (2017)
22. 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)
23. Kuehne, H., Jhuang, H., Garrote, E., Poggio, T., Serre, T.: Hmdb: a large video database for human motion recognition. In: ICCV. pp. 2556–2563. IEEE (2011)
24. Li, J., Wong, Y., Zhao, Q., Kankanhalli, M.S.: Attention transfer from web images for video recognition. In: ACM Multimedia. pp. 1–9 (2017)
25. Li, Y., Wang, N., Shi, J., Hou, X., Liu, J.: Adaptive batch normalization for practical domain adaptation. Pattern Recognition **80**, 109–117 (2018)
26. Liu, H., Wang, J., Long, M.: Cycle self-training for domain adaptation. arXiv preprint arXiv:2103.03571 (2021)
27. Liu, Y., Lu, Z., Li, J., Yang, T., Yao, C.: Deep image-to-video adaptation and fusion networks for action recognition. TIP **29**, 3168–3182 (2019)
28. 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)
29. Ma, S., Bargal, S.A., Zhang, J., Sigal, L., Sclaroff, S.: Do less and achieve more: Training cnns for action recognition utilizing action images from the web. Pattern Recognition **68**, 334–345 (2017)
30. Van der Maaten, L., Hinton, G.: Visualizing data using t-sne. JMLR **9**(11) (2008)
31. Munro, J., Damen, D.: Multi-modal domain adaptation for fine-grained action recognition. In: CVPR. pp. 122–132 (2020)
32. 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)
33. Sahoo, A., Shah, R., Panda, R., Saenko, K., Das, A.: Contrast and mix: Temporal contrastive video domain adaptation with background mixing. In: NeurIPS (2021)
34. Saito, K., Watanabe, K., Ushiku, Y., Harada, T.: Maximum classifier discrepancy for unsupervised domain adaptation. In: CVPR. pp. 3723–3732 (2018)
35. 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)
36. Sun, C., Shetty, S., Sukthankar, R., Nevatia, R.: Temporal localization of fine-grained actions in videos by domain transfer from web images. In: ACM Multimedia. pp. 371–380 (2015)
37. Tanisik, G., Zalluhoglu, C., Ikizler-Cinbis, N.: Facial descriptors for human interaction recognition in still images. Pattern Recognition Letters **73**, 44–51 (2016)
38. Wang, L., Xiong, Y., Lin, D., Van Gool, L.: Untrimmednets for weakly supervised action recognition and detection. In: CVPR. pp. 4325–4334 (2017)
39. Wang, Z., She, Q., Smolic, A.: Action-net: Multipath excitation for action recognition. In: CVPR (2021)
40. Yang, C., Xu, Y., Shi, J., Dai, B., Zhou, B.: Temporal pyramid network for action recognition. In: CVPR (2020)
41. Yang, J., Sun, X., Lai, Y.K., Zheng, L., Cheng, M.M.: Recognition from web data: A progressive filtering approach. TIP **27**(11), 5303–5315 (2018)
42. Yao, B., Jiang, X., Khosla, A., Lin, A.L., Guibas, L., Fei-Fei, L.: Human action recognition by learning bases of action attributes and parts. In: ICCV. pp. 1331–1338. IEEE (2011)

43. Yu, F., Wu, X., Chen, J., Duan, L.: Exploiting images for video recognition: Heterogeneous feature augmentation via symmetric adversarial learning. TIP **28**(11), 5308–5321 (2019)
44. Yu, F., Wu, X., Sun, Y., Duan, L.: Exploiting images for video recognition with hierarchical generative adversarial networks. In: IJCAI (2018)
45. Zhang, J., Han, Y., Tang, J., Hu, Q., Jiang, J.: Semi-supervised image-to-video adaptation for video action recognition. IEEE Trans Cybern **47**(4), 960–973 (2016)
46. Zhang, Y., Deng, B., Jia, K., Zhang, L.: Label propagation with augmented anchors: A simple semi-supervised learning baseline for unsupervised domain adaptation. In: ECCV. pp. 781–797. Springer (2020)
47. Zhuang, B., Liu, L., Li, Y., Shen, C., Reid, I.: Attend in groups: a weakly-supervised deep learning framework for learning from web data. In: CVPR. pp. 1878–1887 (2017)
48. Zou, Y., Yu, Z., Kumar, B., Wang, J.: Unsupervised domain adaptation for semantic segmentation via class-balanced self-training. In: ECCV. pp. 289–305 (2018)
49. Zou, Y., Yu, Z., Liu, X., Kumar, B., Wang, J.: Confidence regularized self-training. In: ICCV. pp. 5982–5991 (2019)