Extended Abstract
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]
Horst Bischof[1]
[1]Institute for 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

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 by leveraging the joint spatial information in images and videos on the one hand and, on the other hand, training 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 on benchmark datasets for image-to-video and mixed-source domain adaptation achieving state-of-the-art results.

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 [6, 24, 25], 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 [8, 11, 14, 23, 26, 29].

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 [7, 9] or perform joint training by combining the web source with annotated target data [5, 18]. To specifically address the domain shift between web images and target videos, some approaches perform class-agnostic domain-invariant feature learning either within [12] or across modalities [16, 27, 28], 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.

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.

Our contributions are: (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.

2. Cycle Domain Adaptation (CycDA)

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:

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 $\phi^I = \{E^I(\cdot;\theta^I_E),\ C^I(\cdot;\theta^I_C),\ D^I(\cdot;\theta^I_D)\}$ and the video model $\phi^V = \{E^V(\cdot;\theta^V_E),\ C^V(\cdot;\theta^V_C)\}$. 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 $I_S = \{(i_j, l(i_j))|_{j=1}^{N^I_S}\}$, where $l(\cdot)$ is the ground truth label of the corresponding image. The unlabeled target video domain is $V_T = \{v_j|_{j=1}^{N^V_T}\}$ and each video v_j has M_j frames, the set of frames of unlabeled target videos $V^F_T = \{\{v^F_{j,m}|_{m=1}^{M_j}\}|_{j=1}^{N^V_T}\}$.

Stage 1 - Class-agnostic Spatial Alignment. We learn the class-agnostic domain alignment between source web images and frames sampled from unlabeled target videos. This reduces the domain gap between the appearance of the web images and target videos even if the classes could be misaligned 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 $\mathcal{L}_{ADD}(I_S, V^F_T)$ on source images and target frames. With the classification loss on source images $\mathcal{L}_{CE}(I_S)$ and the binary cross entropy loss for domain discrimination given as $\mathcal{L}_{ADD}(I_S, V^F_T) = \sum_{i_j, v^F_{j',m}} \log D^I(E^I(i_j;\theta^I_E);\theta^I_D) + \log(1 - D^I(E^I(v^F_{j',m};\theta^I_E);\theta^I_D))$, the overall objective is $\min_{\theta^I_E,\theta^I_C} \mathcal{L}_{CE}(I_S) + \beta \max_{\theta^I_E} \min_{\theta^I_D} \mathcal{L}_{ADD}(I_S, V^F_T)$, where β is the trade-off weight between the two losses. The domain alignment 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 these stages, we use the trained image model ϕ^I from the previous stage to generate pseudo labels for frames of the target videos. We perform frame confidence thresholding and aggregate frame-level predictions into video-level pseudo labels. 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 $\min_{\theta^V_E,\theta^V_C} \sum_{v_j \in \tilde{V}_T} -\hat{l}(v_j) \cdot \log(C^V(E^V(V_j;\theta^V_E);\theta^V))$.. 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. 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 sim-

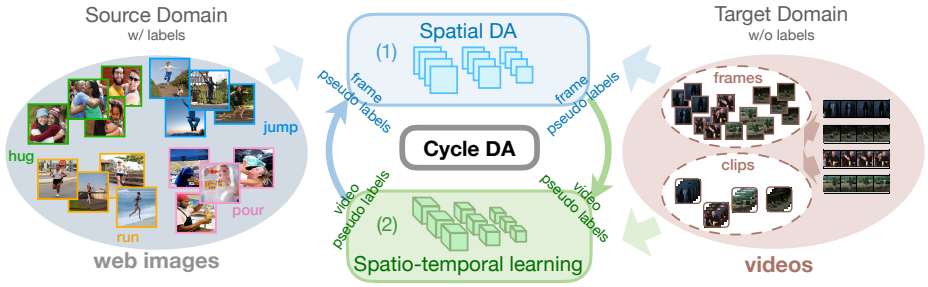

Figure 1. Cycle Domain Adaptation (CycDA) pipeline: we address image-to-video adaption by training a spatial model and a spatio-temporal model alternately, passing pseudo labels to supervise each other in a cycle. The two alternating steps are: (1) domain alignment on the spatial model with pseudo labels from the spatio-temporal model, and (2) training the spatio-temporal model with updated pseudo labels from the spatial model.

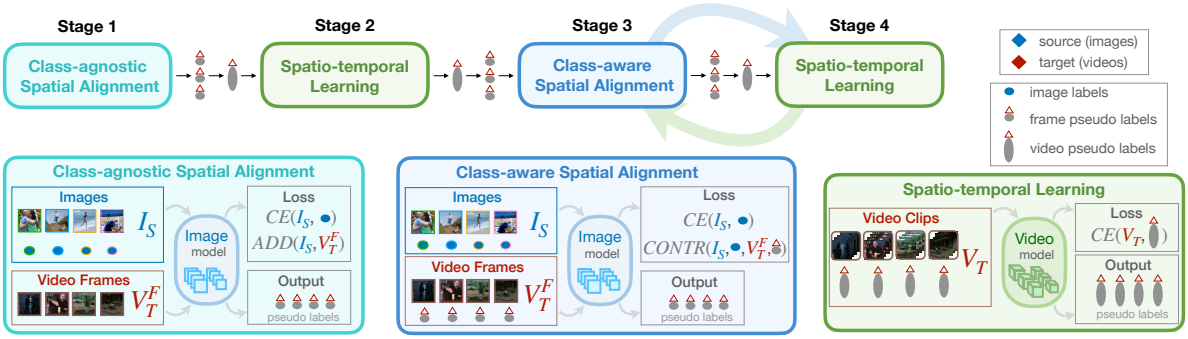

Figure 2. Our CycDA framework alternates between spatial alignment (stage 1 and 3) and spatio-temporal learning (stage 2 and 4). See text for details.

ilarity 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 $\mathcal{L}_{CONTR}(I_S, V_T^F) = -\sum_{z_j^F \in Z_T^F} \log \frac{h(z_j^F, z_{j+}^I)}{h(z_j^F, z_{j+}^I)+h(z_j^F, z_{j-}^I)}$. Following [3], we set $h(u,v) = \exp(\text{sim}(u,v)/\tau)$, where we use the cosine similarity $\text{sim}(u,v) = u^\top v/(\|u\|\|v\|)$ and τ is the temperature parameter. Thus, the objective of stage 3 on the image model is $\min_{\theta_E^I, \theta_C^I} \mathcal{L}_{CE}(I_S) + \mathcal{L}_{CONTR}(I_S, V_T^F)$.

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.

Mixed-source Video Adaptation. Image-to-video DA applies to the case where the source domain consists only of web images. However, other possible settings presume limited amount of annotated videos with the domain shift to the 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 Table 1, we demonstrate that in the context of mixed-source video adaptation, even a limited amount of source videos are sufficient to achieve results competitive to video-to-video adaptation approaches that employ the entire source video dataset.

DA setting		method	video backbone	source data		U→H	H→U
				web image	videos (U or H) in %		
A:	video-to-video	AdaBN [15]	ResNet101	-	100%	75.5	77.4
		MCD [22]	ResNet101	-	100%	74.4	79.3
		TA^{3}N [2]	ResNet101	-	100%	78.3	81.8
		ABG [17]	ResNet101	-	100%	79.1	85.1
		TCoN [20]	ResNet101	-	100%	87.2	89.1
		DANN [10]	I3D	-	100%	80.7	88.0
		TA^{3}N [2]	I3D	-	100%	81.4	90.5
		SAVA [4]	I3D	-	100%	82.2	91.2
		MM-SADA [19]	I3D	-	100%	84.2	91.1
		CrossModal [13]	I3D	-	100%	84.7	92.8
		CoMix [21]	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 1. 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.

Method	Backbone	E→H	S→U	B→U
source only	ResNet18	37.2	76.8	54.8
DANN [10]*	ResNet18	39.6	80.3	55.3
UnAtt [14]	ResNet101	-	-	66.4
HiGAN [28]	ResNet50, C3D	44.6	95.4	-
SymGAN [27]	ResNet50, C3D	55.0	97.7	-
CycDA (1 iteration)	ResNet50, C3D	56.6	98.0	-
DANN [10]+I3D*	ResNet18, I3D	53.8	97.9	68.3
HPDA [1]*	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 2. Results on EADs→HMDB51 (13 classes), Stanford40→UCF101 (12 classes) and BU101→UCF101 (101 classes), averaged over 3 splits. * denotes our evaluation.

3. Experiments

We evaluate on 3 *image-to-video* action recognition benchmark settings: Stanford40 → UCF101 (12 classes), EADs→HMDB51 (13 classes) and BU101→UCF101 (101 classes). Besides, we perform *frame-to-video* adaptation and *mixed-source to video* adaptation on the UCF-HMDB dataset, which is a benchmark for video-to-video DA.

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 2. We report the CycDA performance for the first three iterations. 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.

Frame-to-video DA. We further explore the potential of CycDA on *frame-to-video* adaptation (Table 1) on UCF-HMDB, which is a benchmark for video-to-video DA. Instead of directly using the source videos, we use only one frame from each source video to adapt to target videos. On U→H, CycDA (83.3%) can already outperform *video-to-video* adaptation approaches TA^{3}N [2] (81.4%) and SAVA [4] (82.2%). 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 Mixed-source to video DA (Table 1).

Mixed-source to video DA. For *mixed-source to video* adaptation (Table 1), 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. To validate the efficacy of CycDA, we only sample 50 web images per class as auxiliary training data. 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 comparable to the video-to-video adaptation methods, *i.e.* 82.0% (BU+U→H) and 93.1% (BU+H→U). 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%.

References

[1] Jin Chen, Xinxiao Wu, Yao Hu, and Jiebo Luo. Spatial-temporal causal inference for partial image-to-video adaptation. In *AAAI*, volume 35, pages 1027–1035, 2021. 4

[2] 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*, pages 6321–6330, 2019. 4

[3] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *ICML*, pages 1597–1607. PMLR, 2020. 3

[4] Jinwoo Choi, Gaurav Sharma, Samuel Schulter, and Jia-Bin Huang. Shuffle and attend: Video domain adaptation. In *ECCV*, pages 678–695. Springer, 2020. 4

[5] Haodong Duan, Yue Zhao, Yuanjun Xiong, Wentao Liu, and Dahua Lin. Omni-sourced webly-supervised learning for video recognition. In *ECCV*, pages 670–688. Springer, 2020. 1

[6] Christoph Feichtenhofer. X3d: Expanding architectures for efficient video recognition. In *CVPR*, 2020. 1

[7] Chuang Gan, Chen Sun, Lixin Duan, and Boqing Gong. Webly-supervised video recognition by mutually voting for relevant web images and web video frames. In *ECCV*, pages 849–866. Springer, 2016. 1

[8] Chuang Gan, Chen Sun, and Ram Nevatia. Deck: Discovering event composition knowledge from web images for zero-shot event detection and recounting in videos. In *AAAI*, volume 31, 2017. 1

[9] Chuang Gan, Ting Yao, Kuiyuan Yang, Yi Yang, and Tao Mei. You lead, we exceed: Labor-free video concept learning by jointly exploiting web videos and images. In *CVPR*, pages 923–932, 2016. 1

[10] Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. Domain-adversarial training of neural networks. *JMLR*, 17(1):2096–2030, 2016. 4

[11] Sheng Guo, Weilin Huang, Haozhi Zhang, Chenfan Zhuang, Dengke Dong, Matthew R Scott, and Dinglong Huang. Curriculumnet: Weakly supervised learning from large-scale web images. In *ECCV*, pages 135–150, 2018. 1

[12] Andrew Kae and Yale Song. Image to video domain adaptation using web supervision. In *WACV*, pages 567–575, 2020. 1

[13] Donghyun Kim, Yi-Hsuan Tsai, Bingbing Zhuang, Xiang Yu, Stan Sclaroff, Kate Saenko, and Manmohan Chandraker. Learning cross-modal contrastive features for video domain adaptation. In *ICCV*, pages 13618–13627, 2021. 4

[14] Junnan Li, Yongkang Wong, Qi Zhao, and Mohan S Kankanhalli. Attention transfer from web images for video recognition. In *ACM Multimedia*, pages 1–9, 2017. 1, 4

[15] Yanghao Li, Naiyan Wang, Jianping Shi, Xiaodi Hou, and Jiaying Liu. Adaptive batch normalization for practical domain adaptation. *Pattern Recognition*, 80:109–117, 2018. 4

[16] Yang Liu, Zhaoyang Lu, Jing Li, Tao Yang, and Chao Yao. Deep image-to-video adaptation and fusion networks for action recognition. *TIP*, 29:3168–3182, 2019. 1

[17] Yadan Luo, Zi Huang, Zijian Wang, Zheng Zhang, and Mahsa Baktashmotlagh. Adversarial bipartite graph learning for video domain adaptation. In *ACM Multimedia*, pages 19–27, 2020. 4

[18] Shugao Ma, Sarah Adel Bargal, Jianming Zhang, Leonid Sigal, and Stan Sclaroff. Do less and achieve more: Training cnns for action recognition utilizing action images from the web. *Pattern Recognition*, 68:334–345, 2017. 1

[19] Jonathan Munro and Dima Damen. Multi-modal domain adaptation for fine-grained action recognition. In *CVPR*, pages 122–132, 2020. 4

[20] Boxiao Pan, Zhangjie Cao, Ehsan Adeli, and Juan Carlos Niebles. Adversarial cross-domain action recognition with co-attention. In *AAAI*, volume 34, pages 11815–11822, 2020. 4

[21] Aadarsh Sahoo, Rutav Shah, Rameswar Panda, Kate Saenko, and Abir Das. Contrast and mix: Temporal contrastive video domain adaptation with background mixing. In *NeurIPS*, 2021. 4

[22] Kuniaki Saito, Kohei Watanabe, Yoshitaka Ushiku, and Tatsuya Harada. Maximum classifier discrepancy for unsupervised domain adaptation. In *CVPR*, pages 3723–3732, 2018. 4

[23] Limin Wang, Yuanjun Xiong, Dahua Lin, and Luc Van Gool. Untrimmednets for weakly supervised action recognition and detection. In *CVPR*, pages 4325–4334, 2017. 1

[24] Zhengwei Wang, Qi She, and Aljosa Smolic. Action-net: Multipath excitation for action recognition. In *CVPR*, 2021. 1

[25] Ceyuan Yang, Yinghao Xu, Jianping Shi, Bo Dai, and Bolei Zhou. Temporal pyramid network for action recognition. In *CVPR*, 2020. 1

[26] Jufeng Yang, Xiaoxiao Sun, Yu-Kun Lai, Liang Zheng, and Ming-Ming Cheng. Recognition from web data: A progressive filtering approach. *TIP*, 27(11):5303–5315, 2018. 1

[27] Feiwu Yu, Xinxiao Wu, Jialu Chen, and Lixin Duan. Exploiting images for video recognition: Heterogeneous feature augmentation via symmetric adversarial learning. *TIP*, 28(11):5308–5321, 2019. 1, 4

[28] Feiwu Yu, Xinxiao Wu, Yuchao Sun, and Lixin Duan. Exploiting images for video recognition with hierarchical generative adversarial networks. In *IJCAI*, 2018. 1, 4

[29] Bohan Zhuang, Lingqiao Liu, Yao Li, Chunhua Shen, and Ian Reid. Attend in groups: a weakly-supervised deep learning framework for learning from web data. In *CVPR*, pages 1878–1887, 2017. 1