Unsupervised Domain Adaptation for Spatio-Temporal Action Localization

Nakul Agarwal[1], Yi-Ting Chen[2], Behzad Dariush[2], and Ming-Hsuan Yang[1,3]

[1]University of California, Merced [2]Honda Research Institute USA [3]Google Research

Abstract

Spatio-temporal action localization is an important problem in computer vision that involves detecting where and when activities occur, and therefore requires modeling of both spatial and temporal features. This problem is typically formulated in the context of supervised learning, where the learned classifiers operate on the premise that both training and test data are sampled from the same underlying distribution. However, this assumption does not hold when there is a significant domain shift, leading to poor generalization performance on the test data. To address this problem, we propose an end-to-end unsupervised domain adaptation algorithm for spatio-temporal action localization. We show that significant performance gain can be achieved when spatial and temporal features are adapted separately, or jointly for the most effective results.

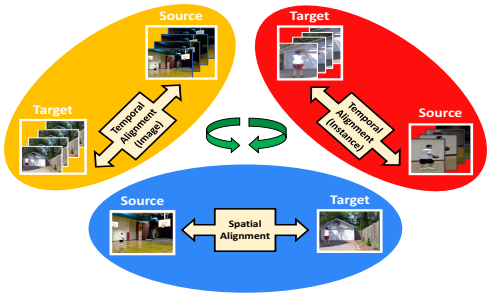

Figure 1: Spatio-temporal action localization requires modeling of spatial features for localization and temporal features for classification. We propose to align both spatial and temporal features for effective adaptation. While the spatial features are only adapted at the image level, the temporal features are aligned both at the image and the instance level. Here, we use the *basketball* action as an example.

1. Introduction

Recently, there has been a significant interest in tackling the spatio-temporal human action localization problem due to its importance in many applications. Although significant advances have been made, existing algorithms generally require a large-scale labeled dataset for supervised learning which is i) non-trivial and not scalable because annotating bounding boxes is expensive and time consuming and ii) do not generalize well when there is a significant domain shift between the underlying distributions in the training and test datasets. This domain shift can be caused by difference in scenarios, lighting conditions or image appearance. In case of videos, the variation in the progression of activity over time can also cause domain shift.

In this work, we focus on the harder problem of generalizing training models to target samples without access to any form of target labels by proposing an end-to-end trainable unsupervised domain adaptation framework based on the Faster R-CNN [9] algorithm for spatio-temporal action localization. To reduce the impact of domain shift, we introduce adaptation modules to jointly align both spatial and temporal features. Specifically, three adaptation modules are proposed: i) for aligning temporal features at the image level, ii) for aligning temporal features at the instance level and iii) for aligning spatial features at the image level. An overview of our proposed approach is shown in Figure 1.

Experimental results demonstrate the effectiveness of the proposed approach in multiple scenarios with domain discrepancies. The contributions of this work are summarized as follows: 1) we present an end-to-end learning framework based on the Faster R-CNN algorithm for unsupervised domain adaptation in the context of spatio-temporal action localization, 2) we design and integrate three domain adaptation modules at the image-level (temporal and spatial) and instance-level (temporal) to alleviate the spatial and temporal domain discrepancy, 3) we design a new experimental setup along with benchmark protocol and perform multiple adaptation experiments and ablation studies to analyze the effect of different adaptation modules and achieve state-of-the-art performance, and 4) we demonstrate that not only does the individual adaptation of spatial and temporal features improve performance, but the adaptation is most effec-

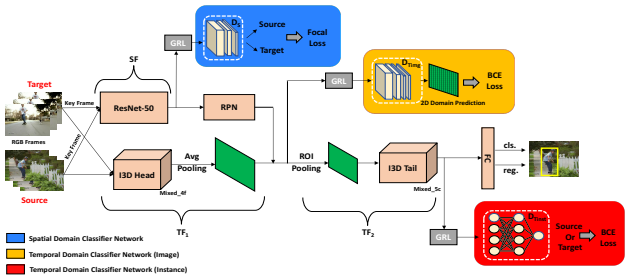

Figure 2: Proposed Network Architecture. The proposed algorithm aligns the distribution of both the spatial and temporal features of source and target domains for adapting actor proposals and action classification respectively.

Previous works [6, 2] focus on adapting actions but only at the whole-clip level. To the best of our knowledge, this work is one of the first to adapt spatio-temporal action localization under the unsupervised setting.

2. Proposed Algorithm

Our framework consists of an action localization model and three different adaptation modules for aligning both spatial and temporal feature distribution. The architecture of the proposed framework is shown in Figure 2.

2.1. Action Localization Model

Our model is based on the Faster R-CNN [9] for end-to-end localization and classification of actions. To model the temporal context, the I3D model [1] is incorporated. The I3D model takes a video V of length T frames and generates the corresponding temporal feature representation using TF, which is the temporal feature extractor. In this work, TF is further decomposed into TF_1 and TF_2. While TF_1 temporally flattens the features from the fused mixed_4f layer of I3D, TF_2 is an action classifier. This results in a compact representation of the entire input sequence.

For the actor proposal generation, we use a 2D ResNet-50 model as the spatial encoder SF on the keyframe K as the input for the region proposal network (RPN), where K is the middle frame of an input clip to I3D. The proposals are generated using the conv4 block of ResNet [5], and are converted to a fixed size feature representation using ROI pooling. This representation is then passed through TF_2, which is an action classifier comprising of the remaining I3D layers up to mixed_5c and an average pooling layer that outputs a feature vector with size $1\times1\times1024$. This feature is used to learn a action classifier and a bounding box regressor.

The loss function of the action localization model is formulated:

$$\mathcal{L}_{act} = \mathcal{L}_{rpn} + \mathcal{L}_{cls} + \mathcal{L}_{reg}, \tag{1}$$

where $\mathcal{L}_{rpn}$, $\mathcal{L}_{cls}$, $\mathcal{L}_{reg}$ are the loss functions for the RPN, final classifier and box regressor respectively. The details regarding these individual loss functions can be found in the original paper [9].

2.2. Adaption in Space and Time

The adaptation process is comprised of two components: i) actor proposal adaptation and ii) action classification adaptation.

Actor Proposal Adaptation. We present a method based on adversarial learning to align the distribution of source and target features for the actor proposal network. Specifically, the spatial domain discriminator D_S is designed to discriminate whether the feature $SF(K)$ is from the source or the target domain. Motivated by [11], the domain classifier is trained to ignore easy-to-classify examples and focus on hard-to-classify examples with respect to the classification of the domain by using the Focal Loss [8]. This prevents strong alignment between global features, which is both difficult and not desirable when there is a considerable domain shift. The loss is based on domain label d of the input image, where $d = 0$ refers to K from the source domain and $d = 1$ refers to K from the target domain. The estimated probability by D_S for the class with label $d = 1$ is denoted by $P \in [0, 1]$, where P is defined as:

$$P = \begin{cases} D_S(SF(K)), & \text{if } d = 1, \\ 1 - D_S(SF(K)), & \text{otherwise.} \end{cases} \tag{2}$$

We formulate the spatial discriminator loss function as:

$$\mathcal{L}_{\mathcal{D}_S} = -\Big(\frac{1}{n_s}\sum_{i=1}^{n_s}(1-P_i^s)^{\gamma}\log(P_i^s) + \frac{1}{n_t}\sum_{j=1}^{n_t}(P_j^t)^{\gamma}\log(1-P_j^t)\Big), \tag{3}$$

where n_s and n_t denote the number of source and target samples in a minibatch respectively, and γ controls the weight on hard to classify examples.

The gradient reversal layer (GRL) [4] is placed between the spatial domain discriminator D_S and spatial feature extractor SF. It helps SF generate domain invariant features $SF(K)$ that fool the discriminator while D_S tries to distinguish the domain.

Action Classification Adaptation. We extend adaptation in the case of images, specifically object detection [3], to videos by proposing to adapt the temporal features at both the image and instance level. While the former focuses on aligning global scene features that serve as context for actions, the latter reduces domain shift between the actor/action dynamics. Specifically, we use TF_1 as a feature extractor for adaptation at the image level and TF_2 for adaptation at the instance level. The TF_1 takes a video clip V

of T frames and generates a compact feature representation $TF_1(V)$ using temporal pooling. We find that adaptation after temporal pooling of features performs well as the actions in our experiments do not vary significantly across datasets in terms of temporal dynamics. This characteristic is also shown in [2] for certain cases where adaptation after temporal pooling performs on par with explicit temporal adaptation modeling. The temporal domain discriminator D_{Timg} then takes $TF_1(V)$ as input and outputs a 2D domain classification map $Q = D_{Timg}(TF_1(V)) \in \mathbb{R}^{H\times W}$. The parameters H and W are determined based on the resolution of V as the spatial strides of TF_1 and D_{Timg} are fixed. We then apply binary cross-entropy (BCE) loss on Q based on the domain label d of the input video V, where $d = 0$ if V belongs to the source domain, and $d = 1$ if V belongs to the target domain. The loss function for D_{Timg} is formulated as:

$$\mathcal{L}_{D_{Timg}} = -\Big(\frac{1}{n_s}\sum_{i=1}^{n_s}\sum_{h,w}(1-d_i)\log\left(1-Q_i^{(h,w)}\right)+ \frac{1}{n_t}\sum_{j=1}^{n_t}\sum_{h,w} d_j \log Q_j^{(h,w)}\Big), \tag{4}$$

where h and w correspond to the spatial indices of an activation in Q.

The instance level representation generated by TF_2 refers to the ROI-based feature vectors before they are fed to the final category classifiers (i.e., the FC layer in Figure 2). The instance level temporal domain classifier D_{Tinst} takes the feature vector $TF_2(TF_1(V))$ as input and outputs a domain classification output for the k-th region proposal in the i-th image as $R_{i,k}$. The BCE loss is used to generate the final output. The corresponding loss function is formulated as:

$$\mathcal{L}_{D_{Tinst}} = -\Big(\frac{1}{n_s}\sum_{i=1}^{n_s}\sum_{k}(1-d_i)\log\left(1-R_{i,k}\right)+ \frac{1}{n_t}\sum_{j=1}^{n_t}\sum_{k} d_j \log R_{j,k}\Big), \tag{5}$$

where $d = 0$ if V belongs to the source distribution and $d = 1$ if V belongs to the target distribution.

2.3. Overall Objective

The overall objective combines losses from the action localization model and the domain adaptation modules. We denote the overall adversarial loss from domain adaptation modules as:

$$\mathcal{L}_{adv}(SF, TF, D) = \mathcal{L}_{\mathcal{D}_S} + \mathcal{L}_{D_{Timg}} + \mathcal{L}_{D_{Tinst}}. \tag{6}$$

For the adaptation task $s \rightarrow t$, given the source video V^s and target video V^t, and by extension their corresponding key frames K^s and K^t respectively, the overall min-max loss function of the proposed framework is defined as the following:

$$\mathcal{L}(V^s, K^s, V^t, K^t) = \mathcal{L}_{act} + \lambda\mathcal{L}_{adv}, \tag{7}$$

where λ is a weight applied to the adversarial loss that balances the action localization loss.

3. Experiments and Analysis

We focus on the scenario of adapting from a smaller annotated domain to a much larger and diverse dataset, which is more challenging than typical settings in the literature. In this work, the target domain is the UCF-101 dataset [12], and the sources are the UCF-Sports [10] and JHMDB [7] sets. Note that the source datasets are much smaller in size and less diverse than the target one. For each adaptation scenario, we show the baseline results of action localization (I3D+RPN) trained on the source data without applying domain adaptation, and a supervised model trained fully on the target domain data (oracle) to illustrate the existing gap between the domains. Then, we train the proposed model on the selected source and target domain to demonstrate the effectiveness of the proposed method. We report intersection-over-union (IoU) performance on both frame-level (frame-mAP) and video-level (video-mAP) at IoU threshold of 0.5.

UCF-Sports → UCF-101. We conduct experiments on common classes from both the datasets, i.e., *Diving*, *Golf-Swing*, *Horse-Riding*, and *Skate-Boarding*. Table 1 show the experimental results on UCF-101 dataset. Since UCF-101 is not a trimmed dataset, we consider background frames during evaluation but not during training to show the effect of adaptation on temporal localization. Significant domain shift and adaptation difficulty can be observed from the baseline and oracle results. A considerable improvement over the baseline is achieved by adapting both spatial or temporal features. For aligning the temporal features, both image level as well as instance level adaptation

Table 1: Frame and video mAP results for adaptation from UCF-Sports to UCF-101 with background frames included. Average Precision (%) is evaluated on target images.

Method	T img	T ins	S img	Div ing	Glf Swg	Hrs Rdg	Skt Bdg	Fr. mAP	Vid. mAP
I3D+RPN				6.9	44.7	30.2	39.0	30.2	18.1
Ours	✓			11.7	51.0	39.3	41.6	35.9	22.6
		✓		11.6	51.1	40.0	42.1	36.2	22.5
			✓	13.3	50.9	50.8	51.5	41.7	22.3
	✓		✓	14.2	51.1	55.5	54.6	43.8	24.0
		✓	✓	12.4	**53.7**	50.5	50.3	41.7	21.6
	✓	✓	✓	**16.9**	51.8	**62.2**	**54.7**	**46.4**	**24.1**
Oracle				83.2	67.9	92.8	91.0	83.7	56.6

yields similar improvement of 5.7% and 6.0% for frame-mAP, and 4.5% and 4.4% for video-mAP respectively.

However, aligning the spatial features, which is responsible for adapting the actor proposals, yields 11.5% (frame-mAP) and 4.2% (video-mAP) improvement over the baseline. The results demonstrate the importance of localizing the action in space, as it is necessary to localize the action first before classification. Finally, we show that the combination of aligning both spatial and temporal features leads to the best results, with performance gains of 16.2% (frame-mAP) and 6.0% (video-mAP) over the baseline. We also observe that the improvement generalizes well across different categories, suggesting that the proposed framework is effective in reducing domain discrepancy across different action classes. Additionally, in order to gauge the individual contribution of the image and instance level adaptation of the temporal features in the final result, we also show ablation results in the table demonstrating that both are required for achieving best performance.

JHMDB → UCF-101. While the UCF-101 dataset is comprised of activities in the sports domain, the JHMDB database consists of videos comprising of everyday activities, although it does contain some sports related sequences. Both datasets have a few common classes: *Basketball*, *Golf-Swing*, *Walk*. Note that the *Walk* from the JHMDB dataset is visually significantly different from the *walking with dog* action in the UCF-101 database, but we still incorporate it in our experiments to increase the number of common classes.

For the results shown in Table 2, we do not consider background frames, but still do temporal localization for *Walk* action which contains few sequences with multiple action instances. The performance gap between baseline and oracle results suggests significant domain shift. We observe similar trends as in the previous case, with considerable improvement obtained by adaptation of either spatial or temporal features for both frame-mAP and video-mAP metrics, and the combination of adapting both spatial and temporal features leading to the best performance gain of 14.2% (frame-mAP) and 17.7% (video-mAP) over the baseline. we also observe that differently from [3], instance level feature alignment combined individually with image level spatial feature adaptation does not yield any improvement. This is because [3] focuses only on spatial feature alignment from the same backbone at image level before RPN and instance level before classification, while we are dealing with both temporal and spatial feature alignment from two separate backbones (i.e., I3D and Resnet-50). Consequently, as shown in the Table 2, temporal feature adaptation at image level is needed, which highlights the importance of our design choice – adaptation for both spatial (image level) and temporal (image and instance level) features.

Table 2: Frame and video mAP results for adaptation from JHMDB to UCF-101. Average precision (%) is evaluated on target images.

Method	T img	T ins	S img	Golf Swg	Bskt Ball	Walk	Fr. mAP	Vid. mAP
I3D+RPN				62.6	38.2	47.2	49.3	51.8
Ours	✓			64.3	40.8	50.6	51.9	56.4
		✓		64.5	40.8	50.8	52.0	56.7
			✓	74.5	56.9	55.3	62.2	69.0
	✓		✓	73.7	56.9	54.4	61.7	64.1
		✓	✓	73.8	58.6	55.5	62.6	68.2
	✓	✓	✓	**75.1**	**59.2**	**56.2**	**63.5**	**69.5**
Oracle				95.7	87.0	90.4	91.0	88.2

References

[1] Joao Carreira and Andrew Zisserman. Quo Vadis, Action Recognition? A New Model and the Kinetics Dataset. In *CVPR*, 2017. 2

[2] Min-Hung Chen, Zsolt Kira, Ghassan AlRegib, Jaekwon Woo, Ruxin Chen, and Jian Zheng. Temporal Attentive Alignment for Large-Scale Video Domain Adaptation. In *ICCV*, 2019. 2, 3

[3] Yuhua Chen, Wen Li, Christos Sakaridis, Dengxin Dai, and Luc Van Gool. Domain Adaptive Faster R-CNN for Object Detection in the Wild. In *CVPR*, 2018. 2, 4

[4] 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), 2016. 2

[5] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep Residual Learning for Image Recognition. In *CVPR*, 2016. 2

[6] Arshad Jamal, Vinay P Namboodiri, Dipti Deodhare, and KS Venkatesh. Deep Domain Adaptation in Action Space. In *BMVC*, 2018. 2

[7] Hueihan Jhuang, Juergen Gall, Silvia Zuffi, Cordelia Schmid, and Michael J Black. Towards Understanding Action Recognition. In *ICCV*, 2013. 3

[8] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal Loss for Dense Object Detection. In *ICCV*, 2017. 2

[9] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. In *NIPS*, 2015. 1, 2

[10] Mikel Rodriguez, Javed Ahmed, and Mubarak Shah. Action MACH a Spatio-temporal Maximum Average Correlation Height filter for action recognition. In *CVPR*, 2008. 3

[11] Kuniaki Saito, Yoshitaka Ushiku, Tatsuya Harada, and Kate Saenko. Strong-Weak Distribution Alignment for Adaptive Object Detection. In *CVPR*, 2019. 2

[12] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. UCF101: A Dataset of 101 Human Actions Classes from Videos in the Wild. *arXiv preprint arXiv:1212.0402*, 2012. 3