Unsupervised Domain Adaptation for Spatio-Temporal Action Localization

Nakul Agarwal[12]
nagarwal@honda-ri.com
Yi-Ting Chen[2]
ychen@honda-ri.com
Behzad Dariush[2]
bdariush@honda-ri.com
Ming-Hsuan Yang[13]
mhyang@ucmerced.edu

[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, we focus on the hard and novel task of generalizing training models to test samples without access to any labels from the latter for spatio-temporal action localization by proposing an end-to-end unsupervised domain adaptation algorithm. We extend the state-of-the-art object detection framework to localize and classify actions. In order to minimize the domain shift, three domain adaptation modules at image level (temporal and spatial) and instance level (temporal) are designed and integrated. We design a new experimental setup and evaluate the proposed method and different adaptation modules on the UCF-Sports, UCF-101 and JH-MDB benchmark datasets. We show that significant performance gain can be achieved when spatial and temporal features are adapted separately, or jointly for the most effective results.

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. Based on the recent benchmark datasets [14, 39, 47] and temporal neural networks [2, 41], numerous algorithms for spatio-temporal action localization have been proposed. Although significant advances have been made, existing algorithms generally require a large-scale labeled dataset for supervised learning which i) is 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

© 2020. The copyright of this document resides with its authors.
It may be distributed unchanged freely in print or electronic forms.

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. Such domain discrepancy causes unfavorable model generalization.

To address problems associated with domain shift, various domain adaptation algorithms have been proposed. Nevertheless, the majority of existing methods focus on images rather than video, catering to problems associated with image classification [10, 25, 32, 43], semantic segmentation [35, 42, 50] and object detection [6, 34]. The ones that do focus on video action understanding can be divided into three categories: whole-clip action recognition, action segmentation, and spatio-temporal action localization. Some progress has been made in this field but only for the former two categories [3, 4, 5, 19, 28], while the latter category remains unattended. Therefore, it is of great interest to develop algorithms for adapting spatio-temporal action localization models to a new domain.

In this work, we focus on the hard problem of generalizing training models to target samples without access to any form of target labels for spatio-temporal action localization by proposing an end-to-end trainable unsupervised domain adaptation framework based on the Faster R-CNN [30] algorithm. To reduce the impact of domain shift, we design and integrate 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. In each module, we train a domain classifier and employ adversarial training to learn domain-invariant features. For aligning the temporal features, both instance-level as well as image-level adaptation are considered. While the former focuses on the actor/action dynamics, the latter incorporates global scene features as context for action classification, which has shown to be effective [40].

Existing video action understanding datasets are not designed for developing and evaluating domain adaptation algorithms in the context of spatio-temporal action localization. To validate the proposed algorithm, we design new experimental settings. We first focus on the scenario of adapting to large scale data using a smaller annotated domain to show that we can harvest more from existing resources. We then provide additional experiments and analysis to study the effect of individual adaptation modules. Extensive experiments and ablation studies are conducted using multiple datasets, i.e., UCF-Sports, UCF-101 and JHMDB. Experimental results demonstrate the effectiveness of the proposed approach for addressing the domain shift of spatio-temporal action localization in multiple scenarios with domain discrepancies.

The contributions of this work are summarized as follows. First, we propose an end-to-end learning framework for solving the novel task of unsupervised domain adaptation in the context of spatio-temporal action localization. Second, 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. Third, we propose a new experimental setup along with benchmark protocol and perform extensive adaptation experiments and ablation studies to analyze the effect of different adaptation modules and achieve state-of-the-art performance. Fourth, we demonstrate that not only does the individual adaptation of spatial and temporal features improve performance, but the adaptation is most effective when both spatial and temporal features are adapted.

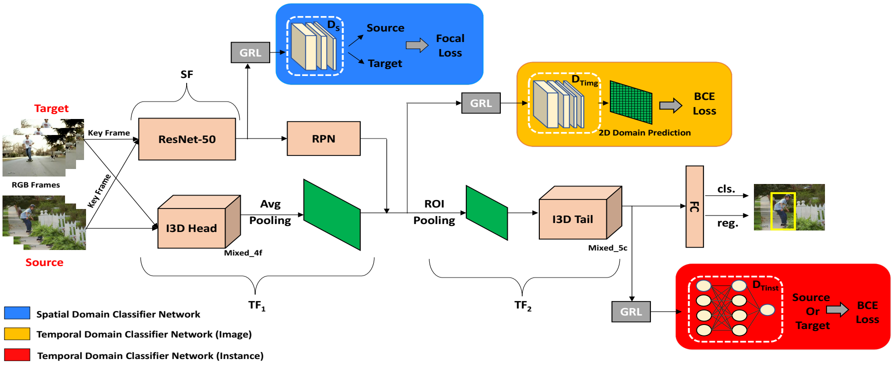


Figure 1: Proposed Network Architecture. The proposed algorithm aligns the distribution of both the spatial and temporal features of source and target domains for adapting actor proposals and action classification respectively. We use a spatial domain classifier network D_s to align the spatial features generated by SF. The temporal features are adapted at the image and instance level using their respective temporal domain classifier networks, i.e., D_{Timg} and D_{Tinst}. Image level features are extracted by TF$_1$ and instance level features are obtained from TF$_2$.

2 Related Work

2.1 Spatio-temporal Action Localization

Spatio-temporal action localization is an active research topic in computer vision. The goal is to localize and classify actions in both space and time. Majority of the existing approaches are supervised and can be categorized as either single frame or multi-frame. Most of the recent methods [13, 14, 29, 33, 37, 45] fall in the former category. These schemes extend object detection frameworks [12, 30] to first generate region proposals and then classify them into actions at the frame level using a two-stream variant which processes both RGB and flow data separately. The backbone of these networks is generally a 3D CNN (e.g., C3D [41] or I3D [2]). The resulting per-frame detections are then linked using dynamic programming [13, 37] or tracking [45]. Some recent approaches, however, aim to jointly estimate localization and classification over several frames [21] or use 3D convolutions to predict short tubes [18]. There has been recent attempts to learn without labels as well [38], where unlabeled data is used to automatically generate labels and train the classifiers.

2.2 Domain Adaptation

Domain adaptation aims to bridge the gap between the source and target data collected from different domains. Recent domain adaptation techniques under both semi-supervised and unsupervised settings have been introduced for image applications [7]. The majority of these methods have been dedicated to applications involving image classification [10, 15, 25, 27, 32, 36, 43], object detection [6, 34], and semantic segmentation [35, 42, 50]. Several unsupervised domain adaptation approaches use adversarial learning on the intermediate feature representations to align the feature distribution between the two domains [1, 6, 10, 44].

In contrast, much less attention has been paid to adapt models for video analysis between domains, and especially for activity understanding. While some progress has been made in this field recently, it is limited to whole-clip action recognition [3, 19, 28] and action segmentation [4, 5]. One reason can be attributed to the fact that a well-organized setting to develop and benchmark the performance of domain adaptation algorithms for spatio-temporal action localization does not exist. Existing datasets, e.g., CMU [22], MSR Actions [49], UCF Sports [31], and JHMDB [20] provide spatio-temporal annotations but only for a small number of short video clips. The DALY [46], UCF-101 [39] and AVA [14] datasets address some of the aforementioned limitations by providing large-scale annotatios for spatio-temporal action localization. However, these datasets have very few overlapping categories amongst them. Additionally, the annotation setting of AVA is different from the other datasets, making it difficult to evaluate domain adaptation algorithms.

To the best of our knowledge, this work is one of the first to adapt spatio-temporal action localization under the unsupervised setting. To evaluate the new task, we propose a new experimental setup and evaluation protocol for future development.

3 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 1.

3.1 Action Localization Model

Our model is based on the Faster R-CNN [30] for end-to-end localization and classification of actions [29]. To model the temporal context, the I3D model [2] is incorporated. The I3D model takes a video V of length T frames and generates the corresponding temporal feature representation using feature extractors TF_1 and TF_2 (see Fig. 1). Here, TF_1 extracts and temporally flattens the image level features from the fused mixed_4f layer of I3D, which has a spatial and temporal stride of 16 pixels and 4 frames, respectively. 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 (see Fig. 1) on the keyframe K as the input for the region proposal network (RPN). We note K is also the middle frame of an input clip to I3D. The proposals are generated using the conv4 block of ResNet [16]. As the spatial stride of the conv4 block is also 16 pixels, we directly use the actor RPN proposals on $TF_1(V)$ and perform ROI pooling to obtain a fixed size representation of 7×7×832. This feature representation is then passed through TF_2, which uses the remaining I3D layers up to mixed_5c and an average pooling layer to output an instance level feature vector of size 1×1×1024. This feature is used to learn an 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 [30].

3.2 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 [34], 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 [24]. 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) [11] 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 [6], 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 although the actions in our experiments vary in terms of temporal dynamics across datasets, the datasets are not explicitly designed to capture that notion. This characteristic is also shown in [3] 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:

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

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 1). 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:

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

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

3.3 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:

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

For the adaptation task $s \to t$, 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.

4 Experiments and Analysis

We propose new experimental settings for developing and evaluating domain adaptation algorithms for spatio-temporal action localization as there is no existing benchmarks. We first focus on the scenario of adapting from a smaller annotated domain to a much larger and diverse dataset. and then provide some additional experiments and ablation studies to highlight the effect of the different adaptation modules used in the proposed approach.

The proposed approach is evaluated on three widely used benchmark datasets for action localization, namely UCF-101 [39], JHMDB [20], and UCF-Sports [31]. These datasets are gathered from different sources (suitable for domain adaptation evaluation) and are also commonly used for adaptation of action recognition [3, 28]. Additionally, their suitability for our experiments is further shown through the results where 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 domain shift between the datasets.

4.1 Datasets and Metrics

UCF Sports. UCF Sports [31] contains various trimmed sports actions collected from broadcast television channels. It includes 10 actions, out of which we use 4 for our experiments

Table 1: Frame and video mAP results for adaptation from UCF-Sports to UCF-101 *with* (left) and *without* (right) background frames.

Method	T img	T ins	S img	Div ing	Glf Swg	Hrs Rdg	Skt Bdg	Fr. mAP	Vid. mAP
I3D+RPN				7.1	56.3	30.7	39.5	33.4	57.1
Ours	✓			12.2	64.6	40.0	41.9	39.7	61.0
		✓		12.2	64.9	40.7	42.3	40.0	61.6
			✓	13.9	64.9	51.5	51.8	45.5	68.9
	✓		✓	14.9	64.1	56.2	54.9	47.5	70.6
		✓	✓	13.0	**68.8**	51.3	50.6	45.9	67.1
	✓	✓	✓	**17.9**	63.3	**63.0**	**55.0**	**49.8**	**73.6**
Oracle				90.4	97.6	94.2	91.0	93.3	99.0

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

which are common with UCF-101: *Diving*, *Golf-Swing*, *Horse-Riding*, *Skate-Boarding*. We use the train/test split as suggested in [23].

UCF-101. This action localization dataset [39] is purely collected from YouTube and contains more than 13000 videos and 101 classes. We use 4 classes that are common with UCF-Sports from a 24-class subset with spatio-temporal annotations provided by [37]. We conduct experiments on the official split 1 as is standard.

JHMDB. JHMDB [20] is collected from sources ranging from digitized movies to YouTube, and consists of 928 trimmed clips over 21 classes. Each action class consists of varying number of clips (up to 40 frames). We use the official split 1 for our experiments, and only use 3 classes which are common with UCF-101: *Shoot Ball*, *Golf*, *Walk*.

Metrics. We use the standard evaluation protocols and report intersection-over-union (IoU) performance using mean average precision (mAP) on both frame-level and video-level using an IOU threshold of 0.5. For frame-level IoU, the PASCAL VOC challenge protocol [9] is used. For video-level IoU, we follow [29] to form action tubes by linking frame-level detections using dynamic programming and calculate 3D IoUs.

Implementation Details. We implement the proposed algorithm in Pytorch. ResNet-50 and I3D networks are initialized with pre-trained models based on ImageNet [8] and Kinetics [48] datasets, respectively. For the proposed adaptation method, we first pre-train the action localization network using the source domain clips, and then fine-tune the network for adaptation. We use different adaptation networks for each of the adaptation modules. More experimental details and results can be found in the supplementary material. The source code and trained models will be made available to the public.

4.2 Adaptation to Large-Scale Data

Adapting a model learned from a small dataset to a large unlabeled domain is more challenging than typical settings in the literature, and is also more useful as annotating large amount of data is infeasible for spatio-temporal action localization. In this work, the target domain is UCF-101, and the sources are UCF-Sports and JHMDB sets. Note that the source datasets are much smaller in size and less diverse than the target one, details of which can be found in the supplementary material.

UCF-Sports → UCF-101. We conduct experiments on the common classes from both the datasets and show the results in Table 1. Since UCF-101 is an untrimmed set, we show results both with and without considering background frames, the latter also requiring temporal localization. Note that we do not use background frames during training, making the latter setting extremely challenging.

Although UCF-Sports is also a sport-oriented dataset like UCF-101, a significant performance gap between the baseline and oracle results is observed, suggesting significant domain

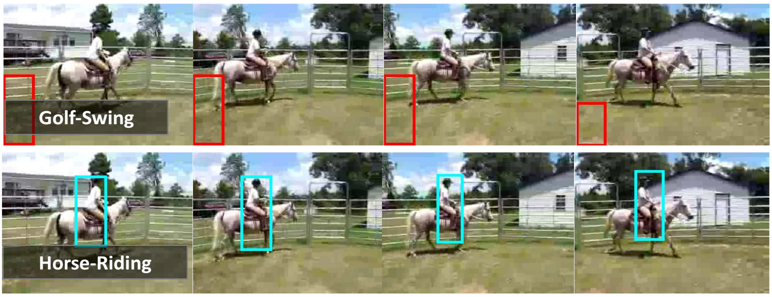


Figure 2: Example clip of *Horse-Riding* action from UCF-101, with baseline model (red) and our best adapted model (cyan) shown and predicted label overlaid.

shift and difficulty for adaptation. For aligning temporal features, both image level as well as instance level adaptation yield similar and considerable improvement over the baseline of 6.3% and 6.6% for frame-mAP, and 3.9% and 4.5% for video-mAP respectively, as shown in Table 1 (left). However, alignment of spatial features, which is responsible for adapting the actor proposals yields 12.1% (frame-mAP) and 11.8% (video-mAP) improvement. 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 16.4% (frame-mAP) and 16.5% (video-mAP). Note that the improvement also generalizes well across different categories, suggesting the effectiveness of the proposed framework in reducing domain discrepancy. Figure 2 shows an example from the UCF-101 dataset, where the baseline model (without adaptation) fails to detect *Horse-Riding* action while the adapted model correctly localizes and classifies the action.

When the background frames are considered in Table 1 (right), we observe similar trends after adding the adaptation modules. The absolute numbers however, are lower, indicating the presence of false positives on the background frames. Although this does not drastically affect the frame-mAP, the video-mAP is considerably affected. This also suggests that an explicit mechanism should be developed to handle background frames during adaptation, especially for temporal localization.

JHMDB → UCF-101. While UCF-101 is comprised of activities in the sports domain, JHMDB consists of videos from everyday activities (some sport-related sequences are also included). Note that from the set of common classes, *walk* action in JHMDB is visually very different from the *walking with dog* action in UCF-101. However, we still incorporate the *walk* action in our experiments to increase the number of common classes. We show the results in Table 2 (left) without considering background frames, but still consider temporal

Table 2: Frame and video mAP results for adaptation from JHMDB to UCF-101 (left) and UCF-101 to JHMDB (right).

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

Method	T img	T ins	S img	Golf Swg	Bskt Ball	Walk	Fr. mAP	Vid. mAP
I3D+RPN				86.6	27.2	38.4	50.7	60.7
Ours	✓			88.5	36.3	42.9	55.9	68.7
		✓		87.1	35.4	42.9	55.1	71.7
			✓	94.9	35.6	55.4	62.0	71.0
	✓	✓	✓	**96.4**	**46.7**	**57.9**	**67.0**	**75.4**
Oracle				96.6	70.5	87.0	84.7	93.4

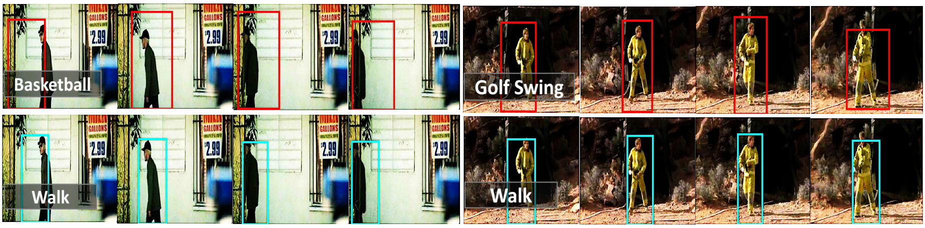


Figure 3: Example clips of *Walk* action from JHMDB, with baseline model (red) and our best adapted model (cyan) shown and predicted label overlaid.

localization for *Walk* action as it has few sequences containing multiple action instances. The performance gap between baseline and oracle results suggests a significant domain shift. A considerable improvement is obtained by adaptation of either spatial or temporal features for both frame and video mAPs, and their combination leads to the best performance gain of 14.2% (frame-mAP) and 17.7% (video-mAP) over the baseline.

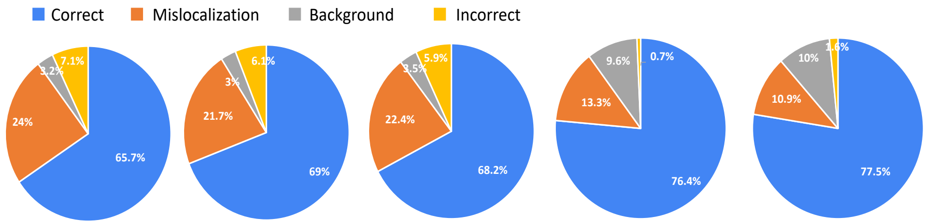
We also observe that differently from [6], instance level feature alignment combined individually with image level spatial feature adaptation does not yield much improvement and performs worse in some cases. This is because [6] focuses only on spatial feature alignment from the same backbone at image 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 (left) and Table 1, 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. The results also suggest that both spatial context and actor/action dynamics are equally important for action classification, as both types of temporal features are required for best performance and yield similar improvement over the baseline.

4.3 Additional Experiments and Analysis

In this section, we study the effect of adapting from a larger annotated domain to a much smaller dataset. We discuss the empirical results and analyze the effects of the individual adaptation modules by studying the classification and localization errors of the different models.

UCF101 → JHMDB. We use UCF-101 and JHMDB as the source and target datasets respectively, with the same set of common classes as before. Even when adapting from a much larger database to a smaller dataset, we observe similar trends in Table 2 (right) as before, with the significant gap between the baseline and oracle results suggesting that even having large amount of annotations does not help much in the case of domain shift. Note that the domain gap mainly comes from two classes: *Basketball* and *Walk*. The baseline performance for *Golf-Swing* is very close to the oracle results due to a significant amount of training labels in UCF-101. However, while *Walk* in UCF-101 contains about 20 times more samples than in JHMDB, the baseline performance is far from the oracle result because of the significant visual differences of the action between the datasets. Specifically, *Walk* action in UCF-101 is always accompanied with a dog in outdoor environments. Due to this, the model trained on UCF-101 (without adaptation) finds it hard to classify *Walk* action on JHMDB, as shown in Figure 3. Adaptation helps alleviate visual differences and improves localization performance. Examples of visual differences can be found in the supplementary material.

Error Analysis on Top Ranked Detections. To study the effect of the individual adapta-



(a) I3D+RPN (b) Ours (T. Img) (c) Ours (T. Ins) (d) Ours (Spatial) (e) Ours (all)

Figure 4: Error analysis of top ranked detections. Fraction of predictions that are correct, mislocalized, are confused with background or incorrectly predicted are shown.

tion modules, we analyze the classification and localization errors from the most confident detections of the model.

We use the UCF-101 → JHMDB experiment for analysis. Since the JHMDB dataset is a small set, we select the top 1000 predictions based on the corresponding predicted confidence score by the baseline model (i.e., I3D+RPN) and our models with various adaptation modules. Motivated by [17, 26], we categorize the detections into four error types: i) **correct**: the detection has an overlap $\in [0.5, 1]$ with the ground-truth; ii) **mis-localized**: the detection has an overlap $\in [0.3, 0.5)$; iii) **background**: the detection has an overlap $\in [0.0, 0.3)$, which means it takes a background as a false positive; and iv) **incorrect**: the detection has a different class than the ground truth. The first three errors are related to the localization error given the detected class is correct, while the last error measures the incorrect classifications. In addition, we also analyze the errors of the bottom 1000 detections in the supplementary material, with the goal to understand the extent of the adaptation effect.

Figure 4 shows that temporal feature alignment at both image and instance level improves the correct detections as well as reduces the mislocalized error. It also reduces the incorrect classifications. The spatial feature alignment, in addition to increasing the correct detections, also considerably reduces the mislocalized error. This can be attributed to that spatial features directly improve the RPN, which is responsible for actor proposal generation. It also reduces the incorrect classification. In addition, we note that there is an increase in the background error, which can be considered as duplicate detections as these are not incorrectly classified. However, as expected, our model with both spatial and temporal features aligned increases the correct detections the most and also gives the least mislocalization error.

5 Conclusion and Future Work

In this paper, we propose a new task and an end-to-end approach for unsupervised domain adaptation for spatio-temporal action localization. Our approach is built by extending the Faster R-CNN algorithm. In order to reduce domain shift, we design and integrate three domain adaptation modules at the image level (temporal and spatial) and instance level (temporal). Experimental results demonstrate that significant performance gain can be achieved when spatial and temporal features are adapted separately, or jointly for the most effective results.

Our experimental setup lacks in large number of overlapping classes and significant temporal variations between the datasets (mentioned in Section 3.2), which is a restriction of the problem space as there does not exist such datasets. Our work is an essential first step to stimulate the community to collectively build large-scale benchmark datasets and algorithms for domain adaptation of spatio-temporal action localization.

Acknowledgment. This work is supported in part by the NSF CAREER Grant #1149783 and Honda Research Institute USA.

References

[1] Konstantinos Bousmalis, George Trigeorgis, Nathan Silberman, Dilip Krishnan, and Dumitru Erhan. Domain Separation Networks. In *NIPS*, 2016. 3

[2] Joao Carreira and Andrew Zisserman. Quo Vadis, Action Recognition? A New Model and the Kinetics Dataset. In *CVPR*, 2017. 1, 3, 4

[3] 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, 4, 5, 6

[4] Min-Hung Chen, Baopu Li, Yingze Bao, and Ghassan AlRegib. Action segmentation with mixed temporal domain adaptation. In *WACV*, 2020. 2, 4

[5] Min-Hung Chen, Baopu Li, Yingze Bao, Ghassan AlRegib, and Zsolt Kira. Action segmentation with joint self-supervised temporal domain adaptation. In *CVPR*, 2020. 2, 4

[6] 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, 3, 5, 9

[7] Gabriela Csurka. *Domain Adaptation for Visual Applications: A Comprehensive Survey*. Springer, 2017. 3

[8] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, 2009. 7

[9] Mark Everingham, SM Ali Eslami, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The Pascal Visual Object Classes Challenge: A Retrospective. *IJCV*, 111(1), 2015. 7

[10] Yaroslav Ganin and Victor Lempitsky. Unsupervised Domain Adaptation by Backpropagation. In *ICML*, 2015. 2, 3

[11] 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. 5

[12] Ross Girshick. Fast R-CNN. In *ICCV*, 2015. 3

[13] Georgia Gkioxari and Jitendra Malik. Finding Action Tubes. In *CVPR*, 2015. 3

[14] Chunhui Gu, Chen Sun, David Ross, Carl Vondrick, Caroline Pantofaru, Yeqing Li, Sudheendra Vijayanarasimhan, George Toderici, Susanna Ricco, Rahul Sukthankar, Cordelia Schmid, and Jitendra Malik. AVA: A Video Dataset of Spatio-temporally Localized Atomic Visual Actions. In *CVPR*, 2018. 1, 3, 4

[15] Philip Haeusser, Thomas Frerix, Alexander Mordvintsev, and Daniel Cremers. Associative Domain Adaptation. In *ICCV*, 2017. 3

[16] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep Residual Learning for Image Recognition. In *CVPR*, 2016. 4

[17] Derek Hoiem, Yodsawalai Chodpathumwan, and Qieyun Dai. Diagnosing Error in Object Detectors. In *ECCV*, 2012. 10

[18] Rui Hou, Chen Chen, and Mubarak Shah. Tube Convolutional Neural Network (T-CNN) for Action Detection in Videos. In *ICCV*, 2017. 3

[19] Arshad Jamal, Vinay P Namboodiri, Dipti Deodhare, and KS Venkatesh. Deep Domain Adaptation in Action Space. In *BMVC*, 2018. 2, 4

[20] Hueihan Jhuang, Juergen Gall, Silvia Zuffi, Cordelia Schmid, and Michael J Black. Towards Understanding Action Recognition. In *ICCV*, 2013. 4, 6, 7

[21] Vicky Kalogeiton, Philippe Weinzaepfel, Vittorio Ferrari, and Cordelia Schmid. Action Tubelet Detector for Spatio-Temporal Action Localization. In *ICCV*, 2017. 3

[22] Yan Ke, Rahul Sukthankar, Martial Hebert, et al. Efficient Visual Event Detection Using Volumetric Features. In *ICCV*, 2005. 4

[23] Tian Lan, Yang Wang, and Greg Mori. Discriminative Figure-centric Models for Joint Action Localization and Recognition. In *ICCV*, 2011. 7

[24] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal Loss for Dense Object Detection. In *ICCV*, 2017. 5

[25] Mingsheng Long, Yue Cao, Jianmin Wang, and Michael Jordan. Learning Transferable Features with Deep Adaptation Networks. In *ICML*, 2015. 2, 3

[26] Pascal Mettes and Cees GM Snoek. Pointly-supervised action localization. *IJCV*, 127 (3):263–281, 2019. 10

[27] Saeid Motiian, Marco Piccirilli, Donald A Adjeroh, and Gianfranco Doretto. Unified Deep Supervised Domain Adaptation and Generalization. In *ICCV*, 2017. 3

[28] Boxiao Pan, Zhangjie Cao, Ehsan Adeli, and Juan Carlos Niebles. Adversarial cross-domain action recognition with co-attention. *arXiv preprint arXiv:1912.10405*, 2019. 2, 4, 6

[29] Xiaojiang Peng and Cordelia Schmid. Multi-region Two-stream R-CNN for Action Detection. In *ECCV*, 2016. 3, 4, 7

[30] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. In *NIPS*, 2015. 2, 3, 4

[31] Mikel Rodriguez, Javed Ahmed, and Mubarak Shah. Action MACH a Spatio-temporal Maximum Average Correlation Height filter for action recognition. In *CVPR*, 2008. 4, 6

[32] Kate Saenko, Brian Kulis, Mario Fritz, and Trevor Darrell. Adapting Visual Category Models to New Domains. In *ECCV*, 2010. 2, 3

[33] Suman Saha, Gurkirt Singh, Michael Sapienza, Philip HS Torr, and Fabio Cuzzolin. Deep Learning for Detecting Multiple Space-Time Action Tubes in Videoss. *arXiv preprint arXiv:1608.01529*, 2016. 3

[34] Kuniaki Saito, Yoshitaka Ushiku, Tatsuya Harada, and Kate Saenko. Strong-Weak Distribution Alignment for Adaptive Object Detection. In *CVPR*, 2019. 2, 3, 5

[35] Swami Sankaranarayanan, Yogesh Balaji, Arpit Jain, Ser Nam Lim, and Rama Chellappa. Learning from Synthetic Data: Addressing Domain Shift for Semantic Segmentation. In *CVPR*, 2018. 2, 3

[36] Ozan Sener, Hyun Oh Song, Ashutosh Saxena, and Silvio Savarese. Learning Transferrable Representations for Unsupervised Domain Adaptation. In *NIPS*, 2016. 3

[37] Gurkirt Singh, Suman Saha, Michael Sapienza, Philip HS Torr, and Fabio Cuzzolin. Online Real-time Multiple Spatiotemporal Action Localisation and Prediction. In *ICCV*, 2017. 3, 7

[38] Khurram Soomro and Mubarak Shah. Unsupervised action discovery and localization in videos. In *ICCV*, 2017. 3

[39] 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. 1, 4, 6, 7

[40] Chen Sun, Abhinav Shrivastava, Carl Vondrick, Kevin Murphy, Rahul Sukthankar, and Cordelia Schmid. Actor-centric Relation Network. In *ECCV*, 2018. 2

[41] Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. Learning Spatiotemporal Features with 3D Convolutional Networks. In *ICCV*, 2015. 1, 3

[42] Yi-Hsuan Tsai, Wei-Chih Hung, Samuel Schulter, Kihyuk Sohn, Ming-Hsuan Yang, and Manmohan Chandraker. Learning to Adapt Structured Output Space for Semantic Segmentation. In *CVPR*, 2018. 2, 3

[43] Eric Tzeng, Judy Hoffman, Ning Zhang, Kate Saenko, and Trevor Darrell. Deep Domain Confusion: Maximizing for Domain Invariance. *arXiv preprint arXiv:1412.3474*, 2014. 2, 3

[44] Eric Tzeng, Judy Hoffman, Kate Saenko, and Trevor Darrell. Adversarial Discriminative Domain Adaptation. In *CVPR*, 2017. 3

[45] Philippe Weinzaepfel, Zaid Harchaoui, and Cordelia Schmid. Learning to Track for Spatio-temporal Action Localization. In *ICCV*, 2015. 3

[46] Philippe Weinzaepfel, Xavier Martin, and Cordelia Schmid. Human Action Localization with Sparse Spatial Supervision. *arXiv preprint arXiv:1605.05197*, 2016. 4

[47] Philippe Weinzaepfel, Xavier Martin, and Cordelia Schmid. Towards Weakly-supervised Action Localization. *arXiv preprint arXiv:1605.05197*, 2, 2016. 1

[48] Karen Simonyan Brian Zhang Chloe Hillier Sudheendra Vijayanarasimhan Fabio Viola Tim Green Trevor Back Paul Natsev Mustafa Suleyman Andrew Zisserman Will Kay, Joao Carreira. The Kinetics Human Action Video Dataset. *arXiv preprint arXiv:1705.06950*, 2017. 7

[49] Junsong Yuan, Zicheng Liu, and Ying Wu. Discriminative Subvolume Search for Efficient Action Detection. In *CVPR*, 2009. 4

[50] Yang Zou, Zhiding Yu, B.V.K. Vijaya Kumar, and Jinsong Wang. Domain Adaptation for Semantic Segmentation via Class-Balanced Self-Training. In *ECCV*, 2018. 2, 3