Exploiting Instance-based Mixed Sampling via Auxiliary Source Domain Supervision for Domain-adaptive Action Detection

Yifan Lu, ETH Zurich
Gurkirt Singh, ETH Zurich
Suman Saha, ETH Zurich
Luc Van Gool, ETH Zurich, KU Leuven

Abstract

We propose a novel domain adaptive action detection approach and a new adaptation protocol that leverages the recent advancements in image-level unsupervised domain adaptation (UDA) techniques and handle vagaries of instance-level video data. Self-training combined with cross-domain mixed sampling has shown remarkable performance gain in semantic segmentation in UDA (unsupervised domain adaptation) context. Motivated by this fact, we propose an approach for human action detection in videos that transfers knowledge from the source domain (annotated dataset) to the target domain (unannotated dataset) using mixed sampling and pseudo-label-based self-training. The existing UDA techniques follow a ClassMix algorithm for semantic segmentation. However, simply adopting ClassMix for action detection does not work, mainly because these are two entirely different problems, i.e., pixel-label classification vs. instance-label detection. To tackle this, we propose a novel action instance mixed sampling technique that combines information across domains based on action instances instead of action classes. Moreover, we propose a new UDA training protocol that addresses the long-tail sample distribution and domain shift problem by using supervision from an auxiliary source domain (ASD). For the ASD, we propose a new action detection dataset with dense frame-level annotations. We name our proposed framework as domain-adaptive action instance mixing (DA-AIM). We demonstrate that DA-AIM consistently outperforms prior works on challenging domain adaptation benchmarks. The source code is available at `https://github.com/wwwfan628/DA-AIM`.

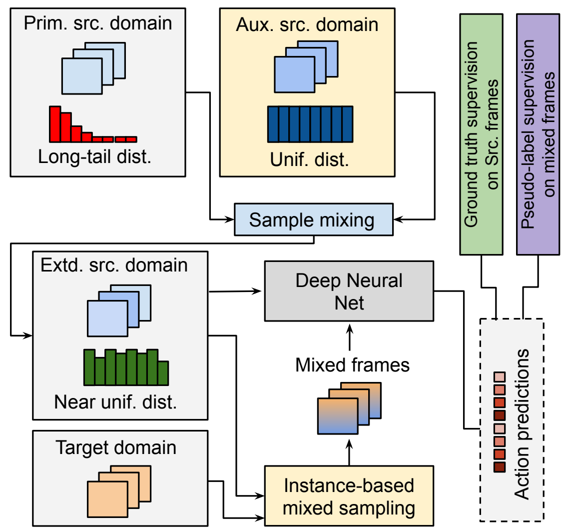

Figure 1: The above diagram illustrates the two main contributions of this work. Firstly, We propose a novel instance-based cross-domain mixed sampling technique designed explicitly for video-based action detection. Unlike the prior UDA method [74], which follows a class-based mixed sampling to generate augmented mixed images, our mixed sampling algorithm randomly samples image patches based on the number of action instances present in the source frames. The output is a set of mixed frames containing instances of the source and target domains. Secondly, we propose to mix auxiliary source domain samples with the primary source domain to create a new extended source domain. This is done to address various problems such as long-tail distribution of the primary source domain, large variability in action instances across domains.

1. Introduction

Over the past few years, we have witnessed tremendous progress in vision-based action detection[36, 55, 83, 17, 39, 2, 66, 67, 59, 4, 29, 33, 60, 79, 88, 54]. This success is largely attributed to the deep neural networks, which demonstrates superior performance in several computer vision tasks. However, these networks require expensive ground truth annotations to be trained appropriately under a supervised-learning setup. Particularly, for action detection, it is highly time-consuming and labor-intensive to generate such a large amount of annotated data [38, 25, 40, 71]. The main reason is that ground truth labels for both action categories and instances are required, i.e. all the action instances in a video frame need to be spatially localized using bounding boxes and these boxes are to be labeled with their respective action categories. As the video duration, the number of videos and action instances increase, the annotation cost rises rapidly, making the labeling process highly impractical and expensive. One standard approach to circumvent this issue is to rely on unsupervised domain adaptation (UDA) [74, 46, 47, 21, 22, 42, 41, 62] in which knowledge transfer is performed by adapting the network trained with the source domain to the target domain. The source domain refers to either synthetic data [56, 57] or publicly available real data [57, 38] for which the ground truth annotations are available. The target domain refers to real data for which ground truths are not accessible.

Prior works [16, 90, 45, 53, 69, 87, 82, 9, 52, 15, 7] mostly focus on domain-adaptive (DA) action recognition which is a simpler problem than DA action detection as the former requires only to solve the action classification without considering the much harder instance localization problem. Agarwal *et al*. [1] propose a DA action detection approach in which domain alignments of spatial and temporal features are performed using GRLs [21]. They introduce two UDA benchmarks which are limited to only three/four sports actions. Since there is no standard UDA benchmark available for action detection, they rely on the sports-related action classes, which are common across different datasets (or domains). Moreover, the datasets used in [1] have low video resolution and are outdated.

In this work, we propose a generic UDA framework that is not limited to certain action categories and can be used for a larger set of action classes, *e.g*. AVA [25]. First, we consider the train set from the AVA-Kinetics [38] dataset as our primary source domain. Since AVA-Kinetics is a large-scale and diversified action detection dataset from YouTube videos, using it as the source domain would allow the model to learn meaningful spatiotemporal representation and better adaption to the target domain. However, it imposes two main challenges. Firstly, AVA-Kinetics has a long-tailed label distribution which biases the model towards certain action categories, resulting in a poor adaptation of under-represented classes. Secondly, there is a large variability in actions (belonging to same action classes) across domains due to factors like differences in capturing devices, backgrounds, temporal motion patterns, appearance. To tackle these problems, we propose to supervise the network using labeled training samples from an auxiliary source domain (ASD) (Fig. 1). ASD alleviates the aforementioned problems by: (a) injecting training samples of under-represented or missing classes into the source domain, and (b) recreating the action scenes to resemble the target domain scenes. For ASD, we create a new action detection dataset with dense ground truth annotations.

We empirically found that the GRL-based approach (similar to [1]) does not show any noticeable improvements in either of our UDA settings (§4.6). Recently, Tranheden *et al*. [74] proposed a UDA method for semantic segmentation, which exhibits superior performance in semantic segmentation task. Their method generates augmented training images following a cross-domain mixed sampling (CDMS) technique. CDMS is suitable for pixel-level prediction (or segmentation) tasks. However, for instance-level (or bounding-box) prediction like action detection, CDMS fails to generate meaningful training samples since these two are entirely different problems, i.e., pixel-label classification vs. instance-label detection. To tackle this issue, we propose a novel action-instance-based mixed sampling technique that combines information across domains based on action instances present in the source domain. For source-to-target knowledge transfer, we adapt the Mean Teacher based self-training [73]. We name our proposed UDA framework as DA-AIM (**d**omain-**a**daptive **a**ction **i**nstance **m**ixing) (Fig. 1). We are the first to propose a DA action detection framework based on cross-domain mixed sampling and self-training. We implement and compare with three state-of-the-art approaches and achieve best results on different UDA benchmarks. We will publish our code and release two new (In-house) dataset used in this work.

2. Related Works

Action Detection is a more challenging problem [19, 88, 67] compared to action recognition [65, 6] problem due to the additional requirement for localisation of actions in a large spatial-temporal search space. Supervised action detection methods [79, 67, 33, 43, 88, 54] has made large strides thanks to large scale datasets like UCF24 [71], AVA [25] and MultiSports [40]. Most of current approaches follow key-frame based approach popularised by SlowFast [19]. There has been more sophisticated approaches, *e.g*. based on actor-context modelling [10, 54], on long-term feature banks [80, 72], and on transformer heads [88, 44]. We will make use of key-frame based SlowFast [19] network as our default action detector because of it's simplicity, competitive performance, and reproducible code base provided on pySlowFast [18], which can be easily extended to include transformer architectures, such as MViTv2 [44]. Apart from fully-supervised methods, there has also been works on pointly-supervised [49] or semi-supervised [36] settings.

Unsupervised Domain Adaptation. The effectiveness of UDA techniques has been studied in different vision tasks including image classification, object detection, semantic segmentation, action recognition and detection. [21, 26, 46, 51, 58, 64, 76] propose methods to tackle DA image classification. DA object detection is studied by [61, 12]. Most DA semantic segmentation methods are based on either adversarial training or self-training. Adversarial training follows a GAN framework [22, 24] to aligns the source and target domains feature distributions at in-

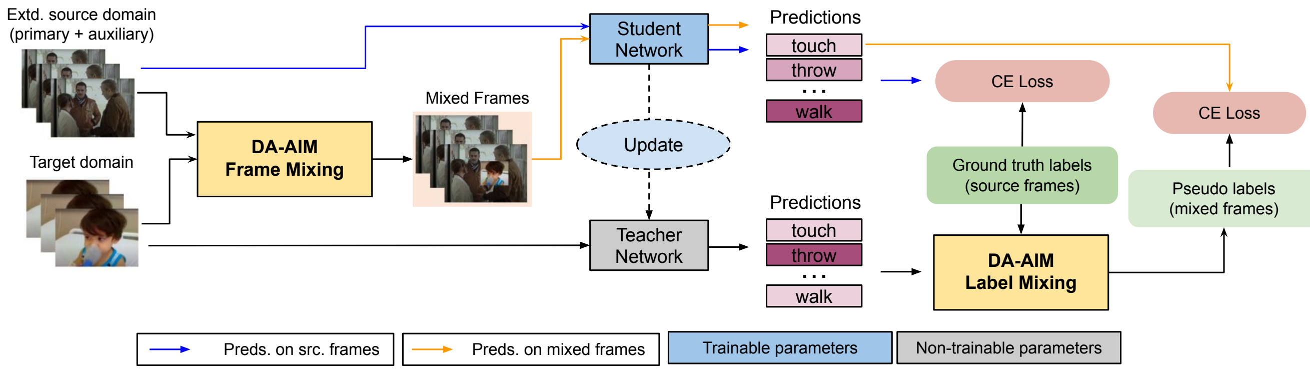

Figure 2: Overview of the proposed DA-AIM framework. The basic building blocks of DA-AIM are (a) training sample mixing, (b) frame mixing, (c) label mixing, and (d) self-training. (a) We first generate an extended (extd.) source domain by mixing training examples of the primary and auxiliary source domains. (b) Next, the frame mixing module generates augmented video frames (or mixed frames) by mixing action instances of the source frame with the target frames. During mixing, spatial and temporal information are considered due to the inherent spatiotemporal nature of actions. The source and mixed frames are then fed to a deep neural network (called the student network). The student network is optimized with action classification losses. Ground truth labels are used to penalize wrong predictions on source frames, and pseudo-labels are used to provide supervision on the mixed frames. (c) Since the mixed frames contain image patches from both source and target domains, the label mixing module generates pseudo-labels based on the inputs from ground truth labels and the teacher network predictions. (d) The teacher network is initialized with the parameters of the student network. Its parameters are non-trainable and updated as the exponential moving average of the parameters of the student network.

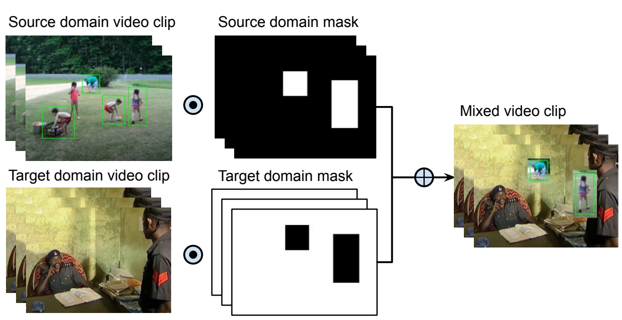

Figure 3: The above diagram illustrates the proposed action-instance-based (AIM) cross-domain mixed sampling.

put [23, 27], output [75, 77], patch [13], or feature level [28, 75]. In self-training, the supervision for target domain comes from pseudo-labels [37] which can be computed offline [63, 84, 91, 92] or online [74, 78, 30]. Consistency regularization [68, 73] or label prototypes [86] formulated on CDMS [74, 89] or data augmentation [3, 14, 48] are used to address training instabilities. In this work, we use online self-training and consistency regularization based on CDMS. Unlike [74, 78, 30, 89], which tackle image-based DA semantic segmentation, we address a video-based DA action detection. [74, 78, 30, 89] use semantic class based CDMS which show poor results in action detection. We propose a novel action instance-based CDMS specifically designed to facilitate video-based action detection.

Mixed sampling. Within-domain and cross-domain mixing have been widely studied for image-based problems [85, 5, 20, 11, 74]. Despite the effectiveness of these algorithms on the image-based problems, mixed sampling has not been studied for video understating tasks. We are the first to propose a novel instance-based CDMS for video action detection.

DA action recognition and detection. There are several methods proposed for single-modal (RGB) [7, 15, 31, 53] or multi-modal (RGB, flow) [52, 70, 35] DA action recognition. [9, 8] propose methods for DA action segmentation. We found only one work [1] that addresses DA action detection using GRL-based adversarial training. [1] propose two UDA benchmarks limited to sports actions. This work has two major limitations. Their proposed UDA setup does not address the long-tail and large variability problems (see §1), and the proposed GRL-based adaptation shows a poor generalization in a UDA setting where the source domain has a long-tailed distribution, and the class-specific actions have large variations across domains. In contrast, our approach addresses these limitations by proposing a new UDA framework in which these problems are alleviated using an auxiliary source domain and a more effective instance-based CDMS and pseudo-labeling techniques.

3. Methodology

In this section, we will introduce the proposed DA-AIM framework. DA-AIM (Fig. 2) can be decomposed into two main steps, namely action-instance-based CDMS (cross-

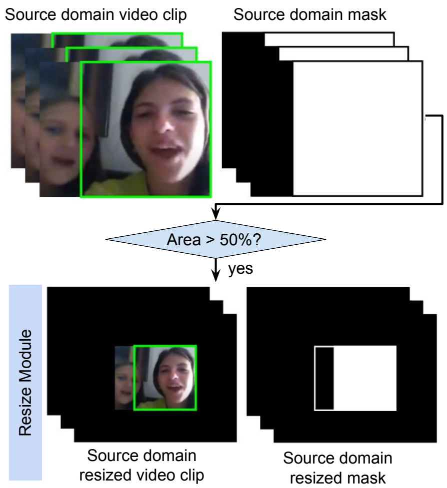

Figure 4: Frames need to be downscaled if source domain's action instance area takes up more than half of the entire frame area. Bounding boxes and the mask are correspondingly adjusted to fit into the resized frames. White represents 1 and black represents 0.

domain mixed sampling) and self-training.

3.1. Action-instance-based CDMS

Fig. 3 illustrates the proposed Action-Instance-based cross-domain Mixed sampling (AIM). Given video clips from the source and target domains, and the corresponding ground truth annotations (i.e., the bounding boxes and their class labels) of the source frames, we randomly sample half of the action instances from the source frame. Since the bounding boxes are created only for the key-frames located in the middle of the clips, considering fast moving actions such as running, we expand each bounding box by 20% when creating the source domain mask. The 3D source domain mask $M \in \{0,1\}^{T\times W\times H}$ is constructed by replicating the 2D mask of the key-frame $M_k \in \{0,1\}^{W\times H}$ in the temporal axis, where M_k is a binary matrix containing 1 for regions where the selected source instance is present and 0 otherwise. only at the places Our mixed video clips can be obtained through:

$$x_M = M \odot x_S + (1-M) \odot x_T, \tag{1}$$

where $x_M, x_S, x_T \in \mathbb{R}^{T\times W\times H}$ represent the mixed video clip, input source and target video clips respectively.

Note that often the videos from the source domain (Kinetics) contain action instances which take most of the image regions, i.e., the instance bounding box has a large spatial overlap with the entire image region. If such a video clip is used for CDMS without action instance resizing, it might lead to imbalance in information across domains. That is, the mixed frames might mostly be occupied with source domain action regions, and there would be too little target regions visible. To address this imbalance issue, we propose to first resize the large action instance in the source frame and then paste it onto the target frame (Fig. 4). More specifically, if the source action instance area takes up more than half of the entire area of the mixed frame, we will downscale the source domain frames by factor 0.5 before mixing. Bounding boxes and the mask are correspondingly adjusted to align with the resized video clip. Given bounding boxes as a tuple (x_1, y_1, x_2, y_2), where (x_1, y_1) corresponds to the top left corner and $(x2, y2)$ corresponds to the bottom right corner, and H, W are the height and width of the video frames, coordinates of bounding boxes after resizing (x'_1, y'_1, x'_2, y'_2) can be expressed as:

$$x'_1 = [\tfrac{W}{4}] + [\tfrac{x_1}{2}], y'_1 = [\tfrac{H}{4}] + [\tfrac{y_1}{2}] \tag{2}$$

$$x'_2 = [\tfrac{W}{4}] + [\tfrac{x_2}{2}], y'_2 = [\tfrac{H}{4}] + [\tfrac{y_2}{2}] \tag{3}$$

where $[\cdot]$ indicates the rounding function to find the nearest integer. The empty borders after resizing are filled with 0. Since target domain action instances might be covered by source domain action instances after mixing, bounding boxes and labels can not be simply concatenated. Due to the possibility of lacking important information to identify the action, if a bounding box from target domain overlaps with any pasted bounding boxes from source domain more than 40% of its area, it is discarded and not included in the loss computation.

3.2. Self-training for UDA

We follow Mean Teacher [73] method for self-training. More formally, the weights of the student network at training step t is defined as θ_t and the weights of the teacher network as θ'_t. At each training step t, weights of the teacher network θ'_t are updated according to Eq.4

$$\theta'_t = \alpha\theta'_{t-1} + (1-\alpha)\theta_t, \tag{4}$$

where α is a smoothing coefficient. In this work, we focus on exclusive actions, which means those actions can not be done at the same time. Consequently, the problem is a single-label classification problem. Hence, the pseudo-label of an action instance is the action class obtaining highest confidence score from the current teacher model.

3.3. Training Optimization

In DA-AIM, the student network parameters θ are trained by minimizing the following loss:

$$\arg\min_\theta \mathcal{L}(\theta) = \arg\min_\theta \mathbb{E}\Big[H\big(f_\theta(X_S, B_S), Y_S\big) + \lambda H\big(f_\theta(X_M, B_M), Y_M\big)\Big] \tag{5}$$

where the expectation is over batches of random variables X_S, B_S, Y_S, X_M, B_M and Y_M. Video clips in X_S are sampled uniformly from the source domain distribution, B_S and Y_S are the corresponding bounding boxes and labels. Furthermore, X_M is the new mixed video clips, B_M and Y_M are mixed bounding boxes and mixed labels. As we focus on exclusive actions and formulate the problem as single-label classification, we use cross-entropy loss H. λ is a hyper-parameter that decides how much the unsupervised part of the loss affects the overall training. Adapted from [74], we use an adaptive schedule for λ, where it is the proportion of instances in the whole unlabeled instances in the mixed video clip, of which the predictions have a confidence above a certain threshold.

3.4. DA-AIM Algorithm

The overall DA-AIM algorithm is summarized in Alg.1. The source-domain and target-domain datasets are referred to as $\mathcal{D}_S$ and $\mathcal{D}_T$. A batch of video clips, bounding boxes and labels, X_S, B_S and Y_S, is sampled from $\mathcal{D}_S$, and another batch of video clips, X_T from $\mathcal{D}_T$. $\widehat{B}_T$ represents bounding boxes of target domain video clips estimated by a pre-trained person detector. The unlabeled video clips X_T and bounding boxes $\widehat{B}_T$ are firstly fed to the teacher network $f_{\theta'}$, from which pseudo-labels $\widehat{Y}_T$ are obtained. Then, the augmented video clips X_M are created by mixing X_S and X_T.The pseudo-labels Y_M and bounding boxes B_M are correspondingly constructed by mixing Y_S, $\widehat{Y}_T$ and B_S, $\widehat{B}_T$. Start from here, the algorithm resembles a supervised learning approach and the process is repeated for a predetermined amount of iterations N.

Algorithm 1 DA-AIM Algorithm

Input: $\mathcal{D}_S$, $\mathcal{D}_T$ (source and target domains),
$f_{\theta'}$, f_θ, θ', θ (teacher, student nets and parameters),
d_p (pretrained person detector).
Output: f_θ (trained student net).
1: Initialize θ and θ' with MiT pretrained weights.
2: **for** $t \leftarrow 1, 2, ..., N$ **do**
3: Randomly sample mini-batches:
$(X_S, B_S, Y_S) \sim \mathcal{D}_S$, $(X_T) \sim \mathcal{D}_T$.
4: Compute bounding boxes: $\widehat{B}_T \leftarrow d_p(X_T)$.
5: Compute pseudo-labels:
$\widehat{Y}_T \leftarrow \mathrm{argmax}\big(f_{\theta'}(X_T, \widehat{B}_T)\big)$.
6: Generate mask M for mixed sampling.
7: Generate the mixed video X_M:
$X_M \leftarrow M \odot X_S + (1 - M) \odot X_T$.
8: Compute pseudo-labels Y_M, and
bounding boxes B_M for X_M:
$Y_M \leftarrow CDMS(Y_S, \widehat{Y}_T)$,
$B_M \leftarrow CDMS(B_S, \widehat{B}_T)$.
9: Forwards pass of student net f_θ:
$\widehat{Y}_S \leftarrow f_\theta(X_S, B_S)$, $\widehat{Y}_M \leftarrow f_\theta(X_M, B_M)$.
10: Compute cross-entropy losses:
$\ell = \mathcal{L}_\mathcal{S}(\widehat{Y}_S, Y_S) + \mathcal{L}_\mathcal{M}(\widehat{Y}_M, Y_M)$.
11: Compute gradient $\nabla_\theta \ell$ by backpropagation.
12: Optimize θ with stochastic gradient descent.
13: Update θ' using EMA (exponential moving average):
$\theta'_t = \alpha\theta'_{t-1} + (1 - \alpha)\theta_t$.
14: **end for**
15: **return** f_θ

4. Experiments and Results

4.1. Datasets

We use four datasets in our experiments: AVA [25], AVA-Kinetics [38], and two in-house labelled datasets, namely InHouseDataset-1 (IhD-1) and InHouseDataset-2 (IhD-2). In this section, we will briefly introduce them and describe how we use them to fit our experiment settings.

AVA [25]: is a dataset with atomic visual action and consists of 430 densely annotated 15-minute video clips with 80 visual actions. In total, roughly $1.62M$ action annotations are provided with the possibility that multiple annotations are made for one action instance, *i.e.* each action instance can perform multiple actions at the same time. We use version $v2.2$ of the annotation files throughout this work. In our experiments, we use AVA as one of the target domain when source domain is AVA-Kinetics

AVA-Kinetics [38]: annotates more than $200k$ videos from Kinetics-400 [34] dataset with AVA action classes and bounding boxes in one key-frame per 10 seconds long video. The main reason for using AVA-Kinetics as primary source domain is that it comes from YouTube and have high diversity compared to AVA which comes from movie clips.

In-House Datasets: we build two in-house datasets using two different scenes. One dataset is recorded in public place with different views of the scene while actors perform one or more actions from action list at a given time. The other dataset is recorded at a private facility to which access is permitted only for limited time and actors are different from former setup due to strict regulations. Going forward, former is named as In-House-Datasets-1 (*IhD-1*). and later (*IhD-2*). Both of these dataset contain three extra classes than AVA-Kinetics or AVA datatset, namely, 'carry-bag', 'drop-bag', and 'leave-bag-unattended'. We will make these datasets publicly available along with the training and evaluation code upon the acceptance of paper.

4.2. Dataset Sampling

We reduce large-scale datasets because of three reasons: (1) action classes needs to be matched to target domain class set, (2) for fair comparison with smaller datasets, (3) for the sake of time and resource consumption. To reduce the large-scale primary source domain dataset (AVA-Kinetics) we set 5000 as the maximum number of training samples for each action class. For the case of insufficient training sam-

Table 1: Overall statistics of datasets used in our experiments. Each sub dataset from large-scale datasets is constructed based on number of classes in target domain and $5k$ limit on number of samples of any of the given classes.

	AVA		**AVA-Kin**		**IhD-2**		**AVA-Kin**		**IhD-1**		**IhD-2**	
	Train	Val	Train	Val	Train	Val	Train	Val	Train	Val	Train	Val
Num.of classes	6		6		3		3		8		8	
Annotations	28,281	89,481	29,009	27,173	441	339	6,686	1,920	18,123	3,415	21,919	3,468
Unique boxes	28,281	89,481	29,009	27,173	441	339	6,686	1,920	18,114	3,415	21,843	3,442
Key-frames	14,248	48,741	15,453	19,205	441	339	6,115	1,779	16,881	2,695	13,974	2,753
Videos	235	64	15,453	19,205	12	7	6,115	1,779	28	7	34	8

Table 2: Ablation study: impact of each operation/module introduced in our DA-AIM framework for action detection. Specifically, impact of resizing (resize), pseudo-labeling (pLabel), and instance-mixing (iMix) modules is shown below.

Operations			AVA-Kinetics → AVA							AVA-Kinetics → IhD-2			
resize	pLabel	iMix	bend/bow	lie/sleep	run/jog	sit	stand	walk	mAP	touch	throw	take a photo	mAP
			33.66	54.82	56.82	**73.70**	**80.56**	**75.18**	62.46	34.12	32.91	27.42	31.48
	✓		30.74	56.20	55.09	73.53	**80.84**	72.44	61.47	29.97	28.10	29.82	29.30
		✓	33.07	55.87	60.69	72.51	79.43	73.05	62.44	33.00	29.79	29.26	30.68
✓		✓	**34.65**	56.50	60.19	70.80	79.17	74.75	62.68	32.27	32.48	30.37	31.71
	✓	✓	32.18	57.70	59.42	**74.03**	80.73	74.38	63.07	33.67	**38.06**	32.83	34.85
✓	✓	✓	33.79	**59.27**	**62.16**	71.67	79.90	**75.13**	**63.65**	**34.38**	35.65	**39.84**	**36.62**

ples owing to the class imbalance inside large scale datasets, the highest possible number of samples from that class will be taken. Regarding validation datasets, there is no restriction on the amount of samples, i.e. we use all the samples from those specific action classes mentioned above during the validation. Overall statistics of datasets used in our experiments is provided in Tab.1. Table contains the statistics of each subset according to the number of target domain classes used in our experiments. More details can be found in the supplementary material. Auxiliary source domain is introduced either when primary source domain does not contain one or more target domain classes or when primary source domain needs help from auxiliary source domain.

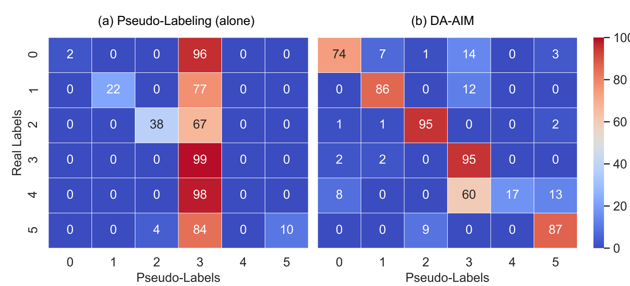

Figure 5: Confusion matrix of pseudo-labels at the end of training for AVA-Kinetics →AVA setup. (a, left) Pseudo-labeling alone for UDA (b, right) Pseudo-labeling within our DA-AIM.

4.3. Baseline and Implementation details

We implement SlowFast [19] with the help of pySlowFast [18] as our supervised baseline on both source and target domain. All of the methods presented in this work uses SlowFastR50 [19] model as backbone model for fair comparison. Since, we use AVA-Kinetics videos as primary source domain, we do not want to show undue bias towards Kinetics [34] dataset, we pretrain SlowFastR50 for video classification task on MiT dataset [50]. Mean Average Precision (mAP) is used as metric to indicate overall performance of various domain adaptation (DA) techniques. We use Stochastic Gradient Descent (SGD) with Nesterov acceleration, and a base learning rate of 1×10^{-2} for baseline experiments while 1.25×10^{-2} for others, which is then decreased using cosine scheduler with final learning rate equal to $1/100$ of base learning rate. Warm-up lasts 1 epoch and starts from $1/10$ of base learning rate. Weight decay is set to 1×10^{-7} and momentum to 0.9. For AVA-Kinetics → AVA experiments, we train on 4 GPUs with batch size 24 for 6 epochs, for all other setups (*e.g.* AVA-Kinetics → IhD-2), we use batch size 8 and train on 2 GPUs for 4 epochs.

4.4. Ablation Studies

We also conduct ablation study to investigate the efficacy of different components of our proposed DA-AIM

Table 3: Evaluation results with IhD-2 dataset as target domain with different source domains.

Source domain	DA-AIM	carryBag	dropBag	leaveBag	stand	take a photo	throw	touch	walk	mAP
IhD-2 (oracle)	✗	54.83	54.61	28.54	99.99	99.48	100.0	27.17	85.25	68.73
AVA-Kin	✗	37.54	7.36	1.14	90.72	96.28	68.40	2.02	88.18	48.96
AVA-Kin	✓	39.97	**9.42**	**1.26**	86.04	83.71	76.88	2.08	**89.83**	48.65
IhD-1	✗	18.06	3.12	0.99	93.17	98.31	98.62	4.18	76.04	49.06
IhD-1	✓	27.75	7.47	1.16	94.88	99.26	97.94	2.70	**81.86**	51.63
AVA-Kin+IhD-1	✗	23.44	3.13	1.09	**97.46**	**99.30**	98.65	3.72	77.21	50.50
AVA-Kin+IhD-1	✓	**42.27**	2.77	1.16	93.45	98.73	**99.01**	**7.55**	74.89	**52.48**

method. We perform ablation study on two setups setups, AVA-Kinetics → AVA and AVA-Kinetics → IhD-2. Results of the same can be found in Tab. 2. Clear message from the above table is that we need to have all the component in place to gain substantial improvement.

Cross-domain instance mixing (iMix) itself can barely promote the model to learn from target domain, as seen in row 3-5 of Tab. 2. Since mixing only utilizes the ground-truth labels to compute final loss, which makes the loss rely heavily on the contents from source domain while contents from target domain only have few impact.

Pseudo-labeling worsen the performance on both source and target domain compared to baseline experiment without any of other DA techniques (see row 2 Tab 2). We observe that the pseudo-labels created by the teacher network tend to be biased towards easy-to-predict classes. Fig.5 (a) illustrate the confusion matrices of pseudo-labels created during the last epoch of training. In AVA-Kinetics → AVA experiment pseudo-labels bias towards class *sit*. Similar phenomenon is identified in earlier works applying pseudo-labelling to UDA for semantic segmentation tasks [91, 74].

The above mentioned drawbacks of cross-domain instance mixing and pseudo-labeling can be redressed by integration with resizing. Taking pseudo-labels into consideration during loss computation push the network to learn domain-invariant features that apply to target domain classification as well. On the other hand, replacing parts of the pseudo-labels by parts of the ground-truth labels incredibly addresses the bias issue of pseudo-labels. The confusion matrices of pseudo-labels created by DA-AIM are present in Fig.5 (b). We observe similar trend in AVA-Kinetics → IhD-2 as well, which we can see in Tab. 2, confusion matrix is provided in the supplementary material.

Resizing is one the important injection into DA-AIM. We verify by comparing results of cross-domain instance mixing (row 3 to row 4) and DA-AIM with and without resizing (row 5 and row 6) that resizing can actually enhance performance on target domain.

4.5. Need for Auxiliary Source Domain

Here we discuss the need for an auxiliary source domain. We need an auxiliary source domain to account for under-represented or missing classes in the primary source domain. It can be observed in Tab. 3, under-represented classes such as 'take photo', 'throw' and 'touch' are highly benefited by the auxiliary source domain supervision. Note the maximum performance gain (**52.48** mAP) is achieved by the model (AVA-Kin+IhD-1), which learns meaningful representations from both primary and auxiliary source domains for adaptation.

4.6. Comparison to State-of-the-art

Here, we compare our DA-AIM with state-of-the-art approaches in Table 4 without adding auxiliary domain. First, we briefly describe each approach, 'baseline' is where SlowFast is trained only on given source domain and tested on target domain. Next, We implement and evaluate four UDA strategies on our datasets: self-supervised learning with rotation prediction (Rotation) [32] or clip-order prediction (Clip-order) [81], adversarial learning with gradient reversal layer (GRL) [1, 21] and our DA-AIM framework.

DA-AIM outperforms other DA techniques on target domain for both AVA-Kinetics → AVA and AVA-Kinetics → IhD-2 benchmarks. Since, our evaluation benchmark are more challenging that of presented in [1], their GRL based approach fails to make any gains (see row second-last in Tab. 4). Simple adaption of image-level approach simply fails in challenging video based unsupervised domain adaptatio action detection, same can be observed in ablation study Section 4.4, where simple adaption of pseudo-labeling fails.

It is important to note that, our DA-AIM consistently improve over other approaches, especially in under represented classes, *e.g.* 'lie/sleep' and 'take a photo'. DA-AIM achieves 63.65 mAP on target domain AVA-Kinetics → AVA benchmark compared with 62.46 mAP of baseline experiment. The improvements of average precision for class *lie/sleep* and class *run/jog* are more than 5%. Meanwhile on AVA-Kinetics → IhD-2 benchmark, DA-AIM increases the

Table 4: Comparison with state-of-the-art methods for UDA. DA-AIM is trained without the supervision of the auxiliary source domain. The “source-only” model is trained on the source domain and evaluated on the target domain without any adaptation. The “oracle model” is trained and evaluated on the target domain.

	AVA-Kinetics → AVA							AVA-Kinetics → IhD-2			
Method	bend/bow	lie/sleep	run/jog	sit	stand	walk	mAP	touch	throw	take a photo	mAP
Oracle model	36.34	67.49	57.74	75.61	84.64	79.26	66.84	37.91	51.76	45.38	45.02
Source-only model	33.66	54.82	56.82	**73.70**	**80.56**	**75.18**	62.46	34.12	32.91	27.42	31.48
Rotation [32]	25.53	58.86	55.05	72.42	79.84	68.49	60.03	30.12	34.58	25.39	30.03
Clip-order [81]	28.24	57.38	56.90	69.54	77.10	74.68	60.64	28.28	32.30	29.93	30.17
GRL [1, 7, 21]	24.99	48.41	59.89	68.68	78.79	71.38	58.69	25.79	**39.71**	28.90	31.46
DA-AIM (ours)	**33.79**	**59.27**	**62.16**	71.67	79.90	75.13	**63.65**	**34.38**	35.65	**39.84**	**36.62**

mAP from 31.48 of baseline experiment to 36.62. There, the improvements of average precision for class *take a photo* exceeds 10%.

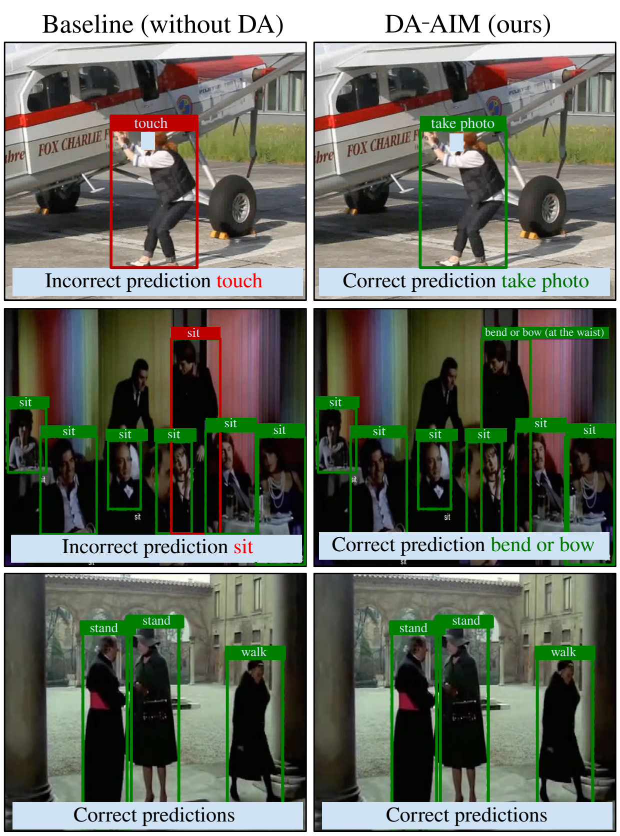

Figure 6: Qualitative results illustrated on key-frames.

Qualitative results: are also provided from our experiments in Fig.6. It shows examples where DA-AIM can identify difficult classes that baseline fails to do or DA-AIM obtains much better confidence scores.

Limitations: there remain limits to be removed and open questions to be answered. We didn’t consider action classes involving more than one action instances at the same time, such as class *talking*. This limit can be removed by treating those action classes particularly during mixing. Moreover, there is still great potential to improve the current performance of DA-AIM. For example, we pasted the action tubes at exactly the same position as it located in the original video clips. If introducing randomness of the pasted positions, there is chance to further avoid overlapping. Oversampling minority classes during mixing may also enhance the performance, especially when datasets are imbalanced.

5. Conclusions

We are the first to propose a DA action detection framework based on cross-domain mixed sampling and self-training. We implemented and systematically analyzed the efficacy of various domain adaptation strategies including self-supervised learning, adversarial learning, self-training and naive cross-domain video mixing. More importantly, we proposed DA-AIM, a novel algorithm tailored for unsupervised domain adaptive action detection. DA-AIM considers the inherent characteristics of action detection and mixes 3D video clips, bounding boxes and labels (ground-truth or pseudo-labels) from source and target domain reasonably. We empirically demonstrated DA-AIM beat other DA techniques on two challenging benchmarks: Kinetics → AVA and Kinetics → IhD-2. Compared with baseline experiment without DA techniques, DA-AIM gives rise to an increase of mAP by 1.2% on Kinetics → AVA benchmark and 5.2% on Kinetics → IhD-2 benchmark. Average precision of class *take a photo* improves over 10%. In addition, we introduced the concept of auxiliary source domain. ASD domain not only help to improve the performance of DA-AIM on classes that are missing in primary source domain but also help other under-represented classes in long-tailed primary source domain.

Acknowledgments. The authors gratefully acknowledge the support by Armasuisse.

References

[1] Nakul Agarwal, Yi-Ting Chen, Behzad Dariush, and Ming-Hsuan Yang. Unsupervised domain adaptation for spatio-temporal action localization. *arXiv preprint arXiv:2010.09211*, 2020.

[2] Alaaeldin Ali and Graham W Taylor. Real-time end-to-end action detection with two-stream networks. In *2018 15th Conference on Computer and Robot Vision (CRV)*, pages 31–38. IEEE, 2018.

[3] Nikita Araslanov and Stefan Roth. Self-supervised augmentation consistency for adapting semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15384–15394, 2021.

[4] Harkirat Singh Behl, Michael Sapienza, Gurkirt Singh, Suman Saha, Fabio Cuzzolin, and Philip HS Torr. Incremental tube construction for human action detection. *arXiv preprint arXiv:1704.01358*, 2017.

[5] David Berthelot, Nicholas Carlini, Ian Goodfellow, Nicolas Papernot, Avital Oliver, and Colin A Raffel. Mixmatch: A holistic approach to semi-supervised learning. *Advances in neural information processing systems*, 32, 2019.

[6] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6299–6308, 2017.

[7] Min-Hung Chen, Zsolt Kira, Ghassan AlRegib, Jaekwon Yoo, Ruxin Chen, and Jian Zheng. Temporal attentive alignment for large-scale video domain adaptation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6321–6330, 2019.

[8] Min-Hung Chen, Baopu Li, Yingze Bao, and Ghassan AlRegib. Action segmentation with mixed temporal domain adaptation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 605–614, 2020.

[9] Min-Hung Chen, Baopu Li, Yingze Bao, Ghassan AlRegib, and Zsolt Kira. Action segmentation with joint self-supervised temporal domain adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9454–9463, 2020.

[10] Shoufa Chen, Peize Sun, Enze Xie, Chongjian Ge, Jiannan Wu, Lan Ma, Jiajun Shen, and Ping Luo. Watch Only Once: An End-to-End Video Action Detection Framework. In *International Conference on Computer Vision (ICCV)*, pages 8178–8187, 2021.

[11] Xiaokang Chen, Yuhui Yuan, Gang Zeng, and Jingdong Wang. Semi-supervised semantic segmentation with cross pseudo supervision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2613–2622, 2021.

[12] Yuhua Chen, Wen Li, Christos Sakaridis, Dengxin Dai, and Luc Van Gool. Domain adaptive faster r-cnn for object detection in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3339–3348, 2018.

[13] Yuhua Chen, Wen Li, and Luc Van Gool. Road: Reality oriented adaptation for semantic segmentation of urban scenes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7892–7901, 2018.

[14] Jaehoon Choi, Taekyung Kim, and Changick Kim. Self-ensembling with gan-based data augmentation for domain adaptation in semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6830–6840, 2019.

[15] Jinwoo Choi, Gaurav Sharma, Samuel Schulter, and Jia-Bin Huang. Shuffle and attend: Video domain adaptation. In *European Conference on Computer Vision*, pages 678–695. Springer, 2020.

[16] N Faraji Davar, Teofilo de Campos, David Windridge, Josef Kittler, and William Christmas. Domain adaptation in the context of sport video action recognition. In *Domain Adaptation Workshop, in conjunction with NIPS*, 2011.

[17] Kevin Duarte, Yogesh Rawat, and Mubarak Shah. Videocapsulenet: A simplified network for action detection. *Advances in neural information processing systems*, 31, 2018.

[18] Haoqi Fan, Yanghao Li, Bo Xiong, Wan-Yen Lo, and Christoph Feichtenhofer. Pyslowfast. `https://github.com/facebookresearch/slowfast`, 2020.

[19] Christoph Feichtenhofer, Haoqi Fan, Jitendra Malik, and Kaiming He. Slowfast networks for video recognition. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6202–6211, 2019.

[20] Geoff French, Timo Aila, Samuli Laine, Michal Mackiewicz, and Graham Finlayson. Semi-supervised semantic segmentation needs strong, high-dimensional perturbations. 2019.

[21] Yaroslav Ganin and Victor Lempitsky. Unsupervised domain adaptation by backpropagation. In *International conference on machine learning*, pages 1180–1189. PMLR, 2015.

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

[23] Rui Gong, Wen Li, Yuhua Chen, Dengxin Dai, and Luc Van Gool. Dlow: Domain flow and applications. *International Journal of Computer Vision*, 129(10):2865–2888, 2021.

[24] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014.

[25] Chunhui Gu, Chen Sun, David A Ross, Carl Vondrick, Caroline Pantofaru, Yeqing Li, Sudheendra Vijayanarasimhan, George Toderici, Susanna Ricco, Rahul Sukthankar, et al. Ava: A video dataset of spatio-temporally localized atomic visual actions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6047–6056, 2018.

[26] Philip Haeusser, Thomas Frerix, Alexander Mordvintsev, and Daniel Cremers. Associative domain adaptation. In *Proceedings of the IEEE international conference on computer vision*, pages 2765–2773, 2017.

[27] Judy Hoffman, Eric Tzeng, Taesung Park, Jun-Yan Zhu, Phillip Isola, Kate Saenko, Alexei Efros, and Trevor Darrell.

Cycada: Cycle-consistent adversarial domain adaptation. In *International conference on machine learning*, pages 1989–1998. Pmlr, 2018.

[28] Judy Hoffman, Dequan Wang, Fisher Yu, and Trevor Darrell. Fcns in the wild: Pixel-level adversarial and constraint-based adaptation, 2016.

[29] Rui Hou, Chen Chen, and Mubarak Shah. Tube convolutional neural network (t-cnn) for action detection in videos. In *Proceedings of the IEEE international conference on computer vision*, pages 5822–5831, 2017.

[30] Lukas Hoyer, Dengxin Dai, and Luc Van Gool. Daformer: Improving network architectures and training strategies for domain-adaptive semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9924–9935, 2022.

[31] Arshad Jamal, Vinay P Namboodiri, Dipti Deodhare, and KS Venkatesh. Deep domain adaptation in action space. In *BMVC*, page 5, 2018.

[32] Longlong Jing and Yingli Tian. Self-supervised spatiotemporal feature learning by video geometric transformations. *arXiv preprint arXiv:1811.11387*, 2(7):8, 2018.

[33] Vicky Kalogeiton, Philippe Weinzaepfel, Vittorio Ferrari, and Cordelia Schmid. Action tubelet detector for spatio-temporal action localization. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4405–4413, 2017.

[34] Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, et al. The kinetics human action video dataset. *arXiv preprint arXiv:1705.06950*, 2017.

[35] 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 *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 13618–13627, October 2021.

[36] Akash Kumar and Yogesh Singh Rawat. End-to-end semi-supervised learning for video action detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14700–14710, 2022.

[37] Dong-Hyun Lee et al. Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks. In *Workshop on challenges in representation learning, ICML*, page 896, 2013.

[38] Ang Li, Meghana Thotakuri, David A. Ross, João Carreira, Alexander Vostrikov, and Andrew Zisserman. The ava-kinetics localized human actions video dataset. *ArXiv*, abs/2005.00214, 2020.

[39] Dong Li, Zhaofan Qiu, Qi Dai, Ting Yao, and Tao Mei. Recurrent tubelet proposal and recognition networks for action detection. In *Proceedings of the European conference on computer vision (ECCV)*, pages 303–318, 2018.

[40] Yixuan Li, Lei Chen, Runyu He, Zhenzhi Wang, Gangshan Wu, and Limin Wang. MultiSports: A Multi-Person Video Dataset of Spatio-Temporally Localized Sports Actions. In *International Conference on Computer Vision (ICCV)*, pages 13536–13545, 2021.

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

[42] Yanghao Li, Naiyan Wang, Jianping Shi, Jiaying Liu, and Xiaodi Hou. Revisiting batch normalization for practical domain adaptation. *arXiv preprint arXiv:1603.04779*, 2016.

[43] Yixuan Li, Zixu Wang, Limin Wang, and Gangshan Wu. Actions as moving points. In *European Conference on Computer Vision (ECCV)*, 2020.

[44] Yanghao Li, Chao-Yuan Wu, Haoqi Fan, Karttikeya Mangalam, Bo Xiong, Jitendra Malik, and Christoph Feichtenhofer. Mvitv2: Improved multiscale vision transformers for classification and detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4804–4814, June 2022.

[45] Yang Liu, Zhaoyang Lu, Jing Li, Tao Yang, and Chao Yao. Deep image-to-video adaptation and fusion networks for action recognition. *IEEE Transactions on Image Processing*, 29:3168–3182, 2019.

[46] Mingsheng Long, Yue Cao, Jianmin Wang, and Michael Jordan. Learning transferable features with deep adaptation networks. In *International conference on machine learning*, pages 97–105. PMLR, 2015.

[47] Mingsheng Long, Han Zhu, Jianmin Wang, and Michael I Jordan. Deep transfer learning with joint adaptation networks. In *International conference on machine learning*, pages 2208–2217. PMLR, 2017.

[48] Luke Melas-Kyriazi and Arjun K Manrai. Pixmatch: Unsupervised domain adaptation via pixelwise consistency training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12435–12445, 2021.

[49] Pascal Mettes and Cees GM Snoek. Pointly-supervised action localization. *International Journal of Computer Vision*, 127(3):263–281, 2019.

[50] Mathew Monfort, Bowen Pan, Kandan Ramakrishnan, Alex Andonian, Barry A McNamara, Alex Lascelles, Quanfu Fan, Dan Gutfreund, Rogerio Feris, and Aude Oliva. Multi-moments in time: Learning and interpreting models for multi-action video understanding. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021.

[51] Saeid Motiian, Marco Piccirilli, Donald A Adjeroh, and Gianfranco Doretto. Unified deep supervised domain adaptation and generalization. In *Proceedings of the IEEE international conference on computer vision*, pages 5715–5725, 2017.

[52] Jonathan Munro and Dima Damen. Multi-modal domain adaptation for fine-grained action recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 122–132, 2020.

[53] Boxiao Pan, Zhangjie Cao, Ehsan Adeli, and Juan Carlos Niebles. Adversarial cross-domain action recognition with co-attention. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 11815–11822, 2020.

[54] Junting Pan, Siyu Chen, Mike Zheng Shou, Yu Liu, Jing Shao, and Hongsheng Li. Actor-Context-Actor Relation Network for Spatio-Temporal Action Localization. In *Com-*

puter Vision and Pattern Recognition (CVPR), pages 464–474, 2021.

[55] Aayush J Rana and Yogesh S Rawat. We don't need thousand proposals: Single shot actor-action detection in videos. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 2960–2969, 2021.

[56] Stephan R Richter, Vibhav Vineet, Stefan Roth, and Vladlen Koltun. Playing for data: Ground truth from computer games. In *European conference on computer vision*, pages 102–118. Springer, 2016.

[57] German Ros, Laura Sellart, Joanna Materzynska, David Vazquez, and Antonio M Lopez. The synthia dataset: A large collection of synthetic images for semantic segmentation of urban scenes. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3234–3243, 2016.

[58] Kate Saenko, Brian Kulis, Mario Fritz, and Trevor Darrell. Adapting visual category models to new domains. In *European conference on computer vision*, pages 213–226. Springer, 2010.

[59] Suman Saha, Gurkirt Singh, and Fabio Cuzzolin. Amtnet: Action-micro-tube regression by end-to-end trainable deep architecture. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4414–4423, 2017.

[60] Suman Saha, Gurkirt Singh, Michael Sapienza, Philip HS Torr, and Fabio Cuzzolin. Deep learning for detecting multiple space-time action tubes in videos. *arXiv preprint arXiv:1608.01529*, 2016.

[61] Kuniaki Saito, Yoshitaka Ushiku, Tatsuya Harada, and Kate Saenko. Strong-weak distribution alignment for adaptive object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6956–6965, 2019.

[62] Kuniaki Saito, Kohei Watanabe, Yoshitaka Ushiku, and Tatsuya Harada. Maximum classifier discrepancy for unsupervised domain adaptation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3723–3732, 2018.

[63] Christos Sakaridis, Dengxin Dai, Simon Hecker, and Luc Van Gool. Model adaptation with synthetic and real data for semantic dense foggy scene understanding. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 687–704, 2018.

[64] Ozan Sener, Hyun Oh Song, Ashutosh Saxena, and Silvio Savarese. Learning transferrable representations for unsupervised domain adaptation. *Advances in neural information processing systems*, 29, 2016.

[65] Karen Simonyan and Andrew Zisserman. Two-stream convolutional networks for action recognition in videos. *Advances in neural information processing systems*, 27, 2014.

[66] Gurkirt Singh, Suman Saha, and Fabio Cuzzolin. Tramnet-transition matrix network for efficient action tube proposals. In *Asian Conference on Computer Vision*, pages 420–437. Springer, 2018.

[67] Gurkirt Singh, Suman Saha, Michael Sapienza, Philip HS Torr, and Fabio Cuzzolin. Online real-time multiple spatiotemporal action localisation and prediction. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3637–3646, 2017.

[68] Kihyuk Sohn, David Berthelot, Nicholas Carlini, Zizhao Zhang, Han Zhang, Colin A Raffel, Ekin Dogus Cubuk, Alexey Kurakin, and Chun-Liang Li. Fixmatch: Simplifying semi-supervised learning with consistency and confidence. *Advances in neural information processing systems*, 33:596–608, 2020.

[69] Sijie Song, Jiaying Liu, Yanghao Li, and Zongming Guo. Modality compensation network: Cross-modal adaptation for action recognition. *IEEE Transactions on Image Processing*, 29:3957–3969, 2020.

[70] Xiaolin Song, Sicheng Zhao, Jingyu Yang, Huanjing Yue, Pengfei Xu, Runbo Hu, and Hua Chai. Spatio-temporal contrastive domain adaptation for action recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9787–9795, June 2021.

[71] 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.

[72] Jiajun Tang, Jin Xia, Xinzhi Mu, Bo Pang, and Cewu Lu. Asynchronous interaction aggregation for action detection. In *European Conference on Computer Vision (ECCV)*, pages 71–87. Springer, 2020.

[73] Antti Tarvainen and Harri Valpola. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. *Advances in neural information processing systems*, 30, 2017.

[74] Wilhelm Tranheden, Viktor Olsson, Juliano Pinto, and Lennart Svensson. Dacs: Domain adaptation via cross-domain mixed sampling. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1379–1389, 2021.

[75] 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 *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7472–7481, 2018.

[76] Eric Tzeng, Judy Hoffman, Ning Zhang, Kate Saenko, and Trevor Darrell. Deep domain confusion: Maximizing for domain invariance. *arXiv preprint arXiv:1412.3474*, 2014.

[77] Tuan-Hung Vu, Himalaya Jain, Maxime Bucher, Matthieu Cord, and Patrick Pérez. Advent: Adversarial entropy minimization for domain adaptation in semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2517–2526, 2019.

[78] Qin Wang, Dengxin Dai, Lukas Hoyer, Luc Van Gool, and Olga Fink. Domain adaptive semantic segmentation with self-supervised depth estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8515–8525, 2021.

[79] Philippe Weinzaepfel, Zaid Harchaoui, and Cordelia Schmid. Learning to track for spatio-temporal action localization. In *Proceedings of the IEEE international conference on computer vision*, pages 3164–3172, 2015.

[80] Chao-Yuan Wu, Christoph Feichtenhofer, Haoqi Fan, Kaiming He, Philipp Krahenbuhl, and Ross Girshick. Long-term feature banks for detailed video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 284–293, 2019.

[81] Dejing Xu, Jun Xiao, Zhou Zhao, Jian Shao, Di Xie, and Yueting Zhuang. Self-supervised spatiotemporal learning via video clip order prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10334–10343, 2019.

[82] Lijin Yang, Yifei Huang, Yusuke Sugano, and Yoichi Sato. Interact before align: Leveraging cross-modal knowledge for domain adaptive action recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14722–14732, 2022.

[83] Xitong Yang, Xiaodong Yang, Ming-Yu Liu, Fanyi Xiao, Larry S Davis, and Jan Kautz. Step: Spatio-temporal progressive learning for video action detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 264–272, 2019.

[84] Yanchao Yang and Stefano Soatto. Fda: Fourier domain adaptation for semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4085–4095, 2020.

[85] Sangdoo Yun, Dongyoon Han, Seong Joon Oh, Sanghyuk Chun, Junsuk Choe, and Youngjoon Yoo. Cutmix: Regularization strategy to train strong classifiers with localizable features. In *International Conference on Computer Vision (ICCV)*, 2019.

[86] Pan Zhang, Bo Zhang, Ting Zhang, Dong Chen, Yong Wang, and Fang Wen. Prototypical pseudo label denoising and target structure learning for domain adaptive semantic segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12414–12424, 2021.

[87] Weichen Zhang, Dong Xu, Jing Zhang, and Wanli Ouyang. Progressive modality cooperation for multi-modality domain adaptation. *IEEE Transactions on Image Processing*, 30:3293–3306, 2021.

[88] Jiaojiao Zhao, Yanyi Zhang, Xinyu Li, Hao Chen, Bing Shuai, Mingze Xu, Chunhui Liu, Kaustav Kundu, Yuanjun Xiong, Davide Modolo, Ivan Marsic, Cees G. M. Snoek, and Joseph Tighe. Tuber: Tubelet transformer for video action detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13598–13607, June 2022.

[89] Qianyu Zhou, Zhengyang Feng, Qiqi Gu, Jiangmiao Pang, Guangliang Cheng, Xuequan Lu, Jianping Shi, and Lizhuang Ma. Context-aware mixup for domain adaptive semantic segmentation. *arXiv preprint arXiv:2108.03557*, 2021.

[90] Fan Zhu and Ling Shao. Enhancing action recognition by cross-domain dictionary learning. In *BMVC*. Citeseer, 2013.

[91] Yang Zou, Zhiding Yu, BVK Kumar, and Jinsong Wang. Unsupervised domain adaptation for semantic segmentation via class-balanced self-training. In *Proceedings of the European conference on computer vision (ECCV)*, pages 289–305, 2018.

[92] Yang Zou, Zhiding Yu, Xiaofeng Liu, BVK Kumar, and Jinsong Wang. Confidence regularized self-training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5982–5991, 2019.