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

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

In this document, we provide supplementary materials for our main paper submission.

S1. Proposed UDA Protocols

Unlike domain-adaptive (DA) semantic segmentation [5, 7, 8], for DA action detection, there is no standard UDA training/validation protocol available [1]. The main reason is the lack of suitable pairs of source and target domains (datasets) which have common action classes. Agarwal *et al*. [1] proposed two UDA protocols (UCF-Sports $\rightarrow$ UCF-101, JHMDB $\rightarrow$ UCF-101) which are limited to only three/four sports-related actions (e.g., "diving", "golf-swing", "horse-riding", "skate-boarding"). Besides, UCF-Sports [1], UCF-101 [1] and JHMDB [1] datasets are quite outdated. In this work, firstly we propose a new UDA protocol *AVA-Kinetics* $\rightarrow$ *AVA* which uses two recent action detection datasets, AVA-Kinetics [10] and AVA [6]. These new datasets are more large-scale and diversified as compared to UCF-Sports, UCF-101 and JHMDB, and thus, would be useful to learn better generalizable representation useful for adaptation task. Moreover, we propose two new action detection datasets, IhD-1 and IhD-2 which allow us two explore several new UDA protocols as shown in Table S1. Please note, our proposed UDA framework allows us to consider a wider range of action classes by leveraging an auxiliary source domain and it is not limited to a certain kind of actions. Our new datasets could be also useful for suspicious action detection. For instance, we introduce a set of new action classes (*carryBag*, *dropBag* and *leaveBag*), where *carryBag*: "person carrying a bag", *dropBag*: "person dropping a bag on the floor", *leaveBag*: "person leaving a bag unattended". These actions quite often performed in a sequnetail manner. Among these three actions, two of them are regular actions, i.e., *carryBag* and *dropBag*. But, *leaveBag* might be a suspicious one.

We hope, our proposed UDA protocols facilitate exploring new research directions in domain-adaptive action detection.

S1.1. Dataset Creation

A real-world setting highly influences our dataset creation process in which access to the target domain data is limited. That is, the target domain videos were captured at a private facility to which access is permitted only for a limited time resulting in a small number of actors and videos. Furthermore, the action categories we are interested in detecting in the target domain are heavily under-represented in the source domain due to the long-tail distribution problem. Another thing to note is that there is a large variability in actions (belonging to the same action categories) across domains (see Fig. S1).

To address these two problems, we propose a new UDA framework that facilities the model training by providing ground-truth supervision from an auxiliary source domain (ASD). ASD alleviates the long-tail distribution by injecting more training samples of those classes which are under-represented in the source domain. The videos of the ASD are captured in a public place that is easily accessible without restriction, allowing us to capture more videos with multiple actors. Since we have access to the unlabeled training data of the target domain and thus, we already know: what are the actions present there, roughly their appearance and motion patterns. We make use of these priors to generate the videos of the ASD. More specifically, we record videos of those action classes which are present in the target domain, and try to resemble (as much as possible) the action scenes of the target domain while recording the videos for ASD. To this end, we create two **i**n-**h**ouse action detection **d**atasets (IhD): (1) IhD-1 and (2) IhD-2.

IhD-1. The videos of IhD-1 are recorded in a public place which is easily accessible without any restrictions. Same action scene is recorded using two cameras to get two different view of the same scene. Fig. S2 illustrates two different views of the same scene. It facilitates adaptation across scenes within the same domain. We keep this setting for future exploration and use only videos from one view in this work. To induce diversity, we use three different backgrounds and five different subjects (actors) (see

Table S1: UDA protocols used in this work for training and evaluation of the proposed domain-adaptive action detection model. ASD: auxiliary source domain, MS: main paper submission. The "Table" column shows the table number in which the experimental results are reported for a particular UDA protocol. The "+" symbol denotes the sample mixing step between the primary and auxiliary source domains. For training, labeled samples from source domain (positioned at the left side of the arrow $\rightarrow$), and unlabeled samples from the target domain (positioned at the right side of the arrow $\rightarrow$) are used. Validation is always done on the target domain validation set.

UDA Protocol	classes	labels	ASD	Table
AVA-Kinetics $\rightarrow$ AVA	6	bend/bow, lie/sleep, run/jog, sit & stand, walk	-	MS:2
AVA-Kinetics $\rightarrow$ IhD-2	3	touch, throw, take a photo	-	MS:2
AVA-Kinetics $\rightarrow$ IhD-2	8	carryBag, dropBag, leaveBag, stand, take a photo, throw, touch, walk	-	MS:3
IhD-1 $\rightarrow$ IhD-2	8	carryBag, dropBag, leaveBag, stand, take a photo, throw, touch, walk	-	MS:3
AVA-Kin+IhD-1 $\rightarrow$ IhD-2	8	carryBag, dropBag, leaveBag, stand, take a photo, throw, touch, walk	IhD-1	MS:3

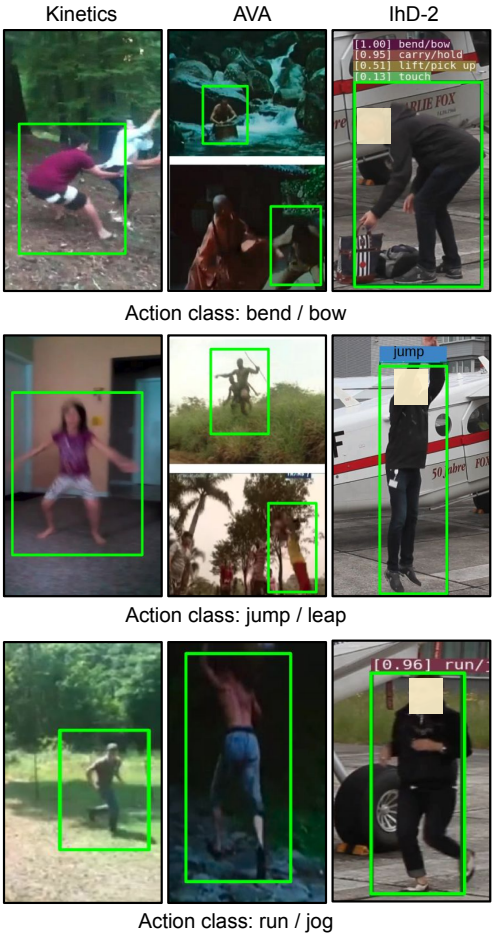

Figure S1: Illustrating large variability of action instances of the same class across three different domains (or datasets): Kinetics (public), AVA (public) and IhD-2 (private). For maintaining anonymity, we cover the subject faces.

Fig. S3) . Moreover, actors change their clothes alternatively in-between two action scenes to bring variations in

Figure S2: Action scenes are recorded using two camera from two different viewing angles. Two sample frames captured from two different views of the same action scene are shown here. Samples belong to the IhD-1 private dataset proposed in this work. For maintaining anonymity, we cover the subject faces.

the appearance. We generate action videos of co-occurring action instances (of same or different action classes) to simulate real-world scenarios. Fig. S4 shows some examples of co-occurring action instances.

IhD-2. The videos of IhD-2 were recorded in a private area and access to the place is limited. That is multiple entries are not allowed and there is a strict time limit to capture some sample videos. Also, due to security reasons, only two subjects or actions are allowed to perform different actions. Due to these limitations, IhD-2 (target domain) has very view train and validation samples with less diversity in the training set.

Video Annotation Process. We use the VoTT (Visual Object Tagging Tool) [11] to annotate the videos of IhD-1 and IhD-2. Two human annotators were assigned for the annotation task. First, the videos are loaded to VoTT and key-frames are selected for annotation. For each video, key-frames are selected at a frame rate of 4 FPS. For instance, video with 30 seconds duration would have 120 key frames. For each key-frame, bounding box annotations and their corresponding class labels are provided. For generating dense frame-level annotation, we guide the annotation process by a YOLO-V5[2] person detector. More specifically, we propagate the key-frame ground truth bounding boxes in time for the regular frames by using a simple track-

Figure S3: Five subjects and three backgrounds are used in IhD-1 dataset. For maintaining anonymity, we cover the subject faces.

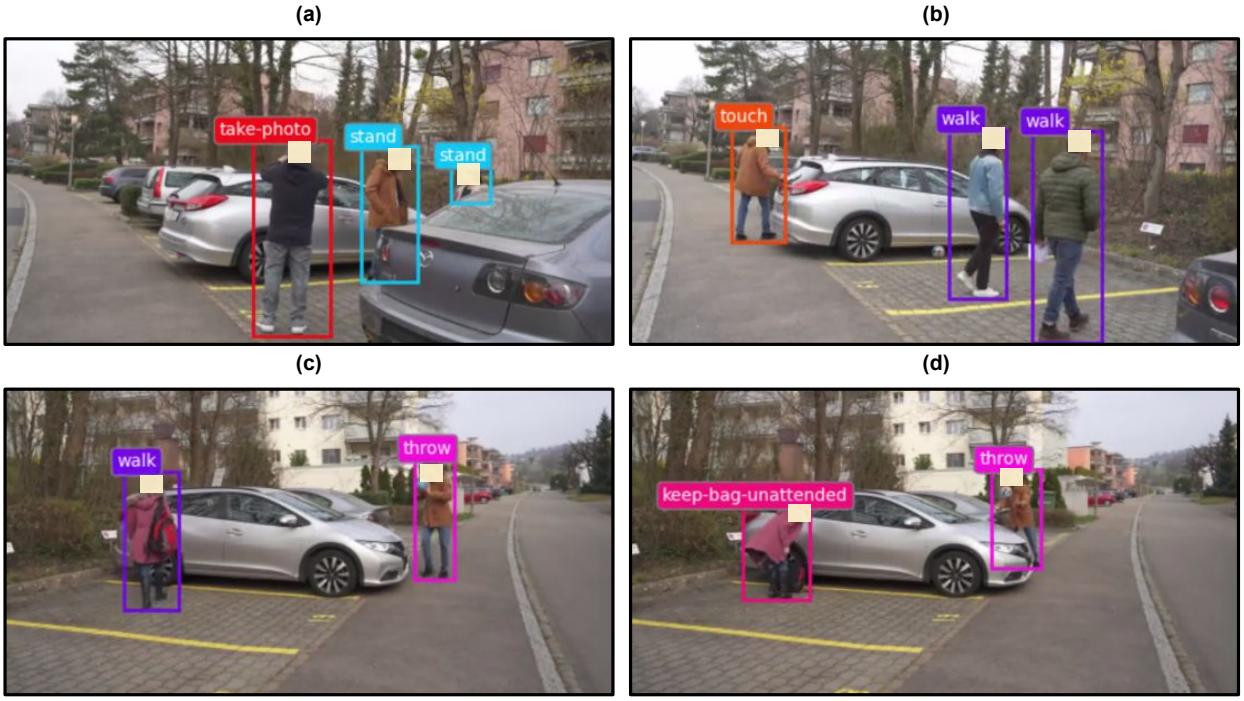

Figure S4: Sample frames from IhD-1 dataset demonstrating co-occurring action instances. (a) "take-photo", "stand"; (b) "touch", "walk", (c) "walk", "throw", and (d) "keep-bag-unattended", "throw". Bounding boxes depict ground truth annotations. Each unique color denote an action class. For maintaining anonymity, we cover the subject faces and any relevant information.

ing algortihm. The tracking algorithm first localize the action instances in the first key-frame using the VoTT ground truth boxes. Next, for each regular frame where there is no ground truth box available, it picks the YOLO-V5 bounding boxes and match them with the previous key-frame's ground truth boxes. The set of best matched boxes are used as ground truth boxes for the current frame. The matching is done based on the intersection-over-union (IoU) scores among the ground truth and YOLO-V5 boxes. For a sanity check of the YOLO-V5 person detector, we run inference

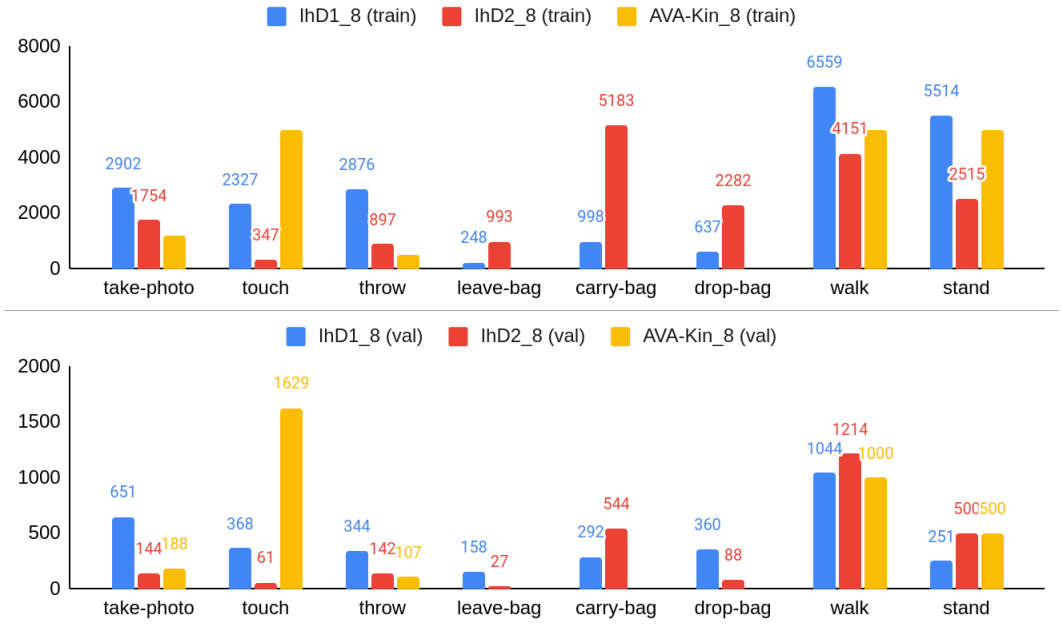

Figure S5: Samples per class for datasets AVA-Kinetics (5 classes) and IhD-1 (8 classes), IhD-2 (8 classes) for respective train and validation sets.

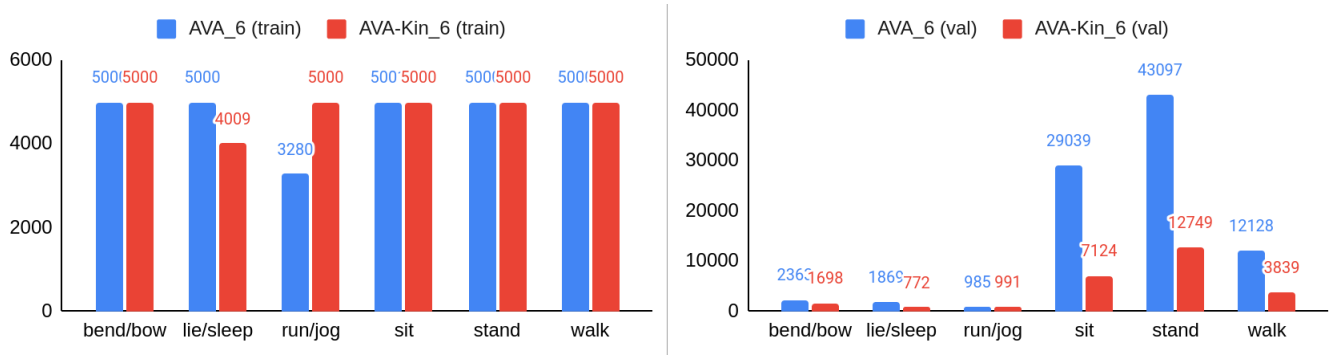

Figure S6: Samples per class for datasets AVA-Kinetics (6 classes) and AVA (6 classes) for respective train and validation sets.

on the AVA-Kinetics validation set and found that the recall to be very high. For both IhD-1 and IhD-2 videos are recorded with a frame rate of 30 FPS. The spatial dimension of the video frames is 920×1080 pixels.

Dataset Statistics. We use the following datasets in this work: (1) AVA-Kinetics (6 classes), (2) AVA (6 classes), (3) AVA-Kinetics (3 classes), (4) IhD-2 (3 classes), (5) AVA-Kinetics (5 classes), (6) IhD-1 (8 classes), and (7) IhD-2 (8 classes). AVA-Kinetics-6, -3, and -5 are the subsets of the original AVA-Kinetics dataset and thus they belong to the same domain. AVA-6 is a subset of the original AVA dataset. IhD-2-3 is the subset of the proposed IhD-2-8 dataset. Fig S5, S6, and S7 show the bar plots depicting the per-class sample distribution for the training and validation sets for these datasets.

Please note in Fig. S5, the number of training samples for classes "take-photo" and "throw" are very less in the source domain (AVA-Kinetics) and restricted target domain (IhD-2). Our proposed ASD helps alleviate these data imbalance issue by injecting labeled training samples for these classes. Although, there are sufficient number of training samples available for class "touch" in the source domain, but due to large variability between the source and target domain's data distribution, the adaptation from the source to target domain is ineffective. Our ASD address this domain shift by generating more training samples of "touch" action

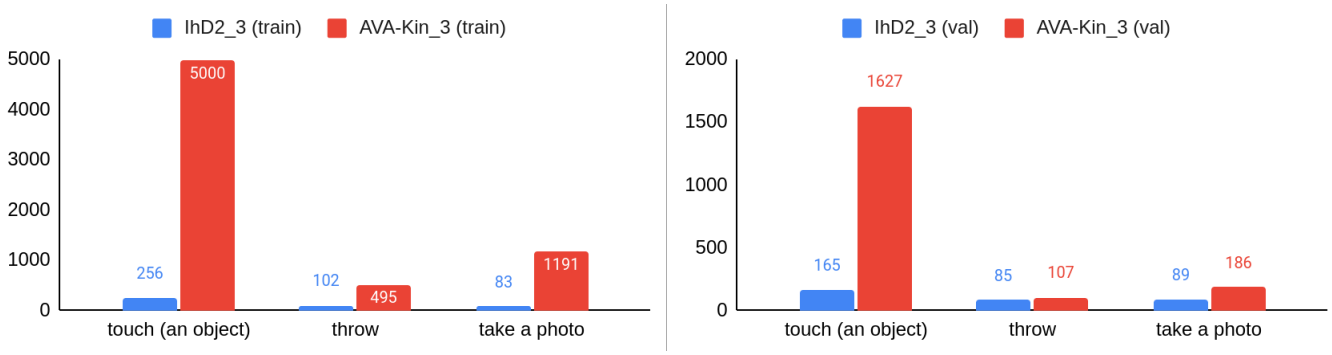

Figure S7: Samples per class for datasets AVA-Kinetics (3 classes) and IhD-2 (3 classes) for respective train and validation sets.

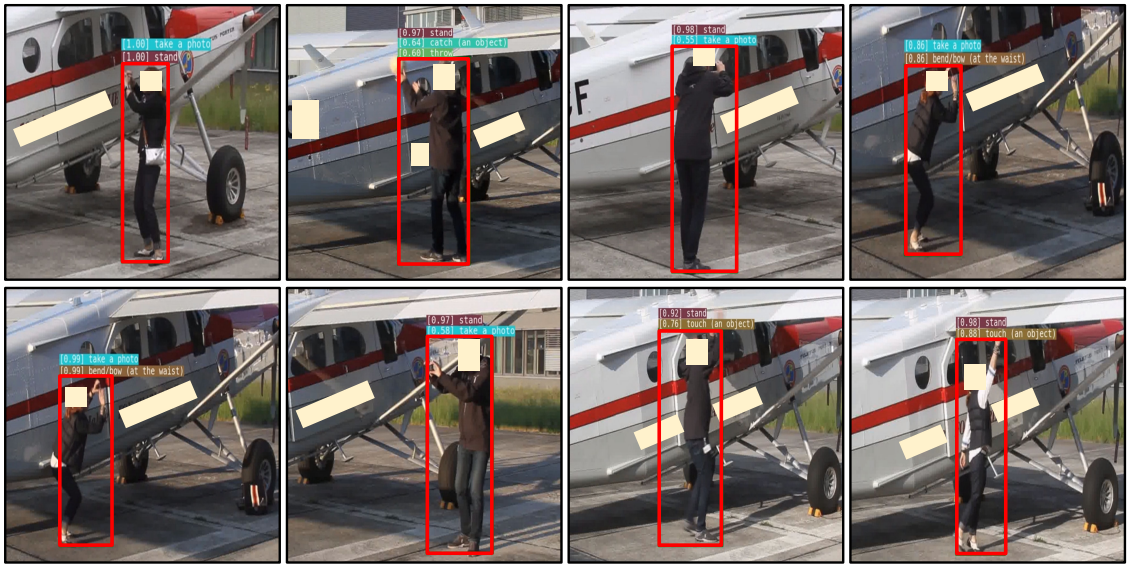

Figure S8: Qualitative DA action detection results of our proposed model trained on UDA protocol AVA-Kin+IhD-1 → IhD-2. Sample detection results are shown on the validation set of IhD-2 dataset. Our DA-AIM can successfully detect action classes such as "take-a-photo", "touch", "throw" and "stand". For maintaining anonymity, we cover the subject faces and any relevant information.

in a setting where the action scenes resembles to the target domain's scenes. Furthermore, for the missing actions such as "carry-bag", "drop-bag" and "leave-bag", our ASD provides more lableled traing samples to provide better supervision to the model. One important thing to note that, although the plot shows more number of training samples for "bag" related actions in the target domain (IhD-2), but these samples are homogeneous (or less diversified). That is, the action scenes in these video frames have limited number of actors, backgrounds due to the fact that the target domain has very limited access. One the other hand, the ASD's samples are more diversified with more number of actors, backgrounds.

S2. Pretrained Weights for UDA

We use SlowFast [4] as our backbone network. There are pretrained weights publicly avaibale for SlowFast at pySlowFast [3]. These pretrained weights are generated by training the SlowFast network on the AVA-Kinetics dataset. Since, we use AVA-Kinetics videos as primary source domain, we do not want to show undue bias towards Kinetics [9] dataset, we pretrain SlowFastR50 for video classification task on MiT dataset [12]. We will make the pretrained weights publicly available upon the acceptance of paper. MiT dataset [12] consits of 305 action/event classes. It has 727,305 training videos and 30,500 testing videos. We train the SlowFast network on MiT using 8 GPUs (GeForce RTX 2080 TI) for 10 days.

Figure S9: Qualitative DA action detection results of our proposed model trained on UDA protocol AVA-Kin+IhD-2 → IhD-1. Sample detection results are shown on the validation set of IhD-1 dataset. Our DA-AIM can successfully detect action classes such as "touch", "throw", "take-a-photo", "stand" and "walk". For maintaining anonymity, we cover the subject faces and any relevant information.

Table S2: Comparison of the source-only model performance with the proposed DA-AIM. The source-only model is trained on the AVA-Kinetics dataset. The DA-AIM is trained following the proposed UDA protocol AVA-Kinetics → IhD-1. Both the models are evaluated on the validation set of IhD-1. Note that the proposed UDA protocol helps improving the action recognition performance for certain classes ("throw", "touch" and "walk") on the unseen target domain samples.

Models	stand	take-photo	throw	touch	walk	mAP
Source-only	**41.0**	**94.3**	46.6	15.3	60.9	51.6
DA-AIM	22.2	94.0	**47.4**	**33.4**	**68.6**	**53.1**

S3. Additional Quantitative Results

S3.1. Effectiveness of the proposed UDA protocol

In this section, we discuss the benefits of the proposed UDA protocol AVA-Kinetics → IhD-1. For this UDA protocol, we have created a new action detection dataset IhD-1. Please refer to S1.1 for information on dataset creation. In Tab. S2, we report the results of the source-only and DA-AIM models. The DA-AIM is trained following the proposed UDA protocol AVA-Kinetics → IhD-1. Note that the proposed UDA protocol helps improving the action recognition performance for certain classes ("throw", "touch" and "walk") on the unseen target domain samples.

S3.2. DA-AIM improves pseudo-labels

In this section, confusion matrices (real-labels vs. pseudo-labels) of different UDA models are presented. Fig. S10 compares the confusion matrices of two different models. The models are trained following the AVA-Kinetics → IhD-2 UDA protocol on three classes. The pseudo-label-only model (Fig. S10**a**) is trained using only pseudo-labels without the DA-AIM. The DA-AIM model (Fig. S10**b**) is trained following the proposed approach. Note that the bias from the pseudo-labels in the pseudo-label-only model is is rectified by the DA-AIM approach.

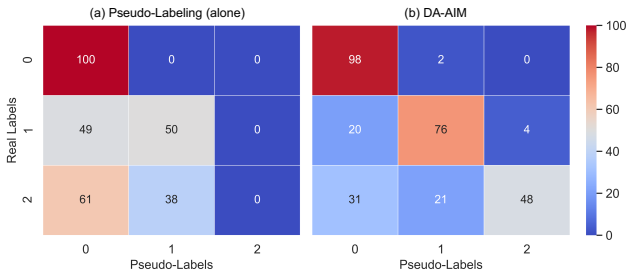

Figure S10: Comparison of confusion matrices (real-labels vs. pseudo-labels) of two different UDA approaches. These two models are trained on AVA-Kinetics →IhD-2. The first model (a) is trained using only pseudo-labels without the proposed adaptation approach. The second model (b) is trained following the proposed DA-AIM approach. Note, our proposed DA-AIM helps improving the quality of the pseudo labels.

However, sometimes due to the presence of the under-represented or missing classes, an auxiliary source domain is required. In Fig. S11, we compare the confusion matrices of pseudo-labels from 4 different models. Significant improvements of pseudo-label accuracy can be observed after introducing auxiliary source domain IhD-1 coupled with the proposed DA-AIM.

S4. Qualitative Results

Fig. S8 and S9 present qualitative DA action detection results of our proposed model trained on UDA protocols AVA-Kin+IhD-1 → IhD-2 and AVA-Kin+IhD-2 → IhD-1 respectively. Sample detection results are shown on the respective target domain's validation frames. Note that, the proposed DA-AIM model can successfully detect action classes such as "touch", "throw", "take-a-photo", "stand" and "walk".

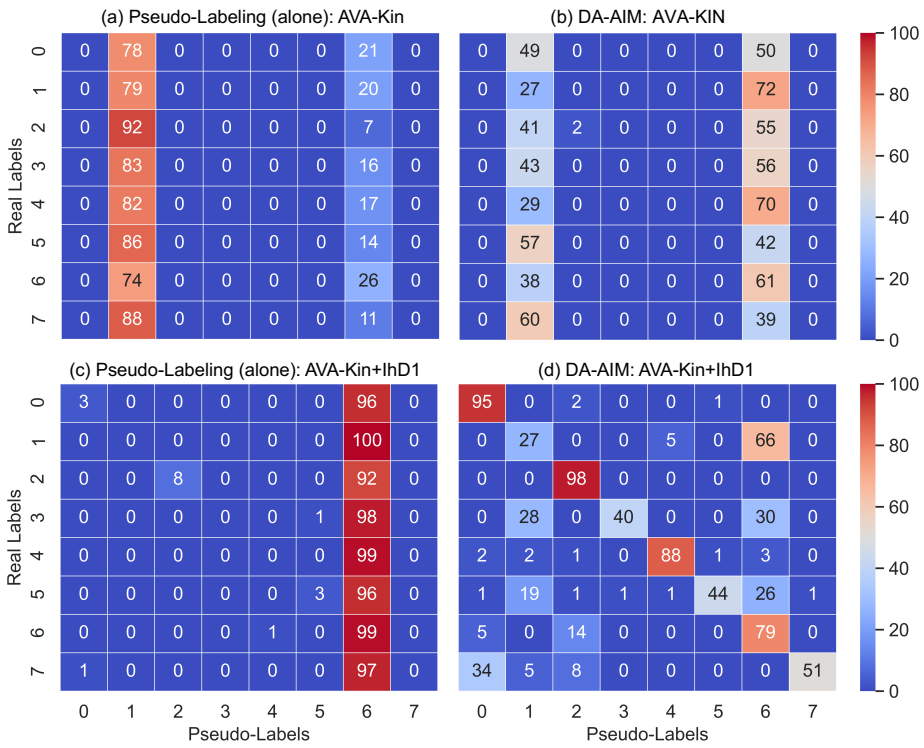

Figure S11: Comparison of confusion matrices (real-labels vs. pseudo-labels) of 4 different models trained on 8 action classes. Models are trained following either AVA-Kinetics $\rightarrow$ IhD-2 (a,b); or AVA-Kinetics $+$IhD-1 $\rightarrow$ IhD-2 (c,d). Note that best quality pseudo-labels are achieved when we perform auxiliary source domain based adaptation. .

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] Glenn Jocher et. al. ultralytics/yolov5: v6.0 - YOLOv5n 'Nano' models, Roboflow integration, TensorFlow export, OpenCV DNN support, Oct. 2021.

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

[4] 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.

[5] 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.

[6] 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.

[7] 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.

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

[9] 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.

[10] 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.

[11] Microsoft. Vott (visual object tagging tool). `https://github.com/microsoft/VoTT`.

[12] 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.