Supplementary Material: 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

1 Implementation Details

1.1 Network Architecture

We describe the architecture of the different adaptation modules in Figure 1 which are used for all our experiments.

1.2 Training Details

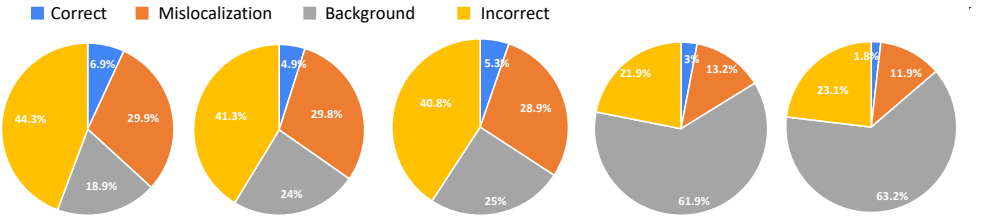
We fix the spatial input resolution to be 320×400 pixels and divide the video clips into overlapping segments with a stride of 1 frame. Given a key frame, we pad the temporal context (circular) to 20 frames. For pre-training the network, we warm-start the learning rate from 0.00001 to 0.001 in 3k steps using linear annealing for stabilizing training and then use cosine learning rate decay [11]. When fine-tuning for adaptation, we use the Stochastic Gradient Descent (SGD) solver and fix the learning rate of the action localization model and

Spatial Domain Classifier
Conv 3 x 3 x 512, stride 1, pad 1
Leaky ReLU
Conv 3 x 3 x 256, stride 1, pad 1
Leaky ReLU
Conv 3 x 3 x 128, stride 1, pad 1
Average Pooling
Fully Connected 128 x 2
Softmax

Temporal Domain Classifier (Image)
Conv 3 x 3 x 512, stride 1, pad 1
Leaky ReLU
Conv 3 x 3 x 256, stride 1, pad 1
Leaky ReLU
Conv 3 x 3 x 128, stride 1, pad 1
Leaky ReLU
Conv 3 x 3 x 64, stride 1, pad 1
Leaky ReLU
Conv 3 x 3 x 1, stride 1, pad 1
Sigmoid

Temporal Domain Classifier (Instance)
Fully Connected 1024 x 1024
ReLU
Dropout
Fully Connected 1024 x 1024
ReLU
Dropout
Fully Connected 1024 x 1
Sigmoid

Figure 1: The architecture of the three adaptation modules.



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

Figure 2: Error analysis of bottom ranked detections. Fraction of predictions that are correct, mislocalized, are confused with background or incorrectly predicted are shown.

domain classifiers to 0.0001 and 0.0005, respectively. All feature layers are jointly updated during training. The networks are trained with two Nvidia V100 GPUs with 16GB memory using a batch size of 8 for both pre-training and fine-tuning for adaptation. During the adaptation process, 4 segments from source domain and 4 from target domain are used in each batch. We empirically fix the value of λ to 0.1 and 0.05 between UCF Sports-UCF 101 and JHMDB-UCF 101 pairs respectively, and γ to 3 similar to [2] for all the experiments. Batch-norm updates are disabled for Resnet-50 but enabled for I3D. We use a non-maximum suppression threshold of 0.5 for post-processing the frame-level detections. Note that we do not use the annotations in the validation set of the source domain for pre-training. For data augmentation, we use horizontal flipping.

UCF Sports → UCF-101. The set of common classes results in about 59k and 2k segments for training from both UCF-101 and UCF-Sports, respectively. We pre-train the model using segments from the UCF-Sports dataset for 60k iterations. Then we add the adaptation modules and fine-tune the whole network for another 5k iterations.

JHMDB → UCF-101. We have about 38k and 3k segments for training from the UCF-101 and JHMDB datasets, respectively. We pre-train the model on the source domain for 30k iterations, and then fine-tune for adaptation for another 10k iterations.

UCF-101 → JHMDB. Same experimental setup is used as above, with source and target datasets interchanged. We pre-train the model on the source domain for 75k iterations, and then fine-tune for adaptation for another 10k iterations.

2 Error Analysis

In addition to providing error analysis on the top ranked detections of the base model (i.e, I3D+RPN) and our models after adaptation in the main paper, we also analyze the errors of the bottom ranked detections in Figure 2.

Even for the bottom ranked detections, we observe that adaptation considerably reduces the fraction of incorrect classifications. This can be especially observed when we adapt the spatial features, which reduces the error by 21.2%. The mislocalization error is also reduced by a considerable margin after adaptation. However, we note that the adaptation does not improve the fraction of correct detections. This happens because whatever fraction is reduced in mislocalization error and incorrect classifications is gained by the background error, resulting in duplicate detections as the predicted class is still correct. This suggests that the adaptation is able to improve action classification but not the localization performance, which is understandable as the model is least confident about these predictions.

3 Example Images from the Datasets

We show example images from the common action videos of UCF-101 and JHMDB datasets in Figure 3. Note that the spatial context and view point variation makes the *basketball* and *golf* actions look very different. We also compare the *Walk* action from JHMDB and the *walking with dog* action from UCF-101 in Figure 4 to demonstrate the visual difference.

Although both the UCF-Sports and UCF-101 are primarily sports-oriented datasets, the actions look quite different because of background scenery and view-point variation. Snapshots of some examples are shown in Figure 5.

4 Qualitative Results

We first provide visual results for UCF-Sports $\rightarrow$ UCF-101 adaptation task. Figure 6 and 7 provide visual comparisons of the baseline model without adaptation (in red) with our best adapted model (in cyan). We show that our approach yields better action detections. We also demonstrate some cases in Figure 8 for which the baseline model fails to predict any action tubes whereas our adapted model correctly localizes and classifies the actions. Finally, we present some failure cases of our adapted model in Figure 9.

Additionally, we also provide results for JHMDB $\rightarrow$ UCF-101 adaptation task. We provide a comparison with the model without adaptation in Figure 10. Figure 11 and 12 demonstrate interesting results and failure cases, respectively.

References

[1] Ilya Loshchilov and Frank Hutter. Sgdr: Stochastic gradient descent with warm restarts. In *ICLR*, 2017. 1

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

Figure 3: Example images of all three common categories between the UCF101 (top) and JHMDB (bottom) datasets. While being the same actions, significant differences in terms of view-point variation and background scenes are observed between the two datasets.

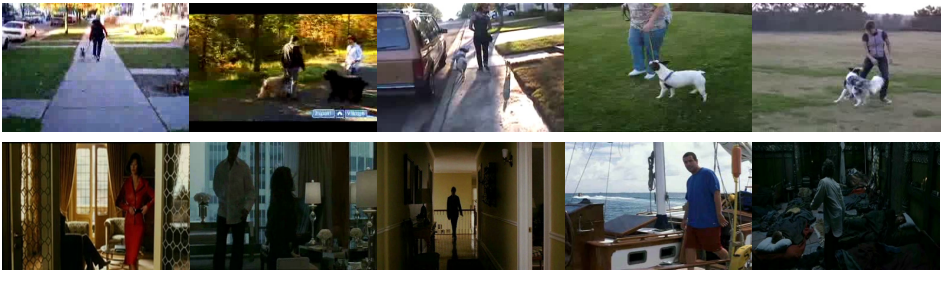


Figure 4: Example images for the *walk* action from UCF-101 (top) and JHMDB (bottom). The *walk* action from the UCF-101 dataset is always accompanied with a dog and is generally carried out in outdoor environments, making it visually very different from its counterpart in the JHMDB dataset.

Figure 5: Example images of the four common categories between the UCF-101 (top) and UCF-Sports (bottom) datasets. Considerable differences in terms of view-point variation and background scenes are observed between the two datasets.

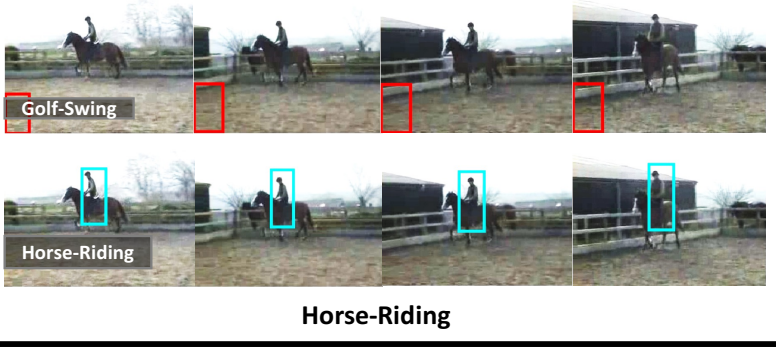


Figure 6: Example clips from the UCF-101 dataset, where we show 4 frames from each video. The highest scoring detected action tube is shown for each clip. The red and cyan box correspond to the region for baseline model and our best adapted model, respectively. The predicted label overlaid on images. The text below the figure is the ground truth class. The examples demonstrate that our adapted model not only provides better spatial localization but also achieves better action classification results.

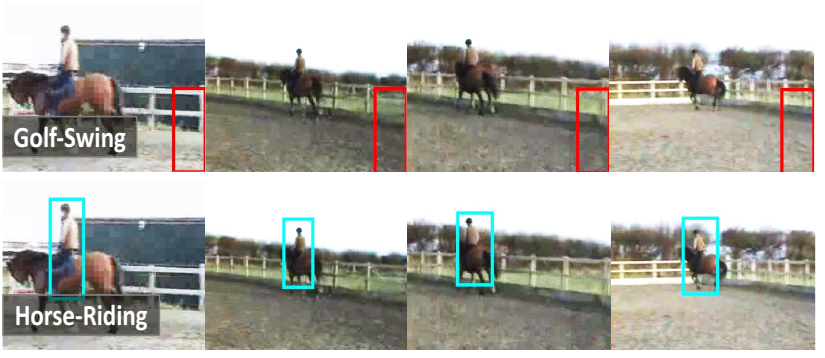


Figure 7: Example clips from the UCF-101 dataset, where we show 4 frames from each video. The highest scoring detected action tube is shown for each clip. The red and cyan box correspond to the region for baseline model and our best adapted model respectively, with the predicted label overlaid. The text below the figure is the ground truth class.

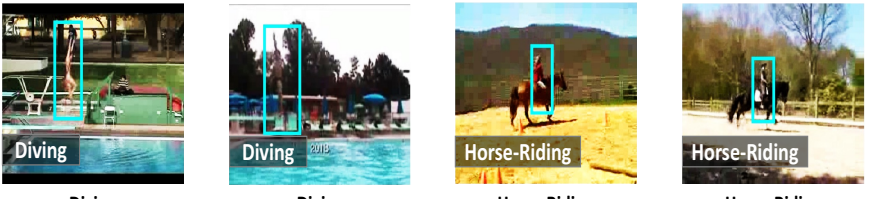


Figure 8: Examples from the UCF-101 dataset, where the baseline model fails to predict any action tubes. Our adapted model (in cyan) correctly localizes and classifies the actions. The text below each figure is the ground truth class.

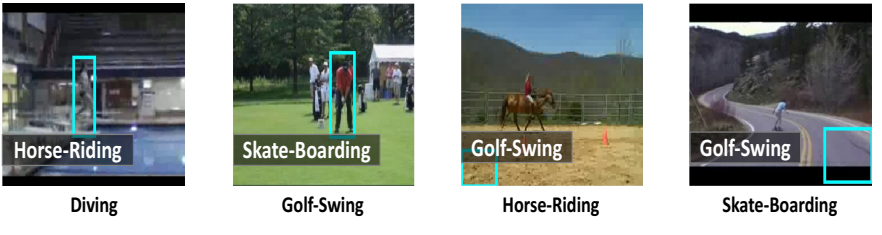
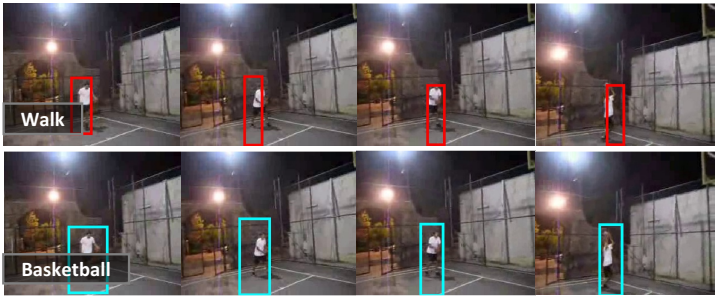
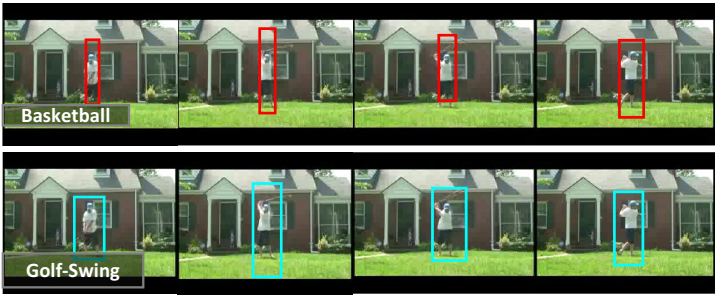


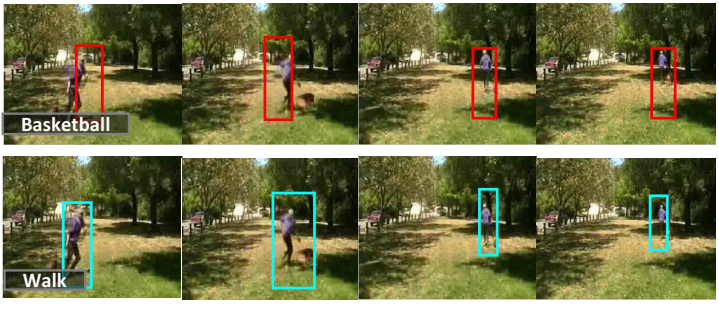
Figure 9: Failure cases of our adapted model (in cyan) on the UCF-101 dataset. The text below the figure is the ground truth class. The first two examples show classification errors. The third and fourth examples illustrate localization and classification errors.



Basketball



Golf-Swing



Walk

Figure 10: Example clips from the UCF-101 dataset, where we show 4 frames from each video. The highest scoring detected action tube is shown for each clip. The red and cyan box correspond to the region for baseline model and our best adapted model respectively, with the predicted label overlaid. The text below the figure is the ground truth class.

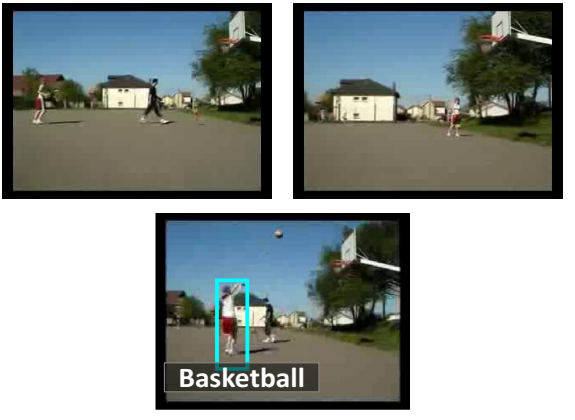


Figure 11: Examples of *Basketball* from the UCF-101 dataset belonging to the same group (i.e., sharing some common features, such as similar background, similar viewpoint, etc). For the two examples in the top row, our adapted model (in cyan) does not give any predictions due to the actor being too small and far away. But our model still correctly localizes and classifies the action in the bottom row. The baseline model (i.e., without adaptation) does not give any predictions for all three examples. The text below the figure is the ground truth class.

Figure 12: Failure case of our adapted model on the UCF-101 for the *Walk* action. Our model (in cyan) is eventually able localize the action, but fails to do so in the initial frames. Even then, it incorrectly classifies the action as *Basketball* due to *Walk* action of JHMDB bieng significantly different from that of UCF-101. The text below the figure is the ground truth class.