Supplementary Material for A Unified Framework for Domain Adaptive Pose Estimation

Donghyun Kim[2⋆], Kaihong Wang[1⋆],
Kate Saenko[1,2], Margrit Betke[1], and Stan Sclaroff[1]

[1]Image and Video Computing, Department of Computer Science, Boston University,
[2]MIT-IBM Watson AI Lab
{donhk, kaiwkh, saenko, betke, sclaroff}@bu.edu

1 Appendix

In this supplementary material, we provide additional training details of our method. In addition to the ablation studies in the main paper, we also provide additional ablation studies on the *SURREAL→Human3.6M* dataset. Finally, we show additional qualitative examples.

1.1 Additional Training Details

We follow settings from AdaIN to train the generator g from Eq. 2 with a content loss and a style loss balanced by style loss weight $\lambda = 0.1$, on images with a resolution of 256×256. Exemplar results are illustrated in Fig. 1. During the training process of our framework, the pre-trained style transfer module will be fixed and perform bidirectional style transfer with a probability of 0.5 in both our supervised and unsupervised learning branch with the content-style trade-off parameter α uniformly sampled from 0 to 1.

Our pose estimation model h is trained with input images with a resolution of 256×256 and output heatmaps with a size of 64×64, with the batch size of 32 in each iteration, following our baselines [1].

As for our adaptive keypoint occlusion, we randomly select keypoints with maximum activation greater than the occlusion threshold τ_{occ} and occlude it with a probability of 0.5. The keypoints will be occluded by a patch from a random position in the same image with the size of 20×20.

1.2 Additional Ablation Studies

In addition to *RHD→H3D* and *SynAnimal→TigDog*, we also present ablation studies on another major benchmark, *SURREAL→Human3.6M* in Table 1. Based on the results we can observe a greater improvement after applying heatmap normalization (the first and the second row), showing the necessity

⋆ Equal Contribution.

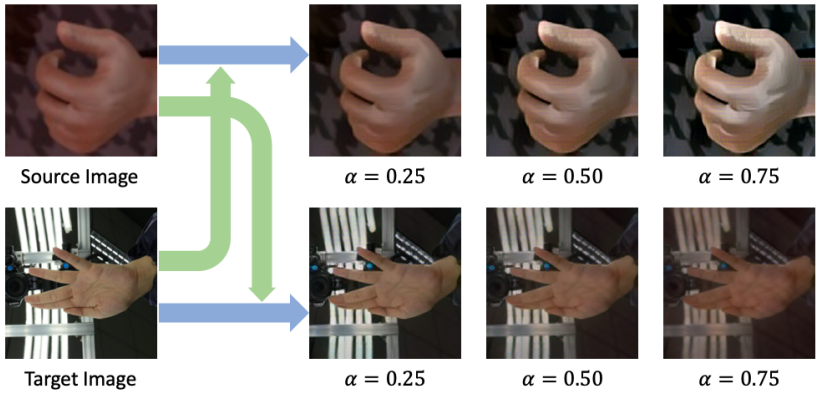

Fig. 1: An illustration of style transfer between source and target domains with different content-style trade-off parameter α. Blue arrows: content. Green arrows: style

Table 1: Ablation studies on *SURREAL→Human3.6M*. Sld: shoulder, Elb: Elbow. MT: Mean Teacher, Norm: Heatmap Normalization, Style: Stylization, Occ: Adaptive Occlusion

Method	Sld	Elb	Wrist	Hip	Knee	Ankle	All
MT	69.8	86.7	75.4	27.5	80.9	83.6	70.6
MT + Norm	76.7	88.6	80.3	50.6	85.2	85.8	77.9
MT + Style	74.8	88.7	79.3	40.1	83.5	85.7	75.4
MT + Norm + Style	75.0	88.2	79.2	49.1	83.8	85.9	76.8
MT + Norm + Style + Occ	78.1	89.6	81.1	52.6	85.3	87.1	79.0

of addressing the drift effect under this scenario. On the other hand, we can also observe fewer improvements (the third and the fourth row) brought by the style transfer module, which coincide with our conclusion from the ablation studies on *RHD→H3D* that the major challenge in human pose estimation tasks comes from the output-level discrepancy instead of the input-level. On that basis, our adaptive keypoint occlusion mechanism further boosts the performance by 2.2 percent points (the last row) and achieves the state-of-the-art performance, which shows the effectiveness of the occlusion mechanism specialized in this task.

1.3 Ablation studies of data augmentation

Tab. 2 presents ablation studies of data augmentation methods on *RHD→H3D*. We compare the performance of our method with different compositions of augmentations commonly used in pose estimation tasks, and we observe that rotation provides the most significant gain.

Table 2: Ablation studies on data augmentation

Translation	Scale	Color	Rotation	Shear	PCK@0.05
✓					53.2
	✓				54.2
		✓			51.7
			✓		77.7
				✓	54.8
✓	✓				54.4
✓	✓	✓			54.7
✓	✓	✓	✓		79.1
✓	✓	✓	✓	✓	79.6

1.4 Additional Qualitative Results & Failure Cases

We provide additional qualitative results in Figures. 2, 3, 4, 5, and 6.

2 Discussion of limitation & future directions

Even though our method gains significant improvements over source-only pretraining, the overall performance is still limited and not comparable to the supervised learning level (target only). Therefore, while we explore only unsupervised DA, semi-supervised DA methods that can leverage a limited amount of target domain annotation to further improve the accuracy will be an interesting future direction. Additionally, while we focus on domain adaptive 2D pose estimation, 3D pose estimation is also a good research direction to explore as it is harder to obtain depth annotations.

3 Learning Animal pose estimation from human

The main challenge under our fully unsupervised settings, if we learn only from a human pose dataset without animals, would be the limited number of shared keypoints because of the anatomical differences between human and animals, which limits the amount of information we can learn from the source human dataset. In SURREAL→Tigdog learning limbs of human and animals, our method achieves 7.9% of accuracy, while the source-only pretraining and RegDA achieves 2.4% and 2.6% respectively. These low accuracies indicate the difficulty of the adaptation in this unsupervised manner when the anatomical differences are significant.

References

1. Jiang, J., Ji, Y., Wang, X., Liu, Y., Wang, J., Long, M.: Regressive domain adaptation for unsupervised keypoint detection. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 6780–6789 (2021)

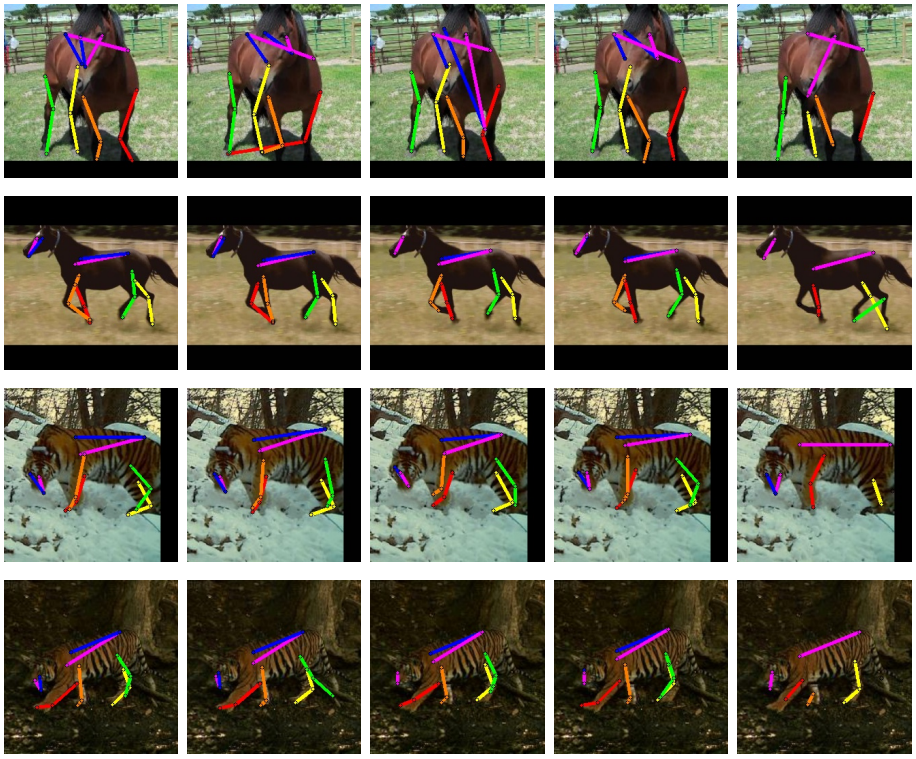

Fig. 2: Additional qualitative results on TigDog. Compared with baselines, our method can more accurately estimate the position of keypoints from different angle (the first row), different motion (the second row), and different animals. It is also worth noting that the position of the legs in the example at the second row is mistakenly annotated in ground-truth, while we can still estimate their actual position. This justifies the motivation of our work that seeks to free pose estimation tasks from the dependence of the laborious and unreliable manual annotation process

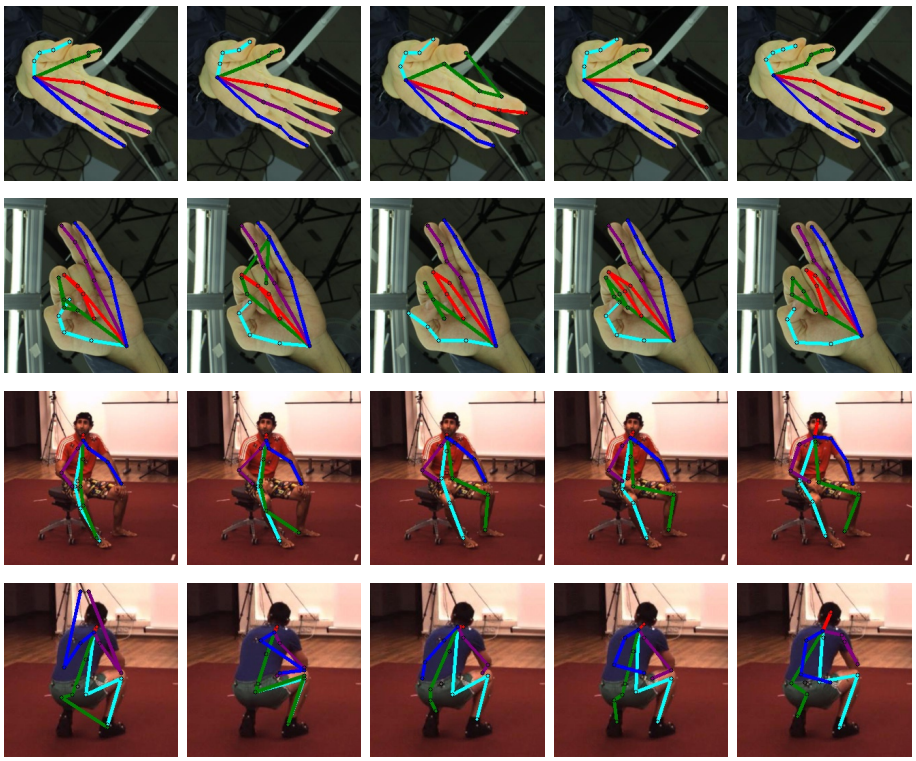

Fig. 3: Additional qualitative results on human pose estimation tasks. We show that our method can better handle the detection of keypoints in diverse poses (the first and the second rows) and from diverse view points (the third and the fourth rows), compared with baselines

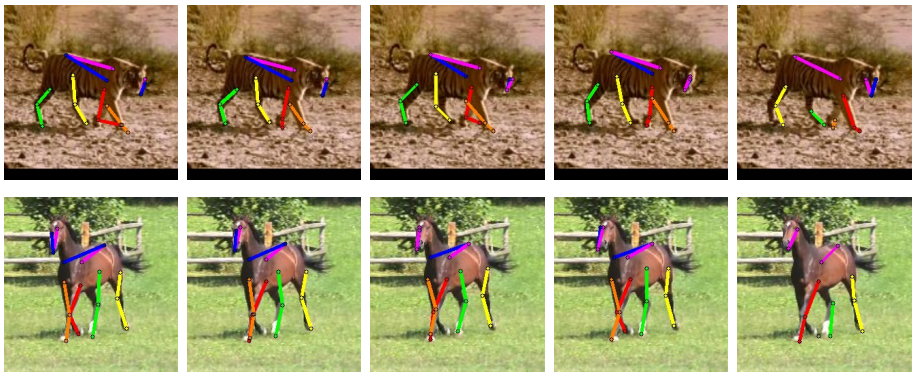

Fig. 4: Failure cases on TigDog. We show that extreme cases in typical pose estimation problems, including distinguishing left and right limbs (the first row) and ambiguous occlusion (the second row), can still be challenges in our method and result in an incorrect prediction

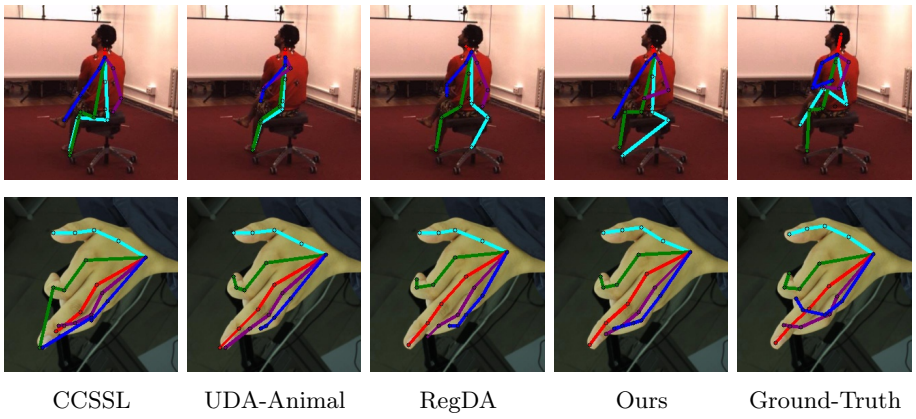

Fig. 5: Failure cases on human pose estimation tasks. Existing difficulties in typical pose estimation tasks still pose a huge challenge to all the baseline methods and ours, especially when ambiguous occlusion happens

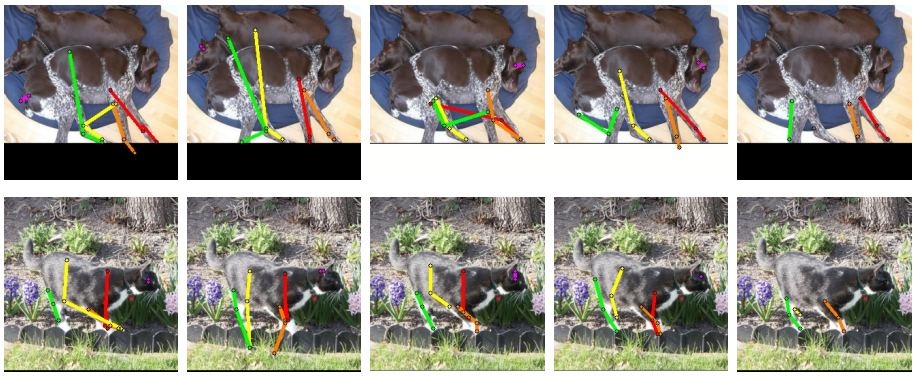

Fig. 6: Qualitative results of generalization to unseen animals and domains. Note that the annotations for occluded keypoints (yellow parts) are not available in ground truth