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

Abstract. While pose estimation is an important computer vision task, it requires expensive annotation and suffers from domain shift. In this paper, we investigate the problem of domain adaptive 2D pose estimation that transfers knowledge learned on a synthetic source domain to a target domain without supervision. While several domain adaptive pose estimation models have been proposed recently, they are not generic but only focus on either human pose or animal pose estimation, and thus their effectiveness is somewhat limited to specific scenarios. In this work, we propose a unified framework that generalizes well on various domain adaptive pose estimation problems. We propose to align representations using both input-level and output-level cues (pixels and pose labels, respectively), which facilitates the knowledge transfer from the source domain to the unlabeled target domain. Our experiments show that our method achieves state-of-the-art performance under various domain shifts. Our method outperforms existing baselines on human pose estimation by up to 4.5 percent points (pp), hand pose estimation by up to 7.4 pp, and animal pose estimation by up to 4.8 pp for dogs and 3.3 pp for sheep. These results suggest that our method is able to mitigate domain shift on diverse tasks and even unseen domains and objects (*e.g.*, trained on horse and tested on dog). Our code will be publicly available at: https://github.com/VisionLearningGroup/UDA_PoseEstimation.

Keywords: Unsupervised Domain Adaptation; Pose Estimation; Semi-supervised Learning; Transfer Learning

1 Introduction

Recent developments in dense prediction tasks, *e.g.*, semantic segmentation [1, 4, 26, 33] or pose estimation [30, 36, 42], are limited by the difficulty in the acquisition of massive datasets [5, 6, 10, 16] due to the expensiveness as well as the unreliability that originates from the annotation phase. In addition, these models often perform poorly under domain shift. In this work, we address the problem of 2D pose estimation in the unsupervised domain adaptation (UDA) setting.

⋆ Equal Contribution.

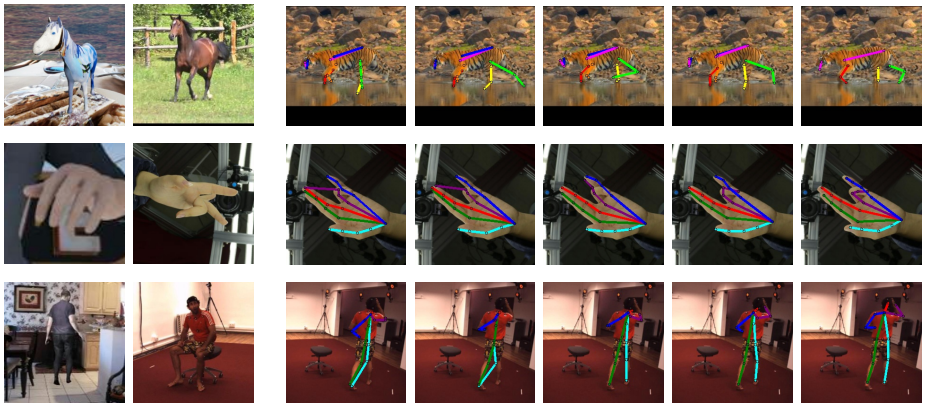


Fig. 1: (a) Top row: An example of high input-level variance in animal pose estimation benchmarks (large color and textual differences). Middle and bottom row: An example of high output-level variance in human and hand pose estimation benchmarks (large pose differences). (b) Visualization of pose estimation results from baselines, our method and ground-truth (GT). Note that both CCSSL and UDA-Animal(UDA-A) are proposed for animal pose estimation, while RegDA is only validated on hand and human pose estimation tasks. Most baseline methods suffer from performance degradation when applied to the other task. In comparison, our unified framework can more accurately estimate poses of hand, human and animal under various scenarios

The UDA setting allows us to train a pose estimation model with supervision from synthetic (source) domains, where data and accurate annotations are much cheaper to acquire, and optimize the model's performance on an unlabeled real (target) domain. Nevertheless, the domain gap between source and target domains due to distributional shift greatly undermines the ability of the model to transfer learned knowledge across different domains. This is a challenge that has been addressed previously for UDA for classificational tasks [14, 25, 27, 34].

Less attention has been paid to using UDA for regression tasks such as 2D pose estimation. Existing works are not generic but specifically target human pose estimation (RegDA [17]) or animal pose estimation (CCSSL [29], UDA-Animal [23]). A reason for this specialization may be the nature of the particular datasets used in those benchmarks. Animal datasets typically show large input-level variance (Fig. 1-(a)top) while human and hand datasets show large output-level variance (Fig. 1-(a)middle and bottom). Therefore, existing UDA methods do not generalize well to different objects of interest, for example, training and testing a human pose estimation model on an animal species or vice versa.

To address the aforementioned problems and keep the framework model-agnostic, we propose to bridge the domain gap via both input-level and output-

level adaptations, *i.e.*, alignments across domains in both the input and the output space of a pose estimation model. In input-level adaptation, we first translate images through a pre-trained style transfer model [15] that can extract similar visual features and bridge the gap between domains. In output-level adaptation, we borrow the architecture of Mean Teacher [8, 37] that enforces consistency in the output space of a student and a teacher model to generate reliable pseudo labels and learn from the unlabeled target domain.

As a typical approach for pose estimation, heatmap regression [38] predicts probabilities of the presence of keypoints in 2D space. However, unlike the output probabilities from other classification models that represent relative significance in the output space and sum to 1, the output heatmaps from a pose estimation model, which learns the task as predicting absolute value, are not normalized. The learning objectives of the student model, guided by the non-normalized output from the teacher model, will then be diverted from learning relative significance in the heatmap to learning absolute values, which is a more challenging task as the output space is no longer constrained. Therefore, the stability of the consistency learning is greatly undermined, and the lack of constraints leads to a problem we identify as a drift effect. Meanwhile, the drifted output heatmap also poses challenges while selecting confident guidance from the teacher model via the confidence thresholding method in Mean Teacher, as it potentially brings in noise that further deteriorates unsupervised learning. Therefore, we propose to normalize the output of the teacher model to make the guidance more stable. Our empirical results demonstrate the importance of this simple yet crucial step to deploy the Mean Teacher model for regression tasks.

In addition to revising consistency learning for the regression task, we design differing self-guiding strategies for student and teacher, developed especially for domain adaptive pose estimation. With style transfer, we generate target-like images from the source images and train a model to minimize the supervised loss with source labels. For the target domain, we generate source-like images from the target images to generate high-quality pseudo-labels from the teacher and give better guidance to the student model. In addition, in the student branch, we adaptively apply an occlusion mechanism, which has shown promising effectiveness especially in pose estimation tasks [7, 19, 43], based on the feedback of the teacher model. This strengthens the robustness of the pose estimation model.

In experiments we validate the effectiveness and generalization ability of our method under various scenarios including hand and human pose estimation as well as animal pose estimation. Our results show significant improvements over the existing domain adaptive pose estimation baselines by up to 4.5 percent point (pp) on hand pose, 7.4 pp on human pose estimation, and 4.8 pp for dog as well as 3.3 pp for sheep on animal pose estimation. Additionally, we present generalization experiments where we test models on unseen datasets or categories (*i.e.*, different animals), and verify the generalization capability. Further sensitivity analysis and ablation studies reveal the relation and interaction between modules and explain the effectiveness of each component of our unified framework. To summarize, our contributions in this work include:

- Unlike prior works, we propose a unified framework for general pose estimation that generalizes well on diverse objects in the pose estimation task.
- We propose a multi-level (*i.e.*, input-level and output-level) alignment method for domain adaptive pose estimation that can effectively address domain gap problems in different levels under different scenarios (*e.g.*, Fig. 1-(a)).
- We address the drifting problem in the Mean Teacher paradigm and facilitate its learning from unlabeled data especially for pose estimation tasks.
- We unified benchmarks from human pose estimation and animal pose estimation in this work and present state-of-the-art performance in general pose estimation, providing a stronger baseline in this line of research.

2 Related Works

2.1 Pose Estimation

Pose estimation has become an active research topic for years. In this paper, we focus on 2D pose estimation. Hourglass [30] is one of the dominant approaches for human pose estimation which applies an encoder-decoder style network with residual modules and finally generate heatmaps. A mean-squared error loss is applied between the predicted heatmap and ground-truth heatmap consisting of a 2D Gaussian centered on the annotated joint location [38]. Xiao *et al.* [42] propose a simple baseline model that combines upsampling and deconvolutional layers without using residual modules. HRNet [36] is proposed to maintain high-resolution in the model and achieves promising results. In this paper, we adopt the architecture of the Simple baseline model [42] following [17] to fairly compare our method with prior domain adaptation algorithms.

2.2 Unsupervised Domain Adaptation

Unsupervised Domain Adaptation (UDA) aims to bridge the domain gap between a labeled source domain and unlabeled target domain. Existing domain adaptation methods utilize adversarial learning [9, 28], minimize feature distances using MMD [11], optimal transport [2], pixel-level adaptation [13], or maximum classifier discrepancy [34] for classification. In addition several other UDA methods have been proposed for dense prediction tasks including semantic segmentation [14, 25, 39, 44] and depth estimation [21, 22, 32]. Compared to other visual tasks, domain adaptation for regression tasks are still not well explored.

2.3 Domain Adaptive Pose Estimation

There are two categories in domain adaptation pose estimation: (1) For human pose estimation, RegDA [17] made changes in MDD [45] for human and hand pose estimation tasks, which measures discrepancy by estimating false predictions on the target domain. (2) For animal pose estimation, pseudo-labeling based approaches have been proposed in [23, 29]. Mu *et al.* [29] proposed invariance and equivariance consistency learning with respect to transformations

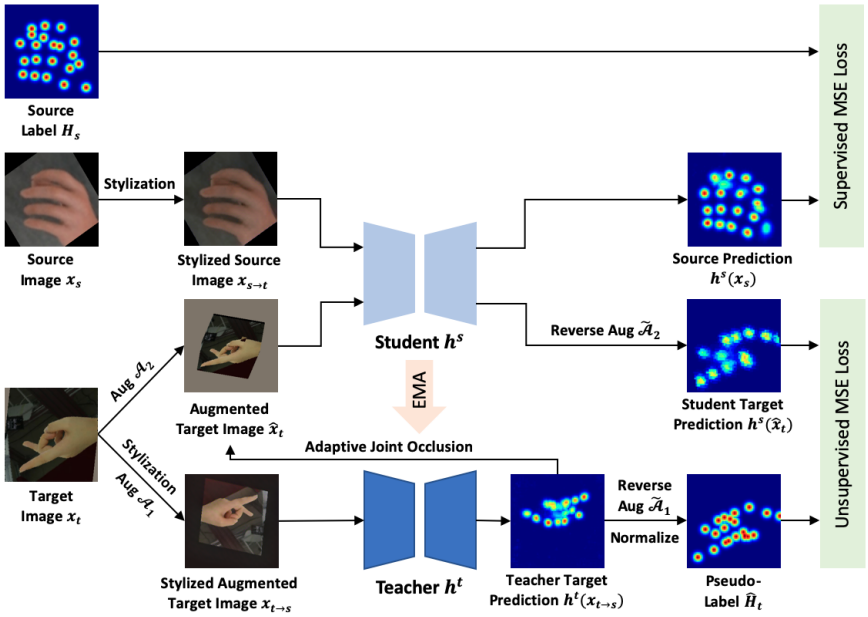


Fig. 2: An overview of our unified framework comprising a supervised branch that learns from source domain data with corresponding annotation, as well as an unsupervised branch that learns from unlabeled target domain data. We perform domain alignment both in the input-level via style-transfer with style references from the opposite domain, and the output-level of the model that guides the training on the target domain with more reliable pseudo-labels. The student model is trained by the combination of two losses, while the teacher model is updated with the exponential moving average weights of the student

as well as temporal consistency learning with a video. Li *et al.* [23] proposed a refinement module and a self-feedback loop to obtain reliable pseudo labels. Besides, WS-CDA [3] leverages human pose data and a partially annotated animal pose dataset to perform semi-supervised domain adaptation. In our experiments, we observed that (1) and (2) do not work well on the other tasks. A likely cause could be that each estimation task has different types of domain shifts, as shown in Fig 1(a). To address this, we propose a unified framework that generalizes well on diverse tasks by utilizing both input-level and out-level cues.

3 Method

3.1 Preliminaries

Given a labeled pose dataset $\mathcal{S} = \{(x_s^i, y_s^i)\}_{i=1}^N$ in source domain consisting of N pairs of images $x_s \in \mathbb{R}^{H\times W\times 3}$ and corresponding annotation heatmap $y_s \in$

$\mathbb{R}^{K\times 2}$ representing the coordinates of K keypoints, as well as an unlabeled pose dataset $\mathcal{T} = \{x_t^i\}_{i=1}^M$ in target domain consisting of M images $x_t \in \mathbb{R}^{H\times W\times 3}$, we aim to learn a 2D pose estimation model h and optimize the performance on the target domain. Typically, the pose estimation model h is pre-trained on the source domain dataset in a supervised manner to learn pose estimation from heatmaps $H_s = L(y_s)$, where $H \in \mathbb{R}^{K\times H'\times W'}$ with the output heatmap size H' and W', generated through the heatmap generating function $L : \mathbb{R}^{K\times 2} \to \mathbb{R}^{K\times H'\times W'}$, with classic MSE loss: $L_{sup} = \frac{1}{N}\sum_{x_s\in\mathcal{S}} ||h(x_s) - H_s||_2$.

3.2 Input-level Alignment via Style Transfer

Different from prior works [13, 14, 40] that adopt adversarial learning, we propose to perform input-level alignments via style transfer for the sake of efficiency and simplicity. We borrow notations from AdaIN [15] and follow its settings and training procedure to extract content features from a content image c and style feature from a style image s through a pre-trained VGG [35] model f. Formally, style transfer is performed with a generator g pre-trained as in AdaIN:

$$T(c, s, \alpha) = g(\alpha t + (1-\alpha)f(c)) \tag{1}$$

where $t = \text{AdaIN}(f(c), f(s))$ is the combination of content and style feature through adaptive instance normalization and α is the content-style trade-off parameter. Exemplar results are illustrated in the appendix. With a fixed AdaIN model, we transform source domain images with styles from target domain $x_{s\to t} = T(x_s, x_t, \alpha)$ and revise the supervised loss above:

$$L_{sup} = \frac{1}{N}\sum_{x_s\in\mathcal{S}} ||h(x_{s\to t}) - H_s||_2 \tag{2}$$

3.3 Output-level Alignment via Mean Teacher

To better exploit information from the unlabeled target domain, we adopt the paradigm of Mean Teacher that trains a student pose estimation model h^s by the guidance produced by its self-ensemble, i.e., the teacher pose estimation model h^t in an unsupervised learning branch. The input image for each model is augmented by $\mathcal{A}_1$ and $\mathcal{A}_2$ stochastically sampled from data augmentation $\mathcal{A}$. While the student h^s is updated according to the supervised loss in Eq. 2 and self-guidance from the teacher h^t, the weights of the latter are updated as the estimated moving average of the former.

On the opposite direction to the supervised learning branch that transforms the source image to the target domain, we also propose to transform the target domain image back to the direction of the source domain where supervised learning happens and bridge the domain gap when generating guidance from the teacher model. Formally, we take a source domain image as the style reference and generate $x_{t\to s} = T(\mathcal{A}_1(x_t), x_s, \alpha)$. After that, we pass the transformed image through the teacher model and get corresponding heatmap $H_t = h^t(x_{t\to s})$.

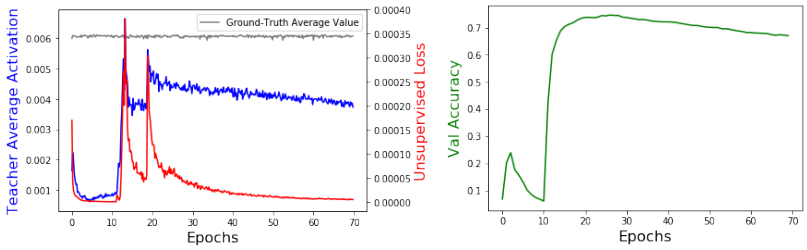


Fig. 3: Drift effect and its influence to the consistency learning. In the left plot, the gray curve represents the averaged value of the ground-truth heatmap. We observe that the averaged activation of teacher's output (blue curve) gradually decreases and drift away from the gray curve while minimizing the unsupervised loss (red curve). This leads to a degradation in accuracy as shown in the right plot

With the generated guidance heatmap from the teacher model, we still need to address the drifting effect that brings in instability in the unsupervised learning, as illustrated in Fig. 3. Technically, we generate pseudo-labels $\hat{H}_t = L(\hat{y}_t)$ with the positions that produce maximum activation $\hat{y}_t = \arg\max_p H_t^{:,p}$ from each keypoints of the guidance heatmap to normalize the heatmap. We also revise the typical thresholding mechanism using a fixed value in Mean Teacher and determine the confidence threshold τ_{conf} with the top $p\%$-th values among maximum activation from each keypoint to exclude noises and further improve the quality of the self-guidance.

In addition to improving the quality of the teacher's prediction, we also seek to challenge the student model by adaptively occluding the input to the student model according to feedback from the teacher. To be more specific, we mask the regions where the teacher model makes confident prediction of a keypoint with activation greater than τ_{occ} via an occlusion operation: $\hat{x}_t = O(\mathcal{A}_2(x_t), \tau_{occ})$, and let the student to learn robust prediction based on its contextual correlation with other keypoints from teacher's pseudo-label after reversing augmentations $\tilde{\mathcal{A}}_1$ and $\tilde{\mathcal{A}}_2$. Overall, the student model h^s will be guided by the normalized heatmap $\hat{H}_t$ via an unsupervised learning loss on keypoints k producing maximum activation $H_t^{k,\hat{y}_t}$ greater than or euqal to threshold τ_{conf}:

$$L_{unsup} = \frac{1}{M} \sum_{x_t \in \mathcal{T}} \sum_{k=0}^{K} \mathbb{1}(H_t^{k,\hat{y}_t} \geq \tau_{conf}) || \tilde{\mathcal{A}}_1(\hat{H}_t^{\,k}) - \tilde{\mathcal{A}}_2(h^s(\hat{x}_t)^k) ||_2 \tag{3}$$

Combining our supervised learning loss from Eq. 2 and unsupervised learning loss from Eq. 3, we present the illustration for the overall pipeline in Fig. 2 and the final learning objectives:

$$L = L_{sup} + \lambda L_{unsup} \tag{4}$$

4 Experiments

To verify the effectiveness and reliability of our method under various pose estimation scenarios (hand, human body, animals), we conducted experiments on benchmark datasets in those domains (Sec. 4.2) and compared our methods with SOTA baselines (Sec. 4.3). We also evaluated our method on domain generalization tasks where we tested our models on unseen domains (*i.e.*, different datasets) and objects (*i.e.*, different animals) (Sec. 4.4). Finally, we present a sensitivity analysis on hyper-parameters and ablation studies to analyze the contribution and interaction between each component in our paradigm (Secs. 4.5 and 4.6).

4.1 Experiment Protocols

We adopted the architecture of Simple Baseline [42] as our pose estimation model for both h^s and h^t, with backbone of pre-trained ResNet101 [12]. Following Simple Baseline and RegDA, we adopted Adam [20] as the optimizer and set the base learning rate as 1e-4. It decreased to 1e-5 at 45 epochs and 1e-6 at 60 epochs, while the whole training procedure consisted of 70 epochs. The batch size was set to 32 and there are in total 500 iterations for each epoch. The confidence thresholding ratio p is 0.5, while the occlusion thresholding value τ_{occ} is 0.9. The momentum η for the update of the teacher model is 0.999 and the unsupervised learning weight was set to 1 to balance the supervised and unsupervised loss to a similar level. Also, the model was only trained by the supervised loss on the source domain for the first 40 epochs. On the basis of augmentation in RegDA, we added rotation (-30°, 30°) and random 2D translation (-5%, 5%) for the input source and target domain images. Finally, it should be noted that we used the same hyper-parameters for all experiments, did not tune the number of training epochs on test sets, and always report the accuracy of models from the last epoch. As for the architecture and optimization procedure of the style transfer model, we follow settings in AdaIN, except that we pre-train the model bidirectionally, i.e., both source and target domain image can be a content or a style image. Additional details can be found in the appendix.

4.2 Dataset

Rendered Hand Pose Dataset [47] (RHD) provides $44k$ synthetic hand images including $41.2k$ training images and $2.7k$ test images along with corresponding 21 hand keypoints annotations. **Hand-3D-Studio** [46] (H3D) is a real-world multi-view indoor hand pose images dataset with $22k$ frames. We follow RegDA's policy to split $3.2k$ frames as the test set. **FreiHAND** [48] includes $44k$ frames of real-world multi-vew hand pose images with more varied pose and view points. It contains $130k$ training image, and we still follow settings in RegDA to select $32k$ test images. **SURREAL** [41] provides more than 6 million synthetic human body pose images with annotations. **Human3.6M** [16] contains 3.6 million frames of real-world indoor human body pose images captured from videos. We follow protocols in [24] and split 5 subjects (S1, S5, S6, S7, S8) as the training

Method	MCP	PIP	DIP	Fin	All
Source only	67.4	64.2	63.3	54.8	61.8
Oracle	97.7	97.2	95.7	92.5	95.8
CCSSL [29]	81.5	79.9	74.4	64.0	75.1
UDA-Animal [23]	82.3	79.6	72.3	61.5	74.1
RegDA [17]	79.6	74.4	71.2	62.9	72.5
Ours	**86.7**	**84.6**	**78.9**	**68.1**	**79.6**

Table 1: Prediction accuracy PCK@0.05 on *RHD→H3D*, i.e., source dataset is RHD, target dataset H3D, for four hand parts and the full hand. Higher values are better

set and 2 subjects (S9, S11) as test set. **Leeds Sports Pose** [18] (LSP) is a real-world outdoor human body pose dataset containing $2k$ images. **Synthetic Animal Dataset** [29] is a synthetic animal pose dataset rendered from CAD models. The dataset contains 5 animal classes, horse, tiger, sheep, hound, and elephant, each with $10k$ images. **TigDog Dataset** [31] includes $30k$ frames from real-world videos of horses and tigers. **Animal-Pose Dataset** [3] provides $6.1k$ real-world images from 5 animals including dog, cat, cow, sheep, and horse.

4.3 Experimental Results

Baselines. We consider the following SOTA baselines: semi-supervised learning based CCSSL [29], UDA-Animal [23], and RegDA [17] under various adaptation tasks. For the sake of fair comparison, we re-train CCSSL and UDA-Animal with the backbone of ResNet-101 as ours, and train CCSSL jointly among all categories in animal pose estimation tasks. Oracle is the performance of a model trained jointly with target 2D annotations following previous works

Metrics. We adopt the evaluation metric of Percentage of Correct Keypoint (PCK) for all experiments and report PCK@0.05 that measures the ratio of correct prediction within the range of 5% with respect to the image size.

Results on Hand Pose Estimation. First, we present the adaption results on the hand pose estimation task *RHD→H3D* on 21 keypoints. We report different anatomical parts of a hand including metacarpophalangeal (MCP), proximal interphalangeal (PIP), distal interphalangeal (DIP), and fingertip (Fin). Our baselines can greatly improve the performance of their pose estimation model on the target domain (Table 1), while UDA-Animal, which is originally proposed for animal pose estimation tasks, achieves a performance of 75.1%. In comparison, our method outperforms all the baseline methods by a noticeable margin of 4.5% and reaches 79.6%.

Results on Human Body Pose Estimation. As for the adaptation in human body pose estimation, we measure the performance of all baselines and ours in the task of *SURREAL→Human3.6M* and *SURREAL→LSP* on 16 keypoints on the human body grouped with different parts, *i.e.*, shoulders, elbow, wrist, hip, knee, and ankle. RegDA can successfully adapt its model closer to the target domain, while CCSSL and UDA-Animal, designed for animal pose estimation, fail to adapt under such scenarios (Table 2). This could probably be because their self-guidance paradigm is more hyper-parameter sensitive and cannot guarantee to generalize to other scenarios, including the high out-level variance (*i.e.*, high pose variance) in human pose estimation. Our method, in contrast, enables effective

Table 2: PCK@0.05 on *SURREAL→Human3.6M* and *SURREAL→LSP*. Sld: Shoulder, Elb: Elbow

Method	*SURREAL→Human3.6M*							*SURREAL→LSP*						
	Sld	Elb	Wrist	Hip	Knee	Ankle	All	Sld	Elb	Wrist	Hip	Knee	Ankle	All
Source only	69.4	75.4	66.4	37.9	77.3	77.7	67.3	51.5	65.0	62.9	68.0	68.7	67.4	63.9
Oracle	95.3	91.8	86.9	95.6	94.1	93.6	92.9	-	-	-	-	-	-	-
CCSSL [29]	44.3	68.5	55.2	22.2	62.3	57.8	51.7	36.8	66.3	63.9	59.6	67.3	70.4	60.7
UDA-Animal [23]	51.7	83.1	68.9	17.7	79.4	76.6	62.9	61.4	77.7	75.5	65.8	76.7	78.3	69.2
RegDA [17]	73.3	86.4	72.8	**54.8**	82.0	84.4	75.6	62.7	76.7	71.1	81.0	80.3	75.3	74.6
Ours	**78.1**	**89.6**	**81.1**	52.6	**85.3**	**87.1**	**79.0**	**69.2**	**84.9**	**83.3**	**85.5**	**84.7**	**84.3**	**82.0**

and robust unsupervised learning via the heatmap normalization which addresses the drift effect and therefore ensures the high quality of the self-guidance.

Results on Animal Pose Estimation. We finally compare our method with the baselines in domain adaptive animal pose estimation under *SynAnimal→TigDog* and *SynAnimal→AnimalPose* as shown in Tables 3 and 4. In *SynAnimal→TigDog*, we follow settings in UDA-Animal and estimate 18 keypoints from different parts including eye, chin, shoulder, hip, elbow, knee, and hoof of horse and tiger shared in the Synthetic Animal and the TigDog datasets. In *SynAnimal→AnimalPose*, we also perform adaptation on the hound and sheep categories for 14 keypoint estimation of eye, hoof, knee, and elbow. For a fair comparison, we run all experiments with the same data augmentation as in CCSSL and UDA-Animal for all tasks, as these augmentations provide crucial improvement (see first and second rows in Table 3). The first row in Table 3 represents the reported [23] source-only performance without augmentations; the second row with augmentation, which, e.g., increases the performance from 32.8% to 71.4% in the horse keypoint estimation (column All).

Among the baseline methods, UDA-Animal achieves the best performance in estimating a horse's pose and approaches the oracle performance from a model trained jointly by the annotated source and target domains. Our method achieves slightly lower performance in the horse set that is close to the oracle level but slightly outperforms UDA-Animal in the tiger set.

In despite of the promising results in *SynAnimal→TigDog*, we observe that UDA-Animal significantly underperforms than RegDA and ours in the AnimalPose dataset from Table 4. This is because *SynAnimal→AnimalPose* is more challenging than *SynAnimal→TigDog* by comparing the accuracy of source only models (32.2% vs. 71.4%). Even though we can still see improvements from the source only with augmentations, CCSSL and UDA-Animal face more noisy pseudo-labels during self-training possibly due to their hyper-parameter sensitivity, so that improvements are marginal. On the contrary, RegDA shows noticeable improvement compared to source only. Our method can handle these challenging settings via heatmap normalization in pseudo-labeling and obtain the best performance in these experiments in both categories.

Table 3: PCK@0.05 on *SynAnimal→TigDog*. Sld: shoulder, Elb: Elbow. Source only* indicates training on only source domain data with strong augmentation

Method	Horse								Tiger							
	Eye	Chin	Sld	Hip	Elb	Knee	Hoof	All	Eye	Chin	Sld	Hip	Elb	Knee	Hoof	All
Source only	49.3	53.5	31.3	53.5	38.7	28.7	18.3	32.8	42.8	32.1	24.2	51.1	32.6	28.1	32.7	33.2
Source only*	87.1	91.4	69.4	76.3	70.1	71.3	61.9	71.4	91.1	86.5	46.5	67.9	44.3	53.1	63.2	60.7
Oracle	92.0	95.8	73.6	90.9	84.4	84.2	79.1	84.1	98.5	97.4	75.1	94.7	74.1	76.0	81.6	82.1
CCSSL [29]	89.3	**92.6**	69.5	78.1	70.0	73.1	65.0	73.1	94.3	91.3	49.5	70.2	53.9	59.1	70.2	66.7
UDA-Animal [23]	86.9	93.7	**76.4**	**81.9**	70.6	**79.1**	**72.6**	**77.5**	98.4	87.2	49.4	**74.9**	49.8	62.0	73.4	67.7
RegDA [17]	89.2	92.3	70.5	77.5	71.5	72.7	63.2	73.2	93.3	92.8	50.3	67.8	50.2	55.4	60.7	61.8
Ours	**91.3**	92.5	74.0	74.2	**75.8**	77.0	66.6	76.4	**98.5**	**96.9**	**56.2**	63.7	**52.3**	**62.8**	**72.8**	**67.9**

Table 4: PCK@0.05 on *SynAnimal→AnimalPose*. Source only* indicates training on only source domain data with strong augmentation

Method	Dog					Sheep				
	Eye	Hoof	Knee	Elb	All	Eye	Hoof	Knee	Elb	All
Source only	39.8	22.8	16.5	17.4	22.0	42.6	31.0	28.2	21.4	29.3
Source only*	26.6	44.0	30.8	25.1	32.2	53.3	63.0	51.5	32.1	49.6
Oracle	88.8	74.9	57.1	51.1	65.1	88.2	84.9	79.9	59.6	76.9
CCSSL [29]	24.7	37.4	25.4	19.6	27.0	44.3	55.4	43.5	28.5	42.8
UDA-Animal [23]	26.2	39.8	31.6	24.7	31.1	48.2	52.9	49.9	29.7	44.9
RegDA [17]	46.8	54.6	32.9	31.2	40.6	**62.8**	68.5	57.0	42.4	56.9
Ours	**56.1**	**59.2**	**38.9**	**32.7**	**45.4**	61.6	**77.4**	**57.7**	**44.6**	**60.2**

4.4 Generalization to Unseen Domains and Objects

So far, we have focused on accuracy in a given target domain, but we may face other types of unseen domains during training in real-world applications. Thus, we compare the generalization capacity of our method with baselines in a domain generalization setting where we test models on unseen domains and objects.

Domain Generalization on FreiHAND. For hand pose estimation, we test models adapted on the RHD→H3D setting with the other real-world hand dataset FreiHAND (FHD). We compare the accuracy on FHD and measure how well each method generalizes on the unseen domain FHD. As presented in Table 5, the test performance on FHD is generally poor compared to the source only and oracle performance, presumably because of the larger domain gap between H3D and FHD. It is worth noticing the performance of CCSSL is lower than the source-only, even if it outperforms that in the *RHD→H3D* setting by a large margin, revealing its lack of generalization capacity to the unseen domain, probably because of the lack of input-level alignment. On the other hand, RegDA and our method show better ability to generalize while ours achieves the best performance under most circumstances.

Domain Generalization on Human3.6M. We test the generalization ability of a model adapted from *SURREAL→LSP* on Human3.6M. It should be noted that LSP contains only 2K images which are very small compared to Hu-

Table 5: Domain generalization experiments on FreiHand (FHD) and Human3.6M. We report PCK@0.05. Fin: Fingertip. Sld: shoulder, Elb: Elbow. Source only indicates training only on RHD or SURREAL while Oracle indicates training only on FHD or Human3.6M

Method	*FreiHand*					*Human3.6M*						
	MCP	PIP	DIP	Fin	All	Sld	Elb	Wrist	Hip	Knee	Ankle	All
Source only	34.9	48.7	52.4	48.5	45.8	51.5	65.0	62.9	68.0	68.7	67.4	63.9
Oracle	92.8	90.3	87.7	78.5	87.2	95.3	91.8	86.9	95.6	94.1	93.6	92.9
CCSSL [29]	34.3	46.3	48.4	44.4	42.6	52.7	76.9	63.1	31.6	75.7	72.9	62.2
UDA-Animal [23]	29.6	46.6	50.0	45.3	42.2	54.4	75.3	62.1	21.6	70.4	69.2	58.8
RegDA [17]	**37.8**	51.8	53.2	47.5	46.9	76.9	80.2	69.7	**52.0**	80.3	80.0	73.2
Ours	35.6	**52.3**	**55.4**	**50.6**	**47.1**	**77.0**	**85.9**	**73.8**	47.6	**80.7**	**80.6**	**74.3**

Table 6: Domain generalization experiments on AnimalPose. We report PCK@0.05. Source only indicates training only on Synthetic Animal

Method	Horse	Dog	Cat	Sheep	Cow	All
Source only	52.2	31.0	14.7	37.5	41.8	33.4
CCSSL [29]	59.8	31.1	16.6	46.4	48.9	37.7
UDA-Animal [23]	**63.2**	32.4	17.6	48.3	53.0	39.8
RegDA [17]	58.4	34.9	17.4	45.1	46.3	39.0
Ours	61.6	**40.7**	**21.6**	**50.1**	**53.5**	**44.0**

man3.6M. Thus, this task is challenging since we use small number of real data for domain generalization. In Table. 5, we show that our method can generalize better than the baselines and achieves 74.3% of accuracy. Our accuracy on the generalization task (74.3%) is also comparable to the baselines performances of *SURREAL→Human3.6M* (*e.g.*, RegDA: 75.6), by using only $2k$ images.

Domain Generalization on AnimalPose. Finally, we evaluate the generalization capacity of models adapted from *SynAnimal→TigDog* and test it on Animal Pose Dataset. It should be noted that models are only trained on horse and tiger images from the Synthetic Animal Dataset and tested on unseen animals (*e.g.*, dog) in Animal Pose Dataset. Based on the results in Table 6, we can also witness an obvious improvement of our method above all the baselines and generalize better on unseen animals from unseen domains.

Qualitative Results. We provide additional qualitative results on generalization in Figs. 4. In Fig. 4, it is clear that the baselines proposed for animal pose estimation do not work well. Our method produces more accurate keypoints compared to baselines. More qualitative results on animal are available in the appendix.

4.5 Sensitivity Analysis

To further validate the robustness and generalization capacity of our method, we conducted sensitivity analysis regarding three major hyper-parameters in our framework, including the confidence thresholding ratio p, occlusion thresholding value τ_{occ}, the momentum η in Mean Teacher on *RHD→H3D*. Additionally, we randomly split a separate validation set with the same size as the test set from the target domain training data to simulate the hyper-parameter tuning

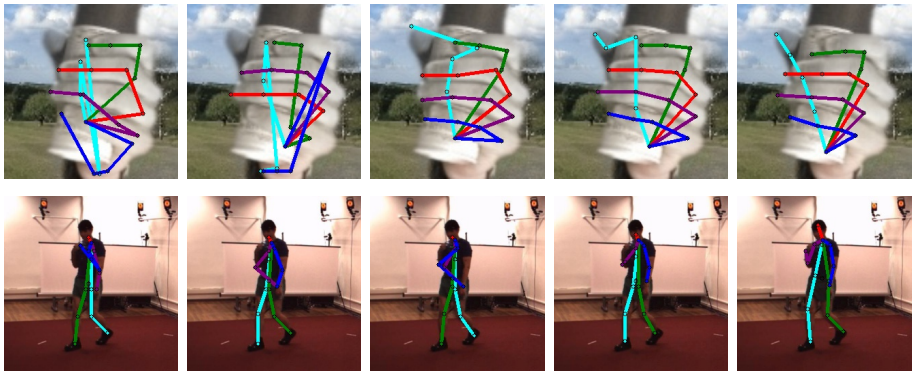
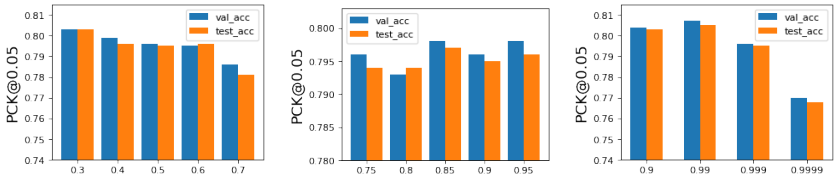


Fig. 4: Qualitative results of generalization to unseen domains



(a) Thresholding Ratio (b) Occlusion Threshold (c) Teacher Momentum

Fig. 5: Sensitivity analysis on the thresholding, occlusion ratio, and momentum. Our method shows stable performance over hyper-parameters

process and avoid directly tuning the test accuracy. Based on the results presented in Fig. 5, we find that our framework works stably under various settings. Meanwhile, we also find that the performance gradually decreases when we have a higher thresholding ratio for pseudo-labels, presumably because it brings in lower confident predictions as pseudo-labels and that deteriorates the unsupervised learning process. Also, we find that a greater teacher momentum is more likely to limit the framework to learn actively and harm the performance. More importantly, we can also learn that the validation accuracy in all experiments is highly correlated with that on the test sets, which also indicates the generalization capacity of our method and the reliability to give indicative clues when tuning hyper-parameters on a separate validation set.

4.6 Ablation Studies

We perform ablation studies in our framework to test their effectiveness and interaction with the rest of the framework. This also justify our other motivations regarding the task and the framework. Experiments are conducted under our

Table 7: Ablation studies on hand & animal pose estimation. Fin: Fingertip. MT: Mean Teacher, Norm: Heatmap Normalization, Style: Stylization, Occ: Adapt. Occlusion

Method	*RHD→H3D*					*SynAnimal→TigDog*							
	MCP	PIP	DIP	Fin	All	Eye	Chin	Sld	Hip	Elb	Knee	Hoof	All
MT	83.5	81.2	74.6	67.3	76.9	92.8	89.2	57.7	73.5	61.3	58.6	66.1	67.0
MT + Norm	86.1	84.4	77.2	67.2	78.8	91.9	89.9	59.3	62.7	60.8	67.6	64.1	68.1
MT + Style	84.6	82.5	76.6	66.9	77.6	95.0	93.8	57.8	74.7	63.5	67.4	67.4	70.4
MT + Norm + Style	86.6	84.4	78.3	68.1	79.1	95.9	94.7	65.7	68.2	64.9	71.7	72.3	73.4
MT + Norm + Style + Occ	86.7	84.6	78.9	68.1	79.6	95.7	94.7	64.1	69.0	64.5	70.7	69.8	72.4

major benchmarks including *RHD→H3D* and *SynAnimal→TigDog*. Additional ablation studies can be found in the appendix.

Based on Table 7, our framework can benefit from the heatmap normalization (denoted by Norm) that stabilizes the drifting effect and enables effective unsupervised learning from pseudo-labels via output-level domain alignment. Nevertheless, experiments on animal adaptation tasks show that such alignment might not be sufficiently helpful. Instead, more improvements are brought by the style transfer module, which confirms our reasoning that input-level variance is the major challenge in this task and can be mitigated by input-level alignments.

Adaptive occlusion can also provide extra focus on learning to detect occluded keypoints, as we can observe from *RHD→H3D*. However such improvements are not reflected in *SynAnimal→TigDog*. Considering the qualitative results in Figs. 1, we conjecture that it is because the improvements in detecting occluded keypoints are not verifiable as their annotations are not available in the real animal dataset and therefore these predictions are not included in the PCK@0.05 evaluation protocol. More ablation studies are available in the appendix.

5 Conclusion

While existing baselines focus on specific scenarios, we propose a unified framework that can be applied to diverse problems of domain adaptive pose estimation including hand pose, human body, and animal pose estimation. Considering the challenges from different types of domain shifts, our method addresses both input and output-level discrepancies across domains and enables a more generic adaptation paradigm. Extensive experiments demonstrate that our method not only achieves state-of-the-art performance under various domain adaptation scenarios but also exhibits excellent generalization capacity to unseen domains and objects. We hope our work can unify branches from different directions and provide a solid baseline for following works in this line of research.

Acknowledgements. This work has been partially supported by NSF Award, DARPA, DARPA LwLL, ONR MURI grant N00014-19-1-2571 associated with AUSMURIB000001 (to M.B.) and by NSF grant 1535797, 1551572, (to. M.B.).

References

1. Badrinarayanan, V., Kendall, A., Cipolla, R.: Segnet: A deep convolutional encoder-decoder architecture for image segmentation. IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI) pp. 2481–2495 (2017)
2. Bhushan Damodaran, B., Kellenberger, B., Flamary, R., Tuia, D., Courty, N.: Deepjdot: Deep joint distribution optimal transport for unsupervised domain adaptation. In: European Conference on Computer Vision (ECCV). pp. 447–463 (2018)
3. Cao, J., Tang, H., Fang, H., Shen, X., Tai, Y., Lu, C.: Cross-domain adaptation for animal pose estimation. In: IEEE International Conference on Computer Vision (ICCV). pp. 9497–9506 (2019)
4. Chen, L., Papandreou, G., Kokkinos, I., Murphy, K., Yuille, A.L.: Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs. IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI) pp. 834–848 (2018)
5. Cordts, M., Omran, M., Ramos, S., Rehfeld, T., Enzweiler, M., Benenson, R., Franke, U., Roth, S., Schiele, B.: The cityscapes dataset for semantic urban scene understanding. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 3213–3223 (2016)
6. Deng, J., Dong, W., Socher, R., Li, L., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 248–255 (2009)
7. Devries, T., Taylor, G.W.: Improved regularization of convolutional neural networks with cutout. arXiv preprint arXiv:1708.04552 (2017)
8. French, G., Mackiewicz, M., Fisher, M.H.: Self-ensembling for visual domain adaptation. In: International Conference on Learning Representations (ICLR) (2018)
9. Ganin, Y., Ustinova, E., Ajakan, H., Germain, P., Larochelle, H., Laviolette, F., Marchand, M., Lempitsky, V.: Domain-adversarial training of neural networks. Journal of Machine Learning Research (JMLR) pp. 2096–2030 (2016)
10. Geiger, A., Lenz, P., Stiller, C., Urtasun, R.: Vision meets robotics: The KITTI dataset. The International Journal of Robotics Research pp. 1231–1237 (2013)
11. Gretton, A., Sejdinovic, D., Strathmann, H., Balakrishnan, S., Pontil, M., Fukumizu, K., Sriperumbudur, B.K.: Optimal kernel choice for large-scale two-sample tests. Advances in Neural Information Processing Systems (NeurIPS) (2012)
12. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 770–778 (2016)
13. Hoffman, J., Tzeng, E., Park, T., Zhu, J., Isola, P., Saenko, K., Efros, A.A., Darrell, T.: Cycada: Cycle-consistent adversarial domain adaptation. In: International Conference on Machine Learning (ICML). pp. 1994–2003 (2018)
14. Hoffman, J., Wang, D., Yu, F., Darrell, T.: Fcns in the wild: Pixel-level adversarial and constraint-based adaptation. arXiv preprint arXiv:1612.02649 (2016)
15. Huang, X., Belongie, S.J.: Arbitrary style transfer in real-time with adaptive instance normalization. In: IEEE International Conference on Computer Vision (ICCV). pp. 1510–1519 (2017)
16. Ionescu, C., Papava, D., Olaru, V., Sminchisescu, C.: Human3.6m: Large scale datasets and predictive methods for 3D human sensing in natural environments. IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI) pp. 1325–1339 (2014)

17. 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)
18. Johnson, S., Everingham, M.: Clustered pose and nonlinear appearance models for human pose estimation. In: British Machine Vision Conference (BMVC). pp. 1–11 (2010)
19. Ke, L., Chang, M., Qi, H., Lyu, S.: Multi-scale structure-aware network for human pose estimation. In: European Conference on Computer Vision (ECCV). pp. 731–746 (2018)
20. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. In: International Conference on Learning Representations (ICLR) (2015)
21. Kundu, J.N., Lakkakula, N., Radhakrishnan, V.B.: Um-adapt: Unsupervised multi-task adaptation using adversarial cross-task distillation. In: IEEE International Conference on Computer Vision (ICCV). pp. 1436–1445 (2019)
22. Kundu, J.N., Uppala, P.K., Pahuja, A., Babu, R.V.: Adadepth: Unsupervised content congruent adaptation for depth estimation. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 2656–2665 (2018)
23. Li, C., Lee, G.H.: From synthetic to real: Unsupervised domain adaptation for animal pose estimation. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 1482–1491 (2021)
24. Li, S., Chan, A.B.: 3D human pose estimation from monocular images with deep convolutional neural network. In: Asian Conference on Computer Vision (ACCV). pp. 332–347. Lecture Notes in Computer Science, Springer (2014)
25. Li, Y., Yuan, L., Vasconcelos, N.: Bidirectional learning for domain adaptation of semantic segmentation. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 6936–6945 (2019)
26. Long, J., Shelhamer, E., Darrell, T.: Fully convolutional networks for semantic segmentation. In: IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015. pp. 3431–3440 (2015)
27. Long, M., Cao, Y., Wang, J., Jordan, M.I.: Learning transferable features with deep adaptation networks. In: International Conference on Machine Learning (ICML). pp. 97–105 (2015)
28. Long, M., Cao, Z., Wang, J., Jordan, M.I.: Conditional adversarial domain adaptation. In: Advances in Neural Information Processing Systems (NeurIPS). pp. 1640–1650 (2018)
29. Mu, J., Qiu, W., Hager, G.D., Yuille, A.L.: Learning from synthetic animals. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 12383–12392 (2020)
30. Newell, A., Yang, K., Deng, J.: Stacked hourglass networks for human pose estimation. In: European Conference on Computer Vision (ECCV). pp. 483–499 (2016)
31. Pero, L.D., Ricco, S., Sukthankar, R., Ferrari, V.: Articulated motion discovery using pairs of trajectories. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 2151–2160 (2015)
32. Rodriguez, A.L., Mikolajczyk, K.: DESC: domain adaptation for depth estimation via semantic consistency. In: British Machine Vision Conference 2020 (BMVC) (2020)
33. Ronneberger, O., Fischer, P., Brox, T.: U-Net: Convolutional networks for biomedical image segmentation. In: Medical Image Computing and Computer-Assisted Intervention, MICCAI. pp. 234–241 (2015)

34. Saito, K., Watanabe, K., Ushiku, Y., Harada, T.: Maximum classifier discrepancy for unsupervised domain adaptation. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 3723–3732 (2018)
35. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. In: International Conference on Learning Representations (ICLR) (2015)
36. Sun, K., Xiao, B., Liu, D., Wang, J.: Deep high-resolution representation learning for human pose estimation. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 5693–5703 (2019)
37. Tarvainen, A., Valpola, H.: Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. In: International Conference on Learning Representations (ICLR) (2017)
38. Tompson, J.J., Jain, A., LeCun, Y., Bregler, C.: Joint training of a convolutional network and a graphical model for human pose estimation. Advances in Neural Information Processing Systems (NeurIPS) (2014)
39. Tsai, Y., Hung, W., Schulter, S., Sohn, K., Yang, M., Chandraker, M.: Learning to adapt structured output space for semantic segmentation. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 7472–7481 (2018)
40. Tzeng, E., Hoffman, J., Saenko, K., Darrell, T.: Adversarial discriminative domain adaptation. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 2962–2971 (2017)
41. Varol, G., Romero, J., Martin, X., Mahmood, N., Black, M.J., Laptev, I., Schmid, C.: Learning from synthetic humans. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 4627–4635 (2017)
42. Xiao, B., Wu, H., Wei, Y.: Simple baselines for human pose estimation and tracking. In: European Conference on Computer Vision (ECCV). pp. 472–487 (2018)
43. Xie, R., Wang, C., Zeng, W., Wang, Y.: An empirical study of the collapsing problem in semi-supervised 2d human pose estimation. In: IEEE International Conference on Computer Vision (ICCV). pp. 11240–11249 (2021)
44. Yang, Y., Soatto, S.: FDA: fourier domain adaptation for semantic segmentation. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 4084–4094 (2020)
45. Zhang, Y., Liu, T., Long, M., Jordan, M.: Bridging theory and algorithm for domain adaptation. In: International Conference on Machine Learning (ICML). pp. 7404–7413. PMLR (2019)
46. Zhao, Z., Wang, T., Xia, S., Wang, Y.: Hand-3D-Studio: A new multi-view system for 3D hand reconstruction. In: IEEE International Conference on Acoustics, Speech and Signal Processing ICASSP. pp. 2478–2482 (2020)
47. Zimmermann, C., Brox, T.: Learning to estimate 3D hand pose from single RGB images. In: IEEE International Conference on Computer Vision (ICCV). pp. 4913–4921 (2017)
48. Zimmermann, C., Ceylan, D., Yang, J., Russell, B.C., Argus, M.J., Brox, T.: Freihand: A dataset for markerless capture of hand pose and shape from single RGB images. In: IEEE International Conference on Computer Vision (ICCV). pp. 813–822 (2019)