Alleviating Human-level Shift : A Robust Domain Adaptation Method for Multi-person Pose Estimation

Xixia Xu
Beijing Key Laboratory of Traffic
Data Analysis and Mining, Beijing
Jiaotong University, Beijing, 100044
19112036@bjtu.edu.cn

Qi Zou*
Beijing Key Laboratory of Traffic
Data Analysis and Mining, Beijing
Jiaotong University, Beijing, 100044
qzou@bjtu.edu.cn

Xue Lin
Beijing Key Laboratory of Traffic
Data Analysis and Mining, Beijing
Jiaotong University, Beijing, 100044
18112028@bjtu.edu.cn

ABSTRACT

Human pose estimation has been widely studied with much focus on supervised learning requiring sufficient annotations. However, in real applications, a pretrained pose estimation model usually need be adapted to a novel domain with no labels or sparse labels. Such domain adaptation for 2D pose estimation hasn't been explored. The main reason is that a pose, by nature, has typical topological structure and needs fine-grained features in local keypoints. While existing adaptation methods do not consider topological structure of object-of-interest and they align the whole images coarsely. Therefore, we propose a novel domain adaptation method for multi-person pose estimation to conduct the human-level topological structure alignment and fine-grained feature alignment. Our method consists of three modules: Cross-Attentive Feature Alignment (CAFA), Intra-domain Structure Adaptation (ISA) and Inter-domain Human-Topology Alignment (IHTA) module. The CAFA adopts a bidirectional spatial attention module (BSAM) that focuses on fine-grained local feature correlation between two humans to adaptively aggregate consistent features for adaptation. We adopt ISA only in semi-supervised domain adaptation (SSDA) to exploit the corresponding keypoint semantic relationship for reducing the intra-domain bias. Most importantly, we propose an IHTA to learn more domain-invariant human topological representation for reducing the inter-domain discrepancy. We model the human topological structure via the graph convolution network (GCN), by passing messages on which, high-order relations can be considered. This structure preserving alignment based on GCN is beneficial to the occluded or extreme pose inference. Extensive experiments are conducted on two popular benchmarks and results demonstrate the competency of our method compared with existing supervised approaches.

CCS CONCEPTS

• **Computing methodologies → Activity recognition and understanding**; *Transfer learning*;

KEYWORDS

Multi-person Pose Estimation; Domain Adaptation; Human-level Knowledge Alignment

ACM Reference Format:
Xixia Xu, Qi Zou, and Xue Lin. 2020. Alleviating Human-level Shift : A Robust Domain Adaptation Method for Multi-person Pose Estimation. In *28th ACM International Conference on Multimedia (MM '20), October 12–16, 2020, Seattle, WA, USA.*. ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3394171.3414040

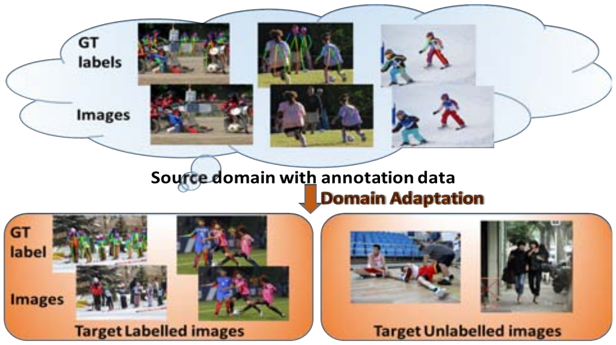


Figure 1: The examples with visual domain gap between the source and target domain.

1 INTRODUCTION

Multi-person pose estimation devotes to locate body parts for multiple persons in 2D image, such as keypoints on the arms, torsos, and the face[38]. It's fundamental to deal with other high-level tasks, such as human action recognition[6] and human-computer interaction[32]. Recently, due to the progress of convolution neural network (CNN)[14], most existing methods[4, 21, 38, 49] have achieved remarkable advances in multi-person pose estimation. However, existing supervised methods cannot generalize well to a novel domain without labels or with sparse labels, especially when the new domain has a different distribution. A natural remedy is the unsupervised domain adaptation (UDA). UDA has been widely applied in computer vision field, such as image classification[24], object detection[37], and semantic segmentation [15]. In these cases, a model trained on a source domain with full annotations is adapted for an unlabeled target domain via minimizing the distribution discrepancy of the features[23] or discriminating the output through

[*] Corresponding author.

MM '20, October 12–16, 2020, Seattle, WA, USA.
© 2020 Association for Computing Machinery.
ACM ISBN 978-1-4503-7988-5/20/10…$15.00
https://doi.org/10.1145/3394171.3414040

adversarial learning[10, 31, 44]. However, UDA for 2D pose estimation has never been explored. And applying the above domain adaptation methods into cross-domain pose estimation cannot guarantee the satisfying performance.

There are three types of adaptation challenges needed to be mitigated for cross-domain pose estimation: **1)** Pose estimation needs fine-grained local features. However, how to adapt these consistent fine-grained human body features across domains remains unexplored. As illustrated in Fig 1, humans in source and target images share much similar semantic representations such as postures, scales and actions, although their surrounding environments are quite different. It can be seen that feature adaptation at the image-level will meet difficulties under such cases. Existing domain adaptation methods widely used in image classification and semantic segmentation[15, 43] typically consider the image as a whole for alignment, while ignore local regions of the object-of-interest. Focusing on such local regions is important for pose estimation. Although some cross-domain object detection methods[53] focus on local objects, they aim at bridging the domain gap at a coarse granularity but not at the fine-grained keypoint level. Moreover, they often consider the domain features separately and neglect the local feature dependency across domains. **2)** Human topological structure is pivotal for adaptation performance especially in extreme poses or occlusions, while it's unrevealed yet. Existing domain adaptation methods for structural output[43, 44] adapt the output distribution and align the global layout across domains for semantic segmentation, which is much different from the human structure alignment in pose estimation. **3)** In semi-supervised setting, the gap between labeled and unlabeled data in the target domain also exists, as shown in Fig 1. Few works[18, 25] consider the intra-domain gap via adapting feature representations simply, let alone exploring the intra-domain structural keypoint relationship.

Aiming at these issues, we propose a novel domain adaptation framework as shown in Fig 2, which consists of three adaptation parts: (1) **C**ross-**A**ttentive **F**eature **A**lignment **(CAFA)**. To explore the similar fine-grained human body features and capture domain-invariant semantics across domains, we innovatively adopt a bidirectional spatial attention module (BSAM) to capture local feature similarity across humans. The local features of source human parts can be encoded in the target domain, and vice versa. It allows us to adaptively capture the consistent fine-grained human body features for adaptation. (2) **I**ntra-domain **S**tructure **A**daptation **(ISA)**. In SSDA setting, we exploit the annotations available in the target domain to learn the structural keypoint information, and adapt the reliable keypoint knowledge to the unlabeled data. Specifically, we align the one-to-one specific keypoint heatmap representation between the labeled and unlabeled data in target domain to augment the latter. (3) **I**nter-domain **H**uman-**T**opology **A**lignment **(IHTA)**. We adopt the recent SemGCN[52] to capture flexible human-topology representations. Besides, a sample selection mechanism is designed to determine which pairs should be aligned to avoid the hard alignment between arbitrary poses. Upon this, we align the human-topology representations to preserve structure-invariant knowledge across domains. The predicted errors in the target domain can be further repaired by the learned structure knowledge (e.g., the structural reasoning helps to infer the occluded or invisible poses).

Up to our knowledge, this is the first attempt that domain adaptation is explored under the multi-person pose estimation task. Our main contributions are as follows.

- A novel CAFA module achieves the fine-grained human body feature alignment and adapts abundant domain-invariant features for accurate pose estimation. We are also the first to investigate the transferability of fine-grained features via exploring the bidirectional spatial feature dependency across domains.
- In SSDA setting, a novel ISA adapts the local keypoint structural knowledge of the labeled to the unlabeled data in the target domain. It encourages the former to augment the keypoint representation of the latter to alleviate intra-domain confusions.
- The IHTA mechanism conducts human topological structure alignment softly to explicitly preserve high-order structure-invariant knowledge across domains. We additionally adopt a semantic graph-based formulation for modeling the human-topology information.
- Comprehensive experiments demonstrate the competency of the proposed method compared with the existing supervised approaches on two benchmark settings, i.e., "MPII to MS-COCO" and "MS-COCO to MPII".

2 RELATED WORK

Supervised Multi-Person Pose Estimation. Recently, multi-pers on pose estimation has aroused a great interest due to the real-life demand. Nowadays, researchers have made painstaking efforts [17, 29, 38, 40] to accelerate its progress. For examples, CASNet[38] improves the feature representation via adopting the spatial and channel-wise attention. HRNet[40] builds a new strong baseline via elaborated network design. However, they are all trained on adequate labeled images. Very few works explore the weakly/semi-supervised study in this field. The PoseWarper[2] leverages the sparse annotated training videos to perform temporal pose propagation and estimation. Although they bring significant improvement to recent benchmarks(e.g., MS-COCO[22] and MPII[1]), it's still hard to apply in practical applications due to the high-cost annotations. In such situation, domain adaptation offers an appealing solution by adapting pose estimator from label-rich source domain to the unlabeled or few labeled target domain.

Domain Adaptation. Domain adaptation utilizes a labeled source domain to learn a model that performs well on an unlabeled or sparse labeled target domain[9, 18, 51]. Most methods tackle UDA by minimizing the distance[23] across two distributions or aligning the output in adversarial learning[45]. For instance, [16] applies the adversarial strategy to align features for semantic segmentation. [48] applies the generator to transfer the source data to the target style for reducing the visual differences. On one hand, these methods simply align the global coarse-level features that benefit the classification task but it's hard to align the fine-grained human pose features. We thus innovatively propose a CAFA module to adapt the consistent fine-grained features across domains. On the other hand, these methods don't need consider the topological structure of the object-of-interest, which is essential for accurate pose estimation. With this derivation, we propose IHTA mechanism to exploit the human-topological relations across domains to better bridge the inter-domain discrepancy.

A plethora of SSDA works[18, 25, 36] have already emerged. For example, the [36] designs a min-max entropy minimization strategy to achieve better adaptation. Some works[18, 25] consider the intra-adaptation bias but they either leverage the labeled data to learn discriminative features like[18] or minimize the entropy similarity between intra-target samples like[25]. Directly applying these methods in pose estimation cannot fully explore the keypoint semantic relationship in intra-target domain. To this end, we propose to align the one-to-one keypoint heatmap vectors in target domain for avoiding the keypoint shift chaotically.

3 METHOD

3.1 Framework Overview

Problem Definition. We denote samples in the source domain and target domain as $X_s = \{x_s^i\}_{i=1}^M$ and $X_t = \{x_t^j\}_{j=1}^N$, where M, N are the sample numbers of each domain. Every source sample in X_s is annotated with corresponding keypoint annotations $Y_s = \{y_s^i\}_{i=1}^M$. In the semi-supervised setting, we have N_1 labeled images, N_2 unlabeled images and the labeled annotations are depicted as $Y_{tl} = \{y_t^j\}_{j=1}^{N_1}$. In unsupervised setting, $N_1 = 0$.

Overview. We adopt the modified SimpleBaseline[49] as our baseline, which utilizes an encoder-decoder architecture to make pose predictions as depicted in Fig 2. During training, given image triplets x_s^i, x_t^j and x_t^k, we generate the corresponding features f_s^i, f_t^j and f_t^k with a feature extractor. Through the CAFA module, we get the adapted features F_s, F_t. Then, we feed them into the estimator and predict the respective keypoint heatmaps. In semi-supervised setting, we put the heatmap outputs into ISA module to get the aligned heatmaps. Then, we turn the heatmaps into the local keypoint features. And we formulate the human pose as graph based on them to align the topological structure via IHTA module to achieve more precise predictions.

3.2 Cross-Attentive Feature Alignment

The fine-grained features are effective for the accurate pose estimation. The goal of CAFA is to adapt more domain-invariant fine-grained human features across domains. Different from previous feature adaptation methods, we capture the related fine-grained feature responses across domains via our BSAM. It explores the local spatial feature dependency across domains rather than simply consider the domain features separately. The fine-grained human features can be well encoded for each domain via exploring the feature interaction in a bidirectional manner.

In specific, we design a source-to-target adaptation (STA) to enhance source human body features by adaptively aggregating the target features based on their similarity. Similarly, we also adopt target-to-source adaptation (TSA) to update target features by aggregating relevant source features. The details of CAFA is depicted in the orange dashed-box in Fig 2.

Given the sample pairs x_s, x_t, we get the feature pairs F_s, F_t and apply two convolution layers to generate A and B, respectively. F_s, F_t are also fed into another convolution layer to obtain S_c, T_c. To determine fine-grained feature dependency between each position in F_s, F_t, an correlation map Φ is formulated as $\Phi = A^T B$, where $\Phi^{(i,j)}$ measure the similarity between i-th position in F_s and j-th position in F_t. To enhance F_s with the similar response from F_t and vice versa, the bidirectional adaptation is defined as follows.

Source-to-Target Adaptation. During the STA, we define the source-to-target spatial correlated map as,

$$\Psi_{s\to t}^{(i,j)} = \frac{exp(\Phi^{(i,j)})}{\sum_{j=1}^{H\times W} exp(\Phi^{(i,j)})}, \tag{1}$$

where $\Psi_{s\to t}^{(i,j)}$ represents the impact of i-th position in F_s to j-th position in F_t. To leverage the fine-grained features with similar spatial responses in the target domain, we update F_s as,

$$F_s^{'} = F_s + \lambda_s T_c \Psi_{s\to t}, \tag{2}$$

where λ_s leverages the importance of target-domain relevant spatial information and original source features. In this way, the target similar feature responses are well encoded in each position of $F_s^{'}$.

Target-to-Source Adaptation. Similarly, we obtain the target-to-source attentive map $\Psi_{t\to s}^{(i,j)}$ in Eq 1. It indicates the impact the j-th position in F_t attends to the i-th position in F_s. F_t is updated by combining the similar fine-grained source-domain repsonses and original target features like in Eq 2. In this manner, $F_s^{'}$ and $F_t^{'}$ enable us to encode more fine-grained features for each domain.

Loss and alignment. Finally, we apply the Maximum Mean Discrepancy **(MMD)**[13] to align $F_s^{'}$ and $F_t^{'}$ across domains in Eq 3.

$$\mathcal{L}_{fd} = \| \frac{1}{M}\sum_{i=1}^{M}\phi(F_{s,i}^{'}) - \frac{1}{N}\sum_{j=1}^{N}\phi(F_{t,j}^{'}) \|_{\mathcal{H}}^2, \tag{3}$$

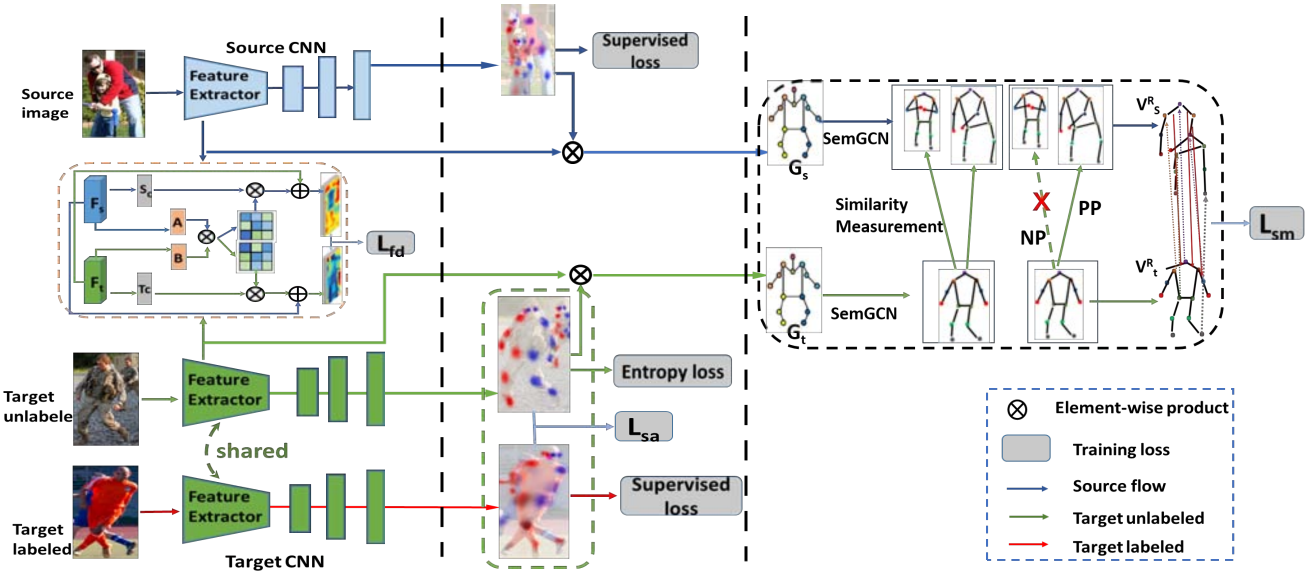
where ϕ is a map operation which projects the domain into a reproducing kernel Hilbert space $\mathcal{H}$ [12]. The arbitrary distribution of features can be represented by the kernel embedding technique. It allows us to learn domain-invariant and fine-grained human representations across domains by minimizing $\mathcal{L}_{fd}$.

3.3 Intra-domain Structure Adaptation

Under the SSDA setting, the **T**arget **L**abeled **(TL)** and **T**arget **U**nlabe led **(TU)** data have a quite potential relationship actually. On one hand, the scales, postures, or appearances of people are varied between them. On the other hand, they subject to a homogeneous distribution and possess similar specific keypoint information. Obviously, the TL is much amenable to acquire more accurate predictions than TU because it contains the detailed annotations. We hypothesize that our model can discover the underlying one-to-one keypoint semantic correspondence across them which benefits recognizing the vague keypoint locations of TU. The TL can provide more explicit guidance to facilitate TU to better rectify the inaccurate localizations (e.g., the confused keypoint locations of baseline with ISA is more explicit in Fig 7). Concretely, we devise ISA module to encourage TL to augment the corresponding keypoint representations of TU via calculating the cosine similarity of their heatmap vectors $y_t'^j$ and $y_t'^k$ as follows,

$$\mathcal{L}_{sa} = \sum_{j=1}^{N_1}\sum_{k=1}^{N_2}\sum_{h=1}^{H} \frac{y_t'^j(h)\cdot y_t'^k(h)}{\| y_t'^j(h) \| \cdot \| y_t'^k(h) \|}, \tag{4}$$

where H is the keypoint number and the $\mathcal{L}_{sa}$ measures the keypoint heatmap similarity of the same category between TL and TU. We



(a) Cross-Attentive Feature Alignment (b) Intra-domain Structure Adaptation (c) Inter-domain Human-Topology Alignment

Figure 2: Overview of the proposed method. The CAFA (in orange dashed line box) is placed at the encoder and learns domain-invariant fine-grained human features. The ISA (in green dashed line box) only works in SSDA setting to explore the keypoint heatmap alignment inside target domain. The IHTA (in a black dashed line box) works at the head of the network to model human pose with SemGCN[52] and further explore human-topology relations across domains to improve accuracy. It's noted that all the components in the dashed line box are only worked in the training phrase.

align them to force TL to guide TU with the exclusive semantic prior via minimizing the $\mathcal{L}_{sa}$.

3.4 Inter-domain Human-Topology Alignment

Although we align the one-order keypoint features of intra-domain, it cannot better conquer large pose discrepancy with severe geometric deformation, especially in heavily occluded cases across domains. However, the domain-invariant human-topology knowledge can provide a reliable guidance for mitigating this issue. We thus propose an IHTA mechanism to preserve this information. Since the human body structure provides the essential constraint information between joints, our IHTA is designed by GCN, which offers an explicit way of modeling the high-order human skeleton structure that is advantageous for capturing the spatial topological information of joints. It makes the cross-domain human-topology adaptation effective and conceivable. The details are as below.

Local Keypoint Feature Extraction. Specifically, we can get a group of semantic local features of keypoints V^{kp} according to the above feature maps F and keypoint heatmaps y_i^{kp} for both domains via an outer product($\otimes$) and a global average pooling in Eq 5.

$$V^{kp} = \{v_i\}_{i=1}^{H} = Global(F \otimes y_i^{kp}). \tag{5}$$

Graph Formulation. Here, we construct an intuitive graph $G = (V, E)$ based on the keypoint local features in Eq 5 for each human pose. V is the node set in G which can be denoted as $V = \{v_i, i = 1, 2, ..., H\}$. $E = \{v_i v_j \mid \text{if } i \text{ and } j \text{ are connected in the human body}\}$ is the edge set which refers to limbs of the human body. The adjacent matrix of G refers to matrix $A = a_{ij}$, with $a_{ij} = 1$ when v_i and v_j are neighbors in G or $i = j$, otherwise $a_{ij} = 0$.

Graph Convolution Network. Our key insight is that human body structure is a natural graph and there exists potential spatial constraint among joints. Hence, we model the human-topology representation via the recent SemGCN[52]. For a graph convolution, propagating features through neighbor joints helps to learn robust local structure and relation information between joints. Meanwhile, the non-local layer[46] helps capture the local and global long-range dependency among nodes to learn more human context information. It enables us to harvest robust human-topology informations, which are essential for learning structure-invariant information across domains.

A graph based convolutional propagation applies to node i in two steps. Firstly, node representations are transformed by a learnable parameter matrix $W \in R^{D_{l+1} \times D_l}$. Second, the transformed node representations are gathered to node i from its neighboring node j, followed by a RELU function. The node features are collected into a matrix $v^{(l)} \in R^{D_l \times H}$. Following the SemGCN[52], a different weighting matrix is applied to each channel d of node features:

$$v^{(l+1)} = \|_{d=1}^{D_{l+1}} \sigma(\vec{w}_d v^{(l)} \varphi_i(M_d \odot A)), \tag{6}$$

where $v^{(l)} \in R^{D_l}$ and $v^{(l+1)} \in R^{D_{l+1}}$ are the node representations before and after l-th convolution respectively, M_d is a set of $M \in R^{H \times H}$, which is a learnable weighting matrix compared with vanilla graph convolution. The weight vectors show the local semantic knowledge of neighboring joints implied in the graph. The $\|$ depicts channel-wise concatenation, and $\vec{w}_d$ is the d-th row of the transformation matrix W. It learns channel-wise weights for edges as priors in the graph (e.g., how one joint influences other body parts in pose estimation) to enhance the graph representations. And φ_i is Softmax nonlinearity which normalizes the input matrix across all choices of node i, $\odot$ is an element-wise operation which returns m_{ij} if $a_{ij} = 1$. A forces that for node i, we only compute the weights of its neighboring nodes j. Hence, the relationship of neighboring nodes are well considered.

Then, aside by the non-local concept[46] and we define the feature updating operation as:

$$v_i^{(l+1)} = v_i^{(l)} + \frac{W_v}{H}\sum_{j=1}^{H} f(v_i^{(l)}, v_j^{(l)}) \cdot g(v_j^{(l)}), \tag{7}$$

where W_v is initialized as zero; f is to compute the affinity between node i and all other j; g computes the node j representation. It computes responses between joints with their features to capture local and global long-range relationships among nodes.

Sample Selection Mechanism. It's unnecessary to conduct a hard alignment for the arbitrary poses in two domains since any two poses have the different shapes and geometric representations originally. Thus, we should predefine the 'similarity standard' for choosing the pose pairs should be aligned across domains to avoid nonsense alignment. Firstly, we calculate the averaged keypoint features similarity between two poses across domains Γ_{sim} in Eq 8,

$$\Gamma_{sim} = \frac{1}{H}\sum_{i=1}^{H}\sqrt{c_i^s c_{i'}^t}|v_i^s - v_{i'}^t|, \tag{8}$$

where the c_i^s and v_i^s indicate the confidence value and keypoint features of i-th keypoint of a source sample and the i' is the corresponding one in target domain. Then, we define a threshold value τ to judge whether the human pairs (m, n) are similar. As shown in Eq 9, if the similarity is above or equals to τ, we view them as positive pairs (PP) otherwise are negative pairs (NP). The value of τ is 0.7 here.

$$Pair_{(m,n)} \in \begin{cases} PP & \text{if } \Gamma_{sim} >= \tau \\ NP & \text{else.} \end{cases} \tag{9}$$

Cross-Graph Topology-Alignment. Based on the above steps, we conduct a cross-graph alignment to align the joint relation information learned by SemGCN[52] across two humans. Rather, given two samples x_s and x_t from both domains respectively, we firstly obtain the updated joint representations via Eq 6, 7. And followed by a 1×1 convolution, we get the final human-topology representations V_s^R, V_t^R. Finally, we choose the positive pairs via Eq 9 and align their V_s^R, V_t^R via Eq 10.

$$\mathcal{L}_{sm} = \sum_{i \in M}\sum_{k \in N_2} cosine(V_{s_i}^R, V_{t_k}^R), \tag{10}$$

where $V_{s_i}^R$ and $V_{t_k}^R$ indicate the topological representation of the i-th sample for each domain. We can learn the generalized high-order structure-invariant representations on both domains by minimizing $\mathcal{L}_{sm}$ to help reason the keypoint locations.

3.5 Optimization

The training of our network is to minimize a weighted combination of the aforementioned loss with respect to their parameter:

$$\mathcal{L}_{pose} = \beta_{sup}\mathcal{L}_{pose}^{sup} + \beta_{da}\mathcal{L}_{pose}^{da}, \tag{11}$$

where the weights β_{sup} and β_{da} are chosen empirically to strike a balance among the model capacity, and prediction accuracy.

Supervised Pose Loss. The supervised loss consists of the L_{pose}^s for labeled source data and the L_{pose}^{tl} for the TL prediction. The mean square error (MSE) is adopted as our regression loss:

$$\mathcal{L}_{pose}^s = \sum_{i=1}^{M}\sum_{h=1}^{H} \| y_{s_i}^h - y_{s_i}'^h \|_2^2, \tag{12}$$

$$\mathcal{L}_{pose}^{tl} = \sum_{j=1}^{N_1}\sum_{h=1}^{H} \| y_{t_j}^h - y_{t_j}'^h \|_2^2, \tag{13}$$

$$\mathcal{L}_{pose}^{sup} = \alpha_{tl}\mathcal{L}_{pose}^{tl} + \mathcal{L}_{pose}^s, \tag{14}$$

where α_{tl} is the trade-off hyperparameter and its value is 0.5, the Eq 13 is only used in semi-supervised setting.

Entropy Loss. Entropy minimization (ENT)[11] is a semi-supervised method assuming that the model is confident about its prediction for the unlabeled data. We adopt it as a regularizer and ensure that it maximally helps the TU achieve better performance in target domain. We add this term to the optimization in Eq 4:

$$\mathcal{L}_{ent}^{tu} = \sum_{k=1}^{N_2} ent(y_t'^k), \tag{15}$$

where $ent(p)$ calculates the entropy of distribution p.

Domain Adaptation Loss. In this case, the domain adaptation loss consists of the above three adaptation losses as follows:

$$\mathcal{L}_{pose}^{da} = \alpha_{sa}\mathcal{L}_{sa} + \alpha_{sm}\mathcal{L}_{sm} + \alpha_{fd}\mathcal{L}_{fd}, \tag{16}$$

where the α_{sa}, α_{sm}, α_{fd} are the weighted balance factors to keep the model effective and their values will be discussed in section 4.4.

4 EXPERIMENTS

4.1 Datasets and Evaluation metric

The proposed method is evaluated on two recent multi-person datasets: MPII Human Pose benchmark[1], COCO 2017 Keypoints Detection dataset[22].

COCO Keypoint Detection consists of the training set (includes $57K$ images), the test-dev set (includes around $20K$ images) and the validation set (includes $5K$ images). The MS-COCO evaluation metrics, OKS-based average precision (AP) and average recall (AR), are used to evaluate the performance. The OKS (object keypoints similarity) defines the similarity between the predicted heatmap and the groundtruth.

MPII Human Pose Dataset consists of images taken from real-world activities with full-body pose annotations. There are about $25K$ images with $40K$ objects, where there are $12K$ objects for testing and the remaining for the training set. We use the standard metric PCKh[1] (head-normalized probability of correct keypoint) score as evaluation. The PCKh@0.5 score is reported in our results, 50% of the head size for normalization.

In our experiment, there are two source-target settings:

1. Source: MS-COCO/ Target: MPII.
2. Source: MPII/ Target: MS-COCO.

4.2 Implementation Details

Network architectures. We adopt SimpleBaseline[49] and HRNet[40] as the pose estimation baseline respectively and the backbone uses the ResNet-50 in default. As for the SemGCN[52], the building block is one residual block[14] built by two SemGConv layers with 128 channels, followed by one non-local layer[46]. This block is repeated four times. All SemGConv layers are followed by Batch Normalization and a RELU activation except the last one.

Data Augmentation. We apply random flip, rotation, and scale in our training stage. The flip value is 0.5. The scale range is ([0.7 ~ 1.3]), and the random rotation range is ([−40°C ~ +40°C]).

Training. We implement all the experiment in PyTorch[34] on a single NVIDIA TITAN XP GPU with 12 GB memory. We select the ResNet-50, 101, 152 as the backbones which are all initialized with the weights of the ImageNet[35] pretrained model. We use Adam optimizer[19] with learning rate 10^{-4}, momentum 0.9 and weight decay 10^{-4} to train our model. We train the model for 150 epochs taking 96 hours for MS-COCO and we resize the images to 256×256 and trained for 135 epochs taking 10 hours for MPII.

It's noted that we don't evaluate the results on MS-COCO test-dev dataset due to it need to test online and unequal keypoint numbers across dataset makes it harder. We unify the MS-COCO keypoint numbers with the MPII finally.

4.3 Domain Adaptation Performance

We compare the performance of our method with the supervised approaches on both domains. For comparison, we take the Direct Transfer (DT) that trained with the source only as the baseline. To further shed light on the effectiveness of our method, we list the results of the model trained without the annotations absolutely (UDA model) and with the few labeled data in target domain (Adaptation model) respectively. Our method brings competitive performance with either data setting.

MPII to MS-COCO. The table 1 reports the results on MPII to MS-COCO. Although the scale of the source domain is much less than the target and the complexity and difficulty are also inferior to it, the UDA model can still achieve 64.4% AP, which is higher 7.7% than baseline although the target domain without any supervision. It powerfully proves that our method is conductive to weaken the influence from the irrelevant source information thereby encouraging relevant knowledge transfer among shared representations even without label. Although we adopt only 40% target labeled data, it works effectively outperforming UDA by 2.1%, indicating that ISA further alleviates the intra-domain bias. Moreover, the accuracy also improves despite we adopt HRNet[40] as the baseline model, which further illustrates our method with pretty generality.

The qualitative visualization is shown in Fig 3. It shows that our method achieves impressive results compared with the baseline, and achieves a comparable performance with the supervised model (e.g., the explicit predictions highlighted with the red circle/box). We also did visual comparisons with other recent pose estimation methods as shown in Fig 4, it can further manifest the robustness of our method.

Table 1: The comparison results on MPII to MS-COCO

Method	Backbone	AP	AP.5	AP.75	AP(M)	AP(L)	AR
Supervised methods (Target only)							
G-RMI[33]	-	65.7	83.1	72.1	61.7	72.5	69.9
MultiPoseNet[20]	-	69.6	86.3	76.6	65.0	76.3	73.5
AE[27]	Hourglass	66.3	86.5	72.7	61.3	73.2	71.5
CSANet[38]	ResNet-50	72.1	-	-	-	-	-
SBN[49]	ResNet-50	70.4	88.6	78.3	67.1	77.2	76.3
HRNet[40]	HRNet-W32	74.4	90.5	81.9	70.8	81.0	79.8
HRNet[40]	HRNet-w48	75.1	90.6	82.2	72.5	81.8	80.4
Direct Transfer (Source only)							
SBN	ResNet-50	56.7	72.3	64.4	54.3	53.4	62.2
HRNet	HRNet-w32	59.1	75.3	66.7	55.4	66.3	64.3
HRNet	HRNet-w48	59.5	75.2	66.8	55.6	66.8	64.9
UDA model (Without ISA module)							
SBN	ResNet-50	64.4	80.4	71.2	62.3	71.4	70.4
HRNet	HRNet-w32	67.2	83.3	74.6	63.4	74.3	72.7
HRNet	HRNet-w48	67.8	83.2	74.3	63.6	74.6	73.4
Adaptation model (Ours)							
Ours(SBN)	ResNet-50	66.5	82.3	73.2	64.1	73.5	72.3
Ours(HRNet)	HRNet-w32	69.2	85.2	76.5	65.6	76.1	74.6
Ours(HRNet)	HRNet-w48	69.9	85.3	76.4	65.7	76.5	75.1

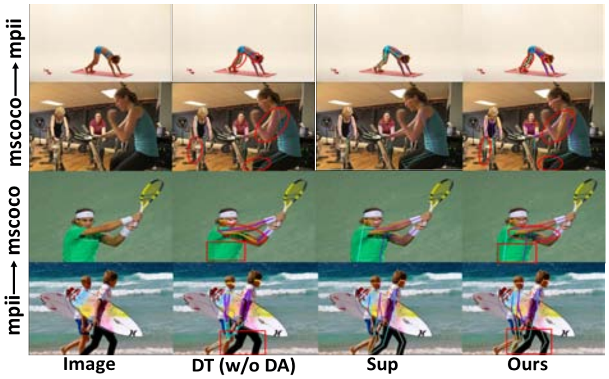


Figure 3: Results comparison of the baseline (DT (w/o DA)) versus the supervised model (Sup) and Ours on both cases.

MS-COCO to MPII. We evaluate the PCKh@0.5 score on MSCOCO to MPII in Tab 2 for comparison. The UDA model achieves 85.2% PCKh@0.5 and improves 7.9% points without annotations than the baseline. It indicates that minimizing the domain distribution difference is essential for tackling pose domain shift. Additionally, 40% labeled data are adopted for target domain (Ours) can further improves PKCh@0.5 to 87.8, which outperforms the baseline 10.5%. It proves that our method ensures reliable knowledge adaptation in different domains and receives the decent localized results. We also test on the HRNet-w32 and obtain the best result 89.4.

As shown in Fig 3, our approach makes more accurate predictions than DT and its supervised model (*e.g.*, the localized joints highlighted with red box) and also achieves a competitive result with the other popular estimation methods like in Fig 4.

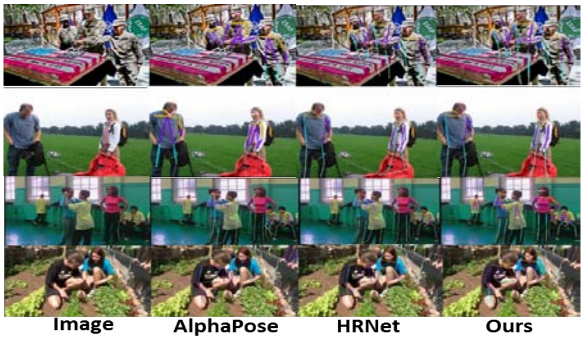


Figure 4: The visual comparisons between existing methods (they are under fully-supervised setting and without adaptation) and Ours, the samples are randomly selected from the MPII/MS-COCO.

Table 2: The comparison results on MS-COCO to MPII

Method	Head	Shoulder	Elbow	Wrist	Hip	Knee	Ankle	Total
Supervised methods (Target only)								
Wei *et al.*[47]	97.8	95.0	88.7	84.0	88.4	82.8	79.4	88.5
Newell *et al.*[28]	98.2	96.3	91.2	87.2	89.8	87.4	83.6	90.9
Sun *et al.*[39]	98.1	96.2	91.2	87.2	89.8	87.4	84.1	91.0
Tang *et al.*[42]	97.4	96.4	92.1	87.7	90.2	87.7	84.3	91.2
Ning *et al.*[30]	98.1	96.3	92.2	87.8	90.6	87.6	82.7	91.2
Luvizon *et al.*[26]	98.1	96.6	92.0	87.5	90.6	88.0	82.7	91.2
Chu *et al.*[7]	98.5	96.3	91.9	88.1	90.6	88.0	85.0	91.5
Chou *et al.*[5]	98.2	96.8	92.2	88.0	91.3	89.1	84.9	91.8
Chen *et al.*[3]	98.1	96.5	92.5	88.5	90.2	89.6	86.0	91.9
Yang *et al.*[50]	98.5	96.7	92.5	88.7	91.1	88.6	86.0	92.0
Ke *et al.*[17]	98.5	96.8	92.7	88.4	90.6	89.3	86.3	92.1
Tang *et al.*[41]	98.4	96.9	92.6	88.7	91.8	89.4	86.2	92.3
SBN *et al.*[49]	98.5	96.6	91.9	87.6	91.1	88.1	84.1	91.5
HRNet *et al.*[40]	98.6	96.9	92.8	89.0	91.5	89.0	85.7	92.3
Direct Transfer (Source only)								
SBN-ResNet50	95.2	89.8	79.7	72.5	75.8	67.8	60.5	77.3
HRNet-w32	95.8	90.3	80.5	74.3	77.6	69.7	62.8	79.6
UDA model (Without ISA module)								
SBN-ResNet50	96.8	96.5	86.1	81.0	86.4	80.2	73.1	85.2
HRNet-w32	95.8	94.5	87.8	83.5	87.5	82.5	79.1	86.4
Adaptation model (Ours)								
Ours(SBN)	97.5	94.4	86.8	82.2	87.2	81.2	73.9	87.8
Ours(HRNet-w32)	97.2	95.6	89.8	84.6	88.7	83.8	80.2	89.4

4.4 Ablation Study for Design Modules

To analyse the effectiveness of each component, we conduct experiments to evaluate their contributions. The SimpleBaseline[49] (ResNet-50) is adopted for our default baseline. The ablated results are illustrated in Table 3, the SL depicts model that both trained and tested on the target domain. DT is stated as above. Ours contains all components.

Does our method really learn fine-grained and domain-invariant human features? Table 3 shows that CAFA delivers a large performance gain than DT in all cases. The result on MPII to MS-COCO achieves 63.4% AP outperforming DT by 6.7% solely with CAFA. The similar improvements can also be observed on MS-COCO to MPII. The PKCh@0.5 score grows to 84.5, which achieves 7.2% higher than DT. This indicates mitigating the feature discrepancy across domains is crucial for addressing domain shift. We also compare CAFA with the other feature adaptation stretagies. The result demonstrates our CAFA achieves outstanding performance even compared with the recent strong alignment strategy[37].

Table 3: Component comparisons using MPII / MS-COCO as the source dataset and MS-COCO / MPII as the target dataset.

Methods	MS-COCO→ MPII	MPII→ MS-COCO
SL(Trained on Target)	**91.5**	**70.4**
DT(Trained on Source)	77.3	56.7
DT+DAN[23]	79.1	57.2
DT+Adversial DA[10]	79.5	58.1
DT+SW Detection[37]	81.6	60.1
DT+CAFA	84.5	63.4
DT+ISA	83.6	62.5
DT+IHTA	85.6	64.8
Ours	87.8	66.5

To illustrate CAFA can adapt the fine-grained human features via BSAM mechanism, we visualize the feature maps on both domains in Fig 5. As shown in red-dashed boxes, many feature maps of baseline are confused or noisy, e.g., the human part regions are ambiguous, or irrelevant regions are activated. On the contrary, our BSAM can effectively discover the similar fine-grained body parts features for each domain and enhance the respective features.

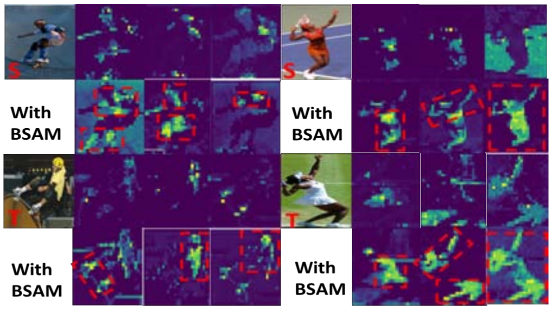


Figure 5: Visualization of features with or without (baseline) our BSAM, best viewed in color.

To further manifest that CAFA can produce domain-invariant features. We visualize the features of MS-COCO to MPII learned from ResNet, DAN, Adversarial DA and Ours respectively with t-SNE[8] in Figure 6 (a)-(d). From left (ResNet) to right (Ours), the source and target domains become more and more indistinguishable. Firstly, the features of "ResNet-50" are not well aligned in both domains. For "DAN"[23], two domains are aligned somewhat, however, the structure of target features is scattered and the shared features are not well aligned. As for "adversarial DA"[10], the target features are well preserved, but the shared features are not well aligned. For ours, the shared feature structure are compact and better aligned while the instance features of each domain are indistinguishable, which clearly evidences CAFA well captures domain-invariant features.

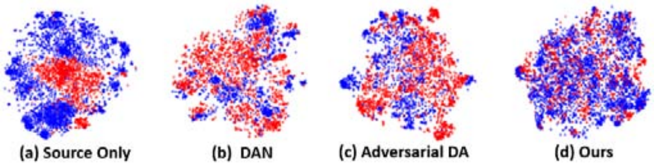


Figure 6: Visualization of features using t-SNE: (a) the baseline (b) DAN (c) adversarial DA (d) Ours. Note that the blue and red points are samples from the source and target domain respectively, best viewed in color.

The effectiveness of the ISA. To better elaborate that the similar pose representations from the labeled data can provide a reliable semantic guidance for the unlabeled, we conduct experiment reporting the result in Tab 3. It verifies our model with ISA achieves substantial improvements over the baseline in two adaptation cases. Concretely, the accuracy increases from 77.3% to 83.6% directly on MS-COCO to MPII and 56.7% to 62.5% on MPII to MS-COCO.

As shown in Fig 7, we notice that the keypoint locations produced by DT are ambiguous and inaccurate. By contrast, the refined heatmaps with the ISA are more accurate and explicit. The ISA can correct subtle localized confusions of unlabeled data aided by the keypoint structural information learned from the labeled data. It exactly demonstrates our ISA well diminishes the intra-domain aliasing.

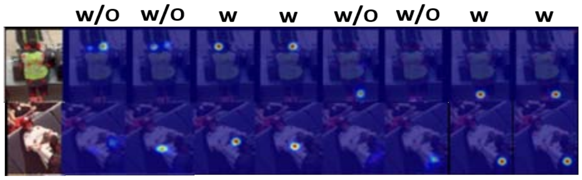


Figure 7: The keypoint heatmap predictions visualization. The 'w/o' depicts the results of DT, 'w' depicts with ISA.

Analysis of proposed IHTA. Table 3 shows that DT model behaves unsatisfying and only achieves poor result. The result with IHTA is improved by 8.3% and 8.1% on the MS-COCO to MPII and MPII to MS-COCO respectively. This obviously proves the effectiveness of the IHTA. Another innovation of IHTA is to formulate the graph model capturing the spatial relation information between joints. The result indirectly proves that adopting the graph model is robust and effective. Compared with the other components, IHTA makes the greatest contribution. It demonstrates the significance of exploring the high-order topological relation across domains.

We also visualize the comparison results in Fig 8, we observe that IHTA can fix some challenging structural errors (e.g., the occluded right ankle in Fig 8(A) and the left hip/shoulder in Fig 8(C)) via the human topology knowledge learned from the source domain.

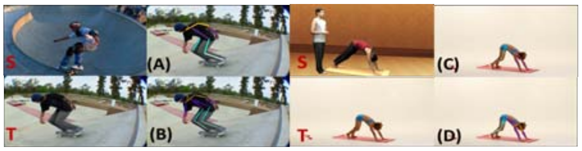


Figure 8: The qualitative comparisons of IHTA. "S" and "T" refer to the source and target image. (A)/(C) represent results generated by DT model. (B)/(D) depict results with IHTA.

Exploring the proper value of the objective weights. Firstly, we analyze the proper value of the objective weights α, β in table 4. We investigate their value by varying in $\{0.3 \sim 0.9\}$. Besides, we testify the result on MPII to MS-COCO and similar robust behavior can also be verified on MS-COCO to MPII. For α, as it varies from 0 to 0.5, the prediction accuracy on MS-COCO increases. It is desirable that when MS-COCO dataset is well-aligned with the MPII, while increasing α much, it preserves more general knowledge leading to negative transfer.

Table 4: Sensitivity of mAP to the hyper-parameter α, β on MPII to MS-COCO.

$\alpha\|\beta$	0.3	0.4	0.5	0.6	0.7	0.8	0.9
mAP(α)	60.0	61.2	**64.8**	64.1	63.5	62.2	61.5
mAP(β)	58.4	62.3	**64.8**	63.8	63.2	62.7	62.0

As for β, the accuracy also shows a growing trend as the value changes from 0.3 to 0.5. It illustrates that our adaptation method benefits adaptation performance improvement. When it's beyond 0.5, it means overemphasizing the contribution of the supervised optimization leads to insufficient adaptation, which results in performance drop. This shows that proper trade-off will enhance effective knowledge transfer across domains.

5 CONCLUSION

In this paper, we propose a novel domain adaptation framework to address the multi-person pose estimation problem. In the feature level, we adopt a cross-attentive feature alignment strategy to learn the well-aligned fine-grained human features for adaptation. To better exploit the human topological structure, we model the human-topology structure via GCN and conduct cross-graph topology alignment across domains to preserve the structure-invariant knowledge. In SSDA setting, we additionally propose to adapt the corresponding keypoint heatmap representations to reduce the intra-domain gap. Extensive experiments show that our method significantly boosts performance of the target domain even with no labels or sparse labels. We also hope our method could inspire more ideas on the cross-domain pose estimation field in the future.

ACKNOWLEDGMENTS

This research is supported by the Fundamental Research Funds for the Central Universities (2020YJS037).

REFERENCES

[1] Mykhaylo Andriluka, Leonid Pishchulin, Peter V Gehler, and Bernt Schiele. 2014. 2D Human Pose Estimation: New Benchmark and State of the Art Analysis. *In Proceedings of the IEEE conference on computer vision and pattern recognition* (2014), 3686–3693.

[2] Gedas Bertasius, Christoph Feichtenhofer, Du Tran, Jianbo Shi, and Lorenzo Torresani. 2019. Learning Temporal Pose Estimation from Sparsely-Labeled Videos. *In Advances in neural information processing systems* (2019), 3021–3032.

[3] Yu Chen, Chunhua Shen, Xiushen Wei, Lingqiao Liu, and Jian Yang. 2017. Adversarial PoseNet: A Structure-Aware Convolutional Network for Human Pose Estimation. *In Proceedings of the IEEE International Conference on Computer Vision* (2017), 1221–1230.

[4] Yilun Chen, Zhicheng Wang, Yuxiang Peng, Zhiqiang Zhang, Gang Yu, and Jian Sun. 2018. Cascaded Pyramid Network for Multi-person Pose Estimation. *In Proceedings of the IEEE conference on computer vision and pattern recognition* (2018), 7103–7112.

[5] Chiajung Chou, Juiting Chien, and Hwanntzong Chen. 2018. Self Adversarial Training for Human Pose Estimation. *Asia Pacific Signal and Information Processing Association Annual Summit and Conference* (2018), 17–30.

[6] Vasileios Choutas, Philippe Weinzaepfel, Jerome Revaud, and Cordelia Schmid. 2018. PoTion: Pose MoTion Representation for Action Recognition. In *In Proceedings of the IEEE conference on Computer Vision and Pattern Recognition.*

[7] Xiao Chu, Wei Yang, Wanli Ouyang, Cheng Ma, Alan L Yuille, and Xiaogang Wang. 2017. Multi-context Attention for Human Pose Estimation. *In Proceedings of the IEEE conference on computer vision and pattern recognition* (2017), 5669–5678.

[8] Laurens Van Der Maaten and Geoffrey E Hinton. 2008. Visualizing Data using t-SNE. *Journal of Machine Learning Research* (2008), 2579–2605.

[9] Yaroslav Ganin and Victor Lempitsky. 2015. Unsupervised Domain Adaptation by Backpropagation. *In Proceedings of the 32th International Conference on Machine Learning* (2015), 1180–1189.

[10] Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, Francois Laviolette, Mario Marchand, and Victor Lempitsky. 2016. Domain-adversarial training of neural networks. *Journal of Machine Learning Research* (2016), 189–209.

[11] Yves Grandvalet and Yoshua Bengio. 2004. Semi-supervised Learning by Entropy Minimization. *In Advances in neural information processing systems* (2004), 529–536.

[12] Arthur Gretton, Karsten M Borgwardt, Malte J Rasch, Bernhard Scholkopf, and Alexander J Smola. 2006. A Kernel Method for the Two-Sample-Problem. *In Advances in neural information processing systems* (2006), 513–520.

[13] Arthur Gretton, Karsten M Borgwardt, Malte J Rasch, Bernhard Scholkopf, and Alexander J Smola. 2012. A kernel two-sample test. *Journal of Machine Learning Research* (2012), 723–773.

[14] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep Residual Learning for Image Recognition. *In Proceedings of the IEEE conference on computer vision and pattern recognition* (2016), 770–778.

[15] Judy Hoffman, Eric Tzeng, Taesung Park, Junyan Zhu, Phillip Isola, Kate Saenko, Alexei A Efros, and Trevor Darrell. 2018. CyCADA: Cycle-Consistent Adversarial Domain Adaptation. *In Proceedings of the 35th International Conference on Machine Learning* (2018), 1989–1998.

[16] Judy Hoffman, Dequan Wang, Fisher Yu, and Trevor Darrell. 2016. FCNs in the Wild: Pixel-level Adversarial and Constraint-based Adaptation. *arXiv preprint arXiv: 1612.02649* (2016).

[17] Lipeng Ke, Mingching Chang, Honggang Qi, and Siwei Lyu. 2018. Multi-Scale Structure-Aware Network for Human Pose Estimation. *In European Conference on Computer Vision. Springer,* (2018), 731–746.

[18] Donghyun Kim, Kuniaki Saito, Taehyun Oh, Bryan A Plummer, Stan Sclaroff, and Kate Saenko. 2020. Cross-domain Self-supervised Learning for Domain Adaptation with Few Source Labels. *arXiv preprint arXiv: 2003.08264* (2020).

[19] Diederik P Kingma and Jimmy Ba. 2015. Adam: A Method for Stochastic Optimization. *In International Conference on Learning Representations* (2015).

[20] Muhammed Kocabas, Salih Karagoz, and Emre Akbas. 2018. MultiPoseNet: Fast Multi-Person Pose Estimation using Pose Residual Network. *In European Conference on Computer Vision. Springer,* (2018), 437–453.

[21] Jiefeng Li, Can Wang, Hao Zhu, Yihuan Mao, Haoshu Fang, and Cewu Lu. 2019. CrowdPose: Efficient Crowded Scenes Pose Estimation and A New Benchmark. *In Proceedings of the IEEE conference on computer vision and pattern recognition* (2019).

[22] Tsungyi Lin, Michael Maire, Serge J Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollar, and C Lawrence Zitnick. 2014. Microsoft COCO: Common Objects in Context. *In European Conference on Computer Vision. Springer,* (2014), 740–755.

[23] Mingsheng Long, Yue Cao, Jianmin Wang, and Michael I Jordan. 2015. Learning Transferable Features with Deep Adaptation Networks. *In Proceedings of the 32th International Conference on Machine Learning* (2015), 97–105.

[24] Mingsheng Long, Zhangjie Cao, Jianmin Wang, and Michael I Jordan. 2018. Conditional Adversarial Domain Adaptation. *In Advances in neural information processing systems* (2018), 1640–1650.

[25] Zelun Luo, Yuliang Zou, Judy Hoffman, and Li Feifei. 2017. Label efficient learning of transferable representations across domains and tasks. *In Advances in neural information processing systems* (2017), 164–176.

[26] Diogo C Luvizon, Hedi Tabia, and David Picard. 2019. Human Pose Regression by Combining Indirect Part Detection and Contextual Information. *Computers & Graphics* 85 (2019), 15–22.

[27] Alejandro Newell, Zhiao Huang, and Jia Deng. 2017. Associative Embedding: End-to-End Learning for Joint Detection and Grouping. *In Advances in neural information processing systems* (2017), 2277–2287.

[28] Alejandro Newell, Kaiyu Yang, and Jia Deng. 2016. Stacked Hourglass Networks for Human Pose Estimation. *In European Conference on Computer Vision. Springer,* (2016), 483–499.

[29] Xuecheng Nie, Jiashi Feng, Jianfeng Zhang, and Shuicheng Yan. 2019. Single-Stage Multi-Person Pose Machines. *In Proceedings of the IEEE International Conference on Computer Vision* (2019), 6951–6960.

[30] Guanghan Ning, Zhi Zhang, and Zhihai He. 2018. Knowledge-Guided Deep Fractal Neural Networks for Human Pose Estimation. *IEEE Transactions on Multimedia* (2018), 1246–1259.

[31] Boxiao Pan, Zhangjie Cao, Ehsan Adeli, and Juan Carlos Niebles. 2020. Adversarial Cross-Domain Action Recognition with Co-Attention. *Association for Advancement of Artificial Intelligence* (2020).

[32] Maja Pantic, Nicu Sebe, Jeffrey F Cohn, and Thomas S Huang. 2005. Affective multimodal human-computer interaction. *In Proceedings of the 13rd ACM international conference on Multimedia. ACM* (2005), 669–676.

[33] George Papandreou, Tyler Zhu, Nori Kanazawa, Alexander Toshev, Jonathan Tompson, Chris Bregler, and Kevin P Murphy. 2017. Towards Accurate Multi-person Pose Estimation in the Wild. *In Proceedings of the IEEE conference on computer vision and pattern recognition* (2017), 3711–3719.

[34] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. 2017. Automatic differentiation in PyTorch. *In Advances in neural information processing systems* (2017).

[35] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael S Bernstein, et al. 2015. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision* 115, 3 (2015), 211–252.

[36] Kuniaki Saito, Donghyun Kim, Stan Sclaroff, Trevor Darrell, and Kate Saenko. 2019. Semi-Supervised Domain Adaptation via Minimax Entropy. *In Proceedings of the IEEE International Conference on Computer Vision* (2019), 8050–8058.

[37] Kuniaki Saito, Yoshitaka Ushiku, Tatsuya Harada, and Kate Saenko. 2019. Strong-Weak Distribution Alignment for Adaptive Object Detection. *In Proceedings of the IEEE conference on Computer Vision and Pattern Recognition* (2019), 6956–6965.

[38] Yu Su and Wang Xu. 2019. Multi-Person Pose Estimation with Enhanced Channel-wise and Spatial Information. *In Proceedings of the IEEE conference on Computer Vision and Pattern Recognition* (2019).

[39] Ke Sun, Cuiling Lan, Junliang Xing, Wenjun Zeng, Dong Liu, and Jingdong Wang. 2017. Human Pose Estimation Using Global and Local Normalization. *In Proceedings of the IEEE International Conference on Computer Vision* (2017), 5600–5608.

[40] Ke Sun, Bin Xiao, Dong Liu, and Jingdong Wang. 2019. Deep High-Resolution Representation Learning for Human Pose Estimation. *In Proceedings of the IEEE conference on Computer Vision and Pattern Recognition* (2019), 5693–5703.

[41] Wei Tang, Pei Yu, and Ying Wu. 2018. Deeply Learned Compositional Models for Human Pose Estimation. *In European Conference on Computer Vision. Springer,* (2018), 197–214.

[42] Zhiqiang Tang, Xi Peng, Shijie Geng, Lingfei Wu, Shaoting Zhang, and Dimitris N Metaxas. 2018. Quantized Densely Connected U-Nets for Efficient Landmark Localization. *In European Conference on Computer Vision. Springer,* (2018), 348–364.

[43] Yihsuan Tsai, Weichih Hung, Samuel Schulter, Kihyuk Sohn, Minghsuan Yang, and Manmohan Chandraker. 2018. Learning to Adapt Structured Output Space for Semantic Segmentation. *In Proceedings of the IEEE conference on Computer Vision and Pattern Recognition* (2018), 7472–7481.

[44] Yihsuan Tsai, Kihyuk Sohn, Samuel Schulter, and Manmohan Chandraker. 2019. Domain Adaptation for Structured Output via Discriminative Patch Representations. *In Proceedings of the IEEE International Conference on Computer Vision* (2019), 1456–1465.

[45] Eric Tzeng, Judy Hoffman, Kate Saenko, and Trevor Darrell. 2017. Adversarial Discriminative Domain Adaptation. *In Proceedings of the IEEE conference on Computer Vision and Pattern Recognition* (2017), 2962–2971.

[46] Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He. 2018. Non-local Neural Networks. *In Proceedings of the IEEE conference on Computer Vision and Pattern Recognition* (2018), 7794–7803.

[47] Shihen Wei, Varun Ramakrishna, Takeo Kanade, and Yaser Sheikh. 2016. Convolutional Pose Machines. *In Proceedings of the IEEE conference on computer vision and pattern recognition* (2016), 4724–4732.

[48] Zuxuan Wu, Xintong Han, Yenliang Lin, Mustafa Gokhan Uzunbas, Tom Goldstein, Sernam Lim, and Larry S Davis. 2018. DCAN: Dual Channel-wise Alignment Networks for Unsupervised Scene Adaptation. *In European Conference on Computer Vision. Springer,* (2018), 518–534.

[49] Bin Xiao, Haiping Wu, and Yichen Wei. 2018. Simple Baselines for Human Pose Estimation and Tracking. *In European Conference on Computer Vision. Springer,* (2018), 472–487.

[50] Wei Yang, Shuang Li, Wanli Ouyang, Hongsheng Li, and Xiaogang Wang. 2017. Learning Feature Pyramids for Human Pose Estimation. *In Proceedings of the IEEE International Conference on Computer Vision* (2017), 1290–1299.

[51] Weichen Zhang, Wanli Ouyang, Wen Li, and Dong Xu. 2018. Collaborative and Adversarial Network for Unsupervised Domain Adaptation. *In Proceedings of the IEEE conference on Computer Vision and Pattern Recognition* (2018), 3801–3809.

[52] Long Zhao, Xi Peng, Yu Tian, Mubbasir Kapadia, and Dimitris N Metaxas. 2019. Semantic Graph Convolutional Networks for 3D Human Pose Regression. *In Proceedings of the IEEE conference on Computer Vision and Pattern Recognition* (2019), 3425–3435.

[53] Xinge Zhu, Jiangmiao Pang, Ceyuan Yang, Jianping Shi, and Dahua Lin. 2019. Adapting Object Detectors via Selective Cross-Domain Alignment. *In Proceedings of the IEEE conference on Computer Vision and Pattern Recognition* (2019), 687–696.