Structure-enriched Topology Learning for Cross-domain Multi-person Pose estimation

Xixia Xu, Qi Zou*, Xue Lin

Abstract—**Human pose estimation has been widely studied with much focus on supervised learning. However, in real applications, a pretrained pose estimation model usually needs be adapted to a novel domain without labels or with sparse labels. Existing domain adaptation methods cannot well deal with it since poses have flexible topological structures and need fine-grained local features. Aiming at the characteristics of human pose, we propose a novel domain adaptation method for multi-person pose estimation (MPPE) to alleviate the human-level shift. Firstly, the training samples of human poses are clustered into groups according to the posture similarity. Within the clustered space, we conduct three adaptation modules: Cross-Attentive Feature Alignment (CAFA), Intra-domain Structure Adaptation (ISA) and Adaptive Human-Topology Adaptation (AHTA). The CAFA adopts a bidirectional spatial attention mechanism to explore fine-grained local feature correlation between two humans, and thus to adaptively aggregate consistent features for adaptation. ISA only works in semi-supervised domain adaptation (SSDA) to exploit semantic relationship of corresponding keypoints for reducing the intra-domain bias. Importantly, we creatively propose an AHTA to enrich human topological knowledge for reducing the inter-domain discrepancy. Specifically, the pose structure and the cross-instance topological relations are modeled via graph networks. This flexible topology learning benefits the occluded or extreme pose inference. Extensive experiments are conducted on two popular benchmarks and additional two challenging datasets. Results demonstrate the competency of our method, which works in unsupervised or semi-supervised modes, compared with the existing supervised approaches.**

Index Terms—**Multi-person Pose Estimation, Domain Adaptation, Adaptive Human-Topolopy Learning**

I. Introduction

Multi-person pose estimation aims to locate body parts of each person from an image, such as keypoints on the arms, torsos, and the face [27], [33], [45]. It's fundamental to deal with other high-level tasks, such as human action recognition [67] and human-computer interaction [11]. Recently, due to the progress of convolution neural network (CNN) [23], most existing methods [8], [31], [56], [70] have achieved remarkable advances in MPPE. 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 [37], object detection [54], and semantic segmentation [24]. 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 [36] or discriminating the output through adversarial learning [18], [47], [62]. However, 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. 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. Most existing domain adaptation methods on image classification and semantic segmentation [24], [61] typically consider the image as a whole for alignment, while ignoring local regions of the object-of-interest. Focusing on such local regions is important for pose estimation. Although some cross-domain object detection methods [83] 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 [61], [62] adapt the output distribution and align the global layout across domains for semantic segmentation, which is much different from the human structure adaptation in MPPE. **3)** In semi-supervised setting, the gap between labeled and unlabeled data in the target domain also exists. Few works consider the intra-domain gap via adapting feature representations simply, let alone exploring the intra-domain structural keypoint relationship. A representative work in such area [38] is for classification, aligning features at the image level.

Aiming at these issues, we propose a novel domain adaptation framework, which consists of a pose clustering preprocessing and three adaptation modules. The pose clustering is mainly leveraged to divide the whole samples into several clusters, in which with similar pose modes according to the learned semantic features. Within the clustered pose space, we conduct our proposed adaptation strategies: (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 creatively adopt a bidirectional spatial attention mechanism to capture local feature similarity across humans. The local features of source human parts can

* Corresponding author

Xixia Xu, Qi Zou and Xue Lin are with School of Computer and Information Technology, Beijing Key Laboratory of Traffic Data Analysis and Mining, Beijing Jiaotong University, Beijing, 100044, China (email : 19112036@bjtu.edu.cn; qzou@bjtu.edu.cn; 18112028@bjtu.edu.cn).

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) **A**daptive **H**uman-**T**opology **A**daptation **(AHTA)**. We adopt the recent SemGCN [82] to capture the flexible human pose structural representations. With such representations as guidance, a cross-instance bidirectional graph is constructed to adaptively explore the human structural relationship across domains. The graph attention mechanism is designed to adaptively learn the beneficial domain-invariant structural information from the correlated instances in opposite domain. It better avoid the hard alignment between the utterly different poses and flexibly benefit the domain-invariant human topological knowledge learning for both domains compared with [72]. Upon this, we reverse the enhanced representations to the former coordinate space and augment the domain-specific human-topology representations to benefit the final predictions. The predicted errors in the target domain can be further repaired by current structural domain-specific knowledge augmented by the opposite domain structural knowledge (*e.g.*, the structural reasoning helps to infer the occluded or invisible joints).

Our main contributions are summarized as follows.

• We creatively propose a robust domain adaptation method for MPPE and discover different modes of pose samples via clustering ahead of time for better alignment. Comprehensive experiments are demonstrated on numerous datasets with various settings, such simple-to-complex and cross-environment scenarios to validate the effectiveness of our method.

• A novel CAFA module achieves the fine-grained human body feature alignment via exploring the bidirectional spatial feature dependency across domains and adapts abundant domain-invariant features for accurate pose estimation.

• In SSDA setting, a novel ISA adapts the local keypoint structural knowledge of the labeled to the unlabeled data in the target domain to alleviate intra-domain confusions.

• The AHTA mechanism explicitly enriches domain-specific human topological knowledge via learning more domain-invariant structural information for both domains. Additionally, we adopt a graph-based formulation for modeling the intra-pose structure and construct a cross-instance graph to adaptively explore the human structural relations.

II. Related Work

Supervised Multi-Person Pose Estimation. Recently, MPPE has aroused a great interest due to the real-life demand. Nowadays, researchers have made painstaking efforts [28], [44], [56], [58] to accelerate its progress. For example, CASNet [56] improves feature robustness via the spatial and channel-wise attention. HRNet [58] 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 on recent benchmarks (*e.g.*, MS-COCO [35] 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 an 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 [17], [80]. Most methods tackle UDA by minimizing the distance [36] across two distributions or aligning the output in adversarial learning [63]. For instance, [25] applies the adversarial strategy to align features for semantic segmentation. [69] 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 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 AHTA mechanism to adaptively exploit the human-topological relations across domains to better bridge the inter-domain discrepancy.

A plethora of SSDA works [38], [53] have already emerged. For example, the [53] designs a min-max entropy strategy to achieve better adaptation. Some works [38] consider the intra-adaptation bias but they simply minimize the entropy similarity between intra-target samples like [38]. 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 heatmaps inside target for avoiding the keypoint shift chaotically.

Domain Adaptation for Pose Estimation. There are a few domain adaptation works related to ours and designed for single person [41], [74], animal [6] or 3D pose estimation [77], [81]. These works are related to ours but their purposes and datasets are completely different from ours. They either fill the gap between synthetic (RGB or depth) images and real ones, or between high-quality images and blurred ones, or between animals and humans. Most importantly, these domain adaptation methods are not specific for pose estimation. They either simply use the adversarial learning to distinguish source samples from the target [6], [41], [77], [81] or directly use classification adaptation techniques [74]. When we tried to apply similar methods on cross-domain MPPE, it achieves unsatisfied performance, for occlusions and clutter backgrounds are more challenging in multi-person scenes, where aligning coarse features will bring negative effects. Although the only Robust DA [72] makes a good start on cross-domain MPPE and achieves good performance, it still has some limitations should be improved. On one hand, it didn't consider the prior of pose diversity before alignment. On the other hand, it conducts human topology adaptation relying on the threshold-based sampling strategy, which is inflexible

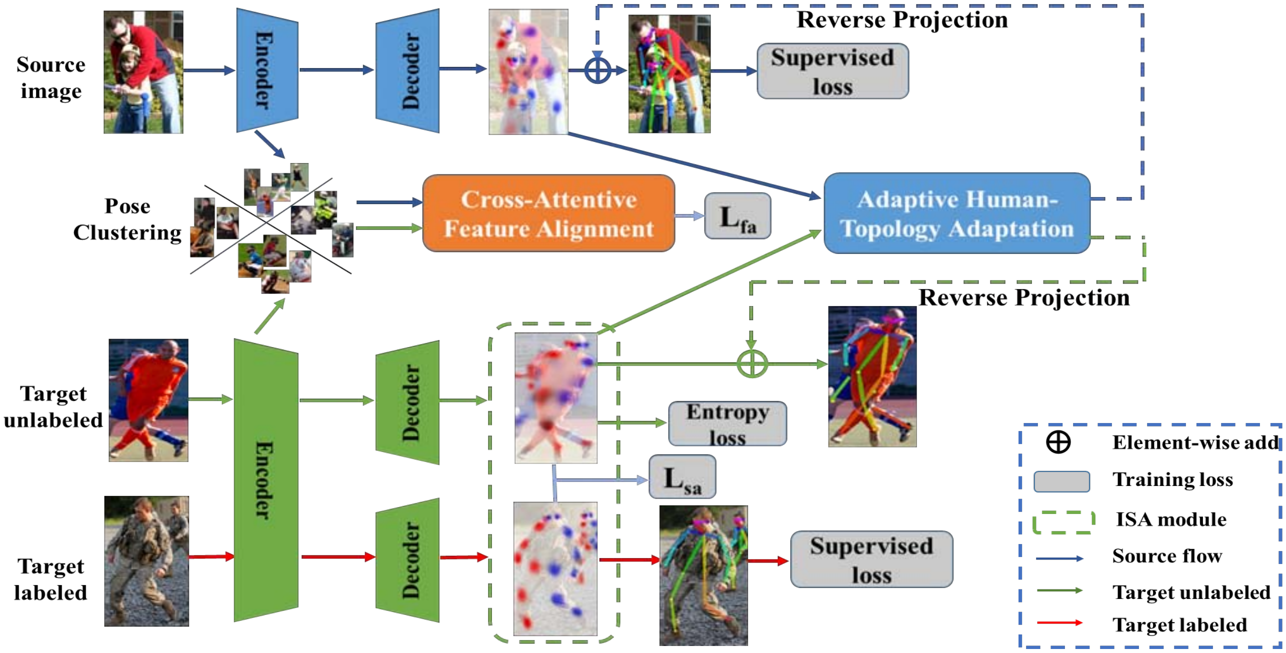

Fig. 1. Overview of the proposed method. Given the input, the pope clustering is applied for the training samples to cluster them into multiple modes. Upon it, the CAFA (in orange box) is further conducted to learn domain-invariant fine-grained local features. The ISA (in green dashed line box) only works in SSDA setting to explore the keypoint heatmap alignment inside target domain. The AHTA (in blue box) models human pose structure learning with SemGCN [82] and further flexibly explores cross-instance human-topology relations to improve accuracy. Noted that all modules are only worked in the training phrase.

for dealing with the variable postures. Aiming at this, we further extend their design in two sides. One is to introduce the pose clustering strategy to project the human poses of training set into a clustered space and conduct adaptation within each space leading to a better domain alignment. Another refers to propose a novel adaptive topological adaptation module to model the intra-pose structural graph and construct a cross-instance graph to adaptively explore the inter-domain human structural relations.

Graph Convolutional Neural Network. Graph Neural Networks (GNNs) are initially introduced in [55] and become a popular tool for efficient message passing and modeling global relations [65]. More recently, GNN models have been applied to model the human body structure. Specifically, ST-GCN [73] was probably the first representative work to adopt graph-based network to model dynamic skeletons for action recognition. Sem-GCN [82] employed graph convolution network (GCN) to regress 3D pose from 2D by capturing both local and global relationships among joints. This work is related to [82], which models the high-order human pose kinematic relationship among joints.

The graph representation for 2D MPPE is not new as well. For example, [51] construct dynamic graphs to tolerate large variations in human pose. And the recent OPEC-Net [50] propose an Image-Guided Progressive GCN to estimate the invisible joints to deal with the occlusions. GPCNN [65] propose a two-stage graph-based framework which adopts a graph pose refinement module to get more accurate results. HGG [26] propose a novel differentiable hierarchical graph grouping method to solve the keypoint assignment for the bottom-up MPPE. However, all above methods are designed for solving the fully-supervised pose estimation but not fit for the cross-domain pose estimation. In this paper, we explore the cross-domain human-topology structural relationship by virtue of graph attention network (GAT) [64] to help each of instance learn the domain-invariant human structural representations from the correlated instance in the opposite domain.

III. Method

A. 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 [70] as our baseline, which utilizes an encoder-decoder architecture to make pose predictions as depicted in Fig. 1. 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. With these representations, we cluster the pose samples of training set into different modes (i.e., multiple postures as sit, stand, lie, squat and so on). Based on the clustered space, we further get the adapted features F_s, F_t via CAFA. 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 heatmap representations. With these features as guidance, we formulate the human pose with graph and harvest the human structural representations via SemGCN

[82]. Based on it, we further construct a cross-instance graph to adaptively adapt the human topological relations via AHTA to achieve more precise predictions.

Pose Clustering. We introduce the pose clustering operation to discover the different pose modes, where each mode encodes a specific visual pattern of postures. The idea behind the clustering operation is rather simple, mainly inspired by the fact that there are complex postures in training samples, if we conduct the alignment directly, the deformations and geometric differences between arbitrary poses will bring about negative effects. Intuitively, the poses with similar visual patterns will be clustered into a common space according to their specific feature relations. Thus, we project the human poses of training set into a clustered space and conduct adaptation within each space leading to a better domain alignment ultimately.

As in Fig. 1, we firstly cluster the human poses of training set into several classes aiming at learning the prototypical pose representations for both domains. Concretely, we apply the K-means clustering on the downsampled features to separate them into C classes (C is 5 in this work). The measurement that decides which class they belong to is up to the cosine similarity of the learned features. We project the human poses into multiple clustered space, in which each cluster possess similar postures or semantic pose knowledge.

B. Cross-Attentive Feature Alignment

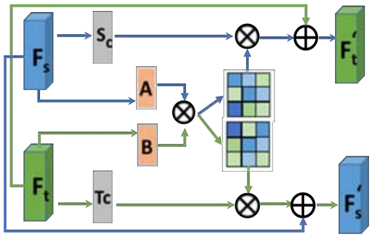

Fig. 2. Cross-Attentive Feature Alignment module.

The fine-grained features are effective for the accurate pose estimation. The CAFA aims to adapt more domain-invariant fine-grained human features across domains. Within the projected clustered space, we seek to adapt more domain-invariant fine-grained human features for each cluster across domains. Different from previous feature adaptation methods, we capture the related fine-grained feature response across domains. It explores the local spatial feature dependency across domains rather than simply consider the domain features separately. The fine-grained 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 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. The F_s, F_t are 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 $\Phi^{(i,j)}$ measures 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 $\Psi_{s\to t}^{(i,j)}$ depicts 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 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 feature are well encoded in 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 responses and original target features in Eq. 2. In this manner, $F_s^{'}$ and $F_t^{'}$ encode more fine-grained features for each domain.

Loss and Alignment. Finally, we apply the Maximum Mean Discrepancy **(MMD)** [21] 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 $\mathcal{H}$ [20]. 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 $\mathcal{L}_{fd}$.

C. Intra-domain Structure Adaptation

Under the SSDA setting, the **T**arget **L**abeled **(TL)** and **T**arget **U**nlabeled **(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. 11). Within the clustered space, we devise an ISA module inside the target domain 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,

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

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

D. Adaptive Human-Topology Adaptation

We further observe that the human topological structure information helps to infer the pose structure in clutter or occluded scenarios. Within the clustered space got in Section III-A, an AHTA mechanism is proposed to enhance the domain-specific human topological information. It mainly consists of two parts: the Intra-Pose Structure Learning (IPSL) and Cross-Instance Topology Adaptation (CITA). Our IPSL adopts GCN to acquire the human structural representations. It offers an explicit way of modeling the high-order human skeleton structure that is advantageous for capturing the spatial topological information of joints. And with the learned features as the node guidance, our CITA further constructs a cross-instance bipartite graph across domains, which adopts GAT to adaptively learn more domain-invariant human structural information. As shown in Fig. 3, the details are described in the following two parts.

1) Intra-Pose Structure Learning:

Local Keypoint Feature Extraction. Specifically, we can get a group of semantic local features V of keypoints according to the learned feature maps and keypoint heatmaps representations for both domains via an outer product and a global average pooling operation.

Intra-Pose Graph Formulation. Here, we construct an intuitive graph $G = (V, E)$ based on the keypoint local features for each human pose. $V = \{v_i, i = 1, 2, ..., H\}$ is the node set which represents the keypoint of human body. $E = \{v_i v_j \mid$ if i and j 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 $a_{ij} = 0$.

Structure Learning via Graph Reasoning. Our key insight is that human body structure is a natural graph and there exists potential spatial constraint among joints. Hence, we model the pose structure representation via the recent SemGCN [82]. For a graph convolution, propagating features through neighbor joints helps to learn robust local structure. Meanwhile, the non-local layer [66] helps capture the dependency among nodes to learn more context information. These form robust pose structural representations, which are essential for exploring the cross-instance domain-invariant topological relations.

Following SemGCN [82], the transformed node representations firstly are gathered to node i from its neighboring node j. The node features are collected into $v^{(l)} \in R^{D_l \times H}$.

$$v^{(l+1)} = \Sigma_{d=1}^{D_{l+1}} \sigma(\vec{w}_d v^{(l)} \varphi_i(M_d \odot A)), \tag{5}$$

where $v^{(l)}$ and $v^{(l+1)}$ 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 learnable to show the local semantic knowledge of neighboring joints implied in the graph. The Σ depicts channel-wise concatenation, and $\vec{w}_d$ learns channel-wise weights for edges as priors (e.g., how one joint influences other body parts in pose estimation) to enhance the graph representations. The φ_i is a Softmax to normalize the input, $\odot$ is an element-wise product which returns m_{ij} if $a_{ij} = 1$. Hence, the relationship between joints are well considered.

Then, aside by the non-local concept [66], we compute responses between joints with their features to capture local and global long-range relationships among nodes and the feature updating process is shown as:

$$v_i^{(l+1)} = v_i^{(l)} + \frac{W_v}{H}\sum_{j=1}^{H} f(v_i^{(l)}, v_j^{(l)}) \cdot g(v_j^{(l)}), \tag{6}$$

where f is to compute the affinity between node i and all other j, g computes the node j representation.

Until now, we have acquired the robust high-level intra-pose structure representations X_{ss} and X_{st} for instances in both domains.

2) Cross-Instance Topology Adaptation:

The proposed Cross-Instance Topology Adaptation (CITA) block aims to adaptively reason the structural similarities between the source pose and the target pose in a bidirectional graph reasoning way. The rationale behind the design of CITA is straightforward since consistent structural patterns of human poses in source and target domains do exist. For example, when there is a human in the source domain playing tennis, it is helpful to infer the similar topological representation of a human playing badminton in the target domain. Therefore, we adopt Graph Attention Networks (GATs) [64] to capture such structural correlations of different instances from the opposite domain. The adaptive cross-instance bipartite graph not only avoids the inflexibility of hand-crafted correlation graphs but also avoids statistical bias caused by imbalanced pose variations in the training set.

Specifically, the CITA is performed between instances across different domains to evolve the structural features by the guidance of correlation of the source poses and target poses via Graph Attention Block. The proposed CITA block contains two symmetrical branches (*i.e.*, S2T branch and T2S branch) as shown in Fig. 3, each node in the source nodes connects all the target nodes; and vice versa. In the following, we mainly describe the detailed modeling process of the S2T branch, and another T2S branch is similar to this.

Cross-Instance Graph Construction. The cross-instance graph takes the above learned source and the target structural representations $X_{ss} \in \mathbb{R}^{M \times H \times D_s}$, $X_{st} \in \mathbb{R}^{N \times H \times D_t}$ as inputs (M, N is the number of instances of source and target domain, H is the keypoint numbers and D is feature dimensions).

Topology Adaptation with Graph Attention. Arbitrary poses from two domains that may have lower structure similarity. To avoid nonsense alignment, we adopt GATs to adaptively attend to more similar poses from different domains. Specifically, we compute their structural pose representation similarity as

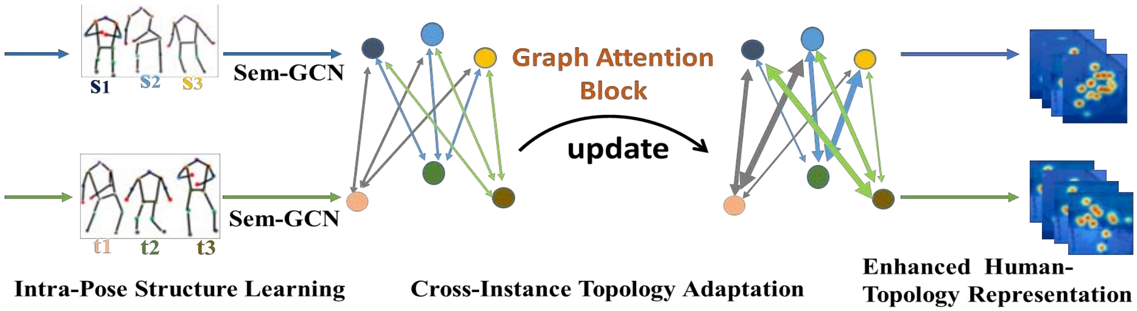
S1
S2
S3
Sem-GCN
t1
t2
t3
Sem-GCN
Graph Attention Block
update
Intra-Pose Structure Learning
Cross-Instance Topology Adaptation
Enhanced Human-Topology Representation

Fig. 3. The Adaptive Human-Topology Adaptation module models the structural learning of human pose with SemGCN [82] and further adaptively explores the cross-instance topological relations across domains by virtue of GAT [64] to enhance the domain-specific structural knowledge.

a guidance matrix $M^{s2t} \in \mathbb{R}^{N \times M}$. For target node x_n in X_{st} and source node x_m in X_{ss}, we can compute the guidance information $M^{s2t}_{m,n}$ that represents the assignment weight of the source node x_m to the target node x_n as follows,

$$M^{s2t}_{m,n} = \frac{exp(\delta(W[x_n||x_m]))}{\sum_{k \in \mathcal{N}_m} exp(\delta(W[x_n||x_k]))}, \tag{7}$$

where $||$ is the concatenation operation, $\mathcal{N}_m$ is the neighborhood of node n (i.e., it contains all the target nodes in our work), $\delta(\cdot)$ is the Sigmoid activation function, and W is weight matrix. After obtaining the M^{s2t}, we can distill the domain-invariant structural representation from the source domain to enhance the target structural representations, according to:

$$\tilde{X}_{st} = X_{st} + M^{s2t} X_{ss} W_{s2t}, \tag{8}$$

where W_{s2t} is a trainable weight matrix, we use a simple sum to melt information from source instances, which may be alternatively replaced by other commutative operators such as mean, max, or concatenate. With the help of the new adjacency matrix M^{s2t}, we effectively reason the pose topological correlation across domains, and incorporate the corresponding source features into the target instance.

Similarly, the T2S branch has learned the human-topological representations $\tilde{X}_{ss}$ from the target domain and augmented the domain-specific human structural features. By combining the S2T and T2S steps, the CITA enables the model to preserve structure-invariant features.

Reverse Projection. After the reasoning, the evolved keypoint representation of humans across domains can be further used to improve the structure representations of each pixel features. We adopt a non-linear transformation θ_t to reverse project the evolved pose structural features of each instance $\tilde{x}_n$ in $\tilde{X}_{st}$ to the original feature space. Finally, we concatenate the reverse features to the former output keypoint features $y_t'^n$ of $n{-}th$ instance to equip it with stronger domain-specific structural knowledge(as shown in Fig. 1) for more precise prediction.

$$\tilde{y'}^{n}_{t} = \theta_t(\phi_t \tilde{x}_n) + \pi_t(y_t'^n), \tag{9}$$

where ϕ_t is a dimension reduction operation and π_t is a dimension transformation function to make the $y_t'^n$ to be compatible with the reversed pose structural features. Similarly, the T2S branch obtains the new source domain enriched keypoint representations $\tilde{y'}^{m}_{s}$ for each instance in the same way for improving the localization accuracy.

E. Optimization

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

$$\mathcal{L}_{pose} = \beta_{sup}\mathcal{L}^{sup}_{pose} + \beta_{da}\mathcal{L}^{da}_{pose}, \tag{10}$$

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

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

$$\mathcal{L}^s_{pose} = \sum_{i=1}^{M}\sum_{h=1}^{H} \| y^h_{s_i} - y'^h_{s_i} \|^2_2, \tag{11}$$

$$\mathcal{L}^{tl}_{pose} = \sum_{j=1}^{N_1}\sum_{h=1}^{H} \| y^h_{t_j} - y'^h_{t_j} \|^2_2, \tag{12}$$

$$\mathcal{L}^{sup}_{pose} = \alpha_{tl}\mathcal{L}^{tl}_{pose} + \mathcal{L}^s_{pose}, \tag{13}$$

where α_{tl} is the trade-off hyperparameter and its value is 0.5, the Eq. 12 is only used in semi-supervised setting.

Entropy Loss. Entropy minimization (ENT) [19] 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:

$$\mathcal{L}^{tu}_{ent} = \sum_{k=1}^{N_2} ent(y_t'^k), \tag{14}$$

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:

$$\mathcal{L}^{da}_{pose} = \alpha_{sa}\mathcal{L}_{sa} + \alpha_{fd}\mathcal{L}_{fd}, \tag{15}$$

where the α_{sa}, α_{fd} are the weighted factors to keep the model effective and their values will be discussed in Sec. IV-D.

IV. Experiments

A. Datasets and Evaluation Metric

The proposed method is evaluated on two recent multi-person datasets, *i.e.*, MPII Human Pose benchmark [1], COCO 2017 Keypoints Detection dataset [35] and additional two crowded and occluded datasets, *i.e.*, CrowdPose dataset [31] and OCHuman dataset [79].

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 adopt OKS-based average precision (AP) and average recall (AR). The OKS (object keypoints similarity) defines the similarity between the predicted heatmap and the groundtruth.

MPII Human Pose Dataset consists of 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 ours, 50% of the head size for normalization.

CrowdPose contains $20K$ images with $80K$ human instances. It divides into three crowding levels by Crowd Index: easy ($0 \sim 0.1$), medium ($0.1 \sim 0.8$) and hard ($0.8 \sim 1$). Its Crowd Index satisfies uniform distribution in $[0, 1]$. CrowdPose dataset aims to promote performance in crowded cases and make models generalize to different scenarios. It uses the same evaluation metric as COCO.

OCHuman dataset contains only difficult cases of occluded and intertwined persons and the average IoU of the bounding boxes is 67%. It consists of 4731 images that comprise a validation and test set with a total of 8110 annotated humans. It has 17 annotated body joint locations for pose estimation. All images are collected from real-world scenarios containing people with challenging poses and viewpoints, various appearances and in a wide range of resolutions. The purpose of this dataset is to examine the limitations of human detection in highly challenging scenarios. Accordingly, it does not contain training samples and is used for evaluating the training model. It's noted that we only test the model trained with the MS-COCO due to only they keep the same keypoint numbers.

In our experiment, there are several source-target settings as below:

1. Source: MS-COCO/MPII Target: MPII/MS-COCO.
2. Source: MPII/CrowdPose Target: CrowdPose/MPII.
3. Source: MS-COCO/ Target: OCHuman.
4. Source: MS-COCO/CrowdPose Target: CrowdPose/MS-COCO.

B. Implementation Details

Network Architectures. We adopt SimpleBaseline [70] and HRNet [58] as the pose estimation baseline respectively and the backbone uses the ResNet-50 in default. As for the SemGCN [82], the building block is one residual block [23] built by two SemGConv layers with 128 channels, followed by one non-local layer [66]. This block is repeated four times. All SemGConv layers are followed by Batch Normalization and a RELU activation except the last one. The GAT block consists of two distinct GCNs; where each GCN is followed by Batch Normalization and RELU units.

Data Augmentation. We apply random flip, rotation, and scale in training. The flip value is 0.5. The scale range is ($[0.7 \sim 1.3]$), and the rotation range is ($[-40^\circ \sim +40^\circ]$).

Training. We implement all experiments in PyTorch [49] on a single NVIDIA TITAN XP GPU with 12 GB memory. We select ResNet-50, 101, 152 as the backbones which are all initialized with the weights of the ImageNet [52] pretrained model. We use Adam optimizer [29] with learning rate 10^{-4}, momentum 0.9 and weight decay 10^{-4}. For the training, we adopt two kinds of input size (*i.e.*, $256 \times 256/192$ or $384 \times 384/288$), and train 150 epochs for MS-COCO, 135 epochs for MPII and 180 epochs for CrowdPose.

It's noted that we don't evaluate the results on MS-COCO test-dev dataset due to it needs to test online and the unequal keypoint numbers across dataset makes it harder. Additionally, we unify the keypoint numbers across different datasets and align with the MPII in our paper. Concretely, we add two kinds of joints ('pelvis', 'thorax') for CrowdPose and ignore the 'nose' for MS-COCO. Concretely, we calculate the distance of the 'lhip' and 'rhip' as the location of 'pelvis'. Similarly, the coordinate of the 'thorax' is regarded as the average distance of the 'lshoulder' and 'rshoulder'. Additionally, the way of evaluation is unchanging for each of them.

C. Domain Adaptation Performance

We compare the performance of ours with the supervised methods 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/CrowdPose to MS-COCO/MS-COCO. The Tab. I reports the results on MPII to MS-COCO and CrowdPose to MS-COCO. For MPII, 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 65.4% AP, which is higher 8.7% 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 1.5%, indicating that ISA further alleviates the intra-domain bias. Moreover, the accuracy also improves when we adopt HRNet [58] as the baseline model, which further illustrates our method with pretty generality. Noted that compared with the only Robust DA [72] baseline, our method achieves consistent improvement in all metrics. It mostly lies in that our AHTA can better adaptively choose the structure similar samples and flexibly preserve richer human-topology features than them.

For CrowdPose, the number of samples are less than MS-COCO but its complexity is much higher than it. When we conduct cross-adaptation across them, our model also achieves 68.1% AP and almost catches up with the supervised model, which is 7.4% AP higher than baseline. It's also still in stable improvement although the pose backbone changes from SBN to the HRNet. This can magnify that the complexity and high intensity of the source domain can benefit more diverse knowledge adaptation.

TABLE I
THE COMPARISON RESULTS ON MPII TO MS-COCO AND CROWDPOSE TO MS-COCO.

Method	Backbone	AP	AP.5	AP.75	AP(M)	AP(L)	AR
Supervised methods (Target only)							
G-RMI [48]	-	65.7	83.1	72.1	61.7	72.5	69.9
AE [42]	HG	66.3	86.5	72.7	61.3	73.2	71.5
MultiPoseNet [30]	-	69.6	86.3	76.6	65.0	76.3	73.5
HigherHRNet [10]	HRNet-w32	66.4	87.5	72.8	61.2	74.2	-
CenterGroup [4]	HRNet-w32	67.6	88.7	73.6	61.9	75.6	-
SIMPLE [78]	HRNet-w32	69.6	89.3	77.9	68.1	77.8	-
CSANet [56]	ResNet-50	72.1	-	-	-	-	-
RSGNet [14]	HRNet-w48	74.7	92.3	82.3	71.4	80.5	79.9
PRTR [32]	HR32	71.7	90.6	79.6	67.6	78.4	78.8
TokenP [34]	HR32	74.0	91.9	81.5	70.6	79.8	79.1
TransP [75]	HR32	73.4	91.6	81.1	70.1	79.3	-
QRPose [71]	HRNet-w32	69.0	89.3	76.0	62.8	77.0	73.6
FCPose [40]	R101	65.6	87.9	72.6	62.1	72.3	-
SBN [70]	ResNet-50	70.4	88.6	78.3	67.1	77.2	76.3
HRNet [58]	HRNet-w48	75.1	90.6	82.2	72.5	81.8	80.4
Direct Transfer (Source only)							
SBN(MPII)	ResNet-50	56.7	72.3	64.4	54.3	53.4	62.2
HRNet(MPII)	HRNet-w48	59.5	75.2	66.8	55.6	66.8	64.9
SBN(CrowdPose)	ResNet-50	60.7	74.3	67.4	53.4	57.4	65.2
HRNet(CrowdPose)	HRNet-w48	62.5	75.2	69.5	54.6	59.5	67.1
UDA model (Without ISA module)							
SBN(MPII)	ResNet-50	65.4	81.4	73.2	63.2	72.5	71.6
HRNet(MPII)	HRNet-w48	68.2	84.2	75.1	64.8	75.2	74.5
SBN(CrowdPose)	ResNet-50	67.3	80.4	73.7	57.3	62.2	71.3
HRNet(CrowdPose)	HRNet-w48	69.5	82.2	75.8	59.6	63.1	73.5
Adaptation model (Ours)							
Robust DA [72]-MPII	ResNet-50	66.5	82.3	73.2	64.1	73.5	72.3
Robust DA [72]-MPII	HRNet-w48	69.9	85.3	76.4	65.6	76.5	75.1
Ours(SBN-MPII)	ResNet-50	66.9	82.8	73.8	64.3	76.5	72.8
Ours(HRNet-MPII)	HRNet-w48	70.2	85.8	76.7	66.1	76.9	75.7
Ours(SBN-CrowdPose)	ResNet-50	68.1	80.8	74.3	58.1	62.8	71.9
Ours(HRNet-CrowdPose)	HRNet-w48	70.2	82.8	76.2	60.5	63.5	74.0

MS-COCO/CrowdPose to MPII/MPII. We evaluate the PCKh@0.5 score on MS-COCO to MPII and CrowdPose to MPII in Tab. II for comparison. For the MS-COCO as the source dataset, the UDA model achieves 85.8% PCKh@0.5 and improves 8.5% 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 88.1%, which outperforms the baseline 10.8%. 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 89.1. Even compared with the recent method [72], our model also has improvements due to the adaptive human topology adaptation strategy on both SBN [70] and HRNet [58]. Noted that the result of Robust DA [72] with HRNet-w32 is lower 0.5% than the result in their paper since that we test their model in single-scale and keep same testing setting with ours for fair comparison.

TABLE II
RESULTS COMPARISON ON MS-COCO TO MPII AND CROWDPOSE TO MPII. * MEANS EXTRA DATA AND LARGER INPUT SIZE ARE USED.

Method	Head	Shoul	Elbow	Wrist	Hip	Knee	Ankle	Total
Supervised methods (Target only)								
Wei *et al.* [68]	97.8	95.0	88.7	84.0	88.4	82.8	79.4	88.5
Newell *et al.* [43]	98.2	96.3	91.2	87.2	89.8	87.4	83.6	90.9
Sun *et al.* [57]	98.1	96.2	91.2	87.2	89.8	87.4	84.1	91.0
Tang *et al.* [60]	97.4	96.4	92.1	87.7	90.2	87.7	84.3	91.2
Ning *et al.* [46]	98.1	96.3	92.2	87.8	90.6	87.6	82.7	91.2
Luvizon *et al.* [39]	98.1	96.6	92.0	87.5	90.6	88.0	82.7	91.2
Chu *et al.* [13]	98.5	96.3	91.9	88.1	90.6	88.0	85.0	91.5
Chou *et al.* [12]	98.2	96.8	92.2	88.0	91.3	89.1	84.9	91.8
Chen *et al.* [9]	98.1	96.5	92.5	88.5	90.2	89.6	86.0	91.9
Yang *et al.* [76]	98.5	96.7	92.5	88.7	91.1	88.6	86.0	92.0
Ke *et al.* [28]	98.5	96.8	92.7	88.4	90.6	89.3	86.3	92.1
Tang *et al.* [59]	98.4	96.9	92.6	88.7	91.8	89.4	86.2	92.3
Zhang *et al.* [78]	97.6	93.2	85.6	76.8	83.9	74.6	71.0	83.3
Bin* [3]	98.9	97.6	94.6	91.2	93.1	92.7	89.1	94.1
Bulat* [5]	98.8	97.5	94.4	91.2	93.2	92.2	89.3	94.1
SBN *et al.* [70]	98.5	96.6	91.9	87.6	91.1	88.1	84.1	91.5
HRNet *et al.* [58]	98.6	96.9	92.8	89.0	91.5	89.0	85.7	92.3
Direct Transfer (Source only)								
SBN(MS-COCO)	95.2	89.8	79.7	72.5	75.8	67.8	60.5	77.3
HRNet-w32(MS-COCO)	95.8	90.3	80.5	74.3	77.6	69.7	62.8	79.6
SBN(CrowdPose)	94.2	88.7	78.5	71.8	74.9	67.2	59.9	76.6
HRNet-w32(CrowdPose)	94.7	89.8	80.1	73.8	77.1	69.1	62.2	78.8
UDA model (Without ISA module)								
SBN(MS-COCO)	96.5	96.1	86.7	81.6	86.8	80.8	73.7	85.8
HRNet-w32(MS-COCO)	96.0	95.0	88.1	84.0	87.5	82.9	79.5	86.8
SBN(CrowdPose)	95.6	95.5	85.6	81.0	84.4	79.5	72.6	84.7
HRNet-w32(CrowdPose)	95.8	94.5	86.4	82.3	85.5	81.0	73.1	85.4
Adaptation model (Ours)								
Robust DA [72]-MS-COCO	97.5	94.4	86.8	82.2	87.2	81.2	73.9	87.8
Robust DA [72]-MS-COCO	97.0	95.5	89.5	84.4	88.3	83.5	80.0	88.9
Ours(SBN-MS-COCO)	97.0	95.0	88.2	84.2	88.0	83.2	80.5	88.1
Ours(HRNet-w32-MS-COCO)	97.2	95.6	89.5	84.5	88.5	83.9	80.1	89.1
Ours(SBN-CrowdPose)	96.0	95.0	86.5	82.0	85.8	81.2	73.2	85.6
Ours(HRNet-w32-CrowdPose)	96.5	95.8	87.0	82.6	86.4	81.6	74.2	86.1

When the CrowdPose as the source domain, the UDA model with SBN achieves 84.7% PCKh, leading to 8.1% gain over the baseline. Equipped with the ISA module, we achieve the best result of 85.6% PCKh score with SBN. When our model utilizes the HRNet as the backbone, we achieve the best performance on CrowdPose to MPII. This also illustrates that our strategy works well on general setting and makes a comparable performance with the existing supervised methods.

MS-COCO/MPII to CrowdPose/CrowdPose. To further illustrate that our method still performs better for extreme crowd scenarios, we also conduct adaptation from MS-COCO/MPII to CrowdPose/CrowdPose. From Tab. III we can observe that our model still behaves better although the target domain is more complex and crowded than the source domain. When the MS-COCO as the source, the baseline has a bad performance and by contrast, our adaptation model achieves 6.6% AP improvement. Additionally, although the scale of MPII is much smaller and simpler than the CrowdPose, it still achieves improvements than baseline by a large margin (*i.e.*, 5.3%). It's noted that we specially validate our method with the HigherHRNet [10] due to it achieves the robust performance on the CrowdPose recently under supervised conditions. When the source is MS-COCO, our model with HigherHRNet achieves 62.5% mAP, it gains over above half of supervised methods and almost up to the supervised result 65.9%. Even when the source is MPII, our model still achieves 59.3%. This powerfully magnifies the effectiveness of our adaptation strategy even facing with very challenging environment.

TABLE III
RESULTS COMPARISON ON MS-COCO TO CROWDPOSE AND MPII TO CROWDPOSE.

Method	AP	AP.5	AP.75	AP(E)	AP(M)	AP(H)
Supervised methods (Target only)						
OpenPose [7]	-	-	-	62.7	48.7	32.3
Mask-RCNN [22]	57.2	83.5	60.3	69.4	57.9	45.8
SBN [70]	60.8	81.4	65.7	71.4	61.2	51.2
RMPE [16]	61.0	81.3	66.0	71.2	61.4	51.1
HigherHRNet [10]	65.9	86.4	70.6	73.3	66.5	57.9
CrowdPose [31]	66.0	84.2	71.5	75.5	66.3	57.4
CenterGroup [4]	67.6	87.7	72.7	73.9	68.2	60.3
OPEC-Net [50]	70.6	86.8	75.6	-	-	-
Direct Transfer (Source only)						
SBN(MS-COCO)	47.2	66.3	51.2	57.2	47.4	37.2
HigherHRNet(MS-COCO)	52.5	73.2	57.4	59.6	53.3	44.2
SBN(MPII)	42.3	62.6	46.5	52.3	42.4	32.5
HigherHRNet(MPII)	46.2	66.3	51.8	53.4	47.5	39.5
UDA model (Without ISA module)						
SBN(MS-COCO)	52.5	71.4	56.8	62.3	52.5	41.9
HigherHRNet(MS-COCO)	61.3	82.3	66.5	68.3	62.5	53.2
SBN(MPII)	46.5	66.4	51.5	56.7	46.6	36.8
HigherHRNet(MPII)	57.8	78.1	62.3	64.5	58.2	49.4
Adaptation model (Ours)						
Ours(SBN-MS-COCO)	53.8	72.6	57.5	63.5	53.8	42.8
Ours(HigherHRNet-MS-COCO)	62.5	83.5	67.4	69.5	63.8	54.6
Ours(SBN-MPII)	47.6	67.8	52.3	57.8	47.8	38.0
Ours(HigherHRNet-MPII)	59.3	79.2	63.5	65.4	59.5	50.6

MS-COCO to OCHuman. It's noted that the OCHuman only has test set and only can be used for testing, and it does not provide public annotations. Therefore, we cannot adopt ISA module to conduct adaptation that regard it as the target domain. Thus, we just show the effectiveness of our UDA model. Additionally, due to the keypoint annotation format of OCHuman only keeps consistent with the MS-COCO dataset, thus we only test the result on MS-COCO to OCHuman. To validate the robustness of the proposed UDA strategy even facing with the extreme occluded and crowded scenes, we evaluate the result in Tab. IV. And we choose to adopt the SBN [70] and HRNet [58] as the baseline respectively. The table indicates that our model achieves improvement of 4.4% and 2.7% AP respectively compared with the baseline and shows our CAFA and AHTA adaptation strategy works better even the target domain is absolutely without labels and exists amount of challenging crowded and occluded humans. Even compared with the best supervised model (OPEC-Net), our SBN-based model generates just slightly lower accuracy and HRNet-based model achieves more promising performance.

Qualitative Comparisons. We carry out the sufficient qualitative comparisons in all cases to validate the effectiveness of our method. In specific, we visualize the adaptation comparison results in Fig. 4, 5, 6 respectively. It shows that our method achieves impressive results compared with the baseline, and achieves a comparable performance with the supervised model (*e.g.*, the result highlighted with the red circle/box).

We also did visual comparisons with other recent supervised pose estimation methods as shown in Fig. 7 on MS-COCO/MPII to MPII/MS-COCO, it can further manifest the robustness of our model. To further manifest the effectiveness of our model on the extreme scenarios, we also randomly select some difficult samples from the OCHuman datasets and conduct some qualitative results in Fig. 8.

TABLE IV
THE COMPARISON RESULTS ON MS-COCO TO OCHUMAN.

Method	mAP	AP.5	AP.75	AP.8	AP.9
Direct Transfer (Source only)					
Mask RCNN [22]	20.2	33.2	24.5	18.3	2.1
SBN [70]	24.1	37.4	26.8	22.6	4.5
CrowdPose [31]	27.5	40.8	29.9	24.8	9.5
OPEC-Net [50]	29.1	41.3	31.4	27.0	12.8
HRNet [58]	26.8	39.6	29.2	25.0	9.3
UDA model (Without ISA module)					
SBN-based	28.5	40.3	29.8	25.6	10.2
HRNet-based	29.5	41.5	31.7	27.4	12.9

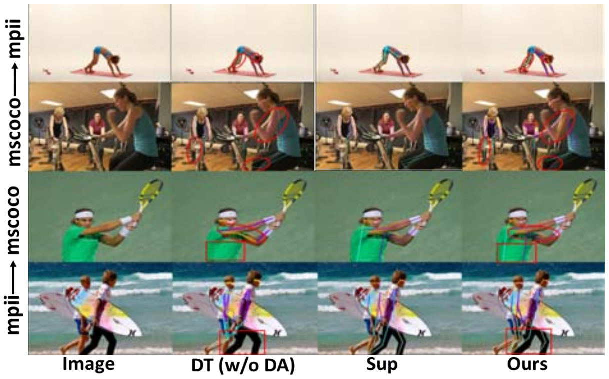

Fig. 4. Results comparison of the baseline (DT (w/o DA)) versus the supervised model (Sup) and Ours on MS-COCO/MPII to MPII/MS-COCO.

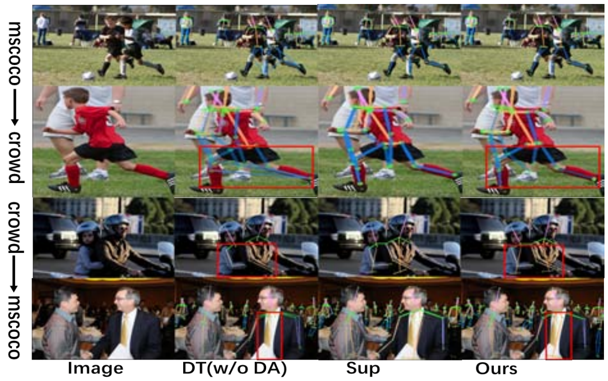

Fig. 5. Results comparison of the baseline (DT (w/o DA)) v.s. the supervised model (Sup) and Ours on MS-COCO/CrowdPose to CrowdPose/MS-COCO.

D. Ablation Study

We conduct ablated experiments on MS-COCO/MPII to MPII/MS-COCO to evaluate the effectiveness of our components. The SimpleBaseline [70] (ResNet-50) is adopted as our default baseline. The results are illustrated in Tab. V, the SL

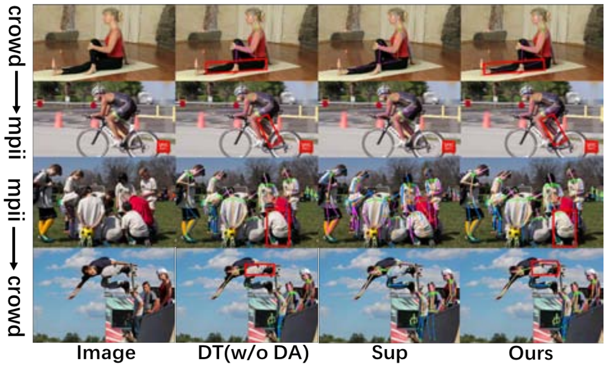

Fig. 6. Results comparison of the baseline (DT (w/o DA)) versus the supervised model (Sup) and Ours on MPII/CrowdPose to CrowdPose/MPII.

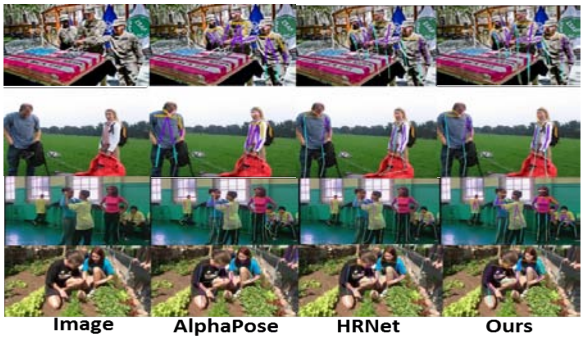

Fig. 7. The visual comparisons between existing methods (they are under fully-supervised setting and without adaptation) and Ours, the samples are randomly selected from the MPII/MS-COCO.

depicts model that both trained and tested on the target domain. DT is stated as above. Ours contains all.

Effectiveness of the pose clustering. To further illustrate the influence of our pose clustering, we also have done the ablated comparison with the baseline as shown in Tab. V. It can be seen that the DPC achieves 1.2% gains compared with the baseline method and further accelerates the better alignment more or less on both cases. Notably, it also effectively benefits the proposed three adaptation mechanisms and helps to improve the prediction result respectively. This indicates that it's meaningful to cluster the different pose modes ahead of time, which avoids some disturbances caused by the arbitrary pose variations during the human-level adaptation to some extent.

Does our method really learn fine-grained and domain-invariant human features? Tab. V shows that CAFA delivers a large performance gain than DT in all cases. The result on MPII to MS-COCO achieves 63.7% AP outperforming DT by 7.0% solely with CAFA. The similar improvements can also be observed on MS-COCO to MPII. The PKCh@0.5 score grows to 84.8%, which achieves 7.5% higher than DT. This indicates that mitigating the feature discrepancy within the clustered space across domains is necessary and effective for addressing domain shift. We also compare CAFA with the other feature adaptation strategies. The result illustrates our CAFA achieves outstanding performance even compared with the recent strong alignment strategy [54].

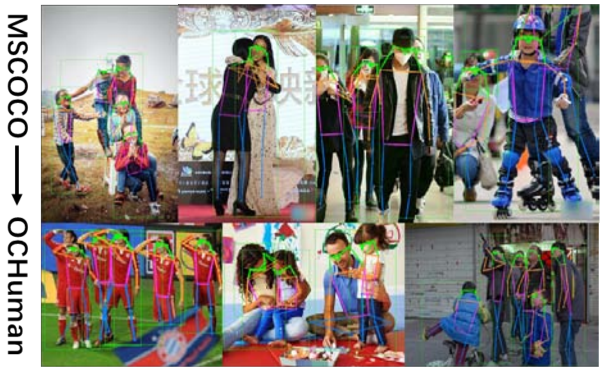

Fig. 8. Qualitative comparison results on MS-COCO to OCHuman.

TABLE V
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+Pose Clustering(DPC)	78.5	57.9
DT+DAN [36]	79.1	57.2
DT+Adversial DA [18]	79.5	58.1
DT+SW Detection [54]	81.6	60.1
DT+CAFA	84.5	63.4
DPC+CAFA	84.8	63.7
DT+IPSL	83.1	62.8
DT+AHTA	86.0	65.0
DPC+AHTA	86.2	65.2
DT+ISA	83.6	62.5
DPC+ISA	83.9	62.8
Ours	88.1	66.9

To illustrate CAFA can adapt the fine-grained human features, we visualize the feature maps on both domains in Fig. 9. As shown in red-dashed boxes, the feature responses produced by our CAFA looks more explicit and complete (*i.e.*, most of the human part responses are activated) and by comparison the 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 CAFA can effectively discover the similar fine-grained body parts features for each domain and enhance the respective features.

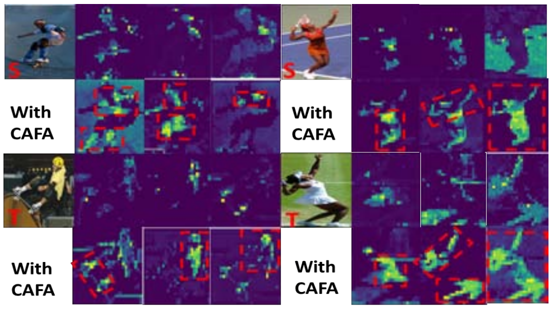

Fig. 9. Feature visualization with or without CAFA, best viewed in color.

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

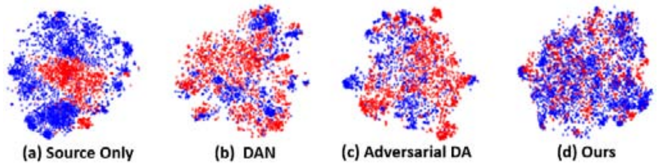

Fig. 10. Visualization of features using t-SNE: (a) the baseline (b) DAN (c) adversarial DA (d) Ours. Note that the blue and red points are samples from the source and target domain respectively, best viewed in color.

Ours respectively with t-SNE [15] in Fig. 10 $(a)-(d)$. 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' [36], 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' [18], 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.

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. V. It verifies our model with ISA achieves substantial improvements over the baseline in two adaptation cases. Concretely, the accuracy increases from 77.3% to 83.9% directly on MS-COCO to MPII and 56.7% to 62.8% on MPII to MS-COCO.

As shown in Fig. 11, 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.

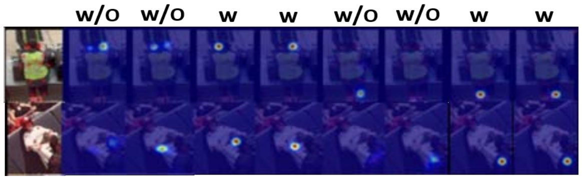

Fig. 11. The keypoint heatmap predictions visualization. The 'w/o' depicts the results of DT, 'w' depicts with ISA.

Analysis of AHTA. Tab. V shows that DT model behaves unsatisfying and only achieves poor result and by comparison, the DT with AHTA is improved by 8.9% and 8.5% on the MS-COCO to MPII and MPII to MS-COCO respectively. This obviously proves the effectiveness of the AHTA. Additionally, the baseline with ISPL can also achieve a better performance and indirectly proves that adopting the graph model to capture the pose structural representations is effective. Additionally, compared with the other components, AHTA makes the greatest contribution. This proves the significance of exploring the cross-instance high-order topological relations across domains.

The AHTA mainly consists of two parts: IPSL and CITA. The former provides an important insurance for the latter. Since unless we acquire sufficient robust human structural information, our CITA can capture better domain-invariant topological representations across instances. Importantly, CITA is a great innovation of AHTA to explore the cross-instance topological relations flexibly, which formulates the high-order topological adaptation via graph attention mechanism.

For further explaining CITA intuitively, we visualize the learned weight map of several instance pairs across domains in Fig. 12 (on the left side). From the learned weight matrix, we observe that if the human structural representations between two instances are of higher similarity, the CITA assigns a larger weight to them, otherwise, assigns a smaller weight. This demonstrates the cross-instance topological similarity can be adaptively captured via CITA mechanism. The specific correlation matrixs of the specific cross-instance are visualized on the right side of the figure. The top one depicts the two instances with a higher similarity (*i.e.*, the weight value of the 's3' and 't7' is 0.78), which shows a relatively consistent and stronger relevance of category-specific keypoint responses. On the contrary, the bottom one represents the examples (i.e., the weight value of the 's6' and 't7' is 0.21) with a lower structural similarity. Their corresponding matrix looks irregular and shows a lower relevance between the feature responses of their keypoint.

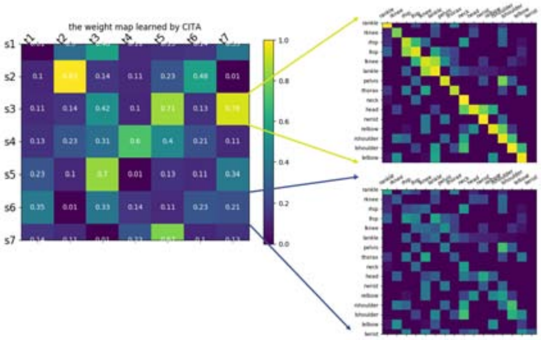

Fig. 12. The visualized weight map to prove the adaptive capability of CITA.

Moreover, we visualize the qualitative results of AHTA in Fig. 13, it shows that the model with AHTA can help infer the human structure cues for each domain (*e.g.*, the occluded right ankle in first two images and the right ankle/knee in last two images) via the learned domain-invariant human topological knowledge assistance from each other.

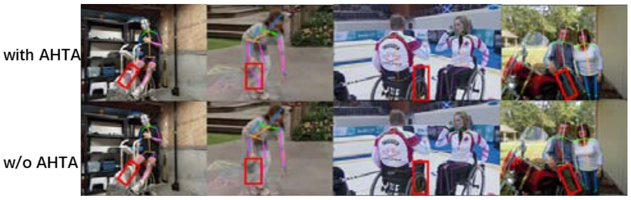

Fig. 13. Qualitative comparisons of AHTA. The top row depicts results of the model with AHTA. The bottom depicts results of baseline without AHTA.

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

TABLE VI
SENSITIVITY OF MAP TO THE HYPER-PARAMETER α, β ON MPII TO MS-COCO.

α || β	0.3	0.4	0.5	0.6	0.7	0.8	0.9
mAP (α)	63.4	64.3	**66.9**	65.4	63.5	62.4	60.1
mAP (β)	62.6	65.5	**66.9**	65.1	62.7	62.9	63.4

As for β, the accuracy also shows a growing trend as the value changes from $0.3 \sim 0.5$. It illustrates that our method benefits 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.

V. CONCLUSION

In this paper, we propose a novel cross domain multi-person pose estimation framework. To achieve precise adaptation estimation, we conduct the pose clustering in shared domains to discover the multiple pose modes with similar postures. Based on the constructed clusters, we firstly adopt a cross-attentive feature alignment to learn the domain-invariant fine-grained local keypoint features. In SSDA setting, we additionally propose to adapt the corresponding keypoint heatmap representations to reduce the intra-domain gap. To better align the human topological structural knowledge, our AHTA firstly models the intra-pose structure via SemGCN and further adaptively augments the domain-specific structure-invariant knowledge via GAT for each domain. Extensive experiments show that our method significantly boosts performance of the target domain even with no labels or sparse labels.

REFERENCES

[1] Mykhaylo Andriluka, Leonid Pishchulin, Peter V Gehler, and Bernt Schiele. 2d human pose estimation: New benchmark and state of the art analysis. *In Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3686–3693, 2014.

[2] Gedas Bertasius, Christoph Feichtenhofer, Du Tran, Jianbo Shi, and Lorenzo Torresani. Learning temporal pose estimation from sparsely-labeled videos. *In Advances in neural information processing systems*, pages 3021–3032, 2019.

[3] Yanrui Bin, Xuan Cao, Xinya Chen, Yanhao Ge, Ying Tai, Chengjie Wang, Jilin Li, Feiyue Huang, Changxin Gao, and Nong Sang. Adversarial semantic data augmentation for human pose estimation. *European Conference on Computer Vision*, pages 606–622, 2020.

[4] Guillem Brasó, Nikita Kister, and Laura Leal-Taixé. The center of attention: Center-keypoint grouping via attention for multi-person pose estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11853–11863, 2021.

[5] Adrian Bulat, Jean Kossaifi, Georgios Tzimiropoulos, and Maja Pantic. Toward fast and accurate human pose estimation via soft-gated skip connections. *2020 15th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2020)*, pages 8–15, 2020.

[6] Jinkun Cao, Hongyang Tang, Haoshu Fang, Xiaoyong Shen, Yuwing Tai, and Cewu Lu. Cross-domain adaptation for animal pose estimation. *In Proceedings of the IEEE International Conference on Computer Vision*, pages 9498–9507, 2019.

[7] Zhe Cao, Gines Hidalgo, Tomas Simon, Shihen Wei, and Yaser Sheikh. Openpose: Realtime multi-person 2d pose estimation using part affinity fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 1–1, 2019.

[8] Yilun Chen, Zhicheng Wang, Yuxiang Peng, Zhiqiang Zhang, Gang Yu, and Jian Sun. Cascaded pyramid network for multi-person pose estimation. *In Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7103–7112, 2018.

[9] Yu Chen, Chunhua Shen, Xiushen Wei, Lingqiao Liu, and Jian Yang. Adversarial posenet: A structure-aware convolutional network for human pose estimation. *In Proceedings of the IEEE International Conference on Computer Vision*, pages 1221–1230, 2017.

[10] Bowen Cheng, Bin Xiao, Jingdong Wang, Honghui Shi, Thomas S Huang, and Lei Zhang. Higherhrnet: Scale-aware representation learning for bottom-up human pose estimation. *In Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, 2020.

[11] Ouk Choi, Young-Jun Son, Hwasup Lim, and Sang Chul Ahn. Co-recognition of multiple fingertips for tabletop human–projector interaction. *IEEE Transactions on Multimedia*, 21(6):1487–1498, 2019.

[12] Chiajung Chou, Juiting Chien, and Hwanntzong Chen. Self adversarial training for human pose estimation. *Asia Pacific Signal and Information Processing Association Annual Summit and Conference*, pages 17–30, 2018.

[13] Xiao Chu, Wei Yang, Wanli Ouyang, Cheng Ma, Alan L Yuille, and Xiaogang Wang. Multi-context attention for human pose estimation. *In Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5669–5678, 2017.

[14] Yan Dai, Xuanhan Wang, Lianli Gao, Jingkuan Song, and Heng Tao Shen. Rsgnet: Relation based skeleton graph network for crowded scenes pose estimation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 1193–1200, 2021.

[15] Laurens Van Der Maaten and Geoffrey E Hinton. Visualizing data using t-sne. *Journal of Machine Learning Research*, pages 2579–2605, 2008.

[16] Haoshu Fang, Shuqin Xie, Yuwing Tai, and Cewu Lu. Rmpe: Regional multi-person pose estimation. *In Proceedings of the IEEE International Conference on Computer Vision*, pages 2353–2362, 2017.

[17] Yaroslav Ganin and Victor Lempitsky. Unsupervised domain adaptation by backpropagation. *In Proceedings of the 32th International Conference on Machine Learning*, pages 1180–1189, 2015.

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

[19] Yves Grandvalet and Yoshua Bengio. Semi-supervised learning by entropy minimization. *In Advances in neural information processing systems*, pages 529–536, 2004.

[20] Arthur Gretton, Karsten M Borgwardt, Malte J Rasch, Bernhard Scholkopf, and Alexander J Smola. A kernel method for the two-sample-problem. *In Advances in neural information processing systems*, pages 513–520, 2006.

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

[22] Kaiming He, Georgia Gkioxari, Piotr Dollar, and Ross Girshick. Mask r-cnn. *In Proceedings of the IEEE International Conference on Computer Vision*, pages 2980–2988, 2017.

[23] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *In Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[24] Judy Hoffman, Eric Tzeng, Taesung Park, Junyan Zhu, Phillip Isola, Kate Saenko, Alexei A Efros, and Trevor Darrell. Cycada: Cycle-consistent adversarial domain adaptation. *In Proceedings of the 35th International Conference on Machine Learning*, pages 1989–1998, 2018.

[25] Judy Hoffman, Dequan Wang, Fisher Yu, and Trevor Darrell. Fcns in the wild: Pixel-level adversarial and constraint-based adaptation. *arXiv preprint arXiv: 1612.02649*, 2016.

[26] Sheng Jin, Wentao Liu, Enze Xie, Wenhai Wang, and Ping Luo. Differentiable hierarchical graph grouping for multi-person pose estimation. *In European Conference on Computer Vision. Springer*, 2020.

[27] Aouaidjia Kamel, Bin Sheng, Ping Li, Jinman Kim, and David Dagan Feng. Hybrid refinement-correction heatmaps for human pose estimation. *IEEE Transactions on Multimedia*, PP(99):1–1, 2020.

[28] Lipeng Ke, Mingching Chang, Honggang Qi, and Siwei Lyu. Multi-scale structure-aware network for human pose estimation. *In European Conference on Computer Vision. Springer,*, pages 731–746, 2018.

[29] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *In International Conference on Learning Representations*, 2015.

[30] Muhammed Kocabas, Salih Karagoz, and Emre Akbas. Multiposenet: Fast multi-person pose estimation using pose residual network. *In European Conference on Computer Vision. Springer,*, pages 437–453, 2018.

[31] Jiefeng Li, Can Wang, Hao Zhu, Yihuan Mao, Haoshu Fang, and Cewu Lu. Crowdpose: Efficient crowded scenes pose estimation and a new benchmark. *In Proceedings of the IEEE conference on computer vision and pattern recognition*, 2019.

[32] Ke Li, Shijie Wang, Xiang Zhang, Yifan Xu, Weijian Xu, and Zhuowen Tu. Pose recognition with cascade transformers. In *CVPR*, pages 1944–1953, 2021.

[33] Miaopeng Li, Zimeng Zhou, and Xinguo Liu. Multi-person pose estimation using bounding box constraint and lstm. *IEEE Transactions on Multimedia*, PP(99):1–1, 2019.

[34] Yanjie Li, Shoukui Zhang, Zhicheng Wang, Sen Yang, Wankou Yang, Shu-Tao Xia, and Erjin Zhou. Tokenpose: Learning keypoint tokens for human pose estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11313–11322, 2021.

[35] Tsungyi Lin, Michael Maire, Serge J Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollar, and C Lawrence Zitnick. Microsoft coco: Common objects in context. *In European Conference on Computer Vision. Springer,*, pages 740–755, 2014.

[36] Mingsheng Long, Yue Cao, Jianmin Wang, and Michael I Jordan. Learning transferable features with deep adaptation networks. *In Proceedings of the 32th International Conference on Machine Learning*, pages 97–105, 2015.

[37] Mingsheng Long, Zhangjie Cao, Jianmin Wang, and Michael I Jordan. Conditional adversarial domain adaptation. *In Advances in neural information processing systems*, pages 1640–1650, 2018.

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

[39] Diogo C Luvizon, Hedi Tabia, and David Picard. Human pose regression by combining indirect part detection and contextual information. *Computers & Graphics*, 85:15–22, 2019.

[40] Weian Mao, Zhi Tian, Xinlong Wang, and Chunhua Shen. Fcpose: Fully convolutional multi-person pose estimation with dynamic instance-aware convolutions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9034–9043, 2021.

[41] Angel Martinezgonzalez, Michael Villamizar, Olivier Canevet, and Jeanmarc Odobez. Investigating depth domain adaptation for efficient human pose estimation. *In European Conference on Computer Vision*, pages 346–363, 2018.

[42] Alejandro Newell, Zhiao Huang, and Jia Deng. Associative embedding: End-to-end learning for joint detection and grouping. *In Advances in neural information processing systems*, pages 2277–2287, 2017.

[43] Alejandro Newell, Kaiyu Yang, and Jia Deng. Stacked hourglass networks for human pose estimation. *In European Conference on Computer Vision. Springer,*, pages 483–499, 2016.

[44] Xuecheng Nie, Jiashi Feng, Jianfeng Zhang, and Shuicheng Yan. Single-stage multi-person pose machines. *In Proceedings of the IEEE International Conference on Computer Vision*, pages 6951–6960, 2019.

[45] Guanghan Ning, Student Member, IEEE, Zhi Zhang, and Student Member. Knowledge-guided deep fractal neural networks for human pose estimation. *IEEE Transactions on Multimedia*, 20(5):1246–1259, 2018.

[46] Guanghan Ning, Zhi Zhang, and Zhihai He. Knowledge-guided deep fractal neural networks for human pose estimation. *IEEE Transactions on Multimedia*, pages 1246–1259, 2018.

[47] Boxiao Pan, Zhangjie Cao, Ehsan Adeli, and Juan Carlos Niebles. Adversarial cross-domain action recognition with co-attention. *Association for Advancement of Artificial Intelligence*, 2020.

[48] George Papandreou, Tyler Zhu, Nori Kanazawa, Alexander Toshev, Jonathan Tompson, Chris Bregler, and Kevin P Murphy. Towards accurate multi-person pose estimation in the wild. *In Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3711–3719, 2017.

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

[50] Lingteng Qiu, Xuanye Zhang, Yanran Li, Guanbin Li, Xiaojun Wu, Zixiang Xiong, Xiaoguang Han, and Shuguang Cui. Peeking into occluded joints: A novel framework for crowd pose estimation. *In European Conference on Computer Vision. Springer*, 2020.

[51] Zhongwei Qiu, Kai Qiu, Jianlong Fu, and Dongmei Fu. Dgcn: Dynamic graph convolutional network for efficient multi-person pose estimation. *National Conference on Artificial Intelligence*, 2020.

[52] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael S Bernstein, et al. Imagenet large scale visual recognition challenge. *International Journal of Computer Vision*, 115(3):211–252, 2015.

[53] Kuniaki Saito, Donghyun Kim, Stan Sclaroff, Trevor Darrell, and Kate Saenko. Semi-supervised domain adaptation via minimax entropy. *In Proceedings of the IEEE International Conference on Computer Vision*, pages 8050–8058, 2019.

[54] Kuniaki Saito, Yoshitaka Ushiku, Tatsuya Harada, and Kate Saenko. Strong-weak distribution alignment for adaptive object detection. *In Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 6956–6965, 2019.

[55] Franco Scarselli, Marco Gori, Ah Chung Tsoi, Markus Hagenbuchner, and Gabriele Monfardini. The graph neural network model. *IEEE Transactions on Neural Networks*, 20(1):61–80, 2009.

[56] Yu Su and Wang Xu. Multi-person pose estimation with enhanced channel-wise and spatial information. *In Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, 2019.

[57] Ke Sun, Cuiling Lan, Junliang Xing, Wenjun Zeng, Dong Liu, and Jingdong Wang. Human pose estimation using global and local normalization. *In Proceedings of the IEEE International Conference on Computer Vision*, pages 5600–5608, 2017.

[58] Ke Sun, Bin Xiao, Dong Liu, and Jingdong Wang. Deep high-resolution representation learning for human pose estimation. *In Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 5693–5703, 2019.

[59] Wei Tang, Pei Yu, and Ying Wu. Deeply learned compositional models for human pose estimation. *In European Conference on Computer Vision. Springer,*, pages 197–214, 2018.

[60] Zhiqiang Tang, Xi Peng, Shijie Geng, Lingfei Wu, Shaoting Zhang, and Dimitris N Metaxas. Quantized densely connected u-nets for efficient landmark localization. *In European Conference on Computer Vision. Springer,*, pages 348–364, 2018.

[61] Yihsuan Tsai, Weichih Hung, Samuel Schulter, Kihyuk Sohn, Minghsuan Yang, and Manmohan Chandraker. Learning to adapt structured output space for semantic segmentation. *In Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 7472–7481, 2018.

[62] Yihsuan Tsai, Kihyuk Sohn, Samuel Schulter, and Manmohan Chandraker. Domain adaptation for structured output via discriminative patch representations. *In Proceedings of the IEEE International Conference on Computer Vision*, pages 1456–1465, 2019.

[63] Eric Tzeng, Judy Hoffman, Kate Saenko, and Trevor Darrell. Adversarial discriminative domain adaptation. *In Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 2962–2971, 2017.

[64] Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Liò, and Yoshua Bengio. Graph attention networks. *International Conference on Learning Representations*, 2018.

[65] Jian Wang, Xiang Long, Yuan Gao, Errui Ding, and Shilei Wen. Graph-pcnn: Two stage human pose estimation with graph pose refinement. *In European Conference on Computer Vision. Springer*, 2020.

[66] Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He. Non-local neural networks. *In Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 7794–7803, 2018.

[67] Ping Wei, Hongbin Sun, and Nanning Zheng. Learning composite latent structures for 3d human action representation and recognition. *IEEE Transactions on Multimedia*, pages 1–1, 2019.

[68] Shihen Wei, Varun Ramakrishna, Takeo Kanade, and Yaser Sheikh. Convolutional pose machines. *In Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4724–4732, 2016.

[69] Zuxuan Wu, Xintong Han, Yenliang Lin, Mustafa Gokhan Uzunbas, Tom Goldstein, Sernam Lim, and Larry S Davis. Dcan: Dual channel-wise alignment networks for unsupervised scene adaptation. *In European Conference on Computer Vision. Springer,*, pages 518–534, 2018.

[70] Bin Xiao, Haiping Wu, and Yichen Wei. Simple baselines for human pose estimation and tracking. *In European Conference on Computer Vision. Springer,*, pages 472–487, 2018.

[71] Yabo Xiao, Dongdong Yu, Xiaojuan Wang, Lei Jin, Guoli Wang, and Qian Zhang. Learning quality-aware representation for multi-person pose regression. *arXiv preprint arXiv:2201.01087*, 2022.

[72] Xixia Xu, Qi Zou, and Xue Lin. Alleviating human-level shift: A robust domain adaptation method for multi-person pose estimation. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 2326–2335, 2020.

[73] Sijie Yan, Yuanjun Xiong, Dahua Lin, and Xiaoou Tang. Spatial temporal graph convolutional networks for skeleton-based action recognition. *National Conference on Artificial Intelligence*, pages 7444–7452, 2018.

[74] Baoyao Yang, Andy J Ma, and Pong C Yuen. Body parts synthesis for cross-quality pose estimation. *IEEE Transactions on Circuits and Systems for Video Technology*, 29(2):461–474, 2019.

[75] Sen Yang, Zhibin Quan, Mu Nie, and Wankou Yang. Transpose: Keypoint localization via transformer. *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11802–11812, 2021.

[76] Wei Yang, Shuang Li, Wanli Ouyang, Hongsheng Li, and Xiaogang Wang. Learning feature pyramids for human pose estimation. *In Proceedings of the IEEE International Conference on Computer Vision*, pages 1290–1299, 2017.

[77] Wei Yang, Wanli Ouyang, Xiaolong Wang, Jimmy Ren, Hongsheng Li, and Xiaogang Wang. 3d human pose estimation in the wild by adversarial learning. *In Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 5255–5264, 2018.

[78] Jiabin Zhang, Zheng Zhu, Jiwen Lu, Junjie Huang, Guan Huang, and Jie Zhou. Simple: Single-network with mimicking and point learning for bottom-up human pose estimation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 3342–3350, 2021.

[79] Songhai Zhang, Ruilong Li, Xin Dong, Paul L Rosin, Zixi Cai, Xi Han, Dingcheng Yang, Haozhi Huang, and Shimin Hu. Pose2seg: Detection free human instance segmentation. *In Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 889–898, 2019.

[80] Weichen Zhang, Wanli Ouyang, Wen Li, and Dong Xu. Collaborative and adversarial network for unsupervised domain adaptation. *In Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 3801–3809, 2018.

[81] Xiheng Zhang, Yongkang Wong, Mohan S Kankanhalli, and Weidong Geng. Unsupervised domain adaptation for 3d human pose estimation. *In Proceedings of the 27rd ACM international conference on Multimedia. ACM*, pages 926–934, 2019.

[82] Long Zhao, Xi Peng, Yu Tian, Mubbasir Kapadia, and Dimitris N Metaxas. Semantic graph convolutional networks for 3d human pose regression. *In Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 3425–3435, 2019.

[83] Xinge Zhu, Jiangmiao Pang, Ceyuan Yang, Jianping Shi, and Dahua Lin. Adapting object detectors via selective cross-domain alignment. *In Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 687–696, 2019.