ORGM: Occlusion Relational Graphical Model for Human Pose Estimation

Lianrui Fu, *Student Member, IEEE*, Junge Zhang, *Member, IEEE*, Kaiqi Huang, *Senior Member, IEEE*

***Abstract*— Articulated human pose estimation from monocular image is a challenging problem in computer vision. Occlusion is a main challenge for human pose estimation, which is largely ignored in popular tree structured models. The tree structured model is simple and convenient for exact inference, but short in modeling the occlusion coherence especially in the case of self-occlusion. We propose an occlusion relational graphical model, which is able to model both self-occlusion and occlusion by the other objects simultaneously. The proposed model can encode the interactions between human body parts and objects, and enables it to learn occlusion coherence from data discriminatively. We evaluate our model on several public benchmarks for human pose estimation, including challenging subsets featuring significant occlusion. The experimental results show that our method is superior to the previous state-of-the-arts, and is robust to occlusion for 2D human pose estimation.**

***Index Terms*— Occlusion, pose estimation, spacial relationship, mixture, graphical model.**

I. Introduction

Articulated human pose estimation from still image is a challenging problem in computer vision. The goal is to determine the human body configuration in 2D space from an input image. It is key to many visual tasks, e.g., action recognition, clothes parsing and human-computer interaction. This task is challenging due to large deformation, illumination, camera viewpoint, cluttered background and occlusion.

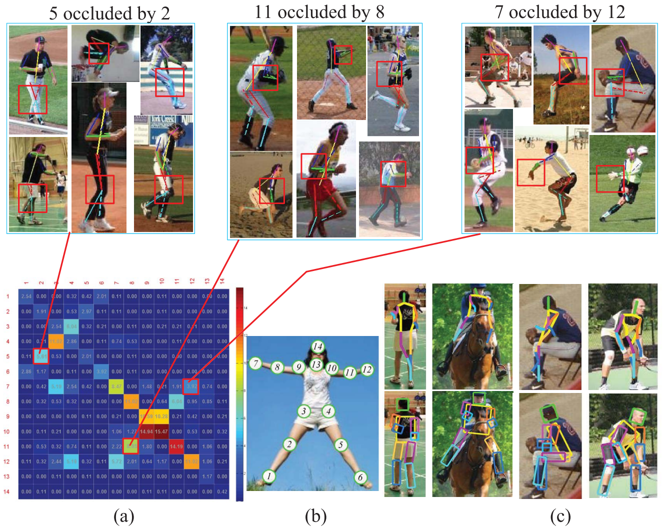

Fig. 1. Occlusions in Leeds Sports dataset. (a) Visualization of the occlusion relation matrix. The color of diagonal squares reflects the probability of other-occlusion for each joint. Hotter color means heavier occlusion. The hotness of each non-diagonal square in row i and column j indicates the probability of joint i being occluded by joint j. The image groups on the top are the instances with the same self-occlusion relationship. (b) The sequence number of body joints in Leeds Sports dataset. (c) Human pose estimation results with occlusions from FMP [1](the first row) and ours(second row).

Recent progress on human pose estimation is ascribed to the pictorial structured model especially simple tree structure [1]–[6]. Although these methods perform well on images with rare occlusion, they may fail when the body parts are occluded by some other body parts (self-occlusion) or the other objects (other-occlusion). Fig. 1(c) depicts that the famous flexible mixtures-of-parts model (FMP) [1] fails under occlusion. The tree structured model is simple, yet fails to model the interaction between unconnected body parts, and the interactions between human body and objects. However, these interactions are important cues for occlusion reasoning. The question is, how can we model such interactions for occlusion reasoning?

There are mainly two types of occlusion for human pose estimation: other-occlusion (occluded by some other objects) and self-occlusion (occluded by some other body parts). Other-occlusion appears when some objects block the view and this will damage the local appearance of body parts and cause failure detection. Take the images of the second column of Fig. 1(c) for example, the FMP fails because the left knee and hip of the rider are occluded by the horse. However, only a few body parts are frequently occluded such as lower arms and legs while head is mostly visible (see the diagonal elements in the occlusion relation matrix).

In contrast to other-occlusion, self-occlusion appears when the body parts occlude each other due to viewpoint or

Manuscript received March 9, 2016; revised August 9, 2016; accepted December 1, 2016. Date of publication December 13, 2016; date of current version January 5, 2017. This work was supported in part by the National Natural Science Foundation of China under Grant 61322209, Grant 61673375, and Grant 61403387, and in part by the International Partnership Program of the Chinese Academy of Science under Grant 173211KYSB20160008. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Nilanjan Ray. *(Lianrui Fu and Junge Zhang contributed equally to this work.) (Corresponding authors: Kaiqi Huang.)*

L. Fu and J. Zhang are with the National Laboratory of Pattern Recognition, Center for Research on Intelligent Perception and Computing, Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China (e-mail: lianrui.fu@nlpr.ia.ac.cn; jgzhang@nlpr.ia.ac.cn).

K. Huang is with the National Laboratory of Pattern Recognition, Center for Research on Intelligent Perception and Computing, Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China, and also with the CAS Center for Excellence in Brain Science and Intelligence Technology, University of Chinese Academy of Sciences, Beijing 100049, China (e-mail: kqhuang@nlpr.ia.ac.cn).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TIP.2016.2639441

pose deformation. In this case, the symmetric body parts on the left and the right side will be easily localized in the same region in 2D image. Self-occlusion is less likely to damage the local appearance of the occludee, yet will cause the ambiguity of pose configuration as there is no interaction between the occluder and the occludee. For example, as seen from the third and fourth column of Fig. 1(c), FMP fails to capture the correct pose configuration under self-occlusion.

Statistics on the Leeds Sport Pose dataset (LSP) [7] show that 47.2% of the images have one more body joints invisible and 16.7% have more than three body joints occluded. Among all these invisible joints, 67.4 other-occluded. Existing works mainly focus on handling other-occlusion, self-occlusion is often ignored or treated in the same manner as other-occlusion. We argue that the occluder in self-occlusion cannot be treated as noise as that in other-occlusion. How can we model both kinds of occlusion in a unified framework simultaneously?

In addition, Convolutional Neural Network (CNN) [8] has been introduced in human pose estimation [9]–[13]. Especially, the approach of Image Dependent Pairwise Relational with Deep Convolutional Neural Network (IDPR-DCNN) [10] is proposed to learn the Image Dependent Pairwise Relational (IDPR) term with CNN and has achieved state-of-the-art performance. However, all these methods above do not consider occlusion handling. How can we take the advantage of strong feature representation of CNN to improve the performance even under occlusion?

Motivated by these above, we propose a novel occlusion relational graphical model (ORGM) which explicitly models both self-occlusion and other-occlusion to improve the robustness. The proposed model can encode the interactions between human body parts and objects, and hence enables it to learn occlusion coherence from data discriminatively. Our ORGM can be considered as a novel extension of tree structured model with occlusion handling. We take the FMP [1] and IDPR-DCNN [10] as baseline methods as they are the most successful tree structured models for human pose estimation.

In this work, there are several main differences from previous studies:

1) Simultaneous modeling for both self-occlusion and other-occlusion. Apparently, there is great difference in mechanism between self-occlusion and other-occlusion. Is it possible to model the two different sources of occlusion in one unified framework? The answer is yes. We show that both self-occlusion and other-occlusion can be modeled simultaneously within our occlusion relational graphical model.

2) Efficient inference for the proposed model. As the proposed model is a graph containing loops, we propose a two-step process to perform inference on the non-tree model. The first step generates candidates based on the "unrolled" tree model. The second step then rescores the candidates based on the scoring function of the full model with loops.

3) Extension of the proposed model with CNN part detectors to achieve state-of-the-art performance for human pose estimation.

We evaluate our model on several public benchmarks for human pose estimation and test on the challenging subsets with significant occlusion. The results verify the effectiveness of the proposed method which obtains competitive accuracy to the previous state-of-the-art methods on public datasets. In particular, our method performs much better than previous methods on those datasets with heavy occlusion.

There are three contributions in this paper:

- We propose an occlusion relational graphical model which is able to model both self-occlusion and occlusion by the other objects simultaneously. The proposed model structure can encode the interactions between human body parts and objects, and hence enables it to learn occlusion coherence from data discriminatively.
- A two-step process is proposed to infer the proposed non-tree model efficiently. It can conduct inference with the same order of speed as tree structured model yet takes the advantage of richer model representation.
- We extend our occlusion relational graphical model with deep CNN model and get superior performance than the pervious state-of-the-arts.

Compared with our previous work [14], the novel contributions of this paper are mainly twofold. One aspect is the extension of occlusion relational graphical model with deep CNN model to achieve superior performance than the pervious state-of-the-arts. The other one is the promotion of our model into a more general ORGM framework including generic model representation, inference and model learning. We also provide more comprehensive experimental results and detailed analysis.

The rest of this paper is organized as follows. In Sec.II, we first review the related studies on occlusion modeling in human pose estimation. Then, we present both generic and specific representations of the proposed model, as well as its inference and learning in Sec.III respectively. To evaluate and analyze the proposed method, we provide detailed experimental results in Sec.IV. Finally, Sec.V summarizes the paper.

II. Related Work

In this section, we will first overview structured model which has dominated the study of human pose estimation for decades. As we address the problem of occlusion handling, we will then review previous work on occlusion modeling in human pose estimation. For the proposed occlusion relational graphical model which contains loops, the inference is intractable. So previous methods on the inference of loopy models are introduced at the end of this section.

A. Structured Model

Previous studies on human pose estimation mainly focus on structured model which incorporates part level constraints and local feature representations.

The most popular modern approaches for human pose estimation is based on the pictorial structured model (PSM) [15]. In the PSM, the human body configuration is represented as a collection of independent parts with pairwise connections. The pairwise part relationships are embodied in tree models [1]–[6], [16], [17], multi-tree model [18], context aware model [19] or loopy models [20]–[24]. Tree models prevail for its simplicity and exact inference. However, they are insufficient in capturing complex spacial relationships among body parts and the message passing tends to break down when occlusion occurs. Loopy models allow more complex relationships among parts, but require approximate inference iteratively. Our occlusion relational graphical model is able to model such interactions among parts with efficient approximate inference.

Feature representation and even middle level feature representations are also important for structured model. The most popular local features for human pose estimation include Shape Context [25], Histogram of Gradients (HOG) [26] and Convolutional Neural Network (CNN) [8]. Strong CNN features are extracted in [9]–[13] and [27]. Poselets [28] and Deformable Part-based Model (DPM) [29] are adopted to generate rich middle level representations with strong pose priors [17], [30]. Some incorporates CNN part detectors and graphical models with either piecewise training [10] or joint training [11]. In contrast to modeling pairwise constraints, some propose [31], [32] to use layered random forest to incorporate rich spatial interactions among multiple parts. However, there is no explicit modeling of occlusion in these approaches.

B. Occlusion Modeling

In terms of handling occlusion of pose estimation, body part visibility is usually modeled as binary variable in either part level or image level. Some previous object detection approaches [33], [34] model occlusion with segmentation of image feature map. Part level occlusion reasoning is frequently used to model more complicated occlusions. For instance, the supervised part model [35] includes visibility variable for each part but imposes no constraints on the visibility of different parts in the model. Similarly, Hejrati and Ramanan [36] extend the flexible mixtures-of-part model [1] with part level occlusion reasoning for 3D car alignment. Desai and Ramanan [37] model the interactions between human and objects which can capture the occlusion relationships. Wang and Mori [18] propose to combine multiple tree framework for occlusion reasoning. The And-Or graph model [38] also incorporates visibility into the part node. The grammar-based model [39] in people detection includes explicit occlusion part templates but enforces more structure in the pattern of occlusion. The strongly supervised deformable model [5], by contrast, tries to sidestep the structure learning problem and automatically learn valid occlusion patterns from data in a non-parametric way. The flexible compositions [6], on the other hand, model visible parts with subtrees and learn occlusion cues with CNNs.

Most of the work above mainly focus on other-occlusion but self-occlusion is often ignored or treated in the same manner as other-occlusion (as noise). There are only a few works trying to model self-occlusion. Sigal and Black [40] propose to use pixel level hidden binary variables for self-occlusion reasoning. Some others try to model self-occlusion in a holistic manner. For instance, Yang and Sundaramoorthi [41] model self-occlusion of pedestrian in a joint shape and appearance tracking framework, Radwan *et al.* [42] treat self-occlusion reasoning as post process using Twin-GP regression method for 2D pose rectification. However, our model learns the part-level occlusion relationships from data and infers the occlusion states of parts explicitly. Our model is more flexible and can encode more complex interactions between parts.

C. Model Inference

In the problem of human pose estimation, tree models are popular for its simplicity and exact inference. In the tree models [1]–[6], [16], [17] the message is passed from the child node to its single parent node sequentially (see Fig. 2(a)), and thus dynamic programming and distance transform [43] can be utilized to compute the best score for each root location efficiently. This is beneficial for the mining of huge amount of hard negative examples and will accelerate the training of tree models.

For those non-tree models which contain loops, the time complexity for exact inference scales exponentially in the size of the largest clique in the graph [44]. Some propose to transform the problem into integer programming [45] or integer quadratic programming [46] and solve it with generalized linear programming solvers. Many other approximate methods, such as Loopy Belief Prorogation [47], Branch and Bound [20] and Dual Decomposition [48], need to iteratively infer on tractable structures many times until converge. Generalized Range Move Algorithm [49] is proposed for efficient optimization of MRFs. For our problem, each part has hundreds or thousands of possible locations and it is impractical to perform exact inference. Moreover, there are a huge number of negative examples, the mining of hard negative examples with iterative approximate inference is very expensive. So efficient inference is also important for human pose estimation, especially for those graphical models with loops.

III. Occlusion Relational Graphical Model

In this section, we will first introduce the generic form of our model, then the specific representation of the proposed occlusion relational graphical model based on FMP [1] and IDPR-DCNN [10], and finally describe the inference and learning procedure of our models with generic formulation.

A. Generic Model

In the pictorial structured model, human body configuration is represented in a conditional random field with a collection of independent parts and pairwise constraints. We denote $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ be the N-node graph, where $\mathcal{V}$ represents the set of body parts, $\mathcal{E}$ denotes the set of constraint edges and $N = |\mathcal{V}|$ is the number of nodes. The position of the i-th part node is denoted as z_i, where $z_i = (x_i, y_i)$. The set of pairwise spacial relationships are denoted as $\boldsymbol{t} = \{t_{ij}, t_{ji} | (i, j) \in \mathcal{E}\}$, where $t_{ij}, t_{ji} \in \{1, \cdots, T_{ij}\}$ reflect the relative position between part i and part j. The pose configuration can be represented as $X = (\boldsymbol{z}, \boldsymbol{t})$, where $\boldsymbol{z} = \{z_i\}$ is the set of part positions.

1) Tree Structured Model: Tree structured model is prevalent for its simplicity and exact inference. In the kinetic tree, the pairwise constraints only exist between adjacent parts.

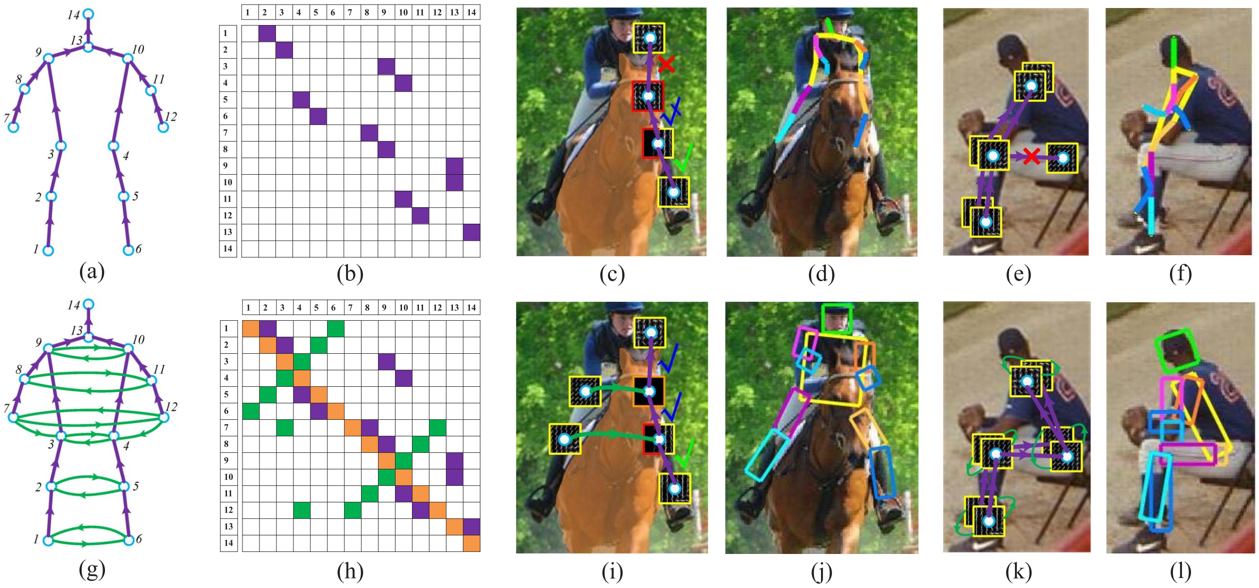

Fig. 2. Comparison of the proposed model structures with respect to that of FMP [1]. (a) Model structure of FMP, which is a kinetic tree. (b) The kinetic constraints represented in adjacent matrix of body parts. (c) The message passing of FMP breaks down under other-occlusion. (d) Failure detection of FMP under other-occlusion. (e) The message passing of FMP breaks down under self-occlusion. (f) Failure detection of FMP under self-occlusion. (g) Model structure of the proposed method. (h) The constraints and occlusion representation in adjacent matrix of body parts. (i) The message passing of the proposed method under other-occlusion. (j) Result of the proposed method under other-occlusion. (k) The message passing of the proposed model under self-occlusion. (l) Result of the proposed method under self-occlusion. The skeletons in (a) and (g) both are in a front view. The charts and pictures are best viewed in color.

Given an input image I, the posterior of pose configuration X of parts is given by

$$P(X|I) \propto \exp\left(\sum_{i\in\mathcal{V}} U(I,z_i) + \sum_{(i,j)\in\mathcal{E}_K} R^K(t_{ij},t_{ji})\right) \quad (1)$$

where $U(I,z_i)$ is the part appearance score of body part detector, $R^K(t_{ij},t_{ji})$ is the deformation score for pairwise kinetic constraints and $\mathcal{E}_K = \mathcal{E}$ denotes the set of kinematic constraints between body parts.

The problem is to maximize the following score function:

$$F(I,X) = \sum_{i\in\mathcal{V}} U(I,z_i) + \sum_{(i,j)\in\mathcal{E}_K} R^K(t_{ij},t_{ji}) \quad (2)$$

2) The Proposed Model: Our occlusion relational graphical model extends the tree-structured model in two aspects. First, part occlusion state is introduced to expand the pose configuration to be $\overline{X} = (\boldsymbol{z},\boldsymbol{t},\boldsymbol{o})$, where $\boldsymbol{o} = \{o_i\}$ with $o_i \in \{0,1,2\}$ to be the occlusion state ("0" for visible, "1" for self-occlusion and "2" for occlusion by the other objects). Second, in contrast to merely considering the kinetic constraints between nearby parts, our model encodes richer contextual information and occlusion relationships between parts.

The goal of our occlusion relational graphical model is to maximize the posterior as follows

$$P(\overline{X}|I) \propto \exp\Bigg(\sum_{i\in\mathcal{V}} U(I,z_i,o_i) + \sum_{(i,j)\in\mathcal{E}_K} R^K(t_{ij},t_{ji},o_i,o_j) + \sum_{(k,l)\in\mathcal{E}_C} R^C(t_{kl},t_{lk},o_k,o_l)\Bigg) \quad (3)$$

where $\mathcal{E}_C$ denotes the set of additional constraints between body parts that are not physically connected , $\mathcal{E} = \mathcal{E}_K \cup \mathcal{E}_C$ is the full set of constraints. For simplicity, we call them to be kinetic edges ($\mathcal{E}_K$) and contextual edges ($\mathcal{E}_C$) respectively. $U(I,z_i,o_i)$ is the local part appearance score considering occlusion states. $R^K(t_{ij},t_{ji},o_i,o_j)$ and $R^C(t_{kl},t_{lk},o_k,o_l)$ represent the deformation scores with occlusion coherence. They not only incorporate spacial deformation (kinetic and contextual respectively), but also encode the occlusion relationships among body parts.

The score function is formulated as

$$F(I,\overline{X}) = \sum_{i\in\mathcal{V}} U(I,z_i,o_i) + \sum_{(i,j)\in\mathcal{E}_K} R^K(t_{ij},t_{ji},o_i,o_j) + \sum_{(k,l)\in\mathcal{E}_C} R^C(t_{kl},t_{lk},o_k,o_l) \quad (4)$$

Note that the Eq. (4) above is only a general formulation which can be recognized as an extension of the generic tree model (Eq. (2)) with occlusion states and contextual constraints. In the following subsections, we will introduce the representations of our ORGM based on the specific baseline tree models of the FMP [1] and IDPR-DCNN [10] respectively.

B. ORGM-FMP

The FMP model simplifies the generic tree model (Eq. (2)) by limiting the spacial constraints for part i to be with its parent only (i.e., $T_{ij} = 1$ if j is child of i). We denote the simplified pose configuration to be $H = \{\boldsymbol{h}_i\}$, where $\boldsymbol{h}_i = (z_i,t_i)$. $t_i \in \{1,\cdots,T\}$ is the part mixture type which reflects the relative position to its parent for part i. The score function of FMP can be formulated as

$$F(I,H) = \sum_{i\in\mathcal{V}} U(I,\boldsymbol{h}_i) + \sum_{(i,j)\in\mathcal{E}_K} R^K(\boldsymbol{h}_i,\boldsymbol{h}_j) \tag{5}$$

Fig. 2 compares the structure of our occlusion relational graphical model ((g) and (h)) and that of FMP [1] ((a) and (b)). The proposed model differs from FMP in two aspects: first, each part of the model contains occlusion state o_i which indicates whether the part is visible, other-occluded (the squares in the diagonal elements in the adjacent matrix of body parts, colored in orange) or self-occluded; second, in contrast to merely considering the kinetic constraints (the purple edges/squares in (a), (b), (g) and (h)) between nearby parts, our model encodes richer interactions between parts (the green edges/squares in (g) and (h)) that are closely related to self-occlusion. For both FMP (a) and our model (g), the purple edges ($\mathcal{E}_K$) represent kinetic constraints. The green edges ($\mathcal{E}_C$) in (g) denote the contextual constraints.

When considering occlusion states, we denote the pose configuration to be $\overline{H} = \{\overline{\boldsymbol{h}}_i\}$ where $\overline{\boldsymbol{h}}_i = (z_i, t_i, o_i)$. The score function of our ORGM-FMP is composed of part appearance score and deformation score as follows:

$$F(I,\overline{H}) = \sum_{i\in\mathcal{V}} U(I,\overline{\boldsymbol{h}}_i) + \sum_{(i,j)\in\mathcal{E}_K\cup\mathcal{E}_C} R(\overline{\boldsymbol{h}}_i,\overline{\boldsymbol{h}}_j) \tag{6}$$

1) Part Appearance Score: The part appearance score is a summation of part filter response and compatibility biases.

$$U(I,\overline{\boldsymbol{h}}_i) = \alpha_i^{t_i}\cdot\phi(I,z_i,o_i) + \beta_i^{t_i}(o_i) \tag{7}$$

where $\alpha_i^{t_i}$ is the part filter parameters and $\beta_i^{t_i}(o_i)$ is the bias term for each mixture type and occlusion state. The part appearance $\phi(I,z_i,o_i)$ is defined as

$$\phi(I,z_i,o_i) = \phi(I,z_i,o_i)\mathbb{1}_{\{0,1\}}(o_i) \tag{8}$$

where $\mathbb{1}_{\{\cdot\}}(\cdot)$ is an indicator function as below:

$$\mathbb{1}_{\{0,1\}}(o_i) = \begin{cases} 1, & if\ o_i = 0,1 \\ 0, & if\ o_i = 2 \end{cases} \tag{9}$$

$\phi(I,z_i)$ is the HOG feature. It indicates that we set the part score to be zero only when it is occluded by some other objects. This differs from those approaches that treat both self-occlusion and other occlusion as noise and prune the local part score. In our method, the pattern of self-occlusion can be captured for further inference even when the body part is invisible (occluded by some other body part).

2) Deformation Score: The deformation score is as follows:

$$R(\overline{\boldsymbol{h}}_i,\overline{\boldsymbol{h}}_j) = \gamma_{ij}^{t_it_j}\cdot\psi(z_i - z_j) + \delta_{ij}^{t_it_j}(o_i,o_j) \tag{10}$$

where $\gamma_{ij}^{t_it_j}$ is the deformation parameters for each pair of connected parts. The part deformation $\psi(z_i - z_j) = \left[dx\ dx^2 dy\ dy^2\right]^T$ is a quadratic function of relative position, where $dx = x_i - x_j$ and $dy = y_i - y_j$, the relative location of part i with respect to j. $\delta_{ij}^{t_it_j}(o_i,o_j)$ encodes the occlusion coherence between body part i and j.

Note that the edges in our model not only contain kinetic constraints between nearby parts ($\mathcal{E}_K$) but also incorporate contextual interactions between non-adjacent parts ($\mathcal{E}_C$) which is helpful for the reasoning of occlusion relationships. As shown in Fig. 2(i), when the left hip of the rider is occluded by the head of the horse, the score of visible parts (on the right leg of the rider) can be passed through green edges (see Fig. 2(g) and (j)). Similarly, when the body parts are occluded by the other parts of the person in Fig. 2(k), the occluder and the occludee can pass the occlusion relationship to each other, so the occluder-occludee part pair can explain the same region without mutual exclusion. However, the FMP model cannot handle these issues and often fails under other-occlusion and self-occlusion (see Fig. 2(d) and (f)).

The unary term $U(I,\overline{\boldsymbol{h}}_i)$ models the appearance of each part i. The appearance varies with view point change, articulation as well as occlusion. To model these variations, $\overline{\boldsymbol{h}}_i = (z_i, t_i, o_i)$ specifies the part appearance with respect to part localization z_i, part mixture type t_i and part occlusion state o_i.

The pairwise term $R(\overline{\boldsymbol{h}}_i,\overline{\boldsymbol{h}}_j)$ models the geometric deformation constraints as well as occlusion relations between body part i and j on an occlusion relational graph $\mathcal{G}$, e.g., the left knee is probably occluded by the right knee while the left and right arms are less likely to occlude each other in Fig. 1(a). However, it is hard to model such subtle relations in a tree structured model such as in [1], [2], and [4].

C. ORGM-IDPR

Though the model structure of IDPR-DCNN [10] is also tree-structured as that of the FMP, it differs from FMP significantly in two aspects. Firstly, it utilizes more powerful CNN as local part detector instead of the HOG-SVM detector. Secondly, the IDPR term is introduced to enhance the spacial deformation constraints of the kinetic edges. In this part, we will demonstrate how to extend our occlusion relational graphical model based on the state-of-the-art method of IDPR-DCNN [10].

For the IDPR-DCNN model, the IDPR term [10] makes the pairwise term in the generic tree model (Eq. (2)) to be image dependent in the following form:

$$F(I,z,\boldsymbol{t}) = \sum_{i\in\mathcal{V}} U(I,z_i) + \sum_{(i,j)\in\mathcal{E}_K} R^K(I,t_{ij},t_{ji}) \tag{11}$$

The full score function of ORGM-IDPR is formulated as

$$\begin{aligned} F(I,z,\boldsymbol{t},\boldsymbol{o}) = &\sum_{i\in\mathcal{V}} U(I,z_i,o_i) \\ &+ \sum_{(i,j)\in\mathcal{E}_K} R^K(I,z_i,z_j,t_{ij},t_{ji},o_i,o_j) \\ &+ \sum_{(k,l)\in\mathcal{E}_C} R^C(z_k,z_l,t_{kl},t_{lk},o_k,o_l) \end{aligned} \tag{12}$$

The three terms on the right hand side of Eq. (12) handle occlusion from different aspects. The first term considers occlusion "locally", i.e., whether each local part is visible, other-occluded or self-occluded. The second term encourages the occlusion states among physically connected parts to be continuous, while the third term reflects the self-occlusion relationship between physically non-connected parts.

1) Part Appearance Score: The part appearance score is represented as

$$U(I, z_i, o_i) = w_i \phi(i|I(z_i); \boldsymbol{\theta})\ \mathbb{1}_{\{0,1\}}(o_i) + b_i(o_i)$$

where w_i is appearance weight parameter, $\phi(i|I(z_i); \boldsymbol{\theta})$ is the image evidence based on the local image patch $I(z_i)$ at location z_i for part i, $\boldsymbol{\theta}$ represents the parameters of convolutional neural network, $\mathbb{1}_{\{\cdot\}}(\cdot)$ is the indicator function as in Eq. (9), and $b_i(o_i)$ is the bias term for occlusion states.

2) Kinetic Pairwise Score: The kinetic pairwise score is represented as

$$\begin{aligned} R^K(I, z_i, z_j, t_{ij}, t_{ji}, o_i, o_j) = {} & \boldsymbol{w}_{ij}^{t_{ij}} \cdot \psi(z_j - z_i - \boldsymbol{r}_{ij}^{t_{ij}}) \\ & + w_{ij}\varphi(t_{ij}|I(z_i); \boldsymbol{\theta}) \cdot \mathbb{1}_{\{0,1\}}(o_i) \\ & + \boldsymbol{w}_{ji}^{t_{ji}} \cdot \psi(z_i - z_j - \boldsymbol{r}_{ji}^{t_{ji}}) \\ & + w_{ji}\varphi(t_{ji}|I(z_j); \boldsymbol{\theta}) \cdot \mathbb{1}_{\{0,1\}}(o_j) \\ & + \delta_{ij}^{t_{ij}t_{ji}}(o_i, o_j) \end{aligned} \tag{13}$$

where $\psi(\Delta z = [\Delta x, \Delta y]) = [\Delta x\ \Delta x^2\ \Delta y\ \Delta y^2]^T$ are the deformation features, $\boldsymbol{r}_{ij}^{t_{ij}}$ is the mean relative position of spacial relationship type t_{ij}. $\varphi(\cdot|\cdot; \boldsymbol{\theta})$ is the Image Dependent Pairwise Relational (IDPR) term [10], $\boldsymbol{w}_{ij}^{t_{ij}}, w_{ij}, \boldsymbol{w}_{ji}^{t_{ji}}, w_{ji}$ are the weight parameters of spacial relationship types, and $\delta_{ij}^{t_i t_j}(o_i, o_j)$ is the bias term which encodes the occlusion coherence between part i and part j. In our model, the IDPR term is ignored when other-occlusion exists as it may become unreliable.

3) Contextual Pairwise Score: The contextual pairwise score is described in the following form

$$\begin{aligned} R^C(z_k, \boldsymbol{o}_k, z_l, o_l, t_{kl}, t_{lk}) = {} & \boldsymbol{w}_{kl}^{t_{kl}} \cdot \psi(z_l - z_k - \boldsymbol{r}_{kl}^{t_{kl}}) \\ & + \boldsymbol{w}_{lk}^{t_{lk}} \cdot \psi(z_k - z_l - \boldsymbol{r}_{lk}^{t_{lk}}) \\ & + \delta_{lk}^{t_{lk}t_{kl}}(o_l, o_k) \end{aligned} \tag{14}$$

where all the terms on the right hand side are in accordance with that in Eq.(13). As the IDPR term only reflects the dependency of image evidence between nearby parts, there is no IDPR term in the contextual pairwise score. So the contextual pairwise score is not image dependent as kinetic pairwise score. The bias term $\delta_{lk}^{t_{lk}t_{kl}}(o_l, o_k)$ reflects long range spacial constraints as well as occlusion coherence (especially for self-occlusion) between non-adjacent parts l and k.

D. Inference

As described above, the structure of our model is a graph which contains loops. Inference on general loopy graph is an NP-hard problem. Many approximate methods, such as Loopy Belief Prorogation [47], Branch and Bound [20] and Dual Decomposition [48], need to iteratively infer on tractable structures many times until converge. However, our model contains large number of parameters and needs to mine huge amount of negative examples. Alternatively Ramanan [50] proposes to use tree-model for generating candidate pose configurations and rescoring the configurations using more complex non-tree constraints. Inspired by this, a two-step process is proposed for the inference. We first unroll the graphical model into a tree model to generate candidate pose hypothesis, and then rescore the candidate pose configurations with graphical model. In the following, we will introduce the inference of the proposed model from a generic perspective.

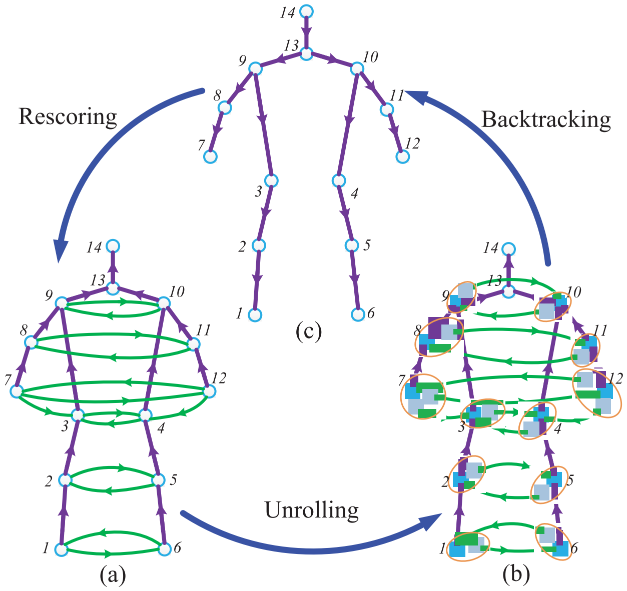

Fig. 3. Inference of the graphical model. (a) The graphical model, (b) The unrolled computation tree for approximate message passing, (c) The nodes backtracked via the tree structure.

We simplify the symbols of the generic score function in Eq. (4) as follows

$$F(I, \overline{X}) = \sum_{i \in \mathcal{V}} U_i(\overline{X}) + \sum_{(i,j) \in \mathcal{E}_K} R_{ij}^K(\overline{X}) + \sum_{(i,j) \in \mathcal{E}_C} R_{ij}^C(\overline{X}) \tag{15}$$

Consequently, the simplified score function of tree structured model is represented as

$$F(I, X) = \sum_{i \in \mathcal{V}} U_i(X) + \sum_{(i,j) \in \mathcal{E}_K} R_{ij}^K(X) \tag{16}$$

1) Model Unrolling: For any part i with out-degree (number of connections pointing to the other parts) $\nu_i > 1$, we generate $\nu_i - 1$ virtual parts and unroll the contextual edges to form a computation tree similar to [51] (See Fig. 3 (b)). As the parent of each virtual part is real part in our model, the effect of model unrolling is equivalent to that of single iteration of loopy belief propagation at the root node. The unrolled tree model is then used for generating and selecting candidate pose hypotheses.

2) Pose Selection: The goal of our graphical model is to maximize the score in Eq. (15), i.e.,

$$\overline{X}_m = \underset{\overline{X}}{\operatorname{argmax}} \sum_{i \in \mathcal{V}} U_i(\overline{X}) + \sum_{(i,j) \in \mathcal{E}_K} R_{ij}^K(\overline{X}) + \sum_{(i,j) \in \mathcal{E}_C} R_{ij}^C(\overline{X}) \tag{17}$$

where $\overline{X}_m$ is the optimal pose configuration with the proposed ORGM. Instead of passing message on a loopy graph, we pass message on the unrolled tree structure to generate root hypothesis. This allows us to employ dynamic programming to pass message from leaf nodes to the root node efficiently.

The optimization over the unrolled model is formulated as:

$$\overline{X}'_m = \underset{\overline{X}}{\operatorname{argmax}} \Bigg[\sum_{i \in \mathcal{V}} U_i(\overline{X}) + \sum_{k \in \mathcal{V}'} U_k(\overline{X}) + \sum_{(i,j) \in \mathcal{E}_K} R^K_{ij}(\overline{X}) + \sum_{(k,l) \in \mathcal{E}_C} R^C_{kl}(\overline{X}) \Bigg] \tag{18}$$

where $\overline{X}'_m$ is the optimal pose configuration of the unrolled tree-structured model. This is equivalent to adding part appearance weights to those nodes that have more connections.

Suppose the number of possible root hypotheses to be L in the test image. We sort them by root score and choose top L_σ hypotheses with the highest score, where $\sigma = L_\sigma / L$ is the ratio of hypotheses selection. In the approach of Ramanan [50], the optimal hypothesis of loopy model is supposed to be in the candidates with high score of the corresponding tree structured model. Similarly, we assume that the optimal hypothesis is included in the selected top-σ hypotheses of the unrolled configurations. The oracle accuracy of this assumption is analyzed in the experimental section (Sec. IV-E).

3) Backtracking and Rescoring: As soon as the top-σ hypotheses of root nodes are determined, the corresponding pose configurations can be obtained by backtracking directly from the root node to the leaf nodes. We only backtrack the child node from real parent node (e.g., node 1 is backtracked from node 2 rather than node 6 in Fig. 3) as the parent near the root node has less freedom of movement and is more reliable. We will recompute the score of the pose configurations with graphical model and rerank the hypothesis to get the optimal pose configuration.

Experimental results in the later sections will show that the performance almost does not change when the ratio $\sigma > 0.01$. We set $\sigma = 0.01$ for all the evaluations.

4) Computation: Let L be the number of possible part locations, N be the number of real parts and N_v be the number of virtual parts. Denote T be the number of mixture types for each part for FMP and ORGM-FMP. For IDPR-DCNN and ORGM-IDPR, we assume that all the pairwise spatial relationships have the same number of types, i.e., $T_{ij} = T_{ji} = T, \forall (i,j) \in \mathcal{E}$. The complexity of message passing is $O((N-1)LT^2)$ with dynamic programming and distance transform [43] for the tree structured models of FMP [1] and IDPR-DCNN [10]. For our occlusion relational graphical model, the complexity becomes $O((N + N_v - 1) L(3T)^2) = O(9(N + N_v - 1)LT^2)$, which is slower than the FMP and IDPR-DCNN yet with the same order. However, the backtracking and rescoring procedure can be much faster as we only process the selected L_σ hypotheses. The average detection speed of ORGM-FMP is 4.5 seconds per image for the LSP dataset on single 3.4 GHz CPU compared with 1.2 seconds per image of the FMP. The IDPR based approaches are tested on single 3.50 GHz CPU and GeForce GTX TITAN GPU, and the average detection speed of ORGM-IDPR is 118.5 seconds per image vs. 63.3 seconds per image of IDPR-DCNN for the LSP dataset.

E. Model Learning

1) Learning Spacial Relationship Types: The spacial relationship types reflect the pairwise relative position between parts. As mentioned in Sec. III-B, in some tree structured models [1], [3], [4], the spacial relationships are limited by only allowing the part to be connected with its parent. So the per-pair spacial relationship types (t_{ij} and t_{ji} for part i and part j) becomes per-part local part mixture types (t_i for part i). In the latent tree model [4], local part mixtures are learnt by clustering part appearance yet without considering the structure of human body. In contrast, the FMP [1] learns the local part mixtures by clustering the relative positions from the parent for each part in the kinetic tree. However, this makes it hard to capture the varying occlusion relationships between non-adjacent parts (especially for lower limbs with more freedom of movements). For our occlusion relational graphical model, many parts have multiple interactions with some other parts. We learn the part mixture types of ORGM-FMP from the relative position from all the parents and children in the graphical model. In this way, we can not only capture the local part deformation but also encode the global pose deformation. This will benefit the localization of occluded parts and the lower level parts which contain much uncertainty of freedom in tree structured models. We use the simple k-means clustering with multiple runs and choose the one with minimum objective function.

For the IDPR-DCNN [10] as well as our ORGM-IDPR, the pairwise spacial relationship types are not simplified. Following the IDPR-DCNN, we cluster the relative position from part i to part j to get the spacial relationship types $t_{ij} = 1, \cdots, T_{ij}$ and use the center of each cluster as the mean relative position $\boldsymbol{r}_{ij}^{t_{ij}}$ of each type in our model of ORGM-IDPR as well. In our experiments we set $T_{ij} = 13$ for all pairwise relations as that in the released model of Chen and Yuille [10] for comparison. Compared with the simplified local mixture types, the pairwise spacial relationship types can model more flexible spacial constraints.

2) Learning Occlusion Coherence: To model spatial coherence among part occlusions, we utilize two sources of occlusion samples. One is from the label of part occlusion states and the other from synthetic occlusion patterns. For the labeled invisible part, we distinguish it as self-occluded if there exist some other visible body part with more than 50% region overlapped with it. The self-occlusion relationships will be captured by the contextual edges in our model. We find that more than half the invisible body parts in Leeds Sports Dataset are occluded by the other body parts due to articulation and viewpoint. And the number of instances with other-occlusion is relatively small. To balance the two different occlusion types in the training sample, we synthesize samples with occlusion by the other objects. We utilize occlusion masks to generate synthetic samples similarly as [52]. And we only use samples without occlusion for synthesization. During training, the occlusion relationship between parts as well as the occlusion pattern are learned and encoded in the model.

3) Learning Model Parameters: Though the model parameters of ORGM-FMP and ORGM-IDPR are different, we integrate the parameter learning of both models in a general formulation with only different pre-training steps.

In the ORGM-FMP, as the HOG features are extracted manually, the local part parameters and global model parameters are learned jointly. However, the large amount of HOG filter parameters make the model parameters converge slowly. We follow the strategy of FMP to pre-train the local part parameters with linear SVM and initialize the model with pre-trained local part parameters.

For the ORGM-IDPR, however, the local part detectors and global model are trained piecewisely. In the training of local part detectors, we follow IDPR-DCNN [10] to train a multi-class DCNN classifier with softmax loss. Denote $\boldsymbol{\theta}$ to be the DCNN parameters, the part evidence and IDPR terms are obtained from the probability output of DCNN. Specifically, $\phi(i, |I(z_i); \boldsymbol{\theta})$ is the probability of the extracted image patch belonging to part i, and $\varphi(t_{ij}, |I(z_i); \boldsymbol{\theta})$ is the probability of the extracted image patch corresponding to spacial relationship type t_{ij}. For comparative comparison, we adopt the same network structure and parameter setting and we advise the readers to refer to [10] for more details.

Although the pre-training steps above are different, the learning of model parameters can be formulated in a unified form. For generic purpose, we denote $\mathcal{J}$ to be the pose configuration where $\mathcal{J} = \overline{H}$ for ORGM-FMP and $\mathcal{J} = \overline{X}$ for ORGM-IDPR. Given the pose configuration $\mathcal{J}$ and the image I, the total score of both Eq. (6) and Eq. (12) can be formulated according to the linear property as follows:

$$F(I, \mathcal{J}) = \boldsymbol{w}_{\mathcal{J}} \cdot \Phi(I, \mathcal{J}) \tag{19}$$

For ORGM-FMP, $\boldsymbol{w}_{\overline{H}}$ is the concatenation of the following weight parameters

$$\boldsymbol{w}_{\overline{H}} = [\alpha_i^{t_i}, \cdots, \beta_i^{t_i}(o_i) \cdots, \gamma_{ij}^{t_i t_j} \cdots, \delta_{ij}^{t_i t_j}(o_i, o_j), \cdots]^T$$

The concatenation weight parameters for ORGM-IDPR is denoted as follows

$$\begin{aligned} \boldsymbol{w}_{\overline{X}} = [\ & w_i, \cdots, b_i(o_i) \cdots, w_{ij}, w_{ji}, \cdots, \boldsymbol{w}_{ij}^{t_{ij}}, \boldsymbol{w}_{ji}^{t_{ji}}, \cdots, \\ & \delta_{ij}^{t_{ij}}(o_i, o_j), \cdots, \boldsymbol{w}_{kl}^{t_{kl}}, \boldsymbol{w}_{lk}^{t_{lk}}, \cdots, \delta_{lk}^{t_{lk}}(o_l, o_k), \cdots\ ]^T \end{aligned}$$

$\Phi(I, \mathcal{J})$ is the concatenation of all the features with the same order. For the bias terms (i.e., $\beta_i^{t_i}(o_i)$ and $\delta_{ij}^{t_i t_j}(o_i, o_j)$ in ORGM-FMP, and $b_i(o_i)$ and $\delta_{ij}^{t_{ij}}(o_i, o_j)$ in ORGM-IDPR), the corresponding dimensions of $\Phi(I, \mathcal{J})$ are set to be 1. For those spacial relationship types t_{ij} (degraded into local part mixture types t_i in ORGM-FMP) and occlusion states o_i that are not activated, the corresponding dimensions of features in $\Phi(I, \mathcal{J})$ are filled with 0.

In this way, the proposed occlusion relational graphical model can be linearly parameterized, allowing efficient training using a large margin objective. The optimization function can be written as:

$$\underset{\boldsymbol{w}_{\mathcal{J}}}{\operatorname{argmin}}\ \frac{1}{2}\boldsymbol{w}_{\mathcal{J}}^T \boldsymbol{w}_{\mathcal{J}} + C \sum_n \max(0, 1 - y_n \langle \boldsymbol{w}_{\mathcal{J}}, \Phi(I, \mathcal{J}) \rangle) \tag{20}$$

where $y_n \in \{1, -1\}$, $y_n = 1$ if $n \in pos$, and $y_n = -1$ if $n \in neg$. This is a standard structural SVM learning problem, which can be solved by the cutting pane solver like SVM$^{\text{struct}}$ [53] or the stochastic gradient descent (SGD) solver. In this paper, we turn to use dual coordinate descent QP solver of [54] as we should meet the requirement of parameters constraints, e.g., the quadratic coefficients of part deformation ($\gamma_{ij}^{t_i t_j}$ for ORGM-FMP, as well as $\boldsymbol{w}_{ij}^{t_{ij}}$, $\boldsymbol{w}_{ji}^{t_{ji}}$, $\boldsymbol{w}_{kl}^{t_{kl}}$ and $\boldsymbol{w}_{lk}^{t_{lk}}$ for ORGM-IDPR) should be negative for generic distance transform [43]. The body part position, visibility and spacial configurations are completely specified during training.

TABLE I

DATASETS USED IN OUR EXPERIMENTS

Dataset	#train	#test	#points	POJ[1]	scene	Pose variation
LSP [7]	1000	1000	14	16.7%	sports	large
PARSE [55]	100	205	14	–	diverse	most upright
FLIC [56]	3987	1016	11	–	feature film	frontal
LSP [7]-sub	1000	468	14	20.7%	sports	large
MPII [57]-sub	1500	698	16	44.1%	diverse	large

[1] POJ = Percentage of Occluded Joints.

IV. EXPERIMENTS

For comprehensive experimental analysis, we will first introduce the datasets, evaluation criteria and implementation details. Then we will present quantitative evaluations on benchmark datasets as well as datasets with heavy occlusions. Finally, diagnostic experiments and discussions are provided for further analysis.

A. Datasets and Criteria

1) Datasets: For comprehensive evaluation on public benchmarks, we firstly evaluate the proposed approach on the popular LSP [7] dataset, and then we test it on the PARSE [55] dataset with the model trained on LSP dataset for generalization ability, finally we evaluate our method on the FLIC dataset [56] with 11 points of upper body annotations from popular Hollywood movies. As this paper intends to address the problem of human pose estimation with occlusion, we specifically design experiments on occluded images for better explaining our approach. We choose subset images with occlusions from LSP and the challenging MPII [57] for detailed analysis of the robustness to occlusion. Tab. I lists the dataset used for evaluation in our work.

2) Criteria: There are three criteria used in the experiments to evaluate the performance of previous human pose estimation approaches: Percentage of Corrected Parts (PCP) [58]–[60], Percentage of Detected Joints (PDJ) [9], [56] and Percentage of Corrected Keypoints (PCK) [60].

a) PCP: The most widely used criterion for human pose estimation is PCP which evaluates the localization accuracy of body parts (sticks of skeleton). It requires the estimated part end points must be within half of the part length from the ground truth part end points. As pointed by Yang and Ramanan [60], some previous work require only the average of the endpoints of a part to be correct (PCP-average), rather than both endpoints (PCP-strict). Moreover, the early PCP implementation [58] selects the best matched output without penalizing false positives. In all our experiments we adopt the most strict measure, i.e., PCP-strict with single output, if not specially specified. For more detailed descriptions on PCP, it is recommended to refer to [58] and [60].

b) PDJ: Though PCP is the initially preferred criterion for evaluation, it has the drawback of penalizing shorter limbs, such as lower arms. Thus PDJ is introduced [9], [56] to measure the detection rate of body joints, where a joint is considered to be detected if the distance between the detected joint and the true joint is less than a fraction of the torso diameter. The torso diameter is usually defined as the distance between opposing joints on the human torso, such as left shoulder and right hip [9]. The Area Under Curve (AUC) metric can be used as the overall evaluation of the PDJ curve.

c) PCK: The PCK measure is very similar to the PDJ criterion. The only difference is that the torso diameter is replaced with the maximum side length of the external rectangle of ground truth body joints. For full body images with extreme pose (especially when the torso becomes very small), the PCK may be more suitable to evaluate the accuracy of part location.

B. Implementation Details

In the experiments, we take the FMP [1] and IDPR-DCNN [10] as baseline. To enable a fair comparison of our models, our ORGM framework uses the same settings of the baseline methods respectively. Both our models and that of the baseline models use 26 parts for full body and 18 parts for upper body. The 1218 non-person images from INRIA person dataset [26] are used as negative training samples. For the FLIC dataset [56], there are only annotations of upper body joints yet without occlusion state. We create 2 other-occluded samples synthetically as in [52] for each image. The joints and edges in the legs are pruned and the occlusion states are limited to model other-occlusion only. The other settings are described respectively in the following.

1) ORGM-FMP: Both FMP and our ORGM-FMP set the number of local part mixture types to be $T = 6$ for each part. HOG features are extracted on grid image with 4×4 pixels for full body models and 8×8 pixels for upper body models. During the pre-training of part detectors, the image patches that have more than 0.6 overlap with groundtruth image patch of each part are extracted as positive samples, and 100 non-person images are used for hard negative mining.

2) ORGM-IDPR: For the training of DCNNs, we utilize the same network architectures as [10] with Caffe [66]. The input patch size is set to be 36×36 pixels for full body and 72×72 pixels for upper body. To overcome overfitting, we follow [10] to augment positive samples by rotating the positive training samples through 360° with step size of 10°.

C. Benchmark Evaluation

1) The Leeds Sport Pose dataset: Tab. II compares the PCP of our model extensively with the state-of-the-art approaches on the LSP dataset with Person-Centric (PC) and Observer-Centric (OC) annotations respectively. For the Person-Centric annotations, the right and the left body parts are marked according to the viewpoint of the person, e.g., the right ankle of a person facing the camera is left in the image, but it is right in the image if the person faces away from the camera. For the Observer-Centric annotations, the right and the left body parts are annotated according to the viewpoint of the observer/camera, e.g., the left limb is always on the left side of the torso in the image. As some previous work use PCP-average for evaluation, we also list the corresponding results for comprehensive comparison. The methods with/wihout deep learning are evaluated separately except that of Toshev and Szegedy [9] for the seek of uniform PCP criterion. All the results are from the authors' papers respectively except for some marked up in the table. The PC and OC results of [1] are from [4] and [64] respectively. For those with released prediction results, we evaluate them with the toolkit of [57]. As some approaches use additional training data (e.g., the Extended Leeds Sport Pose dataset [61] with 10,000 images), we mark them in the table as well.

For the approaches without deep learning, our result of ORGM-FMP is comparable with the pervious state-of-the-arts. Our PC result is better than the previous best result of Pishchulin *et al.* [30] (1.2% on average and 1.9% on the limbs) and even on par with DeepPose of Toshev and Szegedy [9]. Please note that Toshev and Szegedy [9] use additional 10,000 images from the Extended LSP dataset [61] for training. This is due to the huge number of parameters to be learned in the CNN model. Though the average OC result of Pishchulin *et al.* [30] is better than ours, our method still performs better on the limbs which are easy to be occluded. Considering both PC and OC results, the approach of Pishchulin *et al.* [30] performs better in localizing torso and head, this is mainly because they use strong poselet detectors as prior. The Pose Machines of [32] performs better than ours on the limbs except the most challenging lower arms. However, our ORGM-FMP model only uses HOG [26] features while [32] adopts HOG features, Lab color features and gradient magnitude. The model of [30] is built on shape context and HOG features. Compared with our baseline method of FMP [1], our method improves performance by 7.0% for Person-Centric annotation (PCP-average) and 6.9% for Observer-Centric annotation (PCP-strict). The improvement on the limbs is much more significant (10.1% and 8.3% respectively).

When compared with deep learning based methods, our ORGM-IDPR approach performs better than the previous state-of-the-arts. In the evaluation, all the PC results are obtained with models trained with additional Extended LSP dataset, while all the OC results only use LSP dataset for training. For the PC annotation, our ORGM-IDPR model improves the previous best result of Chen and Yuille [10] by 1.3%. And our OC annotation results outperforms that of Chen and Yuille [10] by 3.1%. Especially, our approach is better at detecting legs and arms which are prone to be occluded.

Fig. 5 shows the detection results of our models compared with the baseline method of FMP [1] and IDPR-DCNN [10] as well as the DeepPose [9] approach. The first three rows (above the line in orange) are obtained with models trained with PC

TABLE II

PERCENTAGE OF CORRECT PARTS (PCP) ON LSP DATASET FOR OUR METHOD AS WELL AS THE STATE-OF-THE-ARTS

Method				Head	Torso	Leg		Arm		Avg Limbs	Avg All
						Upper	Lower	Upper	Lower		
Person-Centric	PCP-average	No-Deep	ORGM-FMP	87.5	91.5	74.2	66.8	**62.5**	**41.3**	**61.2**	**66.9**
			Toshev et al. [9]*†	–	–	**77**	**71**	56	38	61	–
			Wang et al. [4]	86.0	91.9	74.0	69.8	48.9	32.2	56.2	62.8
			Johnson et al. [61]†	74.6	88.1	74.5	66.5	53.7	37.5	58.0	62.7
			Tian et al. [3]	**87.8**	**95.8**	69.9	60.0	51.9	32.9	53.7	61.3
			Yang et al. [1]	87.4	92.6	66.4	57.7	50.0	30.4	51.1	58.9
			Dantone et al. [31]	79.2	81.6	66.5	61.0	45.1	24.7	49.3	55.5
			Johnson et al. [7]	62.9	78.1	65.8	58.8	47.4	32.9	51.2	55.1
	PCP-strict	Deep	ORGM-IDPR†	**89.2**	94.1	79.5	**74.9**	**70.0**	**58.6**	**70.7**	**74.9**
			Chen et al. [10]†‡	85.6	**96.0**	77.2	72.2	69.7	58.1	69.3	73.6
			Carreira et al. [62]†‡	84.4	95.3	**81.8**	73.3	66.7	51.0	68.2	72.5
			Fan et al. [12]†‡	86.6	95.4	77.7	69.8	62.8	49.1	64.8	70.1
			Tompson et al. [11]†‡	83.7	90.3	70.4	61.1	63.0	51.2	61.4	66.6
		No-Deep	ORGM-FMP	78.6	83.5	**66.0**	**63.5**	**50.6**	34.9	**53.7**	**59.2**
			Pishchulin et al. [30]	**85.1**	**88.7**	63.6	58.4	46.0	**35.2**	50.8	58.0
			Wang et al. [4]	79.1	87.5	56.0	55.8	43.1	32.1	46.7	54.1
Observer-Centric	PCP-strict	No-Deep	ORGM-FMP	77.7	85.4	75.0	71.9	62.1	**48.8**	64.2	67.7
			Ramakrishna et al. [32]	84.3	88.1	**79.0**	**73.6**	**62.8**	39.5	63.7	67.8
			Kiefel et al. [63]	78.3	84.3	74.5	67.6	54.1	28.3	56.1	61.2
			Pishchulin et al. [30]	**85.6**	**88.7**	78.8	73.4	61.5	44.9	**64.6**	**69.2**
			Eichner et al. [64]	80.1	86.2	74.3	69.3	56.5	37.4	59.4	64.3
			Pishchulin et al. [17]	78.1	87.5	75.7	68.0	54.2	33.9	57.9	62.9
			Yang et al. [1]	77.1	84.1	69.5	65.6	52.5	35.9	55.9	60.8
			Yang et al. [60]	79.3	82.9	70.3	67.0	56.0	39.8	58.3	62.8
			Andriluka et al. [16]	74.9	80.9	67.1	60.7	46.5	26.4	50.2	55.7
		Deep	ORGM-IDPR	**89.8**	**93.9**	**85.3**	**79.8**	**73.0**	**60.7**	**74.7**	**78.1**
			Chen et al. [10]‡	87.8	92.7	82.9	77.0	69.2	55.4	71.1	75.0
			Vasileios et al. [13]	83.2	92.0	79.9	74.3	61.3	40.3	64.0	68.8
			Ouyang et al. [65]	83.1	85.8	76.5	72.2	63.3	46.6	64.6	68.6

*Deep CNNs are used in this work exceptionally.

†Trained with additional Extended Leeds Sport Pose Dataset [61].

‡Evaluated with the released prediction results.

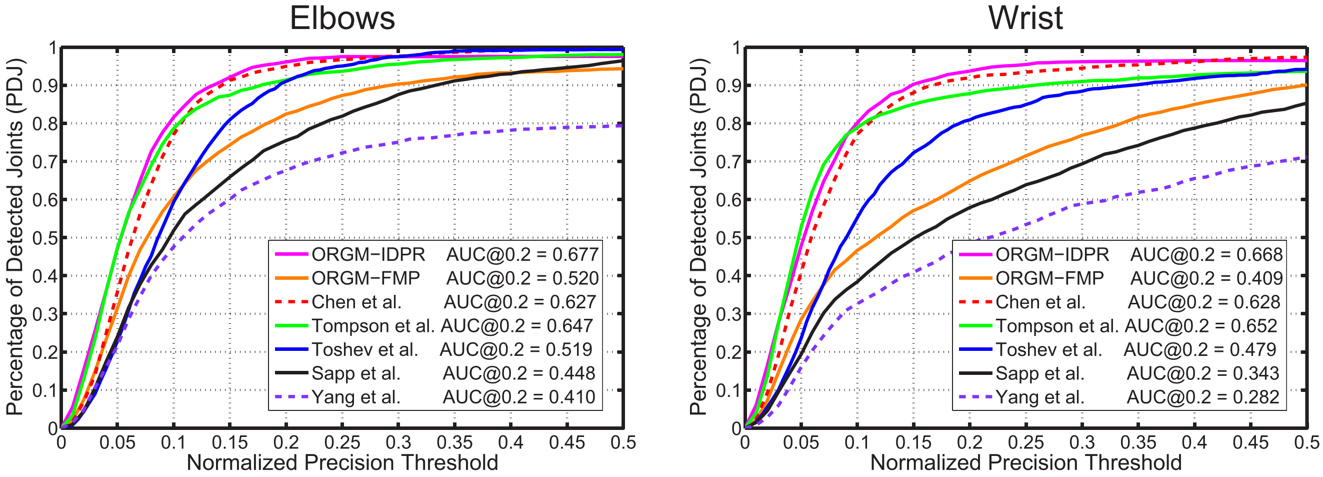

Fig. 4. Test results on FLIC dataset. We compare the localization accuracy for the most challenging parts: elbows and wrists. Best viewed in color.

annotation, while the rest are trained with OC annotation. It shows that the FMP is less robust to large deformation as well as occlusion and may easily lead to double-counting (column 5, 8 and 9). The detection results reflect that the DeepPose model is good at capturing global pose configurations with large deformation (column 1, 3, 4 and 6), yet sometimes locate the body parts inaccurately (column 1, 7, 8 and 9). This problem may lie in two factors: one is the normalization of input image into square to fit into ConvNet [8], the other is that the low resolution of response maps after the last convolutional layer (6×6 for conv5 in AlexNet [8]) may lead to ambiguity of part position. Compared with FMP, our ORGM-FMP model is more robust to deformation and occlusion. The IDPR-DCNN is also robust to large deformation, yet may

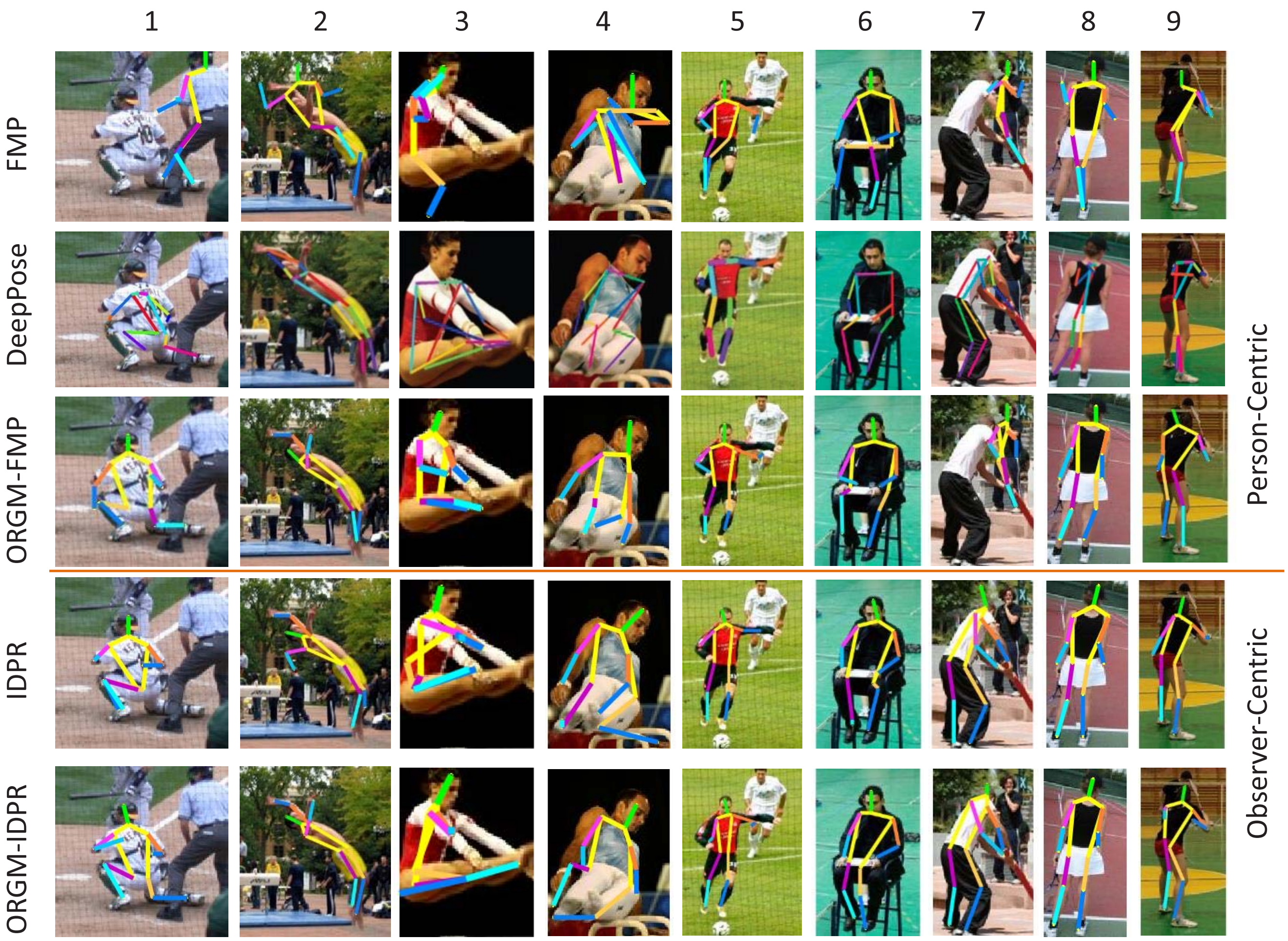

Fig. 5. Comparison of detection results on LSP dataset. The first three rows on the top are the PC results of FMP [1], DeepPose [9] and our ORGM-FMP, while the last two rows on the bottom are the OC results of IDPR [10] and our ORGM-IDPR.

TABLE III

CROSS TEST RESULTS ON PARSE DATASET WITH MODELS TRAINED ON LSP DATASET

Method	Head	Torso	U.Leg	L.Leg	U.Arm	L.Arm	Avg
ORGM-IDPR	89.2	92.8	83.3	**78.7**	**73.0**	61.6	**77.5**
Chen et al. [10]	87.5	**93.1**	78.9	73.1	72.3	**61.8**	75.3
Toshev et al. [9]	–	–	88	75	71	50	–
Ouyang et al. [65]	**89.3**	78.0	**89.3**	72.0	67.8	47.8	71.0
ORGM-FMP	**88.3**	**90.7**	**75.4**	**66.8**	**71.9**	**51.2**	**70.9**
Johnson et al. [7]	76.1	78.1	73.4	65.4	64.7	46.9	66.2
Wang et al. [4]	78.7	88.3	75.2	71.8	60.0	35.9	65.3
Yang et al. [1]	70.0	78.8	66.0	61.1	61.0	37.4	60.0

suffer from double-counting when their is partial occlusion (column 1). However, our ORGM-IDPR method can overcome this issue and is more accurate in localizing lower limbs.

2) Cross Test on PARSE dataset: In order to measure the generalization ability of the proposed model, we test our models on the PARSE dataset as shown in Tab III. Pishchulin et.al's approach [17] adopts the LSP+PARSE training set when evaluated on the PARSE dataset. Both the approaches of Johnson and Everingham [61] and Toshev and Szegedy [9] include 10,000 extra training samples. All the other methods are trained on the 1,000 training images of the LSP dataset. The result of Chen and Yuille [10] is obtained by running their released model, and the other results are from their papers. Our ORGM-FMP improves the accuracy by 10.9% compared with FMP [1], while our ORGM-IDPR improves the accuracy of the IDPR-DCNN method [10] by 2.2%.

3) The FLIC Dataset: Compared with LSP and PARSE datasets, the FLIC dataset features real life scenes and is challenging in the localization of elbows and wrists. We compare our models with the state-of-the-art methods following the PDJ criterion to measure the localization accuracy of elbows and wrists. The result of Sapp and Taskar [56] is derived from the model trained by the authors. We retrain the FMP model of Yang and Ramanan [1] on the FLIC training set and get comparable results as in [56]. As most of the people are not centered in the image in the FLIC dataset, Sapp and Taskar [56] utilize the poselet [67] torso detector for initial detection. Similarly, Toshev and Szegedy [9] adopt a face-based body detector to get a rough estimation. However, the other approaches do not use body detectors for initial detection and restrict the neck of estimated pose to be within the groundtruth bounding box instead.

As shown in Fig. 4, our ORGM-FMP method outperforms MODEC of Sapp and Taskar [56] by 7.2% and 6.6% in AUC@0.2[1] respectively on elbows and wrists. The result

[1] Here AUC@0.2 means the average detection rate for normalized precision threshold to be within 0 $\sim$ 0.2.

TABLE IV

ANALYSIS OF PERFORMANCE ON THE LPS OCCLUDED SUBSET

# occluded joints	1	2	3	4	5
ORGM-IDPR	76.4	71.6	70.0	69.5	64.7
IDPR [10]	72.6	67.9	64.5	67.8	59.5
ORGM-FMP	67.5	62.9	61.4	53.3	43.6
FMP [1]	59.6	52.7	50.3	47.5	39.1
# test images	174	133	105	37	19

shows that the modeling of interactions between physically unconnected parts (e.g., left and right wrists) will benefit the localization of lower arms. Our method improves the baseline of FMP by 11.0% and 12.7% respectively. As the performance of IDPR-DCNN is already very high on the FLIC dataset, the improvement of our ORGM-IDPR approach is relatively small (5.0% and 4.0%) in comparison to the IDPR-DCNN baseline method of Chen and Yuille [10].

D. Experiments on Occlusion

As our approach focuses on the problem of occlusion handling in human pose estimation, we specifically design experiments to test the robustness on occlusion. We select images with occlusion from LSP [7] dataset and the MPII [57] dataset for detailed analysis.

1) Occluded Leeds Sports: We evaluate our method on a subset of the LSP [7] test set consisting of 468 images with one more joints occluded.

Tab. IV shows the performance of our method as well as the baseline under different levels of occlusions. All the models are trained with the 1000 training images of LSP dataset with Observer-Centric annotation. It reflects that the performance of FMP [1] drops quickly with more occluded joints. When the number of occluded joints increases from 1 to 3, the PCP of FMP drops 9.3% while our ORGM-FMP drops 6.1%, and 8.1% vs. 6.4% for IDPR and our ORGM-IDPR. However, the performance of our models drops much slower when there are less than 4 joints occluded.

2) Occluded MPII: Many images in the LSP dataset contain people with sportswear in sports scenarios. We evaluate on a subset of the challenging MPII [57] pose dataset in real life with large pose variation, cluttered background, as well as occlusion. The selected subset consists of 2198 images with severe occlusion (44.1% of the joints) and is suitable for the evaluation of robustness to occlusion. Though PCP is the most frequently used metric for evaluation, it has the drawback of penalizing shorter limbs. For better evaluation of per joint detection, we adopt the PCK criterion for analysis. Fig. 6 illustrates the performance of our models vs. the baseline methods on the Occluded MPII dataset with Observer-Centric annotation. The chart reflects that images with heavier occlusion are much more challenging for most of the approaches. Generally, the IDPR of Chen and Yuille [10] and our ORGM-IDPR are more robust than FMP and ORGM-FMP. The main reason is that the learned CNN features are more discriminative than the handcrafted HOG features. It also reflects that both our models (ORGM-FMP and ORGM-IDPR) performs better than the baseline approaches when there is heavy occlusion.

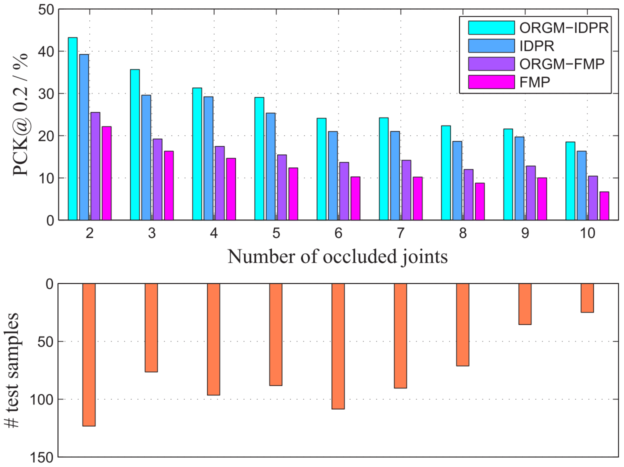

Fig. 6. Analysis of occlusion robustness on the MPII subset for the proposed methods.

TABLE V

THE COMPARISON OF PCP(%) WITH DIFFERENT MODEL STRUCTURES ON LSP DATASET

Model		Head	Torso	U.Leg	L.Leg	U.Arm	L.Arm	Limbs	Avg
FMP	Tree	77.1	84.1	69.5	65.6	52.5	35.9	55.9	60.8
	Tree+O	77.2	84.4	70.9	67.3	54.6	41.2	58.5	62.9
	G+S	77.5	**85.5**	73.2	70.3	59.5	46.2	62.3	66.1
	G+S+O	**77.7**	85.4	**75.0**	**71.9**	**62.1**	**48.8**	**64.2**	**67.7**
IDPR	Tree	87.8	92.7	82.9	77.0	69.2	55.4	71.1	75.0
	Tree+O	87.9	92.5	83.3	78.6	71.3	57.2	72.6	76.1
	G+S	89.3	93.5	84.9	79.5	**73.2**	59.4	74.2	77.7
	G+S+O	**89.8**	**93.9**	**85.3**	**79.8**	73.0	**60.7**	**74.7**	**78.1**

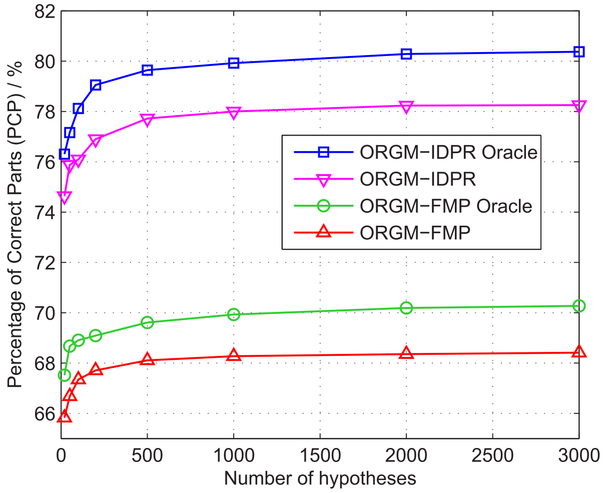

Fig. 7. Analysis of oracle accuracy on the LSP dataset with different number of hypotheses selected per image.

E. Diagnostic Experiments

We design two experiments to better understand the influence of model structures and parameter settings on the performance of our models. We evaluate the parameters on the LSP dataset and take the FMP [1] and IDPR-DCNN [10] as baseline.

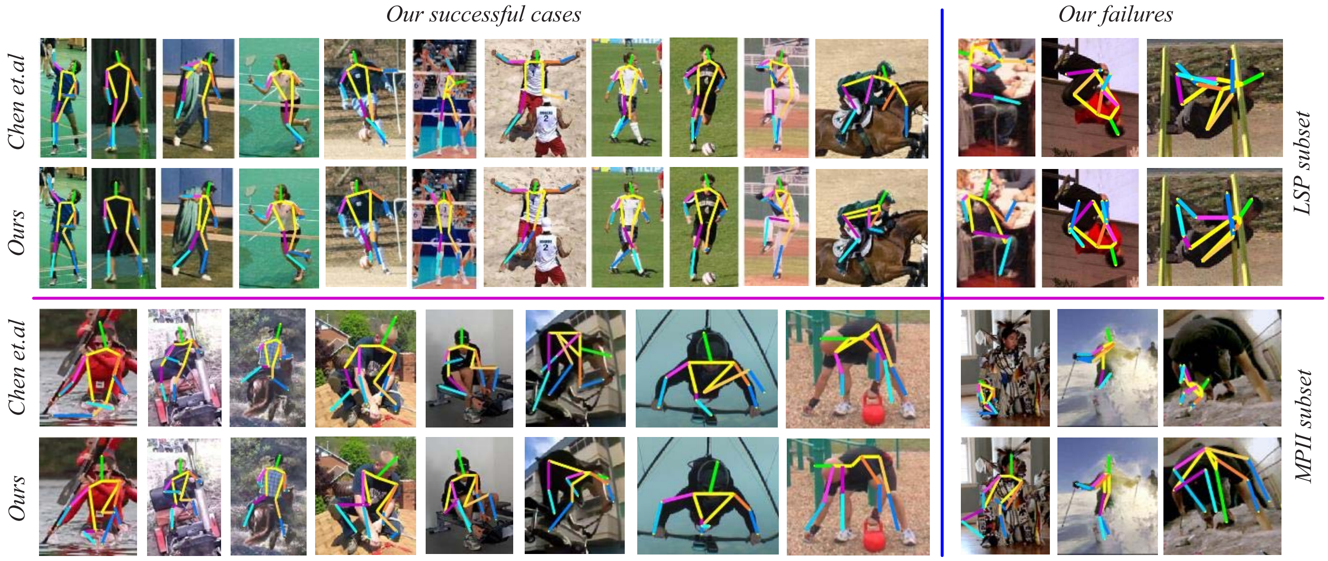

Fig. 8. Detection results on LSP and MPII subset with occlusion. The first and the third rows are the result of Chen and Yuille [10] while the second and the fourth rows are our ORGM-IDPR approach.

1) The Effect of Occlusion Modeling: In terms of occlusion modeling, we consider both self-occlusion and other-occlusion in the proposed model. It is worth analyzing how each feature of the model contributes to the boost of performance.

Tab. V shows the results with different model structures. (1) Tree: the tree structured models of FMP [1] and IDPR [10]. (2) Tree+O: the tree structured model with other-occlusion reasoning only. (3) G+S: our graphical model with self-occlusion handling only. (4) G+S+O: our graphical model with both self-occlusion and other-occlusion reasoning, i.e., ORGM-FMP and ORGM-IDPR. We notice that the localization accuracy of torso and head does not improve since they are rarely occluded.

For the models based on FMP, there is evident improvement in overall performance. It is observed that the introduce of occlusion states is helpful for improving the accuracy of limbs (especially lower arms and legs) which are frequently occluded. For instance, there is 2.6% improvement in PCP of limbs for the Tree+O model vs. the FMP model, and 1.9% for the G+S+O model vs. the G+S model. On the other hand, the edges between non-connected body parts can significantly improve the overall PCP (e.g., 6.4% for G+S compared with FMP and 5.7% for G+S+O compared with Tree+O). This is mainly because the constraints among non-connected parts can eliminate double-counting and improve the PCP of limbs.

For the models based on IDPR, the improvement in performance is less obvious as the baseline performance is relatively high. The improvements also mainly come from performance gain on the limbs.

2) The Influence of Parameter: $\boldsymbol{\sigma}$ In the inference procedure, we unroll the graphical model into a tree for approximate inference. The only parameter that affects this procedure is the ratio of selected candidate of root part nodes. We analyse how different ratios affect the performance of our method. Tab. VI shows that our models can get higher PCP with the increasing of σ. It also reflects that the performance almost does not change when the ratio $\sigma > 0.01$.

TABLE VI
THE INFLUENCE OF σ ON PCP FOR LSP DATASET

σ	0.001	0.002	0.005	0.01	0.02	0.05	0.1
ORGM-FMP	65.8	66.2	67.4	67.7	67.7	67.8	67.9
ORGM-IDPR	74.8	76.5	77.6	78.1	78.1	78.2	78.2

In section III-D, we assume that the optimal hypothesis is included in the selected top-σ hypotheses of the unrolled configurations. Fig. 7 qualitatively analyzes the oracle accuracy and the actual accuracy of our method with different number of hypotheses per image. The oracle accuracy reflects the upper bound of our models with given number of hypotheses selected.

F. Discussion: are CNN Part Detectors Robust to Occlusion?

The experimental results above show that the proposed ORGM-IDPR outperforms the FMP of Yang and Ramanan [1] by a large margin. There are three main factors for the improvement: the strong CNN part detectors compared with HOG+SVM part detectors, the powerful IDPR pairwise term vs. pairwise constraints with spcacial deformation only, the occlusion relational graphical model in contrast to simple tree model without occlusion modeling. Table VII compares the Chen et al.'s method [10] and our ORGM (based on CNN part detectors) with/without IDPR for further analysis. The results of both "No-IDPR" and "IDPR" are from the original paper.

Compared with HOG+SVM part detectors of FMP [1], it seems CNN part detectors may be more robust to partial occlusion due to larger receptive field and stronger feature representation. However, the experimental results in Table VII shows that the performance of CNN part detectors+tree model is lower than the previous state-of-the-art non-CNN method of Pishchulin *et al.* [30] (66.9% vs. 69.2%) and even lower than our ORGM-FMP (67.7%). And our ORGM can improve the performance from 66.9% to 70.6% without IDPR term.

TABLE VII

THE COMPARISON OF PCP(%) FOR CNN PART DETECTORS BASED MODELS ON THE LSP DATASET

Method	Head	Torso	U.Leg	L.Leg	U.Arm	L.Arm	Limbs	Avg
No-IDPR [10]	77.6	88.9	75.6	66.9	63.2	45.6	62.8	66.9
ORGM No-IDPR	78.0	90.1	78.6	72.8	65.5	52.0	67.2	70.6
IDPR [10]	88.7	94.1	84.1	78.2	71.6	57.6	72.9	76.6
ORGM-IDPR	89.8	93.9	85.3	79.8	73.0	60.7	74.7	78.1

Furthermore, it reflects that it is the IDPR term that significantly boosts the performance for the approach of Chen and Yuille [10] (9.7% over that without IDPR). Apparently the combination of CNN part detectors and IDPR are not susceptible to brittle message passing in the presence of occlusions. Actually, the incorporated information is only from neighboring parts with very short range. The simple tree model may still fail without long range message passing. The first two rows of images in Fig. 8 are the detection results on the LSP subset. It shows that when there is partial occlusion or large deformation, the weak part response of one side of the limb inclines to double-counting for the approach of Chen and Yuille [10]. In contrast, our model takes the advantage of long range constraints which can overcome this issue. The last two rows of images in Fig. 8 are the detection results on the MPII subset with severe occlusion. It reflects that our model is more robust under both self-occlusion and other-occlusion as well as large deformation compared with the method of Chen and Yuille [10].

V. CONCLUSION AND FUTURE WORK

In this paper, we have proposed an occlusion relational graphical model to model both self-occlusion and other-occlusion in human pose estimation. The proposed model can capture the complex interactions among parts, enabling occlusion handling, especially self-occlusion. We demonstrate that part level occlusion reasoning is important for human pose estimation as occlusion coherence and stronger structural constraints can be embedded in such model. The experimental results show the superiority of our method compared with the state-of-the-arts. Our method especially obtains promising performance in human pose estimation with occlusion. In the later future, we will try to use CNNs to learn occlusion patterns and occlusion relationships explicitly.

REFERENCES

[1] Y. Yang and D. Ramanan, "Articulated pose estimation with flexible mixtures-of-parts," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2011, pp. 1385–1392.

[2] M. Sun and S. Savarese, "Articulated part-based model for joint object detection and pose estimation," in *Proc. IEEE Int. Conf. Comput. Vis.*, Nov. 2011, pp. 723–730.

[3] Y. Tian, C. L. Zitnick, and S. G. Narasimhan, "Exploring the spatial hierarchy of mixture models for human pose estimation," in *Proc. Eur. Conf. Comput. Vis.*, Oct. 2012, pp. 256–269.

[4] F. Wang and Y. Li, "Beyond physical connections: Tree models in human pose estimation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2013, pp. 596–603.

[5] G. Ghiasi, Y. Yang, D. Ramanan, and C. Fowlkes, "Parsing occluded people," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 2401–2408.

[6] X. Chen and A. Yuille, "Parsing occluded people by flexible compositions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2015, pp. 3945–3954.

[7] S. Johnson and M. Everingham, "Clustered pose and nonlinear appearance models for human pose estimation," in *Proc. Brit. Mach. Vis. Conf.*, 2010, pp. 12.1–12.11.

[8] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2012, pp. 1106–1114.

[9] A. Toshev and C. Szegedy, "DeepPose: Human pose estimation via deep neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 1653–1660.

[10] X. Chen and A. L. Yuille, "Articulated pose estimation by a graphical model with image dependent pairwise relations," in *Proc. Adv. Neural Inf. Process. Syst.*, 2014, pp. 1736–1744.

[11] J. Tompson, A. Jain, Y. LeCun, and C. Bregler, "Joint training of a convolutional network and a graphical model for human pose estimation," in *Proc. Adv. Neural Inf. Process. Syst.*, 2014, pp. 1799–1807.

[12] X. Fan, K. Zheng, Y. Lin, and S. Wang, "Combining local appearance and holistic view: Dual-source deep neural networks for human pose estimation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2015, pp. 1347–1355.

[13] B. Vasileios, R. Christian, C. Gustavo, and N. Nassir, "Robust optimization for deep regression," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2015, pp. 2830–2838.

[14] L. Fu, J. Zhang, and K. Huang, "Beyond tree structure models: A new occlusion aware graphical model for human pose estimation," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2015, pp. 1976–1984.

[15] M. A. Fischler and R. A. Elschlager, "The representation and matching of pictorial structures," *IEEE Trans. Comput.*, vol. C-22, no. 1, pp. 67–92, Jan. 1973.

[16] M. Andriluka, S. Roth, and B. Schiele, "Pictorial structures revisited: People detection and articulated pose estimation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2009, pp. 1014–1021.

[17] L. Pishchulin, M. Andriluka, P. Gehler, and B. Schiele, "Poselet conditioned pictorial structures," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2013, pp. 588–595.

[18] Y. Wang and G. Mori, "Multiple tree models for occlusion and spatial constraints in human pose estimation," in *Proc. Eur. Conf. Comput. Vis.*, Oct. 2008, pp. 710–724.

[19] L. Fu, J. Zhang, and K. Huang, "Context aware model for articulated human pose estimation," in *Proc. IEEE Int. Conf. Image Process.*, Sep. 2015, pp. 991–995.

[20] T.-P. Tian and S. Sclaroff, "Fast globally optimal 2D human detection with loopy graph models," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2010, pp. 81–88.

[21] D. Tran and D. Forsyth, "Improved human parsing with a full relational model," in *Proc. Eur. Conf. Comput. Vis.*, Sep. 2010, pp. 227–240.

[22] B. Sapp, D. Weiss, and B. Taskar, "Parsing human motion with stretchable models," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2011, pp. 1281–1288.

[23] Y. Wang, D. Tran, and Z. Liao, "Learning hierarchical poselets for human parsing," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2011, pp. 1705–1712.

[24] M. Sun, M. Telaprolu, H. Lee, and S. Savarese, "An efficient branch-and-bound algorithm for optimal human pose estimation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2012, pp. 1616–1623.

[25] G. Mori and J. Malik, "Estimating human body configurations using shape context matching," in *Proc. Eur. Conf. Comput. Vis.*, May 2002, pp. 666–680.

[26] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2005, pp. 886–893.

[27] G. Gkioxari, B. Hariharan, R. Girshick, and J. Malik, "Using k-poselets for detecting people and localizing their keypoints," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 3582–3589.

[28] L. D. Bourdev, S. Maji, and J. Malik, "Describing people: A poselet-based approach to attribute classification," in *Proc. IEEE Int. Conf. Comput. Vis.*, Nov. 2011, pp. 1543–1550.

[29] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan, "Object detection with discriminatively trained part-based models," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, no. 9, pp. 1627–1645, Sep. 2010.

[30] L. Pishchulin, M. Andriluka, P. Gehler, and S. Bernt, "Strong appearance and expressive spatial models for human pose estimation," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2013, pp. 3487–3494.

[31] M. Dantone, J. Gall, C. Leistner, and L. V. Gool, "Human pose estimation using body parts dependent joint regressors," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2013, pp. 3041–3048.

[32] V. Ramakrishna, D. Munoz, M. Hebert, J. Bagnell, and Y. Sheikh, "Pose machines: Articulated pose estimation via inference machines," in *Proc. Eur. Conf. Comput. Vis.*, Sep. 2014, pp. 33–47.

[33] X. Wang, T. X. Han, and S. Yan, "An HOG-LBP human detector with partial occlusion handling," in *Proc. IEEE Int. Conf. Comput. Vis.*, Sep. 2009, pp. 32–39.

[34] T. Gao, B. Packer, and D. Koller, "A segmentation-aware object detection model with occlusion handling," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2011, pp. 1361–1368.

[35] H. Azizpour and I. Laptev, "Object detection using strongly-supervised deformable part models," in *Proc. Eur. Conf. Comput. Vis.*, Oct. 2012, pp. 836–849.

[36] M. Hejrati and D. Ramanan, "Analyzing 3D objects in cluttered images," in *Proc. Adv. Neural Inf. Process. Syst.*, 2012, pp. 602–610.

[37] C. Desai and D. Ramanan, "Detecting actions, poses, and objects with relational phraselets," in *Proc. Eur. Conf. Comput. Vis.*, Oct. 2012, pp. 158–172.

[38] B. Rothrock, S. Park, and S.-C. Zhu, "Integrating grammar and segmentation for human pose estimation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2013, pp. 3214–3221.

[39] R. B. Girshick, P. F. Felzenszwalb, and D. A. McAllester, "Object detection with grammar models," in *Proc. Adv. Neural Inf. Process. Syst.*, 2011, pp. 442–450.

[40] L. Sigal and M. J. Black, "Measure locally, reason globally: Occlusion-sensitive articulated pose estimation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2006, pp. 2041–2048.

[41] Y. Yang and G. Sundaramoorthi, "Modeling self-occlusions in dynamic shape and appearance tracking," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2013, pp. 201–208.

[42] I. Radwan, A. Dhall, and R. Goecke, "Monocular image 3D human pose estimation under self-occlusion," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2013, pp. 1888–1895.

[43] P. F. Felzenszwalb and D. P. Huttenlocher, "Pictorial structures for object recognition," *Int. J. Comput. Vis.*, vol. 61, no. 1, pp. 55–79, 2005.

[44] D. Koller and N. Friedman, *Probabilistic Graphical Models: Principles and Techniques*. Cambridge, MA, USA: MIT Press, 2009.

[45] X. Lan and D. P. Huttenlocher, "Beyond trees: Common-factor models for 2D human pose recovery," in *Proc. IEEE Int. Conf. Comput. Vis.*, Oct. 2005, pp. 470–477.

[46] X. Ren, A. C. Berg, and J. Malik, "Recovering human body configurations using pairwise constraints between parts," in *Proc. IEEE Int. Conf. Comput. Vis.*, Oct. 2005, pp. 824–831.

[47] Y. Weiss, "Correctness of local probability propagation in graphical models with loops," *Neural Comput.*, vol. 12, no. 1, pp. 1–41, 2000.

[48] N. Komodakis, N. Paragios, and G. Tziritas, "MRF energy minimization and beyond via dual decomposition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 3, pp. 531–552, Mar. 2011.

[49] K. Liu, J. Zhang, P. Yang, S. Maybank, and K. Huang, "GRMA: Generalized range move algorithms for the efficient optimization of MRFs," *Int. J. Comput. Vis.*, pp. 1–26, 2016.

[50] D. Ramanan, "Part-based models for finding people and estimating their pose," in *Visual Analysis of Humans*. Springer, London, 2011, pp. 199–223. [Online]. Available: http://dx.doi.org/10.1007/978-0-85729-997-0_11

[51] S. C. Tatikonda and M. I. Jordan, "Loopy belief propagation and Gibbs measures," in *Proc. Conf. Uncertainty Artif. Intell.*, Aug. 2002, pp. 493–500.

[52] G. Ghiasi and C. Fowlkes, "Occlusion coherence: Localizing occluded faces with a hierarchical deformable part mode," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 1899–1906.

[53] T. Joachims, T. Finley, and C.-N. J. Yu, "Cutting-plane training of structural SVMs," *Mach. Learn.*, vol. 77, no. 1, pp. 27–59, Oct. 2009.

[54] D. Ramanan, "Dual coordinate solvers for large-scale structural SVMs," *Comput. Sci.*, pp. 1100–1114, 2013. [Online]. Available: http://arxiv.org/abs/1312.1743

[55] D. Ramanan, "Learning to parse images of articulated bodies," in *Proc. Adv. Neural Inf. Process. Syst.*, 2006, pp. 1129–1136.

[56] B. Sapp and B. Taskar, "MODEC: Multimodal decomposable models for human pose estimation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2013, pp. 3674–3681.

[57] M. Andriluka, L. Pishchulin, P. Gehler, and B. Schiele, "2D human pose estimation: New benchmark and state of the art analysis," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 3686–3693.

[58] V. Ferrari, M. J. Mar³n-Jimnez, and A. Zisserman, "Progressive search space reduction for human pose estimation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2008, pp. 1–8.

[59] M. Eichner and M. J. Marín-Jiménez, A. Zisserman, and V. Ferrari, "2D articulated human pose estimation and retrieval in (almost) unconstrained still images," *Int. J. Comput. Vis.*, vol. 99, no. 2, pp. 190–214, Sep. 2012.

[60] Y. Yang and D. Ramanan, "Articulated human detection with flexible mixtures of parts," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 12, pp. 2878–2890, Dec. 2013.

[61] S. Johnson and M. Everingham, "Learning effective human pose estimation from inaccurate annotation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2011, pp. 1465–1472.

[62] J. Carreira, P. Agrawal, K. Fragkiadaki, and J. Malik, "Human pose estimation with iterative error feedback," *CoRR*, Aug. 2015. [Online]. Available: http://arxiv.org/abs/1507.06550

[63] M. Kiefel and P. V. Gehler, "Human pose estimation with fields of parts," in *Proc. Eur. Conf. Comput. Vis.*, Sep. 2014, pp. 331–346.

[64] M. Eichner and V. Ferrari, "Appearance sharing for collective human pose estimation," in *Proc. Asia Conf. Comput. Vis.*, Nov. 2012, pp. 138–151.

[65] W. Ouyang, X. Chu, and X. Wang, "Multi-source deep learning for human pose estimation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 2337–2443.

[66] Y. Jia *et al.* (Jun. 2014). "Caffe: Convolutional architecture for fast feature embedding." [Online]. Available: https://arxiv.org/abs/1408.5093

[67] L. D. Bourdev and J. Malik, "Poselets: Body part detectors trained using 3D human pose annotations," in *Proc. IEEE Int. Conf. Comput. Vis.*, Sep. 2009, pp. 1365–1372.

Lianrui Fu (S'15) received the B.S. and M.S. degrees from Beihang University, Beijing, China, in 2008 and 2013, respectively. He is currently pursuing the Ph.D. degree with the National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing. His research interests include computer vision, object recognition, and human pose estimation.

Junge Zhang (M'14) received the Ph.D. degree in pattern recognition and intelligent systems from the Institute of Automation, Chinese Academy of Sciences, in 2013. In 2013, he joined the Center for Research on Intelligent Perception and Computing as an Assistant Professor. His major research interests include computer vision and pattern recognition. He is a Committee Member of CCF YOCSEF. In 2010 and 2011, he and his group members won the champion of PASCAL VOC challenge on object detection and ranked the second on object classification. He served as the Publicity Chair and the Technical Program Committee Member of several conferences, and the Peer Reviewer of over 10 international journals and conferences.

Kaiqi Huang (SM'09) is currently a Professor with the National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Science, China. He has authored over 140 papers in important international journals and conference, such as the IEEE TPAMI, IJCV, TIP, TSMCB, TCSVT, Pattern Recognition, CVIU, ECCV, CVPR, ICIP, and ICPR. His current research interests include computer vision, pattern recognition, and biological-based vision. He received the Best Student Paper Awards from ACPR 2010, the winner prizes of the detection task in both PASCAL VOC 2010 and PASCAL VOC 2011, the Honorable Mention Prize of the classification task in PASCAL VOC 2011, and the Winner Prize of classification task with additional data in ILSVRC 2014. He has been an Associate Editor of the IEEE TRANSACTIONS ON SYSTEMS, MAN, AND CYBERNETICS: SYSTEMS from 2014 and was the Deputy General Secretary of the IEEE Beijing Section (2006–2008).