CrowdPose: Efficient Crowded Scenes Pose Estimation and A New Benchmark

Jiefeng Li[1], Can Wang[1], Hao Zhu[1], Yihuan Mao[2], Hao-Shu Fang[1], Cewu Lu[1†*]
[1]Shanghai Jiao Tong University, [2]Tsinghua University
{ljf_likit,wangcan123,lucewu}@sjtu.edu.cn
haozhu@zju.edu.cn maoyh16@mails.tsinghua.edu.cn fhaoshu@gmail.com

Abstract

*Multi-person pose estimation is fundamental to many computer vision tasks and has made significant progress in recent years. However, few previous methods explored the problem of pose estimation in crowded scenes while it remains challenging and inevitable in many scenarios. Moreover, current benchmarks cannot provide an appropriate evaluation for such cases. In this paper, we propose a novel and efficient method to tackle the problem of pose estimation in the crowd and a new dataset to better evaluate algorithms. Our model consists of two key components: joint-candidate single person pose estimation (SPPE) and global maximum joints association. With multi-peak prediction for each joint and global association using the graph model, our method is robust to inevitable interference in crowded scenes and very efficient in inference. The proposed method surpasses the state-of-the-art methods on CrowdPose dataset by **5.2** mAP and results on MSCOCO dataset demonstrate the generalization ability of our method. Source code and dataset are available at https://github.com/Jeff-sjtu/CrowdPose*

1. Introduction

Estimating multi-person poses in images plays an important role in the area of computer vision. It has attracted tremendous interest for its wide applications in activity understanding [14, 11], human-object interaction detection [41, 37], human parsing [18, 42] etc. Some works focus on 3D human pose estimation [33, 34, 31]. Currently, most of the 2D methods can be roughly divided into two categories: i) top-down approaches that firstly detect each person and then perform single person pose estimation, or ii) bottom-up approaches which detect each joint and then associate them into a whole person.

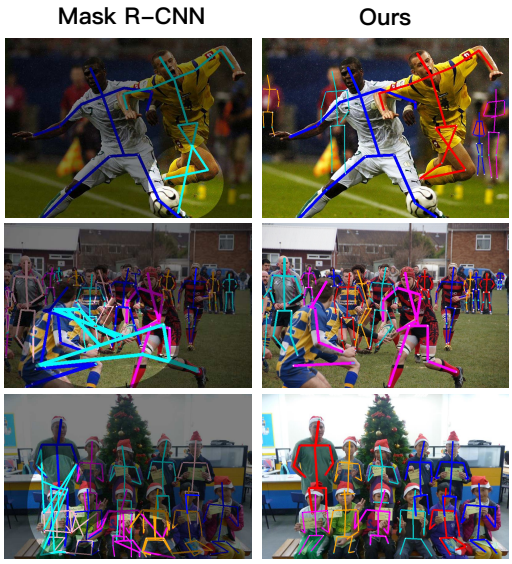


Figure 1. Qualitative comparison of Mask R-CNN and our CrowdPose method in crowded scenes. Though current methods achieve good performance in public benchmarks [13, 15, 20], they fail in crowded cases. There are mainly two types of errors: i) assemble wrong joints into a pose; ii) predict redundant poses in crowded scenes.

To evaluate the performance of multi-person pose estimation algorithms, several public benchmarks were established, such as MSCOCO [15], MPII [13] and AI Challenger [20]. In these benchmarks, the images are usually collected from daily life where crowded scenes appear less frequently. As a result, most of the images in these benchmarks have few mutual occlusion among humans. For example, in MSCOCO dataset (persons subset), 67.01% of the images have no overlapped person. Current methods have obtained encouraging success on these datasets.

However, despite the good performance that current methods have achieved on previous benchmarks, we observe an obvious degradation of their performance in crowded cases. As shown in Figure 1, for the current state-of-the-art methods [40, 27, 23, 26] of both bottom-up and

†Cewu Lu is the corresponding author.
*Cewu Lu is a member of Department of Computer Science and Engineering, Shanghai Jiao Tong University, MoE Key Lab of Artificial Intelligence, AI Institute, Shanghai Jiao Tong University, and SJTU-SenseTime AI Lab.

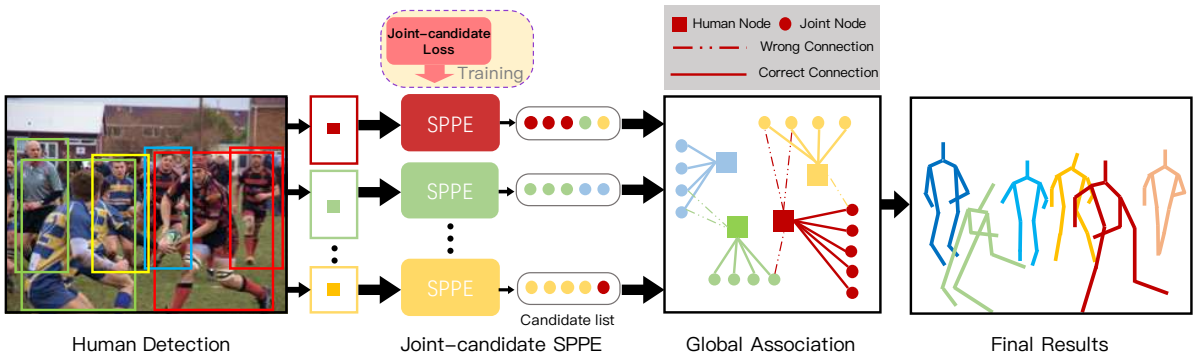


Figure 2. Pipeline of our proposed method. JC SPPE uses joint-candidate loss function during training phase. In inference phase, JC SPPE receives human proposals and generates joint candidates. Then we utilize human proposals and joint candidates to build a person-joint graph. Finally, we associate joints with human proposals by solving the assignment problem in our graph model.

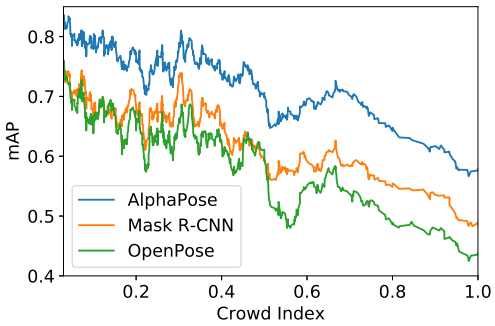


Figure 3. State-of-the-art methods evaluation results on MSCOCO dataset. The x-axis is the *Crowd Index* which we defined to measure the crowding level of an image. Compared to uncrowded scenes, the accuracy (mAP@0.5:0.95) of state-of-the-art methods are about 20 mAP lower in crowded cases.

top-down approaches, their performance decreases dramatically as the crowd level increases (as Figure 3). Few previous methods aimed to tackle the problem of pose estimation in crowded scene, and no public benchmark has been built for this purpose. Meanwhile, crowded scenes are inevitable in many scenarios.

In this paper, we propose a novel method to tackle the problem of pose estimation in a crowd, using a global view to address interference problem. Our method follows the top-down framework, which first detects individual persons and then performs single person pose estimation (SPPE). We propose a joint-candidate SPPE and a global maximum joints association algorithm. Different from previous methods that only predict target joints for input human proposals, our joint-candidate SPPE outputs a list of candidate locations for each joint. The candidate list includes target and interference joints. Then our association algorithm utilizes these candidates to build a person-joint connection graph. At last, we solve the joint association problem in this graph model with a global maximum joints association algorithm. Moreover, the computational complexity of our graph optimization algorithm is the same as the conventional NMS algorithm.

To better evaluate human pose estimation algorithms in crowded scenes and promote the development in this area, we collect a dataset of crowded human poses. We define a *Crowd Index* to measure the crowding level of an image. Images in our dataset have a uniform distribution of *Crowd Index* among $[0, 1]$, which means only an algorithm that performs well on both uncrowded and crowded scenes can achieve a high score in our dataset.

To sum up, the contributions of this paper are as follows: i) we propose a novel method to tackle the crowded problem of pose estimation; ii) we collect a new dataset of crowded human poses to better evaluate algorithms in crowded scenes. We conduct experiments on our proposed method. When using a same ResNet-101 based network backbone, our method surpasses all the state-of-the-art methods by **5.2** mAP on our dataset. Moreover, we replace the SPPE and post-processing steps in the state-of-the-art method with our module and brings 0.8 mAP improvement on MSCOCO dataset. That is, our method can generally work in non-crowded scenes.

2. Related Work

2.1. 2D Pose Estimation Dataset

Pioneer works on 2D human pose estimation dataset on RGB images involve LSP [4], FashionPose [9], PASCAL Person Layout [6], J-HMDB [12], etc. These datasets have contributed to encouraging progress of human pose estimation. However, they only evaluate for single person pose estimation. With the improvement of algorithms, more researchers focus on multi-person pose estimation problems, and several datasets are established, *e.g.*, MPII [13], MSCOCO [15], AI Challenger [20]. In spite of the prevalence of these datasets, they suffer from a low-density problem, which makes the current model overfitted to uncrowded scenes. The performance of the state-of-the-art

methods decreases as the number of human increases.

2.2. Multi-Person Pose Estimation

Part-Based Framework Representative works on the part-based framework [26, 38, 35, 31] are reviewed. Part-based methods detect joints and associate them into a whole person. The state-of-the-art part-based methods are mainly different on their association methods. Cao *et al*. [26] associate joints with a part affinity field and greedy algorithm. Zanfir *et al*. [31] propose a limb scoring network to estimate the connection likelihood of joints and group people by binary integer programming. Papandreou *et al*. [38] detect individual joints and predict relative displacements for association. Kocabas *et al*. [35] propose a multi-task model and assign joints to detected persons by a pose residual network. The joint detectors in part-based approached are relatively vulnerable because they only consider small local regions and output smaller response heatmaps.

Two-Step Framework Our work follows the two-step approach. A two-step approach first detects human proposals [16, 24] and then performs single person pose estimation [19, 22]. The state-of-the-art two-step methods [21, 40, 27, 23] achieve significantly higher scores than the part-based methods. However, the two-step approaches highly depend on human detection results, and it fails in crowded scenes [30]. When people stay close to each other in a crowd, it is improbable to crop a bounding box that only contains one person. Some works [5, 8, 39, 25] in human tracking area use temporal information to fix wrong detection with CNN or RNN [3, 32] module. As a supplement to them, we propose a novel and efficient method that significantly increases pose estimation performance in crowded scenes, which is robust to human detection results.

3. Our Method

The pipeline of our proposed method is illustrated in Figure 2. Human bounding box proposals obtained by human detector are fed into *joint-candidate* (JC) single person pose estimator (SPPE). JC SPPE locates the joint candidates with different response scores on the heatmap (Sec. 3.1). Then our joint association algorithm takes these results and builds a person-joint connection graph (Sec. 3.2). Finally, we solve the graph matching problem to find the best joint association result with a global maximum joints association algorithm (Sec. 3.3).

3.1. Joint-Candidates SPPE

Joint-candidate SPPE receives a human proposal image and outputs a group of heatmaps to indicate human joint locations. Though a human proposal should indicates only one human instance, in the crowded scenarios, we inevitably need to handle a large number of joints from other human instances. Previous works [40, 21, 27] use SPPE to suppress interference joints. However, SPPE fails in crowded scenes because their receptive fields are limited by the input human proposals. To address this problem, we propose joint-candidate SPPE with a novel loss designed in a more global view.

3.1.1 Loss Design

For the i^{th} human proposal, we input its region $\mathbf{R}_i$ into our SPPE network and get the output heatmap $\mathbf{P}_i$. There are two types of joints in $\mathbf{R}_i$, that is, the joints belong to the i^{th} person, and the joints belong to other human instances (not the i^{th} person). We name them as target joints and interference joints respectively.

We adopt heatmap in our loss module, which is widely used in many areas [29, 28] for its pixel-wise supervision and fully convolution structure. Our goal is to enhance target joints response and suppress interference joints response. However, we don't suppress them directly since interference joints for the current proposal can be regarded as target joints for other proposals. Thus, we can leverage interference joints to estimate human poses with other human proposals in a global manner. Therefore, to utilize those two kinds of joint candidates, we output them with different intensities.

Heatmap Loss For the k^{th} joint in the i^{th} person, we denote the target joint heatmap as $\mathbf{T}_i^k$, consisting of a 2D Gaussian $G(\mathbf{p}_i^k|\sigma)$, centered at the target joint location $\mathbf{p}_i^k$, with standard deviation σ.

For interference joints, we denote them as a set Ω_i^k. The heatmap of interference joints is denoted as $\mathbf{C}_i^k$, consisting of a Gaussian mixture distribution $\sum_{p\in\Omega_i^k} G(\mathbf{p}|\sigma)$.

Our proposed loss is defined as,

$$Loss_i = \frac{1}{K}\sum_{k=1}^{K} MSE[\mathbf{P}_i^k, \mathbf{T}_i^k + \mu\mathbf{C}_i^k] \qquad (1)$$

where μ is an attenuation factor ranged in [0,1]. As aforementioned, interference joints will be useful in indicating joints of other human instances. Therefore, we should consider it in a global view by cross-validation. Finally, we have $\mu = 0.5$, which fits our intuition: interference joints should be attenuated but not over-suppressed. The conventional heatmap loss function can be regarded as our special case where $\mu = 0$.

3.1.2 Discussion

A conventional SPPE depends on a high-quality human detection result. Its tasks are locating and identifying target joints according to the given human proposal. If SPPE mistakes interference joints for target joints, it will be an unrecoverable error. Missing joints cannot be restored in the post-processing step like pose-NMS.

Our proposed joints candidate loss is aimed to tackle this limitation. This loss function encourages JC SPPE network to predict multi-peak heatmaps and sets all the possible joints as candidates. In crowded scenes, while conventional SPPE is hard to identify target joints, JC SPPE can still predict a list of joint candidates and guarantee high recall. We leave the association problem to the next procedure, where we have more global information from other JC SPPEs (on other human proposals) to solve it.

3.2. Person-Joint Graph

Due to our joint-candidate mechanism and redundant human proposals from human detector, joint candidates are numerically much greater than the actual joint numbers. To reduce redundant joints, we build a person-joint graph and apply a maximum person-joint matching algorithm to construct the final human poses.

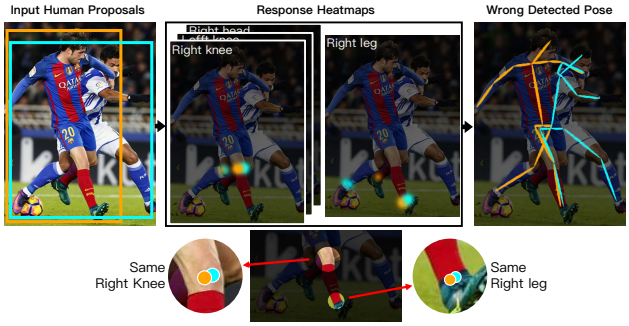


Figure 4. In crowded scenes, human proposals are highly overlapped. Overlapped human proposals tend to predict same actual joint. In this example, if we directly connect the highest response to build final poses, two human proposals will locate same right knee and right leg. Our proposed association algorithm can solve this problem by globally best matching.

3.2.1 Joint Node Building

Since highly overlapped human proposals tend to predict the same actual joint (as Figure 4), we first group these candidates that represent the same actual joint as one joint node.

Thanks to the high-quality joint prediction, candidate joints that indicate the same joint are always close to each other. Thus, we can group them using the following criterion: given two candidate joints located at p_1^k and p_2^k with control deviation δ^k, we label them as the same group, if

$$||p_1^{(k)} - p_2^{(k)}||_2 \leq \min\{u_1^k, u_2^k\}\delta^{(k)}, \tag{2}$$

where u_1^k and u_2^k are the Gaussian response size of two joints on heatmaps, determined by the Gaussian response deviation. $\delta^{(k)}$ is the parameter for controlling deviation of the k^{th} joint, which we directly adopt from MSCOCO keypoint dataset [15]. The reason why we use $\min\{u_1, u_2\}$ rather than a constant threshold is to guarantee that, only if p_1 and p_2 fall into each others' control domain (radii are $u_1^k\delta^k$, $u_2^k\delta^k$) simultaneously, we group them together. One node represents a group of joints that cluster together by the above criterion.

Now, by building a joint group as one node, we have joint node set $\mathcal{J} = \{v_j^k : \text{for } k \in \{1, \ldots, K\}, j \in \{1, \ldots, N_k\}\}$, where N_k is the number of joint nodes of body part k, v_j^k is the j^{th} node of body part k. The total number of joint nodes in $\mathcal{J}$ is $\sum_k N_k$.

3.2.2 Person Node Building

Person nodes represent the human proposals detected by human detector. We denote person node set as $\mathcal{H} = \{h_i : \forall i \in \{1 \ldots M\}\}$, where h_i is the i^{th} person node, and M is the number of detected human proposals.

Ideally, a qualified human proposal tightly bounds a human instance. However, in crowded scenes, this condition is not always satisfied. The human detector will produce many redundant proposals, including truncated and incompact bounding boxes. We will eliminate these low-quality person nodes during global person-joint matching in Sec. 3.3.

3.2.3 Person-Joint Edge

After obtaining the node of both joints and persons, we connect them to construct our person-joint graph. For each person node h_i, JC-SPPE will predict several candidate results of joints. If one of these candidates contributes to the joint node v_j^k, we build an edge $e_{i,j}^k$ between them. The weight of $e_{i,j}^k$ is the response score of that candidate joint, which is denoted as $w_{i,j}^k$. In this way, we can construct the edge set $\mathcal{E} = \{e_{i,j}^k : \forall i, j, k\}$.

The person-joints graph can then be written as:

$$\mathcal{G} = ((\mathcal{H}, \mathcal{J}), \mathcal{E}). \tag{3}$$

3.3. Globally Optimizing Association

From now on, our goal of estimating human poses in the crowd is transformed into solving the above person-joint graph and maximizing the total edge weights. We have our objective function as:

$$\max_d \mathcal{G} = \max_d \sum_{i,j,k} w_{i,j}^{(k)} \cdot d_{i,j}^{(k)} \tag{4}$$

$$s.t. \quad \sum_j d_{i,j}^{(k)} \leq 1, \quad \begin{matrix}\forall k \in \{1, \ldots, K\}, \\ \forall i \in \{1, \ldots, M\}\end{matrix} \tag{5}$$

$$\sum_i d_{i,j}^{(k)} \leq 1, \quad \begin{matrix}\forall k \in \{1, \ldots, K\}, \\ \forall j \in \{1, \ldots, N_k\}\end{matrix} \tag{6}$$

$$d_{i,j}^{(k)} \in \{0, 1\}, \quad \forall i, j, k \tag{7}$$

where $d_{i,j}^{(k)}$ indicates whether we keep the edge $e_{i,j}^{k}$ in our final graph or not. The constraints of Eq. 5 and 6 enforce that each human proposal can only match at most one k^{th} joint.

Note that $\mathcal{G}$ can be decomposed into K sub-graph $\mathcal{G}_k = ((\mathcal{H}, \mathcal{J}^{(k)}), \mathcal{E}^{(k)})$, where $\mathcal{G}_k$ is the sub-graph the only consist of the k^{th} kind of joints. Thus, our objective function can be formulated as

$$\max_{d} \mathcal{G} = \max_{d} \sum_{i,j,k} w_{i,j}^{(k)} \cdot d_{i,j}^{(k)} \tag{8}$$

$$= \sum_{k=1}^{K} (\max_{d^{(k)}} \sum_{i,j} w_{i,j}^{(k)} \cdot d_{i,j}^{(k)}) \tag{9}$$

$$= \sum_{k=1}^{K} \max_{d^{(k)}} \mathcal{G}_k. \tag{10}$$

As shown in Eq. 10, solving the global assignment problem in person-joint graph $\mathcal{G}$ is mathematically equivalent to solving its sub-graph $\mathcal{G}_k$ separately. $\mathcal{G}_k$ is a bipartite graph that composed of person subset and the k^{th} joint subset. For each sub-graph, the updated Kuhn-Munkres algorithm [1] is applied to get the optimized result. By addressing each $\mathcal{G}_k$ respectively, we obtain the final result set $\mathcal{R}$.

Given the graph matching result, if $d_{i,j}^{(k)} = 1$ the weighted center of v_j^k is assigned to the i^{th} human proposal as its k^{th} joint. Here, weighted center means the linear combination of candidate joints coordinate in v_j^k and the weights are their heatmap response scores. In this way, the pose of each human proposal can be constructed. The person nodes that can not match any joint will be removed.

Computational Complexity The inference speed of pose estimation is essential in many applications. We prove that our global association algorithm is as efficient as common greedy NMS algorithms.

As the hereditary property identified by White and Whiteley [2], a graph G is $(k,l)-sparse$ if every nonempty sub-graph X has at most $k|X| - l$ edges, where $|X|$ is the number of vertices in sub-graph X and $0 \leq l < 2k$.

Consider the sub-graph $\mathcal{G}^{(k)} = ((\mathcal{H}, \mathcal{J}^{(k)}), \mathcal{E}^{(k)})$. It represents the connection between human proposals and the k^{th} type of joints. According to our statistics (Fig. 5), every human bounding box covers four persons at most in crowded scenes. Therefore, one person node builds connection edges to 4 joints at most. In other words, our person-joint sub-graph $\mathcal{G}^{(k)}$ is $(4,0) - sparse$ since

$$|\mathcal{E}^{(k)}| \leq 4|\mathcal{G}^{(k)}| - 0. \tag{11}$$

Due to the sparsity of our person-joint graph, we can solve the association problem efficiently. We transform $\mathcal{E}^{(k)}$ into an adjacency matrix M_{e^k} (unconnected nodes refer to 0). According to the work of Carpaneto *et al.* [1], this linear assignment problem for the sparse matrix can be solved in $\mathcal{O}(n^2)$, i.e., $\mathcal{O}((|\mathcal{H}| + |\mathcal{J}^{(k)}|)^2)$. Since we have eliminated the redundant joints and there is a one-to-one correspondence between joints and persons, the expectation of $|\mathcal{J}^{(k)}|$ is equal to $|\mathcal{H}|$. Thus we have $\mathcal{O}((|\mathcal{H}| + |\mathcal{J}^{(k)}|)^2) = \mathcal{O}(|\mathcal{H}|^2)$. Such computation complexity is the same as the complexity of conventional greedy NMS algorithms.

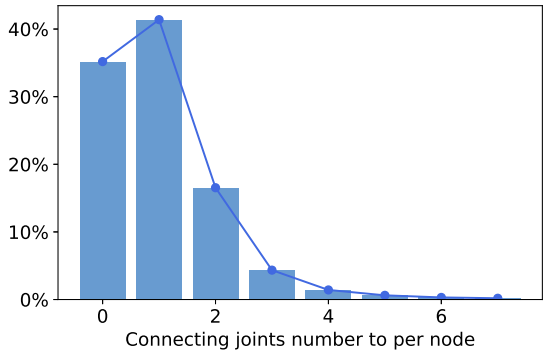


Figure 5. Instance-Joint connection distribution. The x-axis denote the number of human bounding boxes that cover a same joint. This statistical result is based on the ground truth annotations.

3.4. Discussion

Our method adopts the graph-based approach to associate joints with human proposals in a globally optimal manner. Human proposals compete with each other for joint nodes. In this way, unqualified human proposals without dominant human instance would fail to be assigned any joints, since their joint response scores are all relatively low due to missing dominant human instance. Therefore, many redundant and poor human proposals are rejected. In comparison to our approach, conventional NMS is a greedy and instance-based algorithm, which is less effective. Although [10, 17, 27] proposed pose-NMS to utilize pose information, their algorithms are based on instances and cannot tackle the missing joints and wrong assembling problem. Our globally optimizing association method can deal with such situations well.

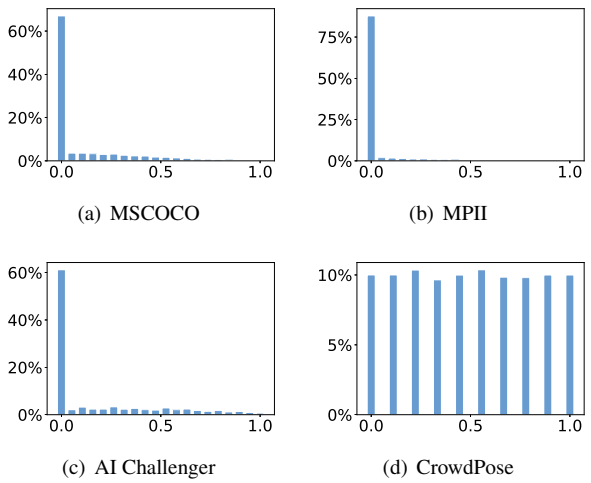
4. CrowdPose Dataset

In this section, we introduce another contribution of our paper, namely, CrowdPose dataset, including crowded scenes definition, data collection process, and the dataset statistics.

4.1. Crowding Level Definition

To build a dataset of crowded human pose, we need to define a *Crowd Index* first, which measure the crowding level in a given image.

Intuitively, the number of persons in an image seems to be a good measurement. However, the principal obstacle to



(a) MSCOCO (b) MPII

(c) AI Challenger (d) CrowdPose

Figure 6. $CrowdIndex$ distributions of current popular datasets and our dataset. Three public datasets are dominated by uncrowded images. Meanwhile, our CrowdPose dataset has a near uniform distribution.

solving crowded cases is not caused by the number of persons, but rather by occlusion in a crowd. Therefore, we need a new *Crowd Index* to indicate crowding level. In the bounding box of the i^{th} human instance, we denote the number of joints that belonging to the i^{th} person and other (not i^{th}) persons as N_i^a and N_i^b respectively. N_i^b/N_i^a is the *crowd ratio* of the i^{th} human instance. Our *Crowd Index* is derived by averaging the crowd ratio of all persons in an image:

$$Crowd\ Index = \frac{1}{n}\sum_{i=1}^{n}\frac{N_i^b}{N_i^a}, \tag{12}$$

where n indicates the total number of persons in the image.

We evaluate the *Crowd Index* distribution of three public benchmarks: MSCOCO (person subset), MPII and AI Challenger. As shown in Figure 6, uncrowded scenes dominate these benchmarks, which leads the SOTA methods only focus on these simple cases and ignore the crowded ones.

4.2. Data Collection

To set up a benchmark that covers various scenes and encourages models to adapt to different kinds of situations, we wish our benchmark covers not only crowded cases but also simple daily life scenes. To achieve that, we first analyze three public benchmarks [15, 13, 20] and divide their images into 20 groups according to *Crowd Index*, ranging from 0 to 1. The step among different groups is 0.05. Then we sample 30,000 images uniformly from these groups in total.

4.3. Image Annotation

Although these images have been annotated, their label formats are not fully aligned. In terms of annotated joints number, MSCOCO has 17 keypoints, while MPII has 16 and AI Challenger annotates 14 keypoints. Meanwhile, human annotators are easier to make mistakes in the crowded cases. Thus, we re-annotate these images by the following steps.

- We annotate 14 keypoints and full-body bounding boxes for persons in 30,000 images.
- We analyze the *Crowd Index* for 30,000 images again with new annotations, and select 20,000 high-quality images.
- We further crop each person in the images, and then annotate the interference keypoints in each box.

We use cross annotation, which means at least two annotators annotate each image. If these two annotations have a large deviation, we regard them as mistakes and re-annotate this image. Finally, we take the average value of each keypoint location to ensure the annotation quality.

4.4. Dataset Statistics

Dataset Size In total, our dataset consists of 20,000 images, containing about 80,000 persons. The training, validation and testing subset are split in proportional to 5:1:4.

Crowd Index Distribution The *Crowd Index* distribution of CrowdPose is shown in Figure 6 (d). Unlike the other datasets, CrowdPose has a uniform distribution of *Crowd Index*. Note that we do not simply force CrowdPose to reach high *Crowd Index*. If a model is only trained on crowded scenes, it may degrade its performance on uncrowded cases due to the bias of training set. Uniform distribution can promote a model to adapt to various scenes.

Average IoU We further calculate the average intersection over union(IoU) of human bounding boxes. It turns out that CrowdPose has an average bounding box IoU of 0.27, while MSCOCO, MPII, and AI Challenger have 0.06, 0.11 and 0.12 respectively.

5. Experiments

In this section, we first introduce datasets and settings for evaluation. Then we report our results and comparisons with state-of-the-art methods, and finally conduct ablation studies on components in our method.

5.1. Datasets

CrowdPose Our proposed CrowdPose dataset contains 20,000 images in total and 80,000 human instances. 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.

MSCOCO Keypoints We also evaluate our method on the MSCOCO Keypoints dataset [15]. It contains over 150,000 instances for training and 80,000 instances for testing. The persons in this dataset overlap less frequently than CrowdPose, and it has a *Crowd Index* centralized near zero.

Figure 7. Qualitative results of our models predictions is presented. Different person poses are painted in different colors to achieve better visualization.

Method	mAP @0.5:0.95	mAP@0.5	mAP @0.75	mAR @0.5:0.95	mAR @0.5	mAR @0.75
Mask R-CNN [23]	57.2	83.5	60.3	65.9	89.5	69.4
AlphaPose [27]	61.0	81.3	66.0	67.6	86.7	71.8
Xiao *et al*. [40]	60.8	81.4	65.7	67.3	86.3	71.8
Ours	**66.0**	84.2	71.5	**72.7**	89.5	77.5

Table 1. Results on CrowdPose test set.

Method	AP_{Easy}	AP_{Medium}	AP_{Hard}	FPS
OpenPose [36]	62.7	48.7	32.3	5.3
Mask R-CNN [23]	69.4	57.9	45.8	2.9
AlphaPose [27]	71.2	61.4	51.1	10.9
Xiao *et al*. [40]	71.4	61.2	51.2	-
Ours	**75.5**	**66.3**	**57.4**	10.1

Table 2. Results on CrowdPose test set. Test set is divided into three part and we report the results respectively. FPS column reports the runtime speed on the whole test set.

5.2. Evaluation Metric

We follow the evaluation metric of MSCOCO, using average precision (AP) and average recall (AR) to evaluate the result. Object keypoint similarity (OKS) plays the same role as the IoU to adopt AP/AR for keypoints detection. We consider **mAP**, averaged over multiple OKS values (.50:.05:.95), as our primary metric. Moreover, we divide the CrowdPose dataset into three crowding levels by *Crowd Index*: easy (0-0.1), medium (0.1-0.8) and hard (0.8-1), to better evaluate our model performance in different crowded scenarios. We use the same keypoint standard deviations as MSCOCO when calculating OKS in all experiments.

5.3. Implementation Details

Our method follows the two-step framework. Since human detector and pose estimation network are not what we focus on, we simply adopt the human detector (YoloV3 [43]) and pose estimation network provided by AlphaPose [27] which is a state-of-the-art two-step method. During the training step, we adopt rotation (± 30), scaling ($\pm 30\%$) and flipping data augmentation. The input resolution is 320×256 and the output heatmap resolution is 80×64. The learning rate is set to 1×10^{-4} and 1×10^{-5} after 80 epochs. Mini-batch size is set to 64, and RMSprop [7] optimizer is used. During testing, the detected human bounding boxes are first extended by 30% along both the height and width directions and then forwarded through the Joint-candidate SPPE. The locations of joint candidates are obtained from the averaged output heatmaps of original and flipped input image. We conduct our experiments on two Nvidia 1080Ti GPUs. Our whole framework is implemented in PyTorch.

For comparison with current state-of-the-art methods [23, 40, 27] on CrowdPose dataset, we retrain them based on the configuration provided by the authors. To be fair, we use ResNet-101 as backbone for all SPPE networks and use the same human detector for [40]. For Mask R-CNN we also use the FPN-based ResNet-101 backbone.

Method	mAP @0.5:0.95	mAR @0.5:0.95
Mask R-CNN [23]	64.8	71.1
AlphaPose [27]	70.1	74.4
Xiao *et al*. [40]	69.8	74.1
Ours	**70.9**	**76.4**

Table 3. Results on MSCOCO test-dev set. We compare the state-of-the-art methods with same detection backbone.

	Method	mAP	mAR
	Ours (SPPE$^+$+Association)	**66.0**	**72.7**
(a)	w/o joint-candidate Loss	61.7	68.7
(b)	Greedy NMS	49.1	63.1
	Parametric Pose-NMS [27]	64.2	71.1

Table 4. Results on CrowdPose test set. mAP and mAR are the average value over multiple OKS values (0.50:0.05:0.95). "w/o X" means without X module in our pipeline. "Greedy NMS" means a conventional NMS method based on proposal scores and IoU.

Same training batch size is used for a fair comparison.

5.4. Results

CrowdPose Quantitative results on CrowdPose test set are given in Table 1. Our method achieves **5.2** mAP higher than state-of-the-art methods. It demonstrates the effectiveness of our proposed method to tackle the problem of pose estimation in crowded scenes. To further evaluate our method in crowded scenes, we report the results on three crowding level in Table 2, *i.e*., uncrowded, medium crowded and extremely crowded. Notably, our method improves **4.1** mAP in uncrowded scenes, while achieves **4.9** and **6.2** mAP higher in medium crowded and extremely crowded scenes separately. This result demonstrates that our method has superior performance in crowded scenes. We present some qualitative results in Figure 7. More results will be given in supplementary file.

MSCOCO We also evaluate our method on MSCOCO dataset to show the generalization ability of our method and results are given in Table 3. Without bells and whistles, our method achieves 70.9 mAP on COCO test-dev set. It brings **0.8** mAP improvements over AlphaPose [27] given the same human detector and SPPE network, which proves that our method can perform general improvement on pose estimation problem. Note that for [40], the results reported in their paper use a strong human detector which is not open-sourced. To make a fair comparison, we report the results using YOLOV3 as human detector.

Inference Speed The runtime speed of fully open-sourced method is tested and shown in Table 2. We obtain the FPS results by averaging the inference time on the test set. To achieve the best performance, we use the most accurate configuration for OpenPose [36], which has an input resolution of 1024×736. As shown in the table, our method achieves 10.1 FPS on the test set, which is slightly slower than AlphaPose [27] but faster than other methods. It proves that our method works the most accurate yet very efficient in crowded cases.

5.5. Ablation Studies

We study different components of our method on CrowdPose test set, as reported in Table 4.

Joints Candidate Loss We first evaluate the effectiveness of our joint-candidate loss. In this experiment, we replace our joint-candidate loss with mean square loss, which is widely used in state-of-the-art pose estimation methods. The experimental result is shown in Table 4(a). The final mAP drops from 66.0% to 61.7%. It proves that our loss function can encourage SPPE to predict more possible joints and resist interference.

Globally Optimizing Association Next, we compare our association algorithm to several NMS algorithms, including bounding box NMS, poseNMS [10, 17] and parametric poseNMS [27]. The experimental results are shown in Table 4(b). We can see that our association algorithm greatly outperforms previous methods. We note that these NMS algorithms are all instance-based. They eliminate redundancy on instance level. Instance-based elimination is not the best solution for pose estimation problem. Because of the complexity of human pose, we need to reduce redundancy on joints level. Meanwhile, all the NMS algorithms are greedy algorithms essentially. Therefore, their results may be not the global optimum. Our association algorithm can give the global optimal result while running as efficient as previous NMS algorithm.

6. Conclusion

In this paper, we propose a novel method to tackle the occlusion problem of pose estimation. By building a person-joint graph based on the outputs of our joint candidates SPPE, we transform the pose estimation problem into a graph matching problem and optimize the results in a global manner. To better evaluate the model performance in crowded scenes, we set up a CrowdPose dataset with a normal distribution of *Crowd Index*. Our proposed method significantly outperforms the state-of-the-art methods in CrowdPose dataset. Experiments on MSCOCO dataset also demonstrate that our method can generalize to different scenarios.

Acknowledgement

This work is supported in part by the National Key R&D Program of China, No. 2017YFA0700800, National Natural Science Foundation of China under Grants 61772332.

References

[1] Giorgio Carpaneto and Paolo Toth. Algorithm for the solution of the assignment problem for sparse matrices. *Computing*, 1983. 5

[2] Walter Whiteley. Some matroids from discrete applied geometry. *Contemporary Mathematics*, 1996. 5

[3] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997. 3

[4] Sam Johnson and Mark Everingham. Clustered pose and nonlinear appearance models for human pose estimation. In *BMVC*, 2010. 2

[5] Jialue Fan, Wei Xu, Ying Wu, and Yihong Gong. Human tracking using convolutional neural networks. *IEEE Transactions on Neural Networks*, 2010. 3

[6] Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (VOC) challenge. In *ICCV*, 2010. 2

[7] Tijmen Tieleman and Geoffrey Hinton. Lecture 6.5-RMSprop: Divide the gradient by a running average of its recent magnitude. *COURSERA: Neural networks for machine learning*, 2012. 7

[8] Irshad Ali and Matthew N. Dailey. Multiple human tracking in high-density crowds. *Image and Vision Computing*, 2012. 3

[9] Matthias Dantone, Juergen Gall, Christian Leistner, and Luc Van Gool. Human pose estimation using body parts dependent joint regressors. In *CVPR*, 2013. 2

[10] Xavier P Burgos-Artizzu, David C Hall, Pietro Perona, and Piotr Dollár. Merging pose estimates across space and time. In *BMVC*, 2013. 5, 8

[11] Hamid Izadinia, Varun Ramakrishna, Kris M. Kitani, and Daniel Huber. Multi-pose multi-target tracking for activity understanding. In *WACV*, 2013. 1

[12] H. Jhuang, J. Gall, S. Zuffi, C. Schmid, and M. J. Black. Towards understanding action recognition. In *ICCV*, 2013. 2

[13] Mykhaylo Andriluka, Leonid Pishchulin, Peter Gehler, and Bernt Schiele. 2D human pose estimation: New benchmark and state of the art analysis. In *CVPR*, 2014. 1, 2, 6

[14] Leonid Pishchulin, Mykhaylo Andriluka, and Bernt Schiele. Fine-grained activity recognition with holistic and pose based features. *GCPR*, 2014. 1

[15] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft COCO: Common objects in context. In *ECCV*, 2014. 1, 2, 4, 6

[16] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster R-CNN: Towards real-time object detection with region proposal networks. In *NIPS*, 2015. 3

[17] Xianjie Chen and Alan L Yuille. Parsing occluded people by flexible compositions. In *CVPR*, 2015. 5, 8

[18] Fangting Xia, Jun Zhu, Peng Wang, and Alan L Yuille. Pose-guided human parsing by an and/or graph using pose-context features. In *AAAI*, 2016. 1

[19] Alejandro Newell, Kaiyu Yang, and Jia Deng. Stacked hourglass networks for human pose estimation. In *ECCV*, 2016. 3

[20] Jiahong Wu, He Zheng, Bo Zhao, Yixin Li, Baoming Yan, Rui Liang, Wenjia Wang, Shipei Zhou, Guosen Lin, Yanwei Fu, et al. AI Challenger: A large-scale dataset for going deeper in image understanding. *arXiv preprint arXiv:1711.06475*, 2017. 1, 2, 6

[21] Yilun Chen, Zhicheng Wang, Yuxiang Peng, Zhiqiang Zhang, Gang Yu, and Jian Sun. Cascaded pyramid network for multi-person pose estimation. *arXiv preprint arXiv:1711.07319*, 2017. 3

[22] Wei Yang, Shuang Li, Wanli Ouyang, Hongsheng Li, and Xiaogang Wang. Learning feature pyramids for human pose estimation. In *ICCV*, 2017. 3

[23] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask R-CNN. In *ICCV*, 2017. 1, 3, 7, 8

[24] Chao Peng, Tete Xiao, Zeming Li, Yuning Jiang, Xiangyu Zhang, Kai Jia, Gang Yu, and Jian Sun. MegDet: A large mini-batch object detector. *arXiv preprint arXiv:1711.07240*, 2017. 3

[25] Yongyi Lu, Cewu Lu, and Chi-Keung Tang. Online video object detection using association lstm. In *The IEEE International Conference on Computer Vision (ICCV)*, Oct 2017. 3

[26] Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Realtime multi-person 2D pose estimation using part affinity fields. In *CVPR*, 2017. 1, 3

[27] Hao-Shu Fang, Shuqin Xie, Yu-Wing Tai, and Cewu Lu. RMPE: Regional multi-person pose estimation. In *ICCV*, 2017. 1, 3, 5, 7, 8

[28] Wenguan Wang, Yuanlu Xu, Jianbing Shen, and Song-Chun Zhu. Attentive fashion grammar network for fashion landmark detection and clothing category classification. In *CVPR*, pages 4271–4280, 2018. 3

[29] Hei Law and Jia Deng. Cornernet: Detecting objects as paired keypoints. In *The European Conference on Computer Vision (ECCV)*, September 2018. 3

[30] Shuai Shao, Zijian Zhao, Boxun Li, Tete Xiao, Gang Yu, Xiangyu Zhang, and Jian Sun. CrowdHuman: A benchmark for detecting human in a crowd. *arXiv preprint arXiv:1805.00123*, 2018. 3

[31] Andrei Zanfir, Elisabeta Marinoiu, Mihai Zanfir, Alin-Ionut Popa, and Cristian Sminchisescu. Deep network for the integrated 3d sensing of multiple people in natural images. In *Advances in Neural Information Processing Systems*, pages 8420–8429, 2018. 1, 3

[32] Bo Pang, Kaiwen Zha, Hanwen Cao, Chen Shi, and Cewu Lu. Deep rnn framework for visual sequential applications. *arXiv preprint arXiv:1811.09961*, 2018. 3

[33] Xiao Sun, Bin Xiao, Fangyin Wei, Shuang Liang, and Yichen Wei. Integral human pose regression. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 529–545, 2018. 1

[34] Hao-Shu Fang, Yuanlu Xu, Wenguan Wang, Xiaobai Liu, and Song-Chun Zhu. Learning pose grammar to encode human body configuration for 3d pose estimation. In *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018. 1

[35] Muhammed Kocabas, Salih Karagoz, and Emre Akbas. MultiPoseNet: Fast multi-person pose estimation using pose residual network. In *ECCV*, 2018. 3

[36] Zhe Cao, Gines Hidalgo, Tomas Simon, Shih-En Wei, and Yaser Sheikh. OpenPose: realtime multi-person 2D pose estimation using Part Affinity Fields, 2018. 7, 8

[37] Hao-Shu Fang, Jinkun Cao, Yu-Wing Tai, and Cewu Lu. Pairwise body-part attention for recognizing human-object interactions. In *The European Conference on Computer Vision (ECCV)*, September 2018. 1

[38] George Papandreou, Tyler Zhu, Liang-Chieh Chen, Spyros Gidaris, Jonathan Tompson, and Kevin Murphy. PersonLab: Person pose estimation and instance segmentation with a bottom-up, part-based, geometric embedding model. *arXiv preprint arXiv:1803.08225*, 2018. 3

[39] Yuliang Xiu, Jiefeng Li, Haoyu Wang, Yinghong Fang, and Cewu Lu. Pose Flow: Efficient online pose tracking. In *BMVC*, 2018. 3

[40] Bin Xiao, Haiping Wu, and Yichen Wei. Simple baselines for human pose estimation and tracking. In *ECCV*, 2018. 1, 3, 7, 8

[41] Yong-Lu Li, Siyuan Zhou, Xijie Huang, Liang Xu, Ze Ma, Hao-Shu Fang, Yan-Feng Wang, and Cewu Lu. Transferable interactiveness prior for human-object interaction detection. *arXiv preprint arXiv:1811.08264*, 2018. 1

[42] Hao-Shu Fang, Guansong Lu, Xiaolin Fang, Jianwen Xie, Yu-Wing Tai, and Cewu Lu. Weakly and semi supervised human body part parsing via pose-guided knowledge transfer. In *CVPR*, 2018. 1

[43] Joseph Redmon and Ali Farhadi. YOLOv3: An incremental improvement. *arXiv preprint arXiv:1804.02767*, 2018. 7