Pose-level Inference for Crowded Scenes Pose Estimation

Anonymous Author(s)
Affiliation
Address
email

Abstract

Multi-person pose estimation in crowded scenes is challenging because many keypoints are invisible due to highly person overlapping. Previous works adopt a keypoint-level inference to detect occluded keypoints, while ignoring the erroneous person detection and prone to be misled by nearby persons. To address these issues, we consider pose as inference objective to better distinguish instances and present a novel Pose-level Inference Network (PINet). Instead of inferring individual keypoint, PINet infers the complete person pose by leveraging both visible person part cues and pose structure prior. It first applies the Part-based Pose Generation (PPG) to infer multiple coarse person poses from different body parts. The coarse pose is hence fed into a Pose Refinement (PR) module, which incorporates pose structure prior and introduces dynamic graph convolution to adaptively aggregate pose features to refine pose. The pose-level inference allows us to address multiple persons simultaneously and is less affected by person overlapping in crowded scenes. Experiments on several crowded scenes pose estimation benchmarks, *i.e.*, OCHuman and CrowdPose demonstrate the superiority of the proposed method.

1 Introduction

Multi-person pose estimation (MPPE) task aims to predict poses for all persons appearing in an image. It performs both person detection and pose keypoint localization. Due to the broad applications in activity understanding, human-object interaction, human parsing, *et al.*, MPPE has attracted increasing attention in recent years. Most MPPE works can be summarized into two categories, *i.e.* i) top-down methods [4, 15, 24, 21], which firstly detect each person and then perform single-person pose estimation (SPPE), and ii) bottom-up methods [1, 17, 11, 14, 8, 6] that first detect each pose keypoint, then assign them into corresponding persons.

Despite the significant progresses achieved with deep convolution neural network (CNN), MPPE is still challenging in crowded scenarios, where occlusion commonly occurs and significantly degrades the accuracy of person detection and keypoint localization. Moreover, for top-down methods, it is hard to get accurate bounding box in crowded scenes to identify different persons. Moreover, bounding boxes cannot distinguish highly overlapped persons and are likely to be suppressed by Non-maximum Suppression (NMS). Bottom-up methods are hard to detect occluded pose keypoints, especially when the same type keypoints are overlapping. And the incomplete keypoints makes it hard to group keypoints into pose. Therefore, existing MPPE methods cannot handle crowded scenes well and directly applying them into crowded scenes will get inferior performance, this is validated in previous work [9, 19].

Researchers have revealed that invisible information could be inferred according to visible information and proper prior knowledge [20, 19, 25]. For instance, to detect occluded person, PRNet [20] first detects person visible parts, then calibrates the anchor with person box ratio prior, which will be used

Submitted to 35th Conference on Neural Information Processing Systems (NeurIPS 2021). Do not distribute.

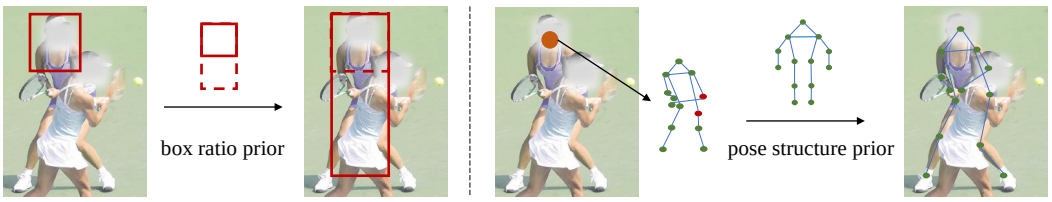

Figure 1: (left) PRNet [20] detect occluded person by firstly detecting visible parts, then use box ratio prior to detect full body box. (right) The proposed Pose-level Inference that progressively infers occluded pose following a similar pipeline.

to detect full body. To detect occluded keypoints, OPEC-Net [19] adopts a keypoint-level inference that utilize a Graph Convolution Network (GCN) [27] to infer occluded keypoints according to visible keypoints and context. However, this method only works on single person and needs bounding box to discriminate persons, thus it also suffers from the above issues. We think that instance discrimination is important to infer occluded keypoints because keypoints may be overlapping in crowded scenes. To address this limitation, we consider the pose (composed by several keypoints of the same person) as inference objective to better distinguish persons and present a novel pose-level inference. Motivated by PRNet, pose-level inference progressively infers the occluded person pose according to visible person parts and pose structure prior, *i.e.*, keypoint connection, which is illustrated in Fig. 1. Inferring pose enjoys several advantages over keypoints. Firstly, we can naturally address multiple persons simultaneously since each pose represents one person, while keypoint cannot. Secondly, pose can be used to better distinguish persons with heavy occlusion than using bounding box [26] and no bounding box or group operation is required for pose-processing, thus we are not affected by above issues in treading MPPE methods and more effective to infer occluded keypoints.

We implement the above idea with the Pose-level Inference Network (PINet) that progressively infer the occluded person pose from coarse to fine. Firstly, we propose Part-based Pose Generation (PPG) that generates coarse person pose from each body parts. Due to the lack of visible part annotation, we divide person into several body parts to ensure that there exists visible parts to generate high quality poses. Secondly, Pose Refinement (PR) adjusts the coarse person pose according to pose structure prior. A dynamic graph convolution is employed to refine pose. Finally, to eliminate low-quality poses inferred from invisible body parts, we further propose Pose Fusion (PF) that select most confident keypoints in each pose to compose the final accurate poses.

We test our method on current crowded scenes pose estimation benckmarks, *i.e.*, OCHuman and CrowdPose. Experiments results show that, our PINet achieves superior performance compared with recent works. For instance, we achieves 58.8% AP on OCHuman, outperforming the recent OPEC-Net [19] and DEKR [3] by 29.7% and 6.8%. On CrowdPose, PINet also achieves 62.0% AP in hard test set (mostly crowded scenes), which is higher than DEKR [3] by 4.5% AP.

In summary, we introduce a novel pose-level inference to deal with crowded scenes pose estimation, which is not affected by the issues in treading MPPE methods. Our contribution can be summarized as follows: i) We firstly propose to estimate occluded pose by inferring with visible information in pose-level. In this way, we can generate complete person pose and handle the missed keypoint and person detection; ii) The Part-based Pose Generation (PPG) and Pose Refinement (PR) are proposed to infer person pose in coarse to fine manner; iii) Our proposed PINet achieves superior performance on current benchmarks, *i.e.*, OCHuman and CrowdPose.

2 Related Work

This work is related with multi-person pose estimation and crowded scenes pose estimation. This section briefly reviews related works in these categories respectively.

Multi-Person Pose Estimation aims to estimate all person pose in an image. Most of related works can be summarized into two categories, *i.e.*, top-down methods and bottom-up methods, respectively. Top-down methods first detect person bounding box by an object detector like yolo, then perform single-person pose estimation in the cropped image or feature map. Mask R-CNN [4] is a typical method that directly adds a keypoint detection branch on Faster R-CNN to use the RoI pooling

feature. G-RMI [15] and the following methods break top-down methods into two steps and use separate models for person detection and pose estimation. SimpleBaseline [24] introduce a strong pose estimation baseline by adding deconvolution to ResNet [5]. Hourglass [12] and HRNet [21] propose new convolution network to generate high quality and high resolution feature map for heatmap estimation, leading to superior performance.

Bottom-up methods first detect identity-free pose keypoints for all the persons in an image and then group them into whole persons. Heatmap is widely adopted for keypoint detection and thus most existing bottom-up methods focus on how to group detected keypoints belonging to the same person. DeepCut [17] is an early work that formulates grouping as an integer linear program, which takes much processing time. OpenPose [1] proposes the part affinity field approach to learn a vector field linking two keypoints. Grouping is done by calculating line integral between two keypoints and group the pair with largest integral. Associative Embedding [11] learns each keypoint with a tag embedding and the tag of the keypoint from the same person are close, thus grouping can be performed by clustering the tag. PersonLab [14] groups keypoints by directly learning a 2D offset field for each pair of keypoints. PifPaf [8] extends OpenPose and PersonLab by using part intensity field to localize body parts and part association field to associate body parts with each other to form full human pose. HGG [6] introduces a differentiable graph clustering to replace traditional grouping operation and obtains superior performance. Several recent works [13, 28] densely regress the person poses from person center, which suffer from the long distance and dense regression problem [13, 22] and center point occlusion.

Crowded Scenes Pose Estimation has rarely been exploited by research community. There are few works to solve multi-person pose estimation when persons are highly overlapping to each other. CrowdPose [9] first construct a large crowded pose estimation benchmark and propose the joint-candidate SPPE with global association to deal with crowded scenes. OPEC-Net [19] refines initial estimated pose to detected invisible keypoint by introducing graph convolution network. However, it only considers keypoint-level inference and cannot handle missed person detection caused by occlusion. In this paper we solve these problems by introducing a novel pose-level inference.

3 Methodology

3.1 Overview

Given an image x with multiple persons in crowded scenes, the goal of crowded scenes pose estimation is to estimate all person pose keypoints, including the visible and invisible keypoints,

$$(\{\mathcal{K}_i^v\}, \{\mathcal{K}_i^{inv}\})_{i=1}^m = \mathrm{MPPE}(x)$$

where $\{\mathcal{K}_i^v\}$ and $\{\mathcal{K}_i^{inv}\}$ denotes the visible and invisible keypoints for i-th person and m represents the number of person in image x. $n = |\{\mathcal{K}_i^v\}| + |\{\mathcal{K}_i^{inv}\}|$ denotes the number of pose keypoints for each person, *i.e.*, $n = 17$ for OCHuman [26] and $n = 14$ for CrowdPose [9].

Unlike previous work that focuses on inferring $\mathcal{K}^{inv}$ with $\mathcal{K}^v$, we propose the PINet that directly infers the complete person pose, *i.e.*, $\mathcal{P} = \{\mathcal{K}^j\}_{j=1}^n$, which naturally contains both visible and invisible keypoints. To achieve this goal, we first adopt the Structure Pose Representation (SPR) [13] that represents a person pose $\mathcal{P}$ as a point and corresponding offsets, which can be denoted as,

$$\mathcal{P}^{SPR} = \{(x, y), (\delta x_j, \delta y_j)_{j=1}^n)\} = \{\mathbf{p}, \mathbf{o}\}, \tag{1}$$

where (x, y) represent the point coordinates and $(\delta x_j, \delta y_j)$ denotes the offset of j-th pose keypoint to the point. Note that the point here is not required to be exactly some pose keypoints, *i.e.*, center, which is different from previous work [13]. Thus inferring a person pose is equivalent to detect the point and regress corresponding offsets.

As stated above, PINet progressively infers the occluded person pose according to visible person parts and pose structure prior. Ideally, the coarse pose can be generated by detecting the visible person parts and regressing offsets. Because there is no available visible part annotation in existing pose estimation benchmarks, we divide person into several body parts and propose the Part-based Pose Generation (PPG) that detects each body parts and regresses corresponding pose offsets. Intuitively, the finer the body partition, the more likely PPG exploits visible parts. Suppose we divide person into s parts, then PPG detects s points and offsets for each person in input image x,

$$\{\{\mathbf{p}_l^i, \mathbf{o}_l^i\}_{l=1}^s\}_{i=1}^m = \mathrm{PPG}(\Phi(x)) \tag{2}$$

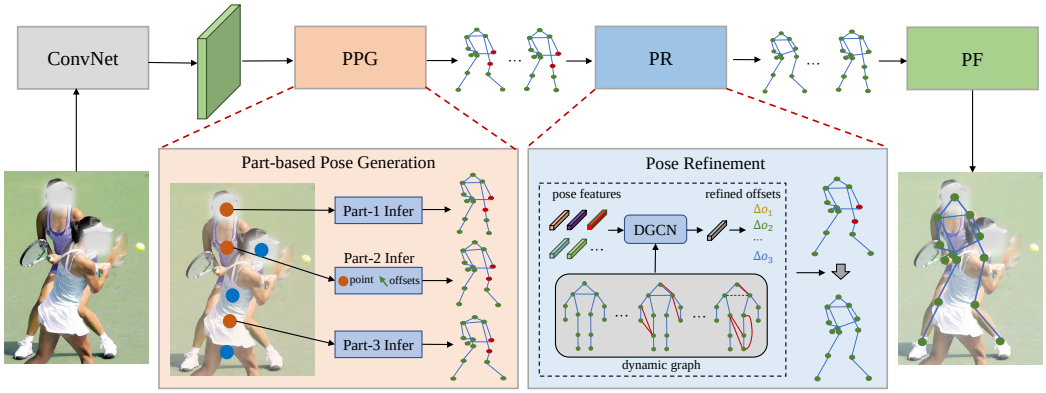

Figure 2: Framework of the proposed Pose-level Inference Network (PINet). We use a CNN to extract the high resolution feature map. Part-based Pose Generation (PPG) infers the whole person pose from different body parts, as denoted by the orange and blue circle. Pose Refinement (PR) refines the coarse person pose with dynamic pose keypoint graph. PF denotes the pose fusion operation that fuses the poses from different parts and output the final pose. The green and red point denotes the visible and invisible pose keypoints. We only show the pose of occluded person. Best viewed in color.

where $\mathbf{p}_l, \mathbf{o}_l$ denote the point and offsets from l-th body parts, $\Phi(\cdot)$ is the backbone. The coarse pose can be obtained by $\mathcal{P}_l^1 = \mathbf{p}_l + \mathbf{o}_l$.

The coarse poses $\mathcal{P}_l$ inferred by PPG can be further refined by Pose Refinement (PR) module with pose structure prior G and global feature map $\mathbf{F} = \Phi(x)$,

$$\mathcal{P}_l^2 = \mathrm{PR}(\mathcal{P}_l^1, \mathbf{F}, G), \tag{3}$$

where $\mathcal{P}_l^2$ denotes the pose refined by PR.

PPG may generate low-quality poses from invisible body parts, which could degrade the performance. To eliminate these low-quality poses, we propose a post-processing operation Pose Fusion (PF) that filter low-quality poses and fuse duplicate poses and output the fused high-quality pose,

$$\{\mathcal{P}_i^3\}_{i=1}^m = \mathrm{PF}(\{\mathcal{P}_i^2\}_{i=1}^t), \tag{4}$$

where $\mathcal{P}_i^3$ denotes the 3rd-phase fused pose, m denotes the final number of estimated pose and t is the number of poses generated by PPG and PR.

Since PF is a pose-processing operation, we learn PINet by training the backbone, PPG and PR module in an end-to-end manner, the overall objective is denoted by,

$$\mathcal{L} = \mathcal{L}_{PPG} + \lambda \mathcal{L}_{PR}, \tag{5}$$

where $\mathcal{L}_{PPG}$ and $\mathcal{L}_{PR}$ denote the loss for two modules and λ is a weight to balance two losses. Fig. 2 illustrates the pipeline of proposed PINet. With the PPG, PR and PF, we can infer high quality person pose even when they are highly overlapping, and are free of missed keypoint and person detection in previous methods. In the following, we will present details of PPG, PR and PF.

3.2 Part-based Pose Generation

As stated above, we divide each person into s body parts according to body skeleton to exploit the visible information. Generally speaking, the finer the body partition, the more likely there exists visible parts. However, small part is hard to regress full person pose offset due to the lack of global person information. Moreover, large s also increases the challenge of dense regression optimization, which is discussed in [22].

Inspired by previous work [10, 18] on pose keypoint and body parts, we design several body partition strategies and show them in Fig. 3. The partition is listed from coarse to fine. For instance, *center* denotes that we divide person into one part, while *keypoint* indicates that a keypoint and its neighbor represents a body part. We also show the occlusion ratio (a partition without any visible parts is

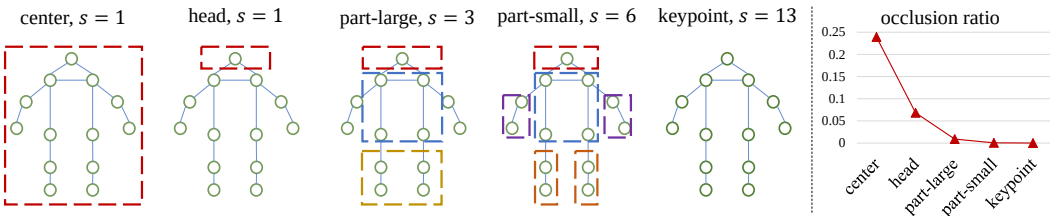

Figure 3: Illustration of different strategies of body partition of PPG: (a) Five different partition strategies, rectangles denote the body parts, and (b) their corresponding occlusion statistics from OCHuman dataset.

regard as occluded) on OCHuman. As shown in Fig. 3, the finer the body partition, the less the occlusion ratio. However, it is also harder for model optimization, which is verified in Sec.4. We empirically find that part-large partition strategy achieves a good tradeoff of these two factors, thus we divide each person into $s = 3$ parts, *i.e.*, upper, middle and bottom parts.

According to above analysis, PPG detects upper, middle and bottom parts for each person and regresses corresponding offsets. Following keypoint-based detection work [28], we detect each part by estimate the part center with heatmap $\mathbf{P}$. Specifically, given the feature map $\mathbf{F} = \Phi(x)$ from backbone, we pass it through the keypoint detection head and offset regression head,

$$\mathbf{P}_l = \mathcal{F}_l(\mathbf{F}), \mathbf{O}_l = \mathcal{R}_l(\mathbf{F}), l \in \{u, m, b\} \tag{6}$$

where $\mathbf{P}_l, \mathbf{O}_l$ represent the point map that indicating the confidence of l-th part and offset maps for corresponding offset, respectively. u, m, b denote the three parts. We adopt the tradeoff L2 regression loss to learn the pointmap and smooth L1 loss to learn the offset map,

$$\mathcal{L}_{point} = ||\mathbf{M} \odot (\mathbf{P} - \mathbf{P}^*)||_2^2, \tag{7}$$

$$\mathcal{L}_{offset} = \frac{1}{|\Omega|} \sum_{i \in \Omega} \frac{1}{S_i} \text{smooth}_{L1}(\mathbf{o}_i - \mathbf{o}_i^*), \tag{8}$$

where $\mathbf{P} = [\mathbf{P}_u; \mathbf{P}_m; \mathbf{P}_b]$ denotes the point maps, $\mathbf{M}$ is a mask to down weight background region, the weight of positions not lying in point or pose keypoint region is 0.1, and others are 1. Ω denotes the index set of part center and its neighbor points on feature map, S_i is the area of i-th person. We generate the ground truth following [13]. The final loss of PPG is the weight sum of above two losses, *i.e.*, $\mathcal{L}_{PPG} = \mathcal{L}_{point} + \lambda_o \mathcal{L}_{offset}$. λ_o is a weight to balance to losses.

To get the detected points and offsets, we first apply the point-NMS to get the local maximum point $\mathbf{p}_l$ in $\mathbf{P}_l$ and sample the corresponding offset map $\mathbf{O}_l$ to get $\mathbf{o}_l$. The part-based design enhances the robustness of PINet to occlusion, which will be verified in Sec. 4.3.

3.3 Pose Refinement

Although PPG can generate complete person pose, it still confronts several issues. Firstly, the part-based generation strategy enhance the robustness of PINet to occlusion, but eliminate the person global information for each part to infer the whole person pose. *i.e.*, the upper body part feature is hard to encode foot due to limited receptive field. Secondly, the occluded pose keypoint $\mathcal{K}^{inv}$ should be inferred by considering keypoint connections, namely pose structure, not only visual information.

Thus we are motivated to propose a Pose Refinement (PR) module to address above issues. For pose $\mathcal{P}_i^1 = \{\mathcal{K}^j\}_{j=1}^n$ generated by point feature f_p in PGG, we supply it with keypoint features f_j to encode the person global information, which can be obtained by bilinear sampling on the feature map $\mathbf{F}$ with keypoint location $\mathcal{K}^j$. To fully utilize pose structure pose prior, we propose a dynamic graph convolution network (DGCN) that considers keypoint connection information to refine and aggregate these keypoint features,

$$X^{l+1} = \sigma(GX^lW), \tag{9}$$

where W is the parameters of DGCN, σ denotes ReLU, X^l and X^{l+1} is the original and updated features, and G is the dynamic graph, which is initialized according to the human body skeleton and

updated as the model parameters. Each row of the initial feature matrix X^0 represents one keypoint. Following [23], we apply max pooling and average pooling to the refined keypoint features to obtain the global features of the pose, which is concatenate with f_p and pass to two fully connected layers to predict the offset $\Delta\mathcal{P}_i^1$ of initial pose $\mathcal{P}_i^1$ and the ground truth. Thus we can obtain the refined pose by $\mathcal{P}_i^2 = \mathcal{P}_i^1 + \Delta\mathcal{P}_i^1$. Similar to the offsetmap loss, we calculate the refinement loss by,

$$\mathcal{L}_{PR} = \frac{1}{|\mathcal{T}|} \sum_{\mathcal{P}_i^1 \in \mathcal{T}} \frac{1}{S_i} \text{smooth}_{L1}(\mathcal{P}_i^1 + \Delta\mathcal{P}_i^1, \mathcal{P}_i^*), \tag{10}$$

where $\mathcal{T}$ is the sampled pose set for PR refinement, S_i is the area of i-th person.

Discussions. Previous OPEC-Net [19] also utilize GCN. However, PR is different from it in several aspects. Firstly, OPEC-Net focus on inferring occluded keypoints and takes keypoint coordinates as input, thus require extra feature adaptation module to extract context information. Our PR focus on extract better person global information with pose structure prior and does not require extra module. Secondly, PR adopts a dynamic graph to refine keypoint features and can capture more keypoint relationship. For example, the skeleton contains no connections between left and right body. However, when the right ankle is occluded, we could refer to left ankle and knee for inference. The dynamic graph is more flexible and can capture these properties with model learning. Fig. 2 illustrates an example of the dynamic graph during training.

3.4 Pose Fusion

PPG may generates low-quality poses from invisible person body parts and duplicate poses, thus a pose-processing operation is required to eliminate them. However, we notice that duplicate poses generated from different body parts are useful, as they perform well on different keypoints due to long distance regression [13, 28]. A pose regressed from upper body parts performs well on nose and eye keypoints, while badly on ankle and knee. The pose regressed from the bottom body part may present a opponent result.

According above analysis, we propose a PF operation that fuse duplicate poses and remove low-quality poses to get a better results. Given the poses inferred by PPG and PR, *i.e.*, $\{\mathcal{P}_i^2\}_{i=1}^t$, PF firsts find the pose set that belong to the same person. Inspired by OKS-NMS [15], we adopt a OKS based matching strategy to identify the poses belonging to the same person. For two poses $\mathcal{P}_1$ and $\mathcal{P}_2$, we calculate their object keypoint similarity by,

$$OKS(\mathcal{P}_1, \mathcal{P}_2) = \frac{1}{n} \sum_{i=1}^{n} \exp(-\frac{||\mathcal{K}_1^i - \mathcal{K}_2^i||_2^2}{(S_1 + S_2)s_i^2}) \tag{11}$$

where $\mathcal{K}^i$ denotes the i-th keypoint, S_1 and S_2 represent the area of two poses, s_i is a per-keypoint constant that controls falloff. If the OKS of two poses is larger than t, then we think they are belong to the same person. According to above match strategy and NMS procedure, we can divide $\{\mathcal{P}_i^2\}_{i=1}^t$ into m group, each group $\{\mathcal{P}_i^2\}_{i=1}^{m_k}$ represent one person pose set. Then PF traverse each pose set $\{\mathcal{P}_i^2\}_{i=1}^{m_k}$ and select n keypoints with maximum score to construct the final person pose $\mathcal{P}_i^3$.

The keypoint score is formed by $s = s_p \cdot s_k$, where s_p is the score of generated pose from $\mathbf{P}$, s_k is the keypoint confidence score which can be easily obtained by learning an extra score map $\mathbf{H}$ for each pose keypoints, *i.e.*, the heatmap in most MPPE methods. A low-quality poses generated by PPG have small s_p, thus will not appear in the final results. Moreover, s_k will up weight the well estimated keypoints and relieve the long regression problem. In implementation we learn the score map with the point map $\mathbf{P}$ and change Eq. 6 to,

$$\mathcal{L}_{point} = ||\mathbf{M} \odot ([\mathbf{P}; \mathbf{H}] - [\mathbf{P}^*; \mathbf{H}^*])||_2^2, \tag{12}$$

where $[\cdot;\cdot]$ denotes concatenation operation.

Discussions. PF is a model-agnostic operation and can also be applied to other regression methods, *i.e.*, [13]. However, it is more effective in our method when duplicate predictions are from different parts, not merely center points. We also validate the effectiveness of PF on other methods in Sec.4.

Table 1: Ablation studies of our proposed PINet on individual components.

(a) Effectiveness of each component in PINet.

PPG	PR	PF	AP	AP^{50}	AP^{75}
			47.6	67.5	53.2
✓			53.1	74.6	60.4
✓	✓		57.9	75.7	65.0
✓		✓	55.8	75.0	62.3
✓	✓	✓	**58.8**	**75.5**	**65.4**

(b) Test different body partition strategies in PPG.

Part partition	AP	AP^{50}	AP^{75}
Center	47.6	67.5	53.2
Head	49.8	70.9	55.6
Part-large	**58.8**	**75.5**	**65.4**
Part-small	50.3	69.4	57.2
Keypoint	40.8	59.0	45.7

(c) Test different refine strategies in PR.

Method	AP	AP^{50}	AP^{75}
No GCN	56.8	74.1	63.0
Static graph	57.7	74.7	64.3
Dynamic graph	**58.8**	**75.5**	**65.4**

(d) Test PF in different methods.

Method	w/o PF	w/ PF
SPM [13]	47.6	48.1
DEKR [3]	52.2	53.0
PINet(w/o PR)	53.1	55.8

4 Experiments

4.1 Datasets and Evaluation Protocol

OCHuman [26] is a recently proposed benchmark to examine the limitations of MPPE in highly challenging scenarios. It consists of 4731 images in total, including 2500 images for validation and 2231 images for testing. The dataset contains only challenging cases of occlusion and the average IoU of the bounding boxes is 67%. Following [19], we train models on `val` set and report the performance on `test` set.

CrowdPose [9] is another dataset that includes many crowded scenes. It contains 10K, 2K and 8K images for `train`, `val` and `test` set. Following [2, 3], we train our models on the `train` and `val` sets and report the results on the `test` set. Since CrowdPose contains both crowded and non-crowded scenes, we also report the performance on crowded test set, namely metric AP^H.

Evaluation Metrics. We follow the standard evaluation metric and use OKS-based metrics for pose estimation. We report average precision scores with different thresholds: AP, AP^{50}, and AP^{75}. On CrowdPose, we also report AP^E, AP^M and AP^H, which refers to the performance under different crowd index.

4.2 Implementation Details

All experiments are implemented on PyTorch [16]. We adopt HRNet-W32 [21] as backbone for all experiments and follow the most configuration of [2, 3]. We set λ and λ_o to 0.03 and t to 0.7 for all experiments. For PR, we sample 120 poses from PPG, each part contains 40 poses to refine.

Training. The input image is resized to 512*512. We use Adam [7] to optimize the model, the learning rate is set to 0.001 for all layers. We train the model for 140 epochs on OCHuman, with learning rate dividing by 10 at 90th, 120th epoch. For CrowdPose, we train model with 300 epochs and divide learning rate by 10 at 200th, 260th epoch. The batch size is set to 20 OCHuman (2 GPUs) and 40 for CrowdPose (4 GPUs). We adopt data augmentation strategies including random rotation (-30,30), scale ([0.75,1.5]), translation ([-40,40]) and flipping (0.5).

Testing. We resize the short side of the images to 512 and keep the aspect ratio between height and width. We adopt single and multi-scale ([0.5,1.0,2.0]) test with flipping following [2, 3].

Comparison methods. To make fair comparison, we report the performance of SPM [28](with SPR), HigherHRNet [2] and DEKR [3] in the same setting, *i.e.*, backbone, training schedule and test setting.

4.3 Ablation Study

Components analysis. We analyze the effectiveness of each proposed components on OCHuman. The results are shown in Table 1 (a). PINet with only PPG achieves 53.1% AP, which is already higher

Table 2: Comparison with state-of-the-art methods on OCHuman `test` set. $*$ denotes our implementation.

Method	Reference	AP	AP^{50}	AP^{75}
Mask RCNN [4]	ICCV2017	20.2	33.2	24.5
SimpleBaseline [24]	ECCV2018	24.1	37.4	26.8
SPPE+ [9]	CVPR2019	27.6	40.8	29.9
OPEC-Net [19]	ECCV2020	29.1	41.3	31.4
SPM$*$ [13]	ICCV2019	45.6	66.1	50.4
HigherHRNet$*$ [2]	CVPR2020	27.7	66.9	15.9
DEKR$*$ [3]	CVPR2021	52.2	69.9	56.6
Ours	this paper	**58.8**	**75.5**	**65.4**

Table 3: GFLOPs and #parameters of the representative top competitors and our approaches. AE-HG denotes associative embedding-Hourglass.

	AE-HG [11]	PersonLab [14]	HigherHRNet [2]	DEKR [3]	PINet
Input size	512	1401	640	512	512
#param.(M)	227.8	68.7	63.8	29.6	29.4
GFLOPs	206.9	405.5	154.3	45.4	42.4

than previous center-based regression work. This indicates that inferring complete pose from visible part is important for occluded pose estimation. PR incorporate pose structure prior to refine the coarse, it remarkably enhances the AP by 4.8%. We also evaluate the propose pose fusion operation, and it boosts the performance by 0.9%, which validates the effectiveness of fusing duplicate predictions.

Analysis on PPG. An important factor of PPG is the choice of body partition strategy. We compare different partition strategies in Fig. 3. Table 1 (b) summarizes the results. We can observe that center part achieves 47.6% AP, while head part obtains 49.8% AP. This is because that center part is usually occluded by other person, thus PPG cannot generate high quality pose. While head and face are discriminative regions, thus is more likely to be visible. The proposed part-large partition achieves the best performance, obtaining 58.8% AP, outperforming center and head points by 8.3% and 6.1% AP, respectively. We also investigate finer point selection, *i.e.*, part-6 and keypoint.However, they work badly due to the hard optimization with dense regression.

Analysis on PR. We also analyze the component of LPR. As shown in Table 1 (c). We first test to directly refine estimated pose by PPG without incorporating pose prior, and it obtains 56.8% AP. Incorporating pose keypoint graph to refine feature achieves 57.7% AP. It indicates that pose structure prior is useful for occluded pose keypoint inference. Moreover, when adding dynamic graph into PR, we can further boost AP to 58.8% AP, which verify the effectiveness of proposed methods.

Analysis on PF. To verify that PF is a model-agnostic operation, we apply it to other regression-based methods, *i.e.*, SPM and DEKR. As shown in Table 1 (d), PF can constantly boost these methods about 0.6 0.8 AP, which demonstrates the effectiveness of PF on regression methods. Moreover, PF is more useful when we regress pose from different parts, and it boost the performance of PINet from 53.1 to 55.8. Note that PR can also relieve the regression problem, thus we omit it here to verify PF.

4.4 Comparison with Recent Works

We compare PINet with recent works on OCHuman [26] and CrowdPose [9]. The results are shown in Table 2 and 4.

Table 2 reports comparisons on OCHuman. We compare top-down methods: Mask RCNN [4], SimpleBaseline [24], SPPE+ [9] and OPEC-Net [19], and bottom-up methods OpenPose [1], SPM [13], HigherHRNet [2] and recent DEKR [3]. On OCHuman, top-down methods achieve relatively low AP. This is due to the issues mentioned in Sec 1. Our PINet makes a pose-level inference, thus achieves the best 58.8% AP on OCHuman in single-scale test, which is higher than all bottom-up methods

Table 4: Comparison with state-of-the-art methods on CrowdPose `test` set. $*$ denotes our implementation. † denotes multi-scale test.

Method	Reference	AP	AP^{50}	AP^{75}	AP^{E}	AP^{M}	AP^{H}
Mask RCNN [4]	ICCV2017	57.2	83.5	60.3	69.4	57.9	45.8
SimpleBaseline [24]	ECCV2018	60.8	81.4	65.7	71.4	61.2	51.2
SPPE+ [9]	CVPR2019	66.0	84.2	71.5	75.5	66.3	57.4
OpenPose [1]	CVPR2017	-	-	-	62.7	48.7	32.3
SPM$*$ [13]	ICCV2019	63.7	85.9	68.7	70.3	64.5	55.7
HigherHRNet$*$ [2]	CVPR2020	62.2	83.7	66.4	69.9	62.7	54.3
HigherHRNet$*$† [2]	CVPR2020	64.2	85.4	68.5	73.2	64.6	55.4
DEKR [3]	CVPR2021	65.7	85.7	70.4	73.0	66.4	57.5
DEKR† [3]	CVPR2021	67.0	85.4	72.4	75.5	68.0	56.9
Ours	this paper	66.4	87.9	71.9	72.9	66.9	59.4
Ours†	this paper	**69.9**	**89.7**	**75.7**	**76.8**	**70.6**	**62.0**

under the same setting. For instance, DEKR obtains 52.2% AP, which is lower than PINet by 6.6% AP.

On CrowdPose dataset, our method also obtains superior performance. As shown in Table 4, PINet achieves 66.4% and 69.9% AP in single-scale and multi-scale test, which is higher than HigherHRNet and DEKR. Moreover, on crowded scenes, PINet achieves 59.4% AP^{H}, which is higher than DEKR by 1.9%. Results on two occluded datasets demonstrate that PINet is effective to deal with crowded scenes pose estimation.

To further investigate the robustness of our method to crowded scenes, we follow previous work [19] to split datasets according to the average IoU of bounding box. Since there are few or no images with IoU smaller than 0.5 on OCHuman, we start from 0.5 and divide the `test` set of OCHuman and CrowdPose into 5 parts and report the performance at each IoU. As shown in Fig. 4, our method constantly outperforms DEKR and SPM, especially at high IoU. We also report the parameters and GFLOPs of existing bottom-up methods, as shown in Table 3. Our method requires less or similar parameters than other methods.

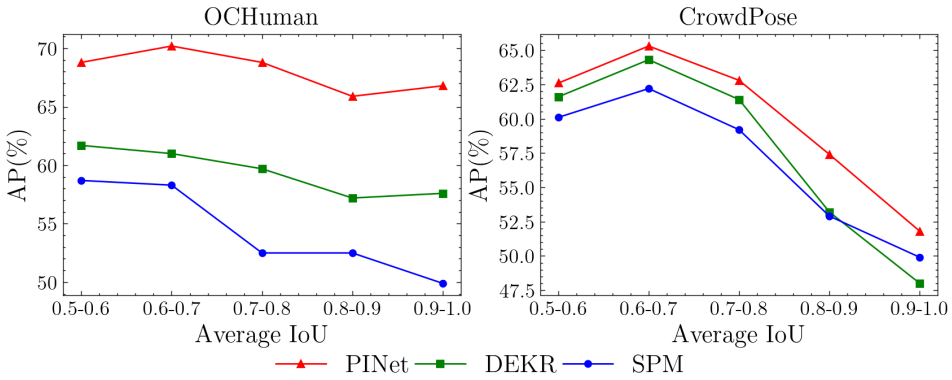

Figure 4: Comparison with DEKR, SPM and our proposed PINet under different IoU threshold. Our method constantly outperforms other methods, especially at highly crowded scenes (high IoU).

5 Discussion and Conclusion

Our proposed PINet has shown considerable improvements in crowded scenes pose estimation. This is achieved by the novel pose-level inference that infers pose by leveraging both body part cues and pose structure prior. However, the performance of crowded scenes is still lower than non-crowded scenes, and PINet directly regresses the pose, which is not as accurate as the heatmap-based methods. Further studies are required to address these problems.

References

[1] Z. Cao, T. Simon, S.-E. Wei, and Y. Sheikh. Realtime multi-person 2d pose estimation using part affinity fields. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7291–7299, 2017.

[2] B. Cheng, B. Xiao, J. Wang, H. Shi, T. S. Huang, and L. Zhang. Higherhrnet: Scale-aware representation learning for bottom-up human pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5386–5395, 2020.

[3] Z. Geng, K. Sun, B. Xiao, Z. Zhang, and J. Wang. Bottom-up human pose estimation via disentangled keypoint regression. *arXiv preprint arXiv:2104.02300*, 2021.

[4] K. He, G. Gkioxari, P. Dollár, and R. Girshick. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969, 2017.

[5] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[6] S. Jin, W. Liu, E. Xie, W. Wang, C. Qian, W. Ouyang, and P. Luo. Differentiable hierarchical graph grouping for multi-person pose estimation. In *Computer Vision – ECCV 2020*, pages 718–734, 2020.

[7] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[8] S. Kreiss, L. Bertoni, and A. Alahi. Pifpaf: Composite fields for human pose estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 11977–11986, 2019.

[9] J. Li, C. Wang, H. Zhu, Y. Mao, H.-S. Fang, and C. Lu. Crowdpose: Efficient crowded scenes pose estimation and a new benchmark. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 10863–10872, 2019.

[10] J. Li, S. Zhang, Q. Tian, M. Wang, and W. Gao. Pose-guided representation learning for person re-identification. *IEEE transactions on pattern analysis and machine intelligence*, 2019.

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

[12] A. Newell, K. Yang, and J. Deng. Stacked hourglass networks for human pose estimation. In *European conference on computer vision*, pages 483–499. Springer, 2016.

[13] X. Nie, J. Feng, J. Zhang, and S. Yan. Single-stage multi-person pose machines. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 6951–6960, 2019.

[14] G. Papandreou, T. Zhu, L.-C. Chen, S. Gidaris, J. Tompson, and K. Murphy. Personlab: Person pose estimation and instance segmentation with a bottom-up, part-based, geometric embedding model. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 269–286, 2018.

[15] G. Papandreou, T. Zhu, N. Kanazawa, A. Toshev, J. Tompson, C. Bregler, and K. Murphy. Towards accurate multi-person pose estimation in the wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4903–4911, 2017.

[16] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. In *Advances in neural information processing systems*, pages 8026–8037, 2019.

[17] L. Pishchulin, E. Insafutdinov, S. Tang, B. Andres, M. Andriluka, P. V. Gehler, and B. Schiele. Deepcut: Joint subset partition and labeling for multi person pose estimation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4929–4937, 2016.

[18] R. Pytel, O. S. Kayhan, and J. C. van Gemert. Tilting at windmills: Data augmentation for deep pose estimation does not help with occlusions. *arXiv preprint arXiv:2010.10451*, 2020.

[19] L. Qiu, X. Zhang, Y. Li, G. Li, X. Wu, Z. Xiong, X. Han, and S. Cui. Peeking into occluded joints: A novel framework for crowd pose estimation. In *Computer Vision – ECCV 2020*, pages 488–504, 2020.

[20] X. Song, K. Zhao, W.-S. C. H. Zhang, and J. Guo. Progressive refinement network for occluded pedestrian detection. In *Proc. European Conference on Computer Vision*, volume 7, page 9, 2020.

[21] K. Sun, B. Xiao, D. Liu, and J. Wang. Deep high-resolution representation learning for human pose estimation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5693–5703, 2019.

[22] A. Varamesh and T. Tuytelaars. Mixture dense regression for object detection and human pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13086–13095, 2020.

[23] Y. Wang, Y. Sun, Z. Liu, S. E. Sarma, M. M. Bronstein, and J. M. Solomon. Dynamic graph cnn for learning on point clouds. *Acm Transactions On Graphics (tog)*, 38(5):1–12, 2019.

[24] B. Xiao, H. Wu, and Y. Wei. Simple baselines for human pose estimation and tracking. In *Proceedings of the European conference on computer vision (ECCV)*, pages 466–481, 2018.

[25] S. Zhang, L. Wen, X. Bian, Z. Lei, and S. Z. Li. Occlusion-aware r-cnn: detecting pedestrians in a crowd. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 637–653, 2018.

[26] S.-H. Zhang, R. Li, X. Dong, P. Rosin, Z. Cai, X. Han, D. Yang, H. Huang, and S.-M. Hu. Pose2seg: Detection free human instance segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 889–898, 2019.

[27] L. Zhao, X. Peng, Y. Tian, M. Kapadia, and D. N. Metaxas. Semantic graph convolutional networks for 3d human pose regression. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3425–3435, 2019.

[28] X. Zhou, D. Wang, and P. Krähenbühl. Objects as points. *arXiv preprint arXiv:1904.07850*, 2019.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes] This paper proposes a novel pose-level inference that can effectively estimated person pose in crowded scenes, which is states in Section 1.
   (b) Did you describe the limitations of your work? [Yes] We describe this in Section 5.
   (c) Did you discuss any potential negative societal impacts of your work? [No] We think our work has no negative societal impact as it does not infringe of people's privacy.
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes] We have read ethics review and our paper conforms to them.
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [N/A]
   (b) Did you include complete proofs of all theoretical results? [N/A]
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [No] We will release the code later.
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes] The datasets and training details are listed in Section 4.1 and 4.2.
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [No] The results is an average results of several experiments. We did not report the error bars because this is not reported in any previous pose estimation work.
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes] Please refer to Section 4.2.
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [Yes] We have cited all of them.
   (b) Did you mention the license of the assets? [No]

(c) Did you include any new assets either in the supplemental material or as a URL? [No] There are no new assets.

(d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [No] All data are free for academic research.

(e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [No] The data is widely reviewed by research committee, there is no personally identifiable information or offensive content.

5. If you used crowdsourcing or conducted research with human subjects...

(a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]

(b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]

(c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]