DecenterNet: Bottom-Up Human Pose Estimation Via Decentralized Pose Representation

Tao Wang*
wangtao@bupt.edu.cn
Beijing University of Posts and Telecommunications
Beijing, China

Lei Jin*†
jinlei@bupt.edu.cn
Beijing University of Posts and Telecommunications
Beijing, China

Zhang Wang
wz2021110383@bupt.edu.cn
Beijing University of Posts and Telecommunications
Beijing, China

Xiaojin Fan
fanxiaojin@bit.edu.cn
Beijing Institute of Technology
Beijing, China

Yu Cheng
e0321276@u.nus.edu
National University of Singapore
Singapore, Singapore

Yinglei Teng
lilytengtt@bupt.edu.cn
Beijing University of Posts and Telecommunications
Beijing, China

Junliang Xing
jlxing@tsinghua.edu.cn
Tsinghua University
Beijing, China

Jian Zhao†
zhaojian90@u.nus.edu
Institute of North Electronic Equipment
Beijing; Intelligent Game and Decision Laboratory
Beijing; Peng Cheng Laboratory
Shenzhen;, China

ABSTRACT

Multi-person pose estimation in crowded scenes remains a very challenging task. This paper finds that most previous methods fail to estimate or group visible keypoints in crowded scenes rather than reasoning invisible keypoints. We thus categorize the crowded scenes into entanglement and occlusion based on the visibility of human parts and observe that entanglement is a significant problem in crowded scenes. With this observation, we propose DecenterNet, an end-to-end deep architecture to perform robust and efficient pose estimation in crowded scenes. Within DecenterNet, we introduce a decentralized pose representation that uses all visible keypoints as the root points to represent human poses, which is more robust in the entanglement area. We also propose a decoupled pose assessment mechanism, which introduces a location map to adaptively select optimal poses in the offset map. In addition, we have constructed a new dataset named *SkatingPose*, containing more entangled scenes. The proposed DecenterNet surpasses the best method on SkatingPose by 1.8 AP. Furthermore, DecenterNet obtains 71.2 AP and 71.4 AP on the COCO and CrowdPose datasets, respectively, demonstrating the superiority of our method. We will release our source code, trained models, and dataset to facilitate further studies in this research direction. Our code and dataset are available in https://github.com/InvertedForest/DecenterNet.

CCS CONCEPTS

• **Computing methodologies** → **Computer vision tasks**; **Computer vision representations**; **Computer vision problems**.

KEYWORDS

neural networks, human pose estimation, single-stage, datasets

ACM Reference Format:
Tao Wang, Lei Jin, Zhang Wang, Xiaojin Fan, Yu Cheng, Yinglei Teng, Junliang Xing, and Jian Zhao. 2023. DecenterNet: Bottom-Up Human Pose Estimation Via Decentralized Pose Representation. In *Proceedings of the 31st ACM International Conference on Multimedia (MM '23), October 29-November 3, 2023, Ottawa, ON, Canada.* ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3581783.3611989

1 INTRODUCTION

2D Multi-Person Pose Estimation (MPPE) aims to detect all persons and locate their 2D keypoints in a given image. Current MPPE methods can be roughly divided into two categories: i) top-down methods [5, 10, 14, 21, 35, 39], which first detect the bounding boxes of all persons in the image with object detectors, and then perform pose estimation for each person. ii) bottom-up methods [7, 8, 11, 28, 44] that sequentially or simultaneously detect all keypoints in the image, and group them to corresponding persons.

As current methods are achieving satisfactory performance in simple scenes, some works begin to pay attention to human pose

*Both authors contributed equally to this research. † is corresponding author.

MM '23, October 29-November 3, 2023, Ottawa, ON, Canada

ACM ISBN 979-8-4007-0108-5/23/10…$15.00
https://doi.org/10.1145/3581783.3611989

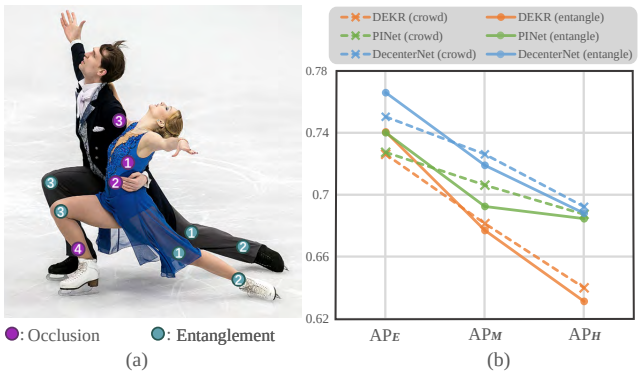

Figure 1: (a) Schematic diagram of occlusion and entanglement. The cyan circles with the same number indicate that the model is more prone to confuse these keypoints of both athletes, leading to wrong positioning or grouping in the entanglement area. The purple circles represent the occluded keypoints of the male, where manual labeling errors often occur, and there is no direct basis for the model to judge. We only label parts of keypoints for clearness. (b) Comparison of AP scores based on Crowd and Entanglement Index. For these typical works, *i.e.*, PINet [37], DEKR [11], and our method DecenterNet, AP_H based on the Entanglement Index are lower than these based on the Crowd Index, indicating that entanglement area is the dominant factor for crowded scenes.

estimation in crowded scenes. CrowdPose [20] has collected a large number of daily images of crowded scenes with human pose annotations to better evaluate algorithms in crowded scenes. Representative works [22, 32, 34, 38, 45] have designed innovative modules for crowded scenes and achieved commendable performance. However, it is observed that previous works often fail to estimate the locations of visible keypoints or group them into one person, rather than reasoning about the locations of invisible keypoints. We arrived at this conclusion by comparing the number of human pose estimation failures caused by visible and invisible keypoints with classic methods [7, 37, 46] on the COCO dataset [24], as shown in Tab. 1. First, we consider the predicted pose to have failed when its OKS (Object Keypoint Similarity) with ground-truth is smaller than 0.5. We then count the number of unmatched predicted poses caused by visible and invisible keypoints, respectively. It is obvious that most of the networks' estimation failures are caused by the prediction errors in visible keypoints in Tab. 1. Therefore, it is more efficient for human pose estimation task to focus on predicting visible keypoints rather than reasoning about invisible keypoints.

To this end, we further categorize the crowded scenes into entanglement and occlusion, as shown in Fig. 1 (a). *Entanglement* refers to the visible part in the crowded area (the overlapping part of two instance bounding boxes), where the model generally incorrectly locates the keypoints or groups the keypoints to the wrong human instance. *Occlusion* is the invisible part, where there are no visual clues to determine the keypoints, and the model can only rely on the context to locate these keypoints. However, popular dataset for crowded scenes, CrowdPose, does not provide visibility flags. Thus, we construct a new figure skating dataset with visibility flags named SkatingPose . We further define a new Entanglement Index

Table 1: The number of matching failures caused by visible keypoints and invisible keypoints.

	Failures (vis)	Failures (invis)	Rate (vis / invis)
HrHRNet [7]	2193	344	**86.4% / 13.6%**
LOGO-CAP [46]	843	90	**90.4% / 9.6%**
PINet [37]	579	64	**90.1% / 9.9%**

(see detailed definition in Sec. 4.2) to measure the entangled level of an image. Compared with the Crowd Index defined in CrowdPose [20], the Entanglement Index can better indicate the main factors of inaccurate pose estimation in crowded scenes as shown in Fig. 1 (b). In more details, we first split our SkatingPose val set into three parts: easy, medium, and hard based on Entanglement Index and Crowd Index. Then we reproduce several representative works [11, 37] and calculate the corresponding AP_E, AP_M, and AP_H. The evaluation shows that the AP scores from Entanglement Index has a steeper downtrend than Crowd Index, which reveals the dominance of the entanglement in the crowded scenes.

To solve the problem of entanglement, we propose an efficient CNN-based DecenterNet with Decentralized Pose Representation (DPR) to focus on the visible keypoints. DPR uses all visible keypoints as the root points to represent human poses, alleviating the occlusion problem of the central point strategy [11]. As DPR will produce more duplicate pose candidates and reduce the performance of the model, we further propose Decoupled Pose Assessment (DPA) to predict the location map that represents the confidence of the predict pose, and help the model select optimal poses adaptively. Moreover, we also propose Limb Disentanglement Learning (LDL) as an auxiliary task to facilitate the model to pay more attention to the direction of limbs in entangled scenes, so as to better help DPR improve the prediction of root points. We evaluate DecenterNet on three datasets, *i.e.*, COCO, CrowdPose, and our SkatingPose. Our DecenterNet achieves 71.2 AP scores on the COCO test-dev 2017 set, outperforming DEKR [11] by 1.9 AP scores. On CrowdPose dataset, DecenterNet still obtains state-of-the-art performance, exceeding DEKR by 4.1 AP scores. On our SkatingPose dataset, we also achieve the best performance compared with our reproduced typical works.

Our contributions are summarized as follows.

- To our best knowledge, we are the first to propose dividing the crowded scenes into entanglement and occlusion, and consider entanglement as a significant factor in crowded scenes. We further construct a new dataset named SkatingPose to better evaluate algorithms in entangled scenes.
- We introduce DecenterNet with a Decentralized Pose Representation (DPR) to deal with pose estimation in entangled scenes. We also propose Decoupled Pose Assessment (DPA) to filter duplicate candidate poses, which further improves DecenterNet's performance.
- Our efficient CNN-based method outperforms previous bottom-up methods on two benchmarks for crowded scenes, *i.e.*, CrowdPose and SkatingPose. Additional experiments on the large-scale dataset, *i.e.*, COCO, further verify the superiority and generalization of DecenterNet.

2 RELATED WORK

The convolutional neural network (CNN) [2, 9, 23, 51–53] solutions to human pose estimation [31, 47] have obtained superior achievement recently. The existing works for the MPPE task can generally be divided into top-down methods and bottom-up methods.

Top-Down Methods. Top-down methods first detect all human instances by an object detector like R-CNN [12], then detect the keypoints of a single person with heatmaps in each human bounding box area. Representative works contain Mask-RCNN [13], G-RMI [30], SimpleBaseline [42], HRNet [36], *etc.* Top-down methods can achieve accurate results at the cost of an additional human instance detector and the repeated single-person keypoint detection for every human bounding box area.

Bottom-Up Methods. Bottom-up methods [15, 16] detect all keypoints and group them into individuals, which can be further divided into two-stage and single-stage methods based on the chronological order of keypoints detection and grouping.

The two-stage methods commonly use keypoint heatmaps to realize the keypoint detection, then group them into individuals by instance information belonging to detected keypoints. Representative works contain PifPaf [18], HrHRNet [7], and CenterGroup [4]. For HrHRNet [7], the individual information is associative embeddings, the associative embeddings of the same individual have much shorter ℓ_2 distances than others. However, the two-stage methods matching each keypoint with corresponding instances will result in extra time complexity, which increases the computational effort.

The single-stage methods [19] finish the keypoint detection and group them into individuals at the same time. Representative works contain CenterNet [55], DEKR [11], and LOGO-CAP [46], *etc.* For example, DEKR [11] represents the keypoint locations with keypoint heatmaps and uses offset map to obtain multiple poses at the same time. Most single-stage methods are accurate and efficient in simple scenes, but they struggle to perform well in crowded scenes due to the root point easily overlapping with others.

Pose Estimation in Crowded Scenes. CrowdPose dataset [20] is collected to better evaluate human pose estimation algorithms in crowded scenes. Alphapose [10] proposes full body pose estimation and pose tracking, which yield high accuracy in heavily crowded scenes. OPEC-Net [32] constructs graph networks to address human pose estimation in the crowded scenes, and collects OCPose dataset which contains images with highly challenging occluded scenes from multiple videos. PINet [37] directly infers the complete pose cues for a person from manually divided body parts. Compared with previous works, images of our SkatingPose contain more entangled scenes with visibility tags. Furthermore, our DecenterNet using all visible keypoints as root points to increase the capability to produce more pose candidates and distinguish instances.

3 METHOD

Overview. Our DecenterNet is illustrated in Fig. 3, the input image is firstly processed by the backbone, *i.e.*, HRNet-W32 or HRNet-W48. Then DecenterNet generates heatmap, offset map, location map, and limb map through their respective branches. Finally, DecenterNet selects poses from the offset map by the location map, and uses the heatmap to obtain scores of these poses.

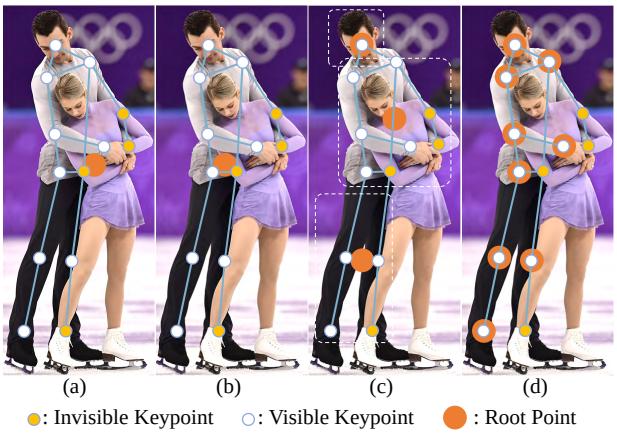

Figure 2: Comparison of different root point settings: (a) central point of visible keypoints, (b) pelvis point, (c) central points of body parts, and (d) visible keypoints (Ours).

3.1 Decentralized Pose Representation

Previous works generally decompose a human pose into a root point and root-based offsets to all keypoints. Specifically, the root point is set at the pelvis [29], the central point of existing keypoints [11], or central points of body parts [37], as illustrated in Fig. 2 (a), (b) and (c). The fine-grained setting like (d) has not been adopted because it will bring more false positive poses and make the result worse [37]. However, when human bodies appear entanglement between each other in crowded scenes, their root points may be occluded mutually and ambiguous about which person to represent. Hence, we propose a Decentralized Pose Representation (DPR) to relieve the entanglement problem in crowded scenes and will introduce Decoupled Pose Assessment (DPA) to solve the false positive problem in Sec. 3.2.

Specifically, DPR uses all visible keypoints as the root points to predict corresponding offsets, as shown in Fig. 2 (d). In this way, the pose of each person is predicted by all visible keypoints, which is more robust in crowded scenes. On the one hand, central point [11] is easy to overlap when two persons are entangled and causes ambiguity about people's identities, while keypoints of the human body are hard to be completely occluded and more discriminated than central point. Using all visible keypoints as the root points will increase the capability to distinguish the entangled persons. On the other hand, fusing predictions from more various positions of root points will produce more comprehensive and robust predictions, as root points are more accurate in predicting nearby keypoints while less accurate for the farther ones [37]. As a result, DPR is simple and efficient in dealing with crowded scenes.

3.2 Decoupled Pose Assessment

As mentioned above, DPR can also lead to more duplicated candidate poses due to more root points, further increasing difficulty of selecting the optimal one. Therefore, we hope that the network can adaptively retain high-quality candidate poses, which can alleviate the false positive pose problem while obtaining more comprehensive candidate poses. So we propose Decoupled Pose Assessment

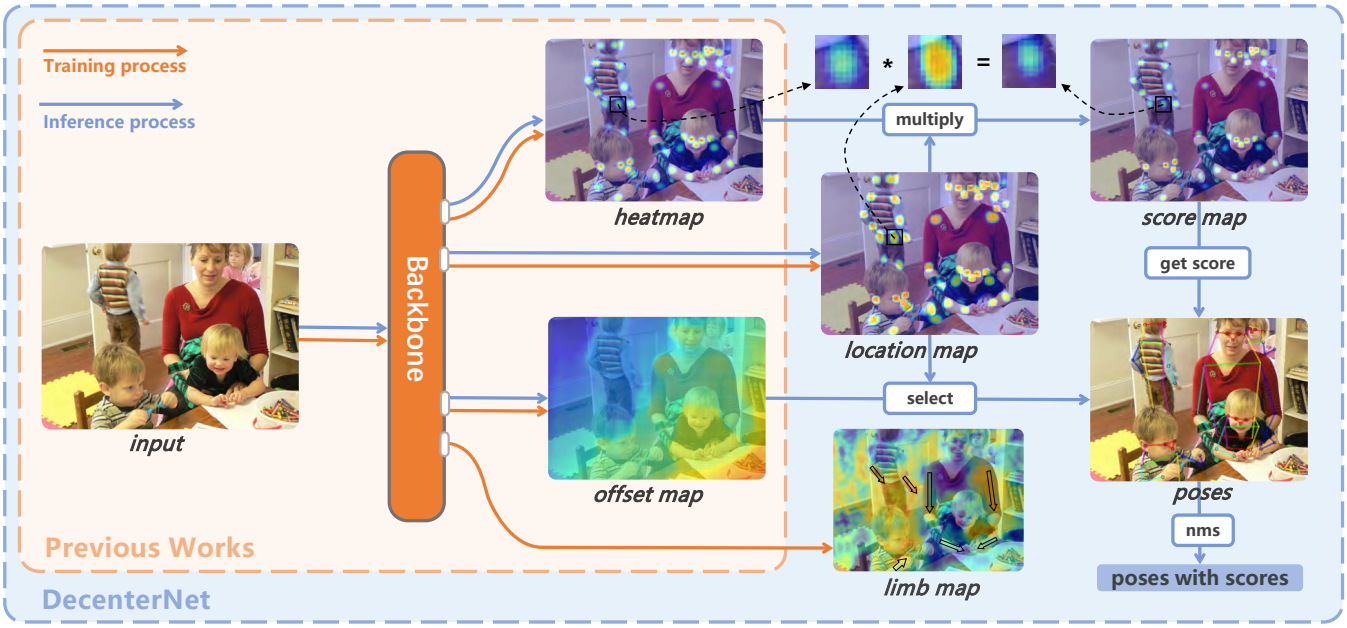

Figure 3: Illustration of the proposed DecenterNet for human pose estimation in crowded scenes (all data from real network output). In the training process, DecenterNet has four outputs, *i.e.*, heatmap, offset map, location map, and limb map. In the inference process (the limb map is not involved in the process), the location map chooses the candidate poses from the offset map, and multiplies with the heatmap to give those candidate poses accurate scores.

(DPA) under this requirement. DPA uses offset map and heatmap to obtain the candidate poses and scores, and introduces a new location map to decouple the root point selection and scoring function of the heatmap as shown in Fig. 3. Specifically, previous works mostly represent the positions of their root points and keypoints by the heatmap, then use the root point heatmap and keypoint heatmap to obtain corresponding poses in the offset map and score these candidate poses, respectively. Nevertheless, the root point heatmap only represents the confidence of the root point's position, not the quality of pose at the corresponding position of the offset map. Therefore, we propose to adopt the location map to replace the root point selection function of the heatmap and further enhance the heatmap's scoring function.

Location Map. Location map has K channels for root points on K keypoint positions. Suppose $(x,\ y)$ denotes the coordinate of a pixel in the location map, and the n^{th} person's k^{th} keypoint's coordinate is $(x_{n,k},\ y_{n,k})$. Then the value $\boldsymbol{C} \in \mathbb{R}^{K\times w\times h}$ in the location map is supervised as follows.

$$\boldsymbol{C}_k(x,y) = \begin{cases} 1 & |x - x_{n,k}| < \alpha, |y - y_{n,k}| < \alpha, \\ 0 & else, \end{cases} \tag{1}$$

where α is used to control the size of the positive sample area around the root points and is a constant set by empirical design(we set $\alpha = 2$ empirically), and $1 \leqslant n \leqslant N$, N is the number of persons with annotations in the given image. In this way we get a root point area for the location map to supervise. The location map is also involved in the supervision process of the offset map, allowing DecenterNet to use the location map to represent the quality of poses in the offset map adaptively.

Offset Map. Offset map has $2 * K$ channels for a person's K 2D coordinates of all keypoints. And the value $\boldsymbol{O} \in \mathbb{R}^{(2*K)\times w\times h}$ in offset map is formulated as follows.

$$\begin{aligned} \boldsymbol{O}_{2^*k+0,x,y} &= x - x_{n,k}, \\ \boldsymbol{O}_{2^*k+1,x,y} &= y - y_{n,k}, \\ s.t. |x - X_n| &< \alpha, |y - Y_n| < \alpha, \end{aligned} \tag{2}$$

where $(X_n,\ Y_n)$ is the coordinate of the n^{th} person's root point. In order to make the location map to represent the quality of the pose in the offset map's corresponding position, the location map is acted as a coefficient for the corresponding position of the offset map's ℓ_1 loss. The predicted location map and the final location map loss are denoted as $\boldsymbol{C}^p \in \mathbb{R}^{K\times w\times h}$ and $\mathcal{L}_c \in \mathbb{R}^1$ respectively. Similarly, we get $\boldsymbol{O}^p \in \mathbb{R}^{K\times w\times h}$ and $\mathcal{L}_o \in \mathbb{R}^1$ for the offset. The formula is expressed as follows.

$$\mathcal{L}_c = \sum_{x=0}^{W}\sum_{y=0}^{H}\sum_{k=0}^{K} \|\boldsymbol{C}^p_{k,x,y} - \boldsymbol{C}_{k,x,y}\|_1^1, \tag{3}$$

$$\mathcal{L}_o = \sum_{x=0}^{W}\sum_{y=0}^{H}\sum_{k_1=0}^{K} (e^{\boldsymbol{C}^p_{k_1,x,y}}) * \sum_{k_2=0}^{2*K} \|\boldsymbol{O}^p_{k_2,x,y} - \boldsymbol{O}_{k_2,x,y}\|_1^1. \tag{4}$$

Hence, the offset loss is controlled by the coefficient $e^{\boldsymbol{C}^p_{k_1,x,y}}$ of the corresponding ℓ_1 offset loss, which means higher weights are applied to pixels with higher predicted values where the estimations are more confident. Then we can get reliable poses from the offset map by the location map.

Heatmap. Obtaining pose scores from keypoint heatmap is a common practice [11]. We follow this practice except that we use the

visible keypoint heatmap multiplied by the location map to enhance the score function. For the heatmap specifically, 2D Gaussian distribution is adopted at the location of keypoints, the value $\boldsymbol{H} \in \mathbb{R}^{(2*K)\times w\times h}$ in the heatmap is $\exp(-\|(x,y)-(x_k^j,\ y_k^j)\|_2^2/\sigma^2)$ and the heatmap loss $\mathcal{L}_{\text{h}} \in \mathbb{R}^1$ is generated by the ℓ_1 distance between the predicted heatmap and the ground-truth heatmap. For the location map, as we define in Eqn. (1), the values converge to 1 in the region around root points, *i.e.*, keypoints in our work. So the location map can be used to limit the range of the heatmap by the multiplication operation, finally we can get a more accurate map named score map as shown in Fig. 3.

To sum up, DPA adaptively obtains the candidate poses by the location map and scores these poses with the keypoint heatmap enhanced by the location map. This method is able to ignore other poses with low confidence in the entangled scenario and accurately select the optimal pose.

3.3 Limb Disentanglement Learning

Assigning a root point on the offset map to the corresponding person is essential for DPR and DPA to work properly. We tend to let the model focus on understanding the link structure of human body through an appropriate multi-task learning, namely Limb Disentanglement Learning (LDL). To be specific, we utilize the limb map which can benefit the network's understanding of limb structure in the training process. Suppose a person has M limbs, the value $\boldsymbol{B} \in \mathbb{R}^{(2*M)\times w\times h}$ in the limb map is displayed as follows.

$$\begin{aligned} \boldsymbol{B}_{2^*m+0,x,y} &= cos(\theta_{n,m}), \\ \boldsymbol{B}_{2^*m+1,x,y} &= sin(\theta_{n,m}), \end{aligned} \tag{5}$$

where θ denotes the counterclockwise rotation angle of the limb vector with respect to the x-axis, and (x,y) is is constrained to a line of one-pixel width on the corresponding n^{th} person's m^{th} limb. Given the predicted limb map $\boldsymbol{O^p} \in \mathbb{R}^{K\times w\times h}$, we can get the ℓ_1 loss of the limb map.

$$\mathcal{L}_{\text{b}} = \sum_{x=0}^{W}\sum_{y=0}^{H}\sum_{m=0}^{2*M} \|\boldsymbol{B}^{\boldsymbol{p}}_{m,x,y} - \boldsymbol{B}_{m,x,y}\|_1^1. \tag{6}$$

Discussion. The formula of LDL is inspired by PAF [18, 54]. However, there are two differences between LDL and PAF. First, in LDL, only the limbs are estimated instead of the whole body, as the limb joints are usually close to each other which are more ambiguous for estimation. Second, LDL serves as an auxiliary loss to improve performance in our framework, which is discarded in the inference process. On the contrary, in PAF, it is used as a grouping clue which is essential in the inference process.

3.4 Training and Inference

In summary, DecenterNet finally applies four losses in the training process. The final loss $\mathcal{L}_{\text{total}}$ is formulated as follows.

$$\mathcal{L}_{\text{total}} = \lambda_c\mathcal{L}_{\text{c}} + \lambda_h\mathcal{L}_{\text{h}} + \lambda_o\mathcal{L}_{\text{o}} + \lambda_b\mathcal{L}_{\text{b}}, \tag{7}$$

where λ_c, λ_h, λ_o, λ_b and $\mathcal{L}_{\text{c}}$, $\mathcal{L}_{\text{h}}$, $\mathcal{L}_{\text{o}}$, $\mathcal{L}_{\text{b}}$ are the coefficients and losses of heatmap, offset map, location map, and limb map, respectively. During training, we set λ_c, λ_h, λ_o, λ_b = 1, 1, 0.1, 0.1 experimentally and obtains promising results.

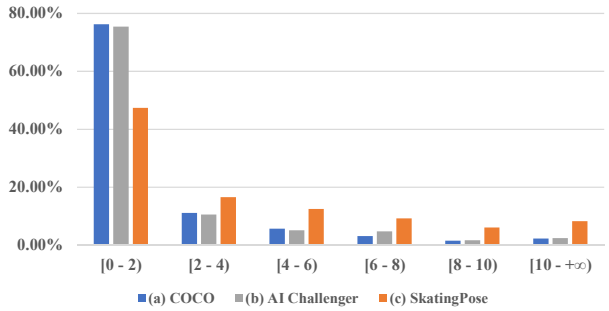

Figure 4: Distribution of images among multiple datasets related to Entangelement Index. The abscissa is Entanglement Index, and the ordinate is the proportion within the range of Entanglement Index. (a) COCO [24]; (b) AI Challenger [41]; (c) SkatingPose (Ours).

In the inference process, given an image with multiple persons, DecenterNet outputs the location map, the offset map, and the heatmap. The channel numbers are K, $2*K$, and $2*K$, respectively (the limb map is only used in the training process). First, we perform the 3×3 MaxPool2d operation in the location map and select the top T largest points in descending order for each channel (we set $T = 8$ empirically). The values of these $T*K$ points represent the quality scores of the $T*K$ poses at the corresponding positions of the offset map. Second, we get the score map by multiplying the visible keypoint heatmap with the location map, and keypoint confidences of each pose can be obtained at the corresponding position in the score map. In the end, we get $T*K$ candidate poses with scores, put these poses with scores via NMS, and maintain at most 30 candidates for one image.

4 SKATINGPOSE DATASET

Motivation. We construct the SkatingPose dataset for two main reasons: (1) most images in the existing human pose datasets, *e.g.*, COCO [24], MPII [3], AI Challenger [41] depict simple scenes but lack intensive entanglement scenes; (2) Crowdpose [20] and MPII [3] dataset have no visibility flags which are used in both the entanglement evaluation and our DecenterNet processing. Considering the large number of entangled scenes and diverse types of background and clothing in figure skating sports, we finally choose to create the figure skating dataset to better evaluate the performance of the model in entangled scenes.

4.1 Data Collection and Annotation

In addition to collecting images with the web crawler, we also collect some images from the competition videos released by the International Skating Union [1]. Finally, we remove the duplicate images through the pHash algorithm [50] and get more than 17,000 original images about figure skating. We process the batch of images by manual screening (removing non-realistic images such as cartoons) and annotation, and then divide them into the training set with 8,113 images and the test set with 7,000 images.

4.2 Entanglement Index

As we have observed that entanglement is a significant factor for inaccurate recognition in crowded scenes, we propose Entanglement Index to measure the entanglement level of an image.

Given an image with N persons, the n_1^{th} human instance will have a primitive external polygon consisting of visible keypoints, which is considered to be the human body area. Considering the visible keypoints of another person close to the human body also have grouping errors, we will expand the polygon P times to take the surrounding visible keypoints into account, we set multiple expansions as (0:0.03:0.2) empirically, and $P = 6$. Suppose V_{i,n_1,n_2} visible keypoints of the n_2^{th} person fall into the p^{th} polygon of the n_1^{th} person, Entanglement Index is formulated as follows.

$$Index = \frac{1}{P}\sum_{n_1=0}^{N}\sum_{p_1=0}^{P}\sum_{n_2=0}^{N} V_{i,n_1,n_2}, \quad s.t.\ n_1! = n_2. \tag{8}$$

Through the above formula, we assign a weight of 1 to the visible keypoints of other people that fall within the primitive polygon, *i.e.*, the human body area. We also decrease the weight of visible keypoints of other people in proximity to the body based on their distance from it. This allows us to accurately quantify the entanglement level of a given image.

4.3 Advantage

As shown in Fig. 4, we count the distribution of Entanglement Index among different datasets that have visibility flags. We can notice that about 80% images of COCO [24] and AI Challenger [41] only have little entanglement scenes. And SkatingPose has more entanglement scenes compared with the two datasets. The multiple advantages of the SkatingPose dataset are summarized as follows.

- SkatingPose dataset contains a large number of entanglement and occlusion scenes due to the frequent interaction between people. Additionally, the annotation of pictures is relatively complete, which can better evaluate the model performance in crowded scenes.
- Human poses are more various and comprehensive due to the professionality in figure skating sports, while the background and clothing are more diverse and extensive than realistic scenes. In this way, the scenes in SkatingPose have a wider distribution than ordinary human pose datasets.
- SkatingPose dataset can also be applied to the development of athlete pose estimation in figure skating competitions, providing a dataset basis for the application of artificial intelligence in sports events.

5 EXPERIMENTS

In this section, we first briefly illustrate datasets, evaluation metrics, and implement details about our DecenterNet in Sec. 5.1. Then we compare DecenterNet with representative works on three datasets, *i.e.*, COCO, CrowdPose, and SkatingPose, in Sec. 5.2. Finally, we analyze on the validity of different modules through ablation study and discuss the model efficiency in Sec. 5.3.

5.1 Experiment Setup

Datasets and Evaluation Metric. The proposed DecenterNet for multi-person pose estimation task is evaluated on three datasets: the SkatingPose dataset we build for the entangled scenes, the COCO dataset [24], and the CrowdPose [20] dataset.

We construct SkatingPose dataset to evaluate the model's ability to untangle bodies in a crowded skating scenario. The SkatingPose dataset mainly contains figure skating images and some team skating images, and these skating images contain numerous body-entangled scenes due to the athletes' frequent interaction with others. In this dataset, there are 15,113 images (8,113 in training set and 7,000 in test set) and 32,008 human instances with 17 keypoint coordinates and their visibility flags.

The COCO dataset [24] contains over 150,000 training instances and 80,000 testing instances labeled with keypoints. Each person has 17 keypoint coordinates and their visibility flags that DecenterNet needs for training. For the evaluation metric, referring to the official evaluation metric of the COCO Keypoint Challenge [24], we report Average Precision (AP) and the Average Recall (AR) with different thresholds and different instance sizes.

The CrowdPose [20] dataset is more likely to occur crowded scenes on each image statistically compared to the COCO dataset. The CrowdPose dataset contains 20,000 images in total and 80,000 human instances, and each person has 14 keypoints but no visibility flag (we have addressed this issue as much as possible by the instance segmentation algorithm Mask2Former [6]).

The CrowdPose dataset and SkatingPose dataset both follow the evaluation metric of the COCO pose estimation task. In addition, we also propose the Entanglement Index for the limb entangled scenes, which is used to measure the entanglement of each image. We divide the SkatingPose test set evenly into 3 difficulty levels according to the Entanglement Index: easy (AP_E), medium (AP_M), and hard (AP_H), to better evaluate our model performance in different entangled scenarios.

Training. Our settings are consistent with the baseline used in DEKR [11]. For the data augmentation, we adopt random affine transform and random flip in the input image with traditional settings, and crop the image to the size of 512 for HRNet-W32 backbone [36] and the size of 640 for HRNet-W48 backbone [36] . We use Adam [17] as the optimizer. For the COCO dataset and SkatingPose dataset, we set learning rate to 1×10^{-3} and train the model for 140 epochs, the learning rate is divided by 10 at 90th and 120th epoch. For the CrowdPose dataset, we set learning rate to 3×10^{-3} and train the model for 300 epochs, the learning rate is divided by 10 at 200th and 260th epoch. Our model is trained by automatic mixed precision [26] on 8 NVIDIA 2080Ti GPUs at the batch size of 14 per GPU.

Testing. During the test, we resize the images' short side to 512 for HRNet-W32 backbone [36] and 640 for HRNet-W48 backbone [36] keeping the aspect ratio between height and width. We also flip the input to enhance DecenterNet's performance. It is noted that we only adopt the single-scale test. The multi-scale test [7] is a time-consuming and unpractical method to improve performance, which is not the focus of our paper. Besides, DPR can partly alleviate the scale problem with multiple predictions.

Table 2: Comparison with state-of-the-arts on COCO val 2017 set and test-dev 2017 set [24]. It is noted that we only present single-scale results. Because multi-scale test is not the focus of our paper and DPR can partly alleviate the body scale problem with multiple predictions.

Method	Backbone	COCO val 2017					COCO test-dev 2017				
		AP	$AP_{0.5}$	$AP_{0.75}$	AP_M	AP_L	AP	$AP_{0.5}$	$AP_{0.75}$	AP_M	AP_L
Top-down Methods											
Mask-RCNN (ICCV'17) [13]	ResNet-50-FPN	64.2	86.6	69.7	58.7	73	63.1	87.3	68.7	57.8	71.4
Lite-HRNet (CVPR'21) [48]	Lite-HRNet-30	70.4	88.7	77.7	**76.2**	**92.8**	69.7	90.7	77.5	66.9	75.0
HRNet (CVPR'19) [36]	HRNet-W48	**76.3**	**90.8**	**82.9**	72.3	83.4	**75.5**	**92.5**	**83.3**	**71.9**	**81.5**
Bottom-up Methods											
PifPaf (CVPR'19) [18]	ResNet-152	67.4	86.9	73.8	63.1	74.1	66.7	87.8	73.6	62.4	72.9
CenterNet (CVPR'19) [55]	Hourglass	64.0	85.6	70.2	59.4	72.1	63.0	86.8	69.6	58.9	70.4
PINet (NIPS'21) [37]	HRNet-W32	67.4	-	-	-	-	66.7	-	-	-	-
HrHRNet (CVPR'20) [7]	HrHRNet-W32	67.1	86.2	73	61.5	76.1	66.4	87.5	72.8	61.2	74.2
DEKR (CVPR'21) [11]	HRNet-W32	67.2	86.3	73.8	61.7	77.1	66.6	87.6	73.5	61.2	75.6
	HRNet-W48	70.3	87.9	76.8	66.3	78	69.3	89.1	76.7	65.3	76.4
CenterGroup (ICCV'21) [4]	HRNet-W32	69.0	87.7	74.4	59.9	75.3	67.6	88.7	73.6	61.9	75.6
	HRNet-W48	71.0	88.7	76.5	63.1	75.2	69.6	**89.7**	76.0	64.9	76.3
LOGO-CAP (CVPR'22) [46]	HRNet-W32	69.6	87.5	75.9	64.1	78.0	68.2	88.7	74.9	62.8	76.0
	HRNet-W48	72.2	**88.9**	**78.9**	**68.1**	78.9	70.8	**89.7**	77.8	**66.7**	77.0
DecenterNet (Ours)	HRNet-W32	70.1	86.9	75.9	64.3	78.9	69.0	87.9	75.8	63.3	77.3
	HRNet-W48	**72.4**	87.8	78.1	67.9	**79.4**	**71.2**	89.0	**78.1**	**66.7**	**77.8**

Table 3: Comparison with state-of-the-arts on SkatingPose test set.

Method	SkatingPose Dataset			
	AP	AP_E	AP_M	AP_H
CenterNet-DLA34 (CVPR'19) [55]	25.6	32.4	24.9	18.9
HrHRNet-W32 (CVPR'21) [7]	45.9	50.7	44.5	41.9
AdaptivePose-W32 (AAAI'22) [44]	66.2	73.1	65.6	60.8
LOGO-CAP-W32 (CVPR'22) [46]	67.3	73.6	67.2	61.6
DEKR-W32 (CVPR'21) [11]	68.3	74.1	67.7	63.1
PINet-W32 (NIPS'21) [37]	70.6	74.0	69.2	68.5
DecenterNet-W32 (Ours)	**72.4**	**76.6**	**71.9**	**68.8**

Table 4: Comparison with state-of-the-arts on CrowdPose test set.

Method	CrowdPose Dataset			
	AP	AP_E	AP_M	AP_H
Top-down Methods				
Mask RCNN (ICCV'17) [13]	57.2	69.4	57.9	45.8
HRNet-W32 (CVPR'19) [36]	67.5	77.0	68.7	55.3
Bottom-up Methods				
SPM (ICCV'21) [29]	63.7	70.3	64.5	55.7
HrHRNet-W48 (CVPR'20) [7]	65.9	73.3	66.5	57.9
CenterGroup-W48 (ICCV'21) [4]	67.6	73.9	68.2	60.3
DEKR-W32 (CVPR'21) [11]	65.7	73.0	66.4	57.5
DEKR-W48 (CVPR'21) [11]	67.3	74.6	68.1	58.7
PINet-W32 (NIPS'21) [37]	68.9	75.4	69.6	61.5
AdaptivePose-W48 (AAAI'22) [44]	69.2	76.7	70.0	60.9
DecenterNet-W32 (Ours)	69.3	76.8	70.0	60.8
DecenterNet-W48 (Ours)	**71.4**	**78.6**	**72.1**	**63.0**

5.2 Main Results

SkatingPose Dataset. We compare DecenterNet with other bottom-up methods on the SkatingPose dataset. As shown in Tab. 3, our DecenterNet achieves the state-of-the-art results, and its AP scores is 46.8 AP above CenterNet [55], 4.1 AP above DEKR [11], and 1.8 AP above PINet [37]. Due to the large number of crowded scenes in SkatingPose test set, it may be hard for the HigherHRNet [7] relying on Associative Embedding [27] to match the appropriate embedding for each dense keypoint. Then HigherHRNet only gets 47.3 AP score on SkatingPose, which is far worse than DecenterNet. Since most images in the test set contain relatively crowded figure skating scenes, the gap between the AP calculated by different models is large, which can better reflect the ability of different models to entangled scenes.

COCO Dataset. In Tab. 2, we compare DecenterNet with state-of-the-art methods on COCO val 2017 and test-dev 2017 sets [24]. Due to the additional human instance detector and repeated single-person keypoints detection, top-down methods, *e.g.*, Mask-RCNN [13], and HRNet [36], Lite-HRNet [48], alleviate the instance scale variance, but DecenterNet can achieve comparable or even better AP scores by our single-stage framework. Compared with bottom-up methods, *e.g.*, DEKR [11], CenterGroup [4], and LOGO-CAP [46], our DecenterNet achieves 69.0 AP scores with the HRNet-W32 and 71.2 AP scores with the HRNet-W48, obtaining state-of-the-art performance on the COCO test-dev 2017 set. Specially, due to our proposed DPR and DPA, DecenterNet exceeds DEKR [11] by 2.9/2.4 AP scores with the equivalent parameters.

CrowdPose Dataset. We further compare DecenterNet with state-of-the-art methods on CrowdPose dataset, which has more crowded

Table 5: Ablation study on DPR, DPA, and LDL. † denotes using all keypoints (visible and invisible) as the root points. * denotes using the visible keypoint heatmap multiplied by the location map to enhance the score map.

DPR	DPA	LDL	AP	AP_M	AP_L	AR
			67.3	62.0	76.7	72.9
√			68.2	62.4	77.2	72.9
√†			67.5	62.1	76.9	72.9
	√		68.4	62.5	77.8	73.2
		√	67.7	62.2	76.9	72.9
√	√		69.3	63.9	78.3	73.8
√	√*		69.7	64.1	78.5	74.0
√	√*	√	70.1	64.3	78.9	74.2

scenes to show the disentanglement performance of our DecenterNet. However, CrowdPose dataset has no detailed visibility flags which DecenterNet used to supervise the location map and the offset map, so we try to adopt Mask2Former [6] to fix the problem (Please refer to the supplementary materials for details).

As shown in Tab. 4, we first make comparisons with the top-down methods, *e.g.*, Mask RCNN [13], HRNet-W32 [36]. These works need to detect human instances with bounding boxes, so the crowded scenarios with many overlapping people are tough for the top-down methods. And our DecenterNet achieves better results on the CrowdPose dataset. Compared with the bottom-up methods, *e.g.*, PINet [37], DEKR-W32 [11], CenterGroup-W48 [4], we still outperform these works even though the repaired visibility flags of the CrowdPose dataset only have a correct rate of about 73%. DecenterNet exceeds DEKR [11] by 3.6/4.1 AP scores with the equivalent parameters.

5.3 Empirical Experiments

Ablative Study. To show contributions of each proposed main component, *i.e.*, Decentralized Pose Representation (DPR), Decoupled Pose Assessment (DPA), and Limb Disentanglement Learning (LDL), we further perform an ablation study on the COCO val 2017 set with the HRNet-W32 as backbone, and the results are shown in Tab. 5. We choose COCO dataset because its large scale can produce more reliable result.

The first row in Tab. 5 is the baseline result without DPR (the baseline uses the center point of visible keypoints as the root point), DPA, and LDL, which achieves 67.3 AP scores on COCO val 2017 set. We first employ DPR, DPA, and LDL separately to update the baseline, and the baseline are improved by 0.7, 1.1, and 0.4 AP scores respectively. These improvements are far from the final result, because there are a lot of mutual blessing between each component. We add both DPR and DPA to the baseline, now this method's offset map and location map have the same center of supervision area which is the locations around keypoints. In this way we obtain an improvement of 1.0 AP scores and our model can be further increased by 0.4 through multiplying the visible keypoint heatmap and location map as the score map. Finally, with LDL handling the entangled scenes, our DecenterNet gets 70.1 AP scores on the COCO val 2017 set.

Discussion on Visible Keypoints. Furthermore, we also try to use all visible keypoints and invisible keypoints to be the root points for DPR, and this method obtains 67.5 AP scores which drops 0.7 AP scores compared with only using visible keypoints. Using visible keypoints as the root points is more inclined to predict the pose on the visible segmentation of human body in the final predicted offset map as shown in Fig. 3. The experiment demonstrates that it is necessary to distinguish keypoint visibility for the human pose estimation in the crowd, and paying attention to visible keypoints is more efficient due to the uncertainty of the invisible keypoints.

Table 6: Efficiency comparisons between DecenterNet and other methods on the COCO val 2017 set [24]. #params(M) and GFLOPs are the results of a single run of the model, and AP scores are the results with flipped test.

	Input size	#params(M)	GFLOPs	AP
FCPose [25]	800	60.3	256.7	65.6
HrHRNet [7]	512	29.6	48.1	67.1
DEKR [11]	512	28.6	45.4	67.2
PETR [46]	800	61.2	274.5	68.5
LOGO-CAP [46]	512	36	112.7	69.6
DecenterNet (Ours)	512	29.7	48.5	70.1

Model Efficiency Analysis. In Tab. 6, we compare the model efficiency between our DecenterNet and other state-of-the-art methods on the parameter size (#params) and computation complexities (FLOPs). The backbone of PETR [33] and FCPose [25] is ResNet-101+FPN, and the backbone of HrHRNet [7], DEKR [11], LOGO-CAP [46], and DecenterNet are all HRNet-W32. Apparently, DecenterNet has comparable parameter size and FLOPs with HrHRNet and DEKR, but achieves 2.9-3.0 higher AP scores on the COCO val 2017 set.

6 CONCLUSION

This paper presents DecenterNet for multi-person pose estimation in crowded scenes. We represent the human pose with a decentralized pose representation that uses all visible keypoints as the root points, making our model more robust in the crowd. To address the issue of false positives, we introduce a location map to replace the point selection function of the heatmap and improve its scoring function. To better evaluate model performance in entangled scenes, we construct the SkatingPose dataset and design an Entanglement Index to measure the entangled level. Finally, our DecenterNet outperforms state-of-the-art bottom-up methods on the SkatingPose, CrowdPose, and COCO datasets while maintaining its efficiency.

ACKNOWLEDGMENT

The paper is supported by National Nature Fund No.62102039, and No.62006244, and the Fundamental Research Funds for the Central Universities, and Young Elite Scientist Sponsorship Program of China Association for Science and Technology YESS20200140, and Young Elite Scientist Sponsorship Program of Beijing Association for Science and Technology BYESS2021178, and Natural Science Foundation of China under Grant No. 62222606.

REFERENCES

[1] Rusa Agafonova. 2019. International skating union versus European Commission: is the European sports model under threat? *The International Sports Law Journal* 19, 1 (2019), 87–101.

[2] Md Zahangir Alom, Tarek M Taha, Christopher Yakopcic, Stefan Westberg, Paheding Sidike, Mst Shamima Nasrin, Brian C Van Esesn, Abdul A S Awwal, and Vijayan K Asari. 2018. The history began from alexnet: A comprehensive survey on deep learning approaches. *arXiv preprint arXiv:1803.01164* (2018).

[3] Mykhaylo Andriluka, Leonid Pishchulin, Peter Gehler, and Bernt Schiele. 2014. 2d human pose estimation: New benchmark and state of the art analysis. In *CVPR*. 3686–3693.

[4] Guillem Brasó, Nikita Kister, and Laura Leal-Taixé. 2021. The center of attention: Center-keypoint grouping via attention for multi-person pose estimation. In *ICCV*. 11853–11863.

[5] Yilun Chen, Zhicheng Wang, Yuxiang Peng, Zhiqiang Zhang, Gang Yu, and Jian Sun. 2018. Cascaded pyramid network for multi-person pose estimation. In *CVPR*. 7103–7112.

[6] Bowen Cheng, Ishan Misra, Alexander G Schwing, Alexander Kirillov, and Rohit Girdhar. 2022. Masked-attention mask transformer for universal image segmentation. In *CVPR*. 1290–1299.

[7] Bowen Cheng, Bin Xiao, Jingdong Wang, Honghui Shi, Thomas S Huang, and Lei Zhang. 2020. Higherhrnet: Scale-aware representation learning for bottom-up human pose estimation. In *CVPR*. 5386–5395.

[8] Yu Cheng, Bo Wang, Bo Yang, and Robby T Tan. 2021. Monocular 3D multi-person pose estimation by integrating top-down and bottom-up networks. In *CVPR*. 7649–7659.

[9] Xiaochuan Fan, Kang Zheng, Yuewei Lin, and Song Wang. 2015. Combining local appearance and holistic view: Dual-source deep neural networks for human pose estimation. In *CVPR*. 1347–1355.

[10] Hao-Shu Fang, Jiefeng Li, Hongyang Tang, Chao Xu, Haoyi Zhu, Yuliang Xiu, Yong-Lu Li, and Cewu Lu. 2022. Alphapose: Whole-body regional multi-person pose estimation and tracking in real-time. *IEEE TPAMI* (2022).

[11] Zigang Geng, Ke Sun, Bin Xiao, Zhaoxiang Zhang, and Jingdong Wang. 2021. Bottom-up human pose estimation via disentangled keypoint regression. In *CVPR*. 14676–14686.

[12] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. 2014. Rich feature hierarchies for accurate object detection and semantic segmentation. In *CVPR*. 580–587.

[13] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. 2017. Mask r-cnn. In *ICCV*. 2961–2969.

[14] Nan Jiang, Kuiran Wang, Xiaoke Peng, Xuehui Yu, Qiang Wang, Junliang Xing, Guorong Li, Jian Zhao, Guodong Guo, and Zhenjun Han. 2021. Anti-UAV: A large multi-modal benchmark for UAV tracking. *arXiv preprint arXiv:2101.08466* (2021).

[15] Lei Jin, Xiaojuan Wang, Xuecheng Nie, Luoqi Liu, Yandong Guo, and Jian Zhao. 2022. Grouping by center: Predicting centripetal offsets for the bottom-up human pose estimation. *IEEE TMM* (2022).

[16] Lei Jin, Xiaojuan Wang, Xuecheng Nie, Wendong Wang, Yandong Guo, Shuicheng Yan, and Jian Zhao. 2023. Rethinking the Person Localization for Single-Stage Multi-Person Pose Estimation. *IEEE TMM* (2023).

[17] Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* (2014).

[18] Sven Kreiss, Lorenzo Bertoni, and Alexandre Alahi. 2019. Pifpaf: Composite fields for human pose estimation. In *CVPR*. 11977–11986.

[19] Jin Lei, Chenyang Xu, Xiaojuan Wang, Yabo Xiao, Yandong Guo, Xuecheng Nie, and Jian Zhao. 2022. Single-Stage is Enough: Multi-Person Absolute 3D Pose Estimation. *CVPR* (2022).

[20] Jiefeng Li, Can Wang, Hao Zhu, Yihuan Mao, Hao-Shu Fang, and Cewu Lu. 2019. Crowdpose: Efficient crowded scenes pose estimation and a new benchmark. In *CVPR*. 10863–10872.

[21] Qun Li, Ziyi Zhang, Fu Xiao, Feng Zhang, and Bir Bhanu. 2022. Dite-HRNet: Dynamic Lightweight High-Resolution Network for Human Pose Estimation. In *IJCAI*. 1095–1101.

[22] Qun Li, Ziyi Zhang, Feng Zhang, and Fu Xiao. [n. d.]. HRNeXt: High-Resolution Context Network for Crowd Pose Estimation. ([n. d.]).

[23] Hongzhou Lin and Stefanie Jegelka. 2018. Resnet with one-neuron hidden layers is a universal approximator. *NIPS* 31 (2018).

[24] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *ECCV*. Springer, 740–755.

[25] Weian Mao, Zhi Tian, Xinlong Wang, and Chunhua Shen. 2021. Fcpose: Fully convolutional multi-person pose estimation with dynamic instance-aware convolutions. In *CVPR*. 9034–9043.

[26] Paulius Micikevicius, Sharan Narang, Jonah Alben, Gregory Diamos, Erich Elsen, David Garcia, Boris Ginsburg, Michael Houston, Oleksii Kuchaiev, Ganesh Venkatesh, et al. 2017. Mixed precision training. *arXiv preprint arXiv:1710.03740* (2017).

[27] Alejandro Newell, Zhiao Huang, and Jia Deng. 2017. Associative embedding: End-to-end learning for joint detection and grouping. *NeuIPS* 30 (2017).

[28] Xuecheng Nie, Jiashi Feng, Junliang Xing, and Shuicheng Yan. 2018. Pose partition networks for multi-person pose estimation. In *ECCV*. 684–699.

[29] Xuecheng Nie, Jiashi Feng, Jianfeng Zhang, and Shuicheng Yan. 2019. Single-stage multi-person pose machines. In *ICCV*. 6951–6960.

[30] George Papandreou, Tyler Zhu, Nori Kanazawa, Alexander Toshev, Jonathan Tompson, Chris Bregler, and Kevin Murphy. 2017. Towards accurate multi-person pose estimation in the wild. In *CVPR*. 4903–4911.

[31] Leonid Pishchulin, Mykhaylo Andriluka, Peter Gehler, and Bernt Schiele. 2013. Poselet conditioned pictorial structures. In *CVPR*. 588–595.

[32] Lingteng Qiu, Xuanye Zhang, Yanran Li, Guanbin Li, Xiaojun Wu, Zixiang Xiong, Xiaoguang Han, and Shuguang Cui. 2020. Peeking into occluded joints: A novel framework for crowd pose estimation. In *ECCV*. Springer, 488–504.

[33] Dahu Shi, Xing Wei, Liangqi Li, Ye Ren, and Wenming Tan. 2022. End-to-end multi-person pose estimation with transformers. In *CVPR*. 11069–11078.

[34] Juil Sock, Kwang In Kim, Caner Sahin, and Tae-Kyun Kim. 2018. Multi-task deep networks for depth-based 6d object pose and joint registration in crowd scenarios. *arXiv preprint arXiv:1806.03891* (2018).

[35] Ke Sun, Cuiling Lan, Junliang Xing, Wenjun Zeng, Dong Liu, and Jingdong Wang. 2017. Human pose estimation using global and local normalization. In *ICCV*. 5599–5607.

[36] Ke Sun, Bin Xiao, Dong Liu, and Jingdong Wang. 2019. Deep high-resolution representation learning for human pose estimation. In *CVPR*. 5693–5703.

[37] Dongkai Wang, Shiliang Zhang, and Gang Hua. 2021. Robust Pose Estimation in Crowded Scenes with Direct Pose-Level Inference. *NeuIPS* 34 (2021), 6278–6289.

[38] Junjie Wang, Zhenbo Yu, Zhengyan Tong, Hang Wang, Jinxian Liu, Wenjun Zhang, and Xiaoyan Wu. 2022. OCR-Pose: Occlusion-Aware Contrastive Representation for Unsupervised 3D Human Pose Estimation. In *ACMMM* (Lisboa, Portugal) *(MM '22)*. Association for Computing Machinery, New York, NY, USA, 5477–5485. https://doi.org/10.1145/3503161.3547780

[39] Qingzhong Wang, Pengfei Zhang, Haoyi Xiong, and Jian Zhao. 2021. Face. evolve: A high-performance face recognition library. *arXiv preprint arXiv:2107.08621* (2021).

[40] Haiping Wu, Bin Xiao, Noel Codella, Mengchen Liu, Xiyang Dai, Lu Yuan, and Lei Zhang. 2021. Cvt: Introducing convolutions to vision transformers. In *ICCV*. 22–31.

[41] J. Wu, H. Zheng, B. Zhao, Y. Li, B. Yan, R. Liang, W. Wang, S. Zhou, G. Lin, and Y. Fu. 2017. AI Challenger : A Large-scale Dataset for Going Deeper in Image Understanding. In *ICME*.

[42] Bin Xiao, Haiping Wu, and Yichen Wei. 2018. Simple baselines for human pose estimation and tracking. In *ECCV*. 466–481.

[43] Yabo Xiao, Kai Su, Xiaojuan Wang, Dongdong Yu, Lei Jin, Mingshu He, and Zehuan Yuan. 2022. QueryPose: Sparse Multi-Person Pose Regression via Spatial-Aware Part-Level Query. (2022).

[44] Yabo Xiao, Xiao Juan Wang, Dongdong Yu, Guoli Wang, Qian Zhang, and HE Mingshu. 2022. Adaptivepose: Human parts as adaptive points. In *AAAI*, Vol. 36. 2813–2821.

[45] Lumin Xu, Ruihan Xu, and Sheng Jin. 2020. Hieve acm mm grand challenge 2020: Pose tracking in crowded scenes. In *ACMMM*. 4689–4693.

[46] Nan Xue, Tianfu Wu, Gui-Song Xia, and Liangpei Zhang. 2022. Learning Local-Global Contextual Adaptation for Multi-Person Pose Estimation. In *CVPR*. 13065–13074.

[47] Yi Yang and Deva Ramanan. 2011. Articulated pose estimation with flexible mixtures-of-parts. In *CVPR*. 1385–1392.

[48] Changqian Yu, Bin Xiao, Changxin Gao, Lu Yuan, Lei Zhang, Nong Sang, and Jingdong Wang. 2021. Lite-hrnet: A lightweight high-resolution network. In *CVPR*. 10440–10450.

[49] Fisher Yu, Dequan Wang, Evan Shelhamer, and Trevor Darrell. 2018. Deep layer aggregation. In *CVPR*. 2403–2412.

[50] Christoph Zauner. 2010. Implementation and benchmarking of perceptual image hash functions. (2010).

[51] Jian Zhao, Yu Cheng, Yi Cheng, Yang Yang, Fang Zhao, Jianshu Li, Hengzhu Liu, Shuicheng Yan, and Jiashi Feng. 2019. Look across elapse: Disentangled representation learning and photorealistic cross-age face synthesis for age-invariant face recognition. In *AAAI*, Vol. 33. 9251–9258.

[52] Jian Zhao, Junliang Xing, Lin Xiong, Shuicheng Yan, and Jiashi Feng. 2020. Recognizing profile faces by imagining frontal view. *IJCV* 128 (2020), 460–478.

[53] Jian Zhao12, Jianshu Li, Fang Zhao, Shuicheng Yan13, and Jiashi Feng. 2017. Marginalized CNN: Learning deep invariant representations. (2017).

[54] C. Zhe, T. Simon, S. E. Wei, and Y. Sheikh. 2017. Realtime multi-person 2d pose estimation using part affinity fields. In *CVPR*.

[55] Xingyi Zhou, Dequan Wang, and Philipp Krähenbühl. 2019. Objects as points. *arXiv preprint arXiv:1904.07850* (2019).

A DECENTERNET

A.1 Performance Discussion

Although we presented numerous performance comparisons with bottom-up methods in the manuscript, there are still some additional comparisons and details that need to be included to further support the efficiency and high performance of our DecenterNet.

Transformer-based Methods. We fellow previous CNN-based works and did not report the comparisons with transformer-based methods, *e.g.*, QueryPose [43], PETR [33], *etc*, in the manuscript. First, their backbone, Swin-L, contains about 197M parameters, while CNN-based methods' backbone only contains 29.7M/65M parameters for HRNet-W32 and HRNet-W48 respectively. Second, we use images resized to 512/640 pixels for shorter sides as input to compare with other CNN-based methods fairly, while the transformer-based methods use 800 pixels instead, which would be too complex to discuss in the manuscript. Thirdly, according to our practice, the GPU memory usage of both transformer-based methods mentioned above is about 14 times higher than that of DecenterNet. which makes it hard to deploy these methods in real-world applications. However, after retraining DecenterNet-W48 with an input size of 800, the same as PETR, DecenterNet is still able to slightly outperform PETR, as shown in Tab. 7.

Table 7: Comparison with transformer-based bottom-up method PETR [33] on COCO, CrpwdPose dataset. We also present the params(M) and FLOPs(G) to show the efficiency of DecenterNet.

	Input Size	Backbone	params (M)	FLOPs (G)	AP
COCO 2017 test-dev set					
PETR (CVPR'22)	800	Swin-L	197	475	70.5
DecenterNet	640	HR-w48	65	103	**71.2**
CrowdPose test set					
PETR (CVPR'22)	800	Swin-L	197	475	71.6
DecenterNet	800	HR-w48	65	152	**71.7**

Additionally, all results are reported without multi-scale test because it is a rough and time-consuming way by scaling the input image multiple times to increase the final accuracy, which is not the focus of our effective approach.

A.2 Post-Processing: NMS Algorithm

We remove duplicated/similar poses via NMS algorithm in post-processing like most bottom-up methods. Suppose there are k candidate poses. The NMS algorithm first calculates the distance between all poses and obtains a k x k distance matrix. Poses whose distances exceed the threshold in each row of the matrix are removed, leaving only duplicates in each row. For each row, only the pose with the highest score is kept, and any pose suppressed by this high score pose is ignored for the entire matrix. Finally, poses with scores less than the threshold are deleted to obtain retained poses. It is evident that the better the score evaluation, the more effectively NMS algorithm handles duplication problems (The introduction of DPA enhances the quality of pose scoring).

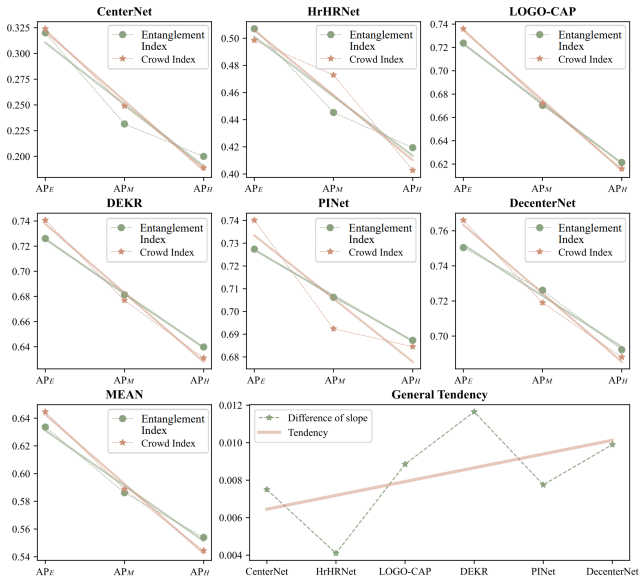

Figure 5: Comparison of AP score distributions based on Crowd Index and Entanglement Index. In the square subgraphs, the orange and green lines represent the fitted lines of the AP distribution of Entanglement Index and Crowd Index, respectively. The rectangular subgraph, named General Tendency, with the values above representing the difference in slope of the fitted lines between Crowd Index and Entanglement Index, and the orange line with an upward trend represents their fitted line.

A.3 DecenterNet with Other Backbones

Limited by computing resources, we chose an efficient CNN network, namely HRNet, to conduct all the experiments of DecenterNet. We will publish results with other backbones, such as CVT [40] and DLA [49] on GitHub soon.

A.4 Failure Cases

Like many other human pose estimation networks, DecenterNet also makes the following mistakes: **Miss error:** Failure in joint localization. **Swap error:** Confusion between the joints of different individuals, resulting in incorrect joint identification. **Inversion error:** Confusion between joints within a person, such as left and right hands. **Jitter error:** Small localization error of a joint. **Score error:** Inappropriate score for a person. An accurate pose with a lower score can be removed by a less accurate pose during NMS. **False Positive error:** A non-existent person is detected. **False Negative error:** An existing person is not detected.

We utilized coco-analyze tool to analyze the baseline and DecenterNet. All results were produced with HRNet-w32 as the backbone on the COCO 2017 val dataset. Tab. 8 shows how much the AP score could increase if these errors were fully corrected. (Note that the sum of each column is not equal to 100 due to the way AP is calculated).

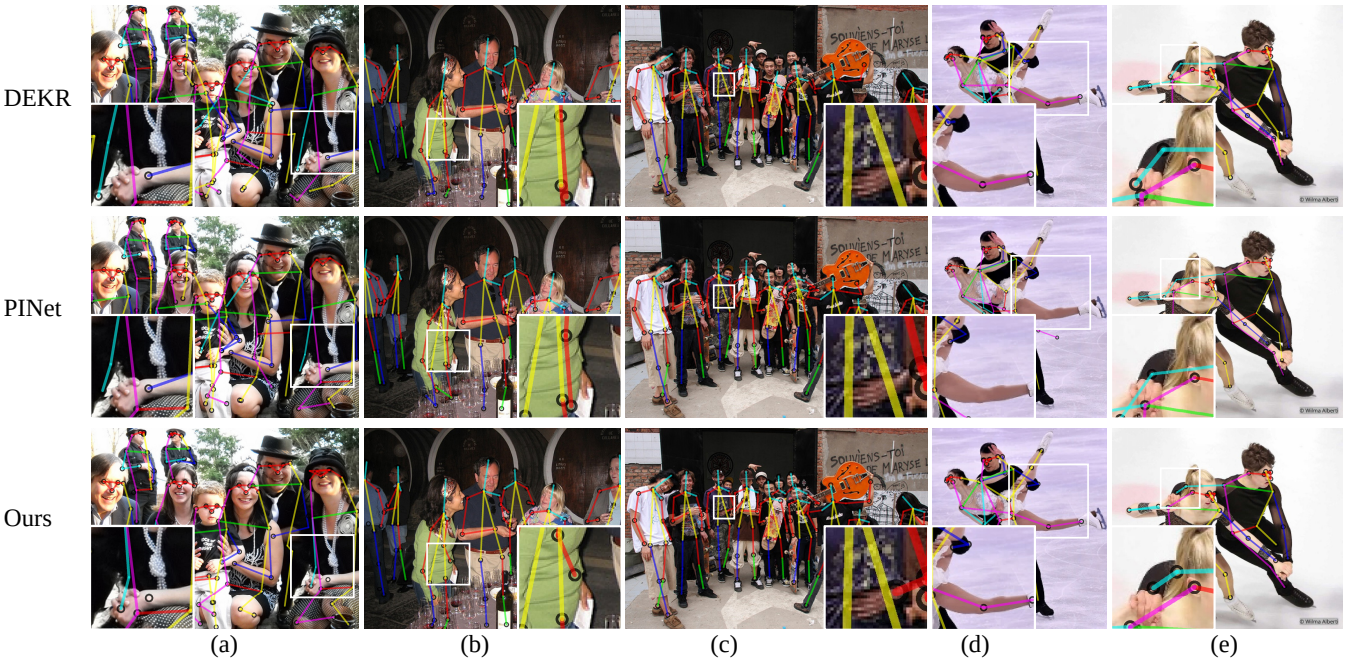

Figure 6: Visual comparison with state-of-the-art methods. The following is a detailed description of the keypoints in question. (a) The end of the cyan limb corresponds to the hidden right wrist of the woman, which is an invisible keypoint. (b) The end of the red limb corresponds to the hidden left wrist of the woman, which is an invisible keypoint. (c) The end of the red limb corresponds to the hidden left wrist of the man, which is an visible keypoint. (d) The keypoints of the right leg of the female athlete and the left elbow of the male athlete, and they are all visible keypoints. (e) Male athlete's right wrist keypoint, and it is a visible keypoint.

Table 8: The impact of various errors on AP score of pose estimators.

Name	Baseline	DecenterNet
origin AP	67.3	70.1
Miss	8.62	6.24
Swap	0.86	0.86
Inversion	2.47	1.93
Jitter	6.56	5.40
Score	5.12	3.52
FP	5.37	3.72
FN	27.22	23.86

A.5 Visual Comparison

To better demonstrate the performance of DecneterNet in crowded scenes, we make a visual comparison with state-of-the-art methods, *i.e.*, DEKR [11] and PINet [37], as shown in Fig. 6.

B DATASET

B.1 Entanglement Index

We have already demonstrated the superiority of Entanglement Index over Crowd Index [20] in crowded scenarios in Fig. 1 of the manuscript. Due to the paper length limit, we show more comparisons here. As we can see in Fig. 5, the fitted line of Entanglement Index always has a greater angle of inclination than Crowd Index, which proves that entanglement is the dominant factor for estimation errors in the crowded scenes.

In the General Tendency subgraph(the rectangular subgraph at the bottom of Fig. 5), higher values represent a larger proportion of entanglement errors made by the model in crowded scenes. In the ascending order of recent works' performance (including ours), we find that the overall trend is progressively increasing. This phenomenon proves that if you want to achieve higher capabilities than the current methods in crowded scenes, paying more attention to entanglement may be more efficient.

B.2 CrowdPose Repairment

The visibility flags in the popular CrowdPose dataset do not appear to be entirely correct via our visualization. However, DecenterNet requires these flags as the root points to represent human poses. Thus, we used the state-of-the-art algorithm Mask2Former [6] in instance segmentation to obtain masks of human instances and match them with keypoints to get visibility flags. However, the accuracy of this method is about 73% based on manual annotations of 50 images due to poor performance of Mask2Former in incentive crowded scenes. Additionally, there are some recognition errors and mismatches between segmentations and poses. Despite the deficiency above, our method still outperforms bottom-up methods and achieves better results on the repaired CrowdPose dataset.