Bottom-Up 2D Pose Estimation via Dual Anatomical Centers for Small-Scale Persons

Yu Cheng[a,1], Yihao Ai[a,1,*], Bo Wang[b], Xinchao Wang[a], Robby T. Tan[a]

[a]*Electrical and Computer Engineering, National University of Singapore, Singapore*
[b]*CtrsVision, USA*

Abstract

In multi-person 2D pose estimation, the bottom-up methods simultaneously predict poses for all persons, and unlike the top-down methods, do not rely on human detection. However, the SOTA bottom-up methods' accuracy is still inferior compared to the existing top-down methods. This is due to the predicted human poses being regressed based on the inconsistent human bounding box center and the lack of human-scale normalization, leading to the predicted human poses being inaccurate and small-scale persons being missed. To push the envelope of the bottom-up pose estimation, we firstly propose multi-scale training to enhance the network to handle scale variation with single-scale testing, particularly for small-scale persons. Secondly, we introduce dual anatomical centers (i.e., head and body), where we can predict the human poses more accurately and reliably, especially for small-scale persons. Moreover, existing bottom-up methods use multi-scale testing to boost the accuracy of pose estimation at the price of multiple additional forward passes, which weakens the efficiency of bottom-up methods, the core strength compared to top-down methods. By contrast, our multi-scale training enables the model to predict high-quality poses in a single forward pass (i.e., single-scale testing). Our method achieves 38.4% improvement on bounding box precision and 39.1% improvement on bounding box recall over the state of the art (SOTA) on the challenging small-scale persons subset of COCO. For the human pose AP evaluation, we achieve a new SOTA (71.0 AP) on the COCO test-dev set with the single-scale testing. We also achieve the top performance (40.3 AP) on the OCHuman dataset in cross-dataset evaluation.

---

[*]Corresponding author. Email: e0724394@u.nus.edu
[1]Co-author

1. Introduction

Human 2D pose estimation is a fundamental research topic in computer vision and has a broad impact on a great number of applications such as human action understanding, person image generation, augmented reality, and motion capture [1, 2, 3, 4, 5]. Significant progress has been made by employing deep learning in recent years [6, 7, 8, 9, 10, 11, 12, 13, 14], yet multi-person 2D pose estimation from a single image is still an open problem.

There are two paradigms addressing the problem: top-down and bottom-up. The major difference between the two is the use of human detection and the follow-up sequential processes. The top-down methods [15, 16, 17, 18, 14] employ human detection to extract and crop each person into individual image patches, and then apply pose estimation to each patch sequentially. As each image patch is normalized, the scale variation of each person is largely reduced, which is suitable for convolutional neural network (CNN) training. While with the help of human detection, the top-down methods enjoy higher accuracy than the bottom-up methods for human pose estimation, but it is still an open problem for human detection in multi-person settings. Moreover, existing top-down methods process each person's image patch sequentially, which makes them inapplicable and unscalable in crowded scenarios.

The bottom-up methods [13, 12, 19, 20, 21, 22, 23, 24, 25] process the whole image at once, which simultaneously estimate all possible keypoints belonging to every person, and then group the keypoints together to form individual persons' skeletons. In the bottom-up paradigm, one forward pass is able to obtain all possible keypoints from an input image, and the follow-up grouping step associates the keypoints into individual skeletons, which is more efficient than the sequential patch-by-patch process of the top-down methods.

However, since the input image is processed in one forward pass and no person-wise image normalization is utilized, the bottom-up multi-person pose estimation methods fail to handle human scale variation, leading to missing or wrongly predicted poses for small-scale persons, which is a fundamental problem in bottom-up multi-person pose estimation [21]. Image pyramid for heatmap prediction [20] and scale-aware representation using high-resolution
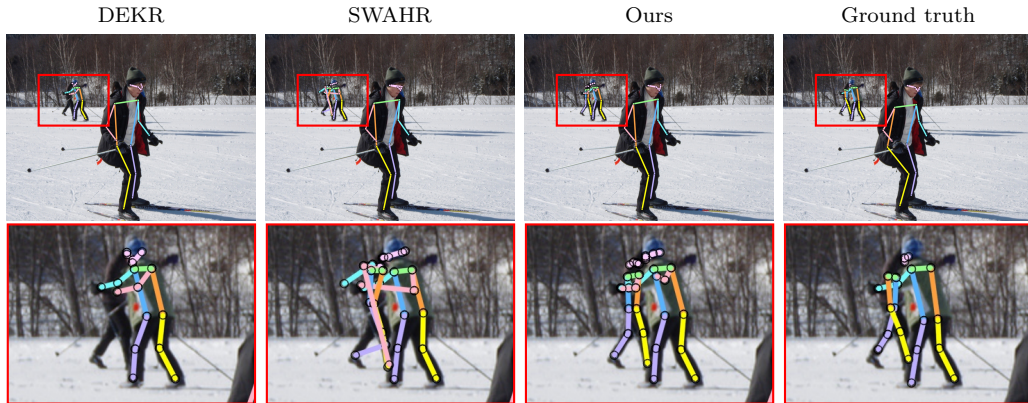
2

Figure 1: Representative results of the proposed method compared to two SOTA methods (DEKR [22] and SWAHR [23]) on COCO validation set. The first row shows each result on the input image, where the red bounding box indicates a zoom-in area. The second row shows the enlarged zoom-in area. All results are based on single-scale testing.

feature pyramid [21] have been developed to mitigate the problem. Unfortunately, the existing bottom-up methods still constantly miss small-scale persons. Missing small-scale persons is a common problem in the SOTA bottom-up methods as the network lacks the capability of handling scale variation. Multi-scale testing is used in the SOTA bottom-up methods [21, 22] to mitigate the problem, where multiple additional forward passes are needed, which clearly slows down inference.

To solve the problem, we propose to use multi-scale training to enhance the network's capability in handling persons across different scales. This multi-scale training allows the model to generate a more robust estimation in a single forward pass. It also yields two advantages: (1) Since the annotations in training data contain limited scales of persons (mainly including medium and large scales), our multi-scale training will broaden the scale variation of existing annotations and enhances the robustness against scale changes, especially the small-scale persons. (2) The multi-scale property of network structure is specifically utilized to handle scale variation. As shown in Fig.2, our multi-scale training utilizing different kernel sizes for different human scales is more capable of dealing with the scale variation than the simple scale augmentation approach, where only one heatmap with a fixed kernel size is used. Moreover, our multi-scale training can better handle the scale variation than the existing methods that use only a single scale of heatmap [23] or a fixed kernel size [21].

Besides solving the scale variation, our multi-scale training significantly improves the pose estimation for small-scale persons. However, the predicted poses of small-scale persons are more sensitive to the same deviation from the ground truth than that of large or medium-scale persons. Hence, to further improve the pose estimation for small-scale persons, it is important to ensure the deviation from the ground truths of the predicted poses is small, which is the problem of the SOTA bottom-up methods [22, 25]. The unreliable human center(s) (bounding box center) used to detect each person in these methods cause their predicted poses to be inaccurate, particularly for small-scale persons. Specifically, because of the human center(s) (i.e., single [22] or multiple [25]) that are not defined anatomically are relatively arbitrary, the subsequent regression of the offsets of each human keypoint to the center(s) becomes inaccurate, resulting in error-prone human pose estimation.

To solve the aforementioned problem, we propose to use human anatomical centers to represent a person instead of the bounding box center(s) used in the existing methods [22, 25] for improving pose estimation accuracy. In particular, we found that directly regressing keypoints' location as offsets to the human bounding box center is not accurate because the location of the bounding box center is relatively arbitrary (i.e., not defined based on human anatomy). Therefore, we propose to use two anatomical landmarks (i.e., head and body centers) as human centers instead of the arbitrary center (i.e. bounding box center). Based on the two anatomical centers, we estimate two human poses for each person to increase the chance to capture a person, and then merge the two poses into one based on pose similarity that considers both spatial configuration and image appearance similarity.

Fig. 1 shows a representative image with large scale variation, and the multi-person pose estimation results of our method, two SOTA bottom-up methods (DEKR [22] and SWAHR [23]), and the ground-truth human poses. This example clearly demonstrates that our method can correctly predict human poses of different scales. In contrast, the SOTA methods fail to predict (DEKR) or produce wrong human poses (SWAHR) for small-scale persons. More qualitative results are provided in Fig. 7 and 8 to illustrate that our method consistently outperforms the SOTA methods in multi-person pose estimation, particularly for small-scale persons. Our major contributions are summarized in the following:

- We propose to estimate dual anatomical centers, and then fuse the obtained poses together by pose confidence and similarity, replacing
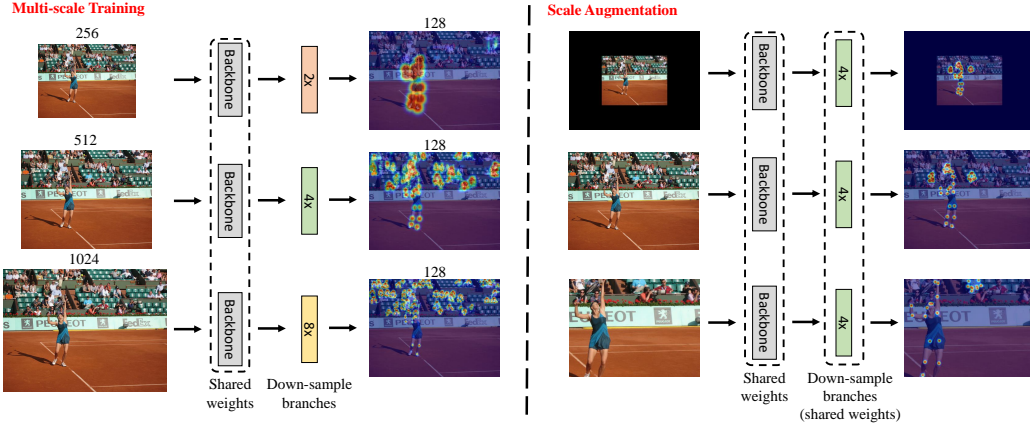
4

Figure 2: Difference between scale augmentation and our proposed multi-scale training. Our multi-scale training utilizes different kernel sizes and multiple branches where each of them is responsible for a certain scale of persons, which makes it more adaptive and can better handle the scale variation compared to scale augmentation.

the unreliable human bounding box center estimation.

- We introduce a multi-scale training framework, which handles persons across different scales, especially for small-scale persons.

- We achieve a new SOTA 71.0 AP on the COCO test-dev set with single-scale testing, and **38.4%** and **39.1%** improvements on bounding box precision and recall respectively over the state-of-the-art methods on the challenging small-scale persons subset of COCO validation set.

2. Related Work

Multi-person pose estimation has been developed rapidly with the help of deep learning in recent years. Earlier works propose to use CNN for human pose estimation from image [26, 6] or video [7, 27, 17]. Several popular network structures are developed including stacked hourglass networks [8], OpenPose [28, 11, 13], and HRNet [14]. Different keypoints grouping approaches are developed for multi-person pose estimation such as DeepCut [9, 10], associative embedding [12], part affinity fields [11], and PifPaf [29].

Top-Down Methods With the fast development in object detection [30, 31, 32, 15], top-down human pose estimation methods are proposed, where human detection is performed first to crop and normalize image patch for each

person, and then pose estimation is applied to each image patch sequentially to predict the human pose [15, 16, 17, 18, 14]. Top-down methods achieve high accuracy, which benefits from human detection and image normalization. As a result, each image patch contains one target person, and the scale of the person is normalized, which is suitable for network training. However, these methods require human detection and sequential person-by-person pose estimation, resulting in slow inference speed for images with multiple persons. In fact, the inference time is linear to the number of persons in an image.

Bottom-Up Methods Bottom-up methods simultaneously detect all keypoints in an image without using human detection [13, 12, 19, 21, 22]. Compared to top-down methods, bottom-up methods process the whole image at once (i.e., one forward pass), thus, it can achieve fast inference. However, the efficiency brought by one forward passing affects the accuracy as persons from large to small scale are processed together at the same resolution, which makes the network difficult to train and perform poorly on small-scale persons. The bottom-up paradigm can be realized in a two-stage manner, first to detect all keypoints at once by taking the whole image as input, and then group the keypoints together to form individual human poses [12, 28, 11, 13, 29].

Recently, DEKR [22] developed a multi-head framework, one head for keypoints' heatmap estimation, the other for human bounding box center and offset to replace grouping. But it also relies on multi-scale testing to boost accuracy. CGNet [24] proposes an attention mechanism to cluster each keypoint to its corresponding center, which helps in occlusion cases but still cannot handle small-scale persons. PINet [25] proposes to divide the human bounding box into three parts and estimate each part separately, which makes it more robust to body-part occlusion. These human bounding box center-based methods [22, 25] that use bounding box center-based offset (i.e., keypoints to the center) regression show promising results. However, the bounding box only covers the visible part of a person, leading the bounding box center to be located arbitrarily over a person. This arbitrary location is used as the human center in the existing methods [22, 25] to regress the offset between each keypoint and the center, which increases the difficulty of the network to reliably and accurately learn the offset, thus, hurting the accuracy of the pose estimation, regardless one bounding box center is used like DEKR [22] or multiple bounding box centers are used like PINet [25].
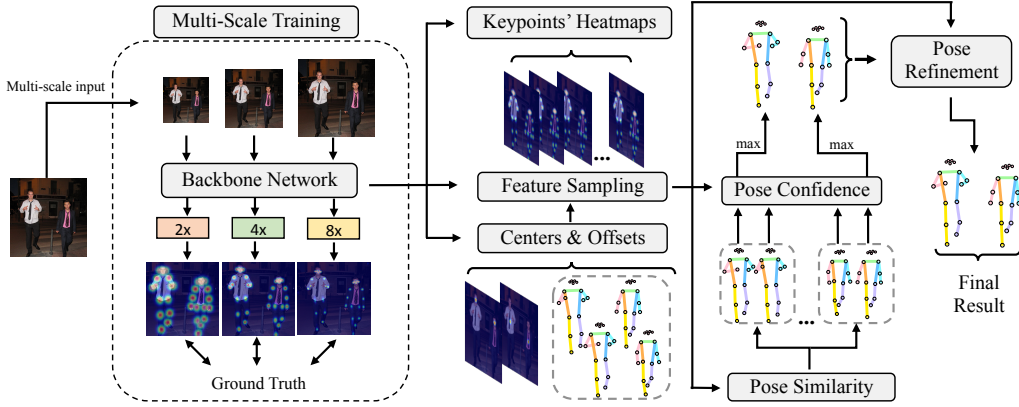
6

Figure 3: An overview of the proposed framework. Our method is composed of a backbone network with multi-scale training and two pose estimation branches, one for keypoints' heatmaps, the other for dual human centers and their offsets. Other major components include: 1) a pose similarity network that estimates pose similarity to group poses according to their identities. 2) a pose confidence network that predicts the confidence of a predicted pose, which is used to select the most confident pose out of a group of candidates. 3) a pose refinement network that merges the information from the predicted poses and multi-scale features for final human poses.

Without human detection and normalization like in top-down methods, handling human scale variation is a fundamental problem in the bottom-up methods. Earlier works tackle the problem by proposing feature pyramid [12, 20], or skip connections [33]. Recent work [34] proposes to regress human poses based on pre-set anchors to deal with scale variation. HigherHRNet [21] proposes an upsampling module to generate multi-scale heatmaps and aggregate them at inference time. SWAHR [23] introduces an adaptive Gaussian kernel to represent the scale of a person, which enables the model to be more robust to scale variance and labeling ambiguities. The proposed solutions may help to enhance the pose estimation accuracy if small-scale persons are detected in the first place, however, the resolution for those persons is mostly insufficient, leading to many small-scale persons being missed.

3. Proposed Method

Fig. 3 is an overview of the proposed framework. Our backbone network takes multi-scale images as input and produces multi-scale heatmaps. Two major branches are built on top of the backbone network to handle the

7

obtained multi-scale feature maps: One branch is to estimate the keypoints' heatmaps, and the other branch is to estimate two human centers and their corresponding offsets (i.e., keypoint locations with respect to the centers). As we use two human centers for each person, resulting in two predicted human poses from the offsets for each person, the offset regression network may also generate false positives (i.e. duplicated human poses for the same person). Thus, the number of the obtained human poses usually is more than the number of persons in the image.

To select a unique pose for each person, we propose a pose similarity network to make image appearance features from the same person close to each other, which is used to predict the similarity to determine if several poses belong to the same person. Furthermore, it is necessary to pick one out of several candidate poses. To this end, we introduce a pose confidence score that measures how close a predicted pose is to its corresponding ground truth. With the pose confidence score, we can pick the pose with the highest score from each pose group. Finally, the poses, the heatmaps, and the multi-scale features are fed into a pose refinement network to obtain the refined poses.

3.1. Multi-Scale Training

In bottom-up human pose estimation, as the whole image is processed with one forward pass to find all keypoints belonging to all possible persons, the resolution of small-scale persons is mostly insufficient to produce accurate pose estimation. Existing bottom-up methods resort to the multi-scale testing to boost their accuracy [21, 22], which involves multiple additional forward passes of images with different resolutions at test time. Thus, the improvement of the accuracy is at the cost of slowing the processing speed, where efficiency is one of the key advantages of the bottom-up methods over the top-down methods.

To address this problem, we propose to perform multi-scale training, which utilizes multi-scale images as input for our network in the training stage. On the one hand, such an approach grants and enhances the model's robustness to the scale variation by broadening the scale diversity. In the multi-scale training process, the network takes the image with different sizes as input and obtains the heatmaps with different kernel sizes as output. This process makes each size of heatmap respond to different sizes of persons, i.e. large kernel size responds to larger persons and small kernel responds to

8

smaller persons, which makes the model better capable of handling human-scale variation. The proposed multi-scale training takes a longer training time in exchange for better performance, where only one forward pass (i.e., single-scale testing) is needed at a time for desirable accuracy (e.g., 71.0 AP on the COCO test-dev is achieved, which is a new SOTA.). The general idea of the multi-scale training is illustrated in the left part of Fig. 2. Note HRNet backbone can take different sizes of input [21, 22].

Two Branches of Pose Estimation As shown in Fig. 3, in our framework, multi-person pose estimation is achieved using two branches. The first branch is to regress the heatmaps of each keypoint, which does not provide the identity information of the keypoints, and just estimates all possible keypoints for every person. The second branch is to estimate human centers and the offset of each keypoint with respect to each center. The estimated offsets form individual poses P, but with duplication (i.e., partially due to the two human centers for each person, and partially due to the offset network's false positive estimates).

3.2. Dual Anatomical Centers for Human Localization

To obtain individual poses, existing bottom-up methods use either grouping (e.g., associative embedding [12], Part Affinity Fields [13]) or human center detection (e.g., human bounding box center and offset [22]). Compared with the grouping strategy, where the process is performed at a second stage after all keypoints are identified, estimating the human center and keypoints' offset to the center is more efficient, since the process is in parallel to the regression of the keypoints' heatmaps. However, we consider that in the context of multi-person pose estimation, human anatomical keypoints as centers are more consistently defined in visual context across different scenarios than the human detection bounding box center used in the existing method [22].

There are two benefits of using dual anatomical keypoints-based centers. First, the bounding box center is an arbitrary location depending on the visible part of the person [22], which is difficult for a network to learn. On the contrary, keypoints such as head or body center (i.e., the interpolation of shoulders and hips), have clear anatomical definitions, and thus their locations can be inferred based on the visual context. Hence, If a network is trained based on the dual anatomical centers, the network can learn more reliable and consistent centers. Second, the anatomical centers-based offsets can be learned more reliably and accurately than the bounding box center-

9

based ones. Due to the location of the bounding box centers' being relatively arbitrary, the corresponding offsets are more random, which is difficult to learn by the network. In contrast, the offsets of the anatomical centers can be learned more accurately due to human anatomy being well-defined, and therefore enabling the model to produce more accurate human pose prediction.

Once the two centers for each person are detected, one person may have two centers and the same number of sets of corresponding offsets. The network may also produce duplicated predictions, resulting in false positives in evaluation. To remove such duplication, we propose a pose similarity network to learn the appearance feature embedding of the predicted human pose. Given M predicted poses $\{P^1, ..., P^M\}$ in an image, we assign their identity based on the closest ground truth, then generate positive and negative pose pairs (e.g., if two poses belong to the same person, they form a positive pose pair; otherwise, that's a negative pose pair). The poses that are not close to any ground truth, are excluded from generating the pose pairs.

In training, the network takes a pair of poses' corresponding image features and heatmaps as input, and outputs two pose feature vectors f^i and f^j. The cosine similarity of the two feature vectors is computed, and the network is trained with a binary label which is 1 if they belong to the same person and -1 otherwise. After the training, if two poses' features are similar to each other, the feature vectors should be close to each other. To compute the similarity of two human poses, we use a feature-based distance $D_{app}(f^i, f^j)$ from the pose similarity network, where f^i is the embedded appearance feature of pose P^i, together with Gaussian distance of pose spatial configuration $D_{sp}(P^i, P^j)$, which is defined as:

$$D_{sp}(P^i, P^j) = \frac{1}{N} \sum_{k=1}^{N} \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{1}{2}\left(\frac{P_k^i - P_k^j}{\sigma}\right)^2\right) \tag{1}$$

where P^i and P^j are a pair of human poses, P_k indicates the kth keypoint of the pose P, N is the number of keypoints, σ is the variance parameter of the Gaussian distance. The pose similarity between a pair of two poses is computed using $D_{pose}(P^i, P^j) = D_{app}(f^i, f^j)D_{sp}(P^i, P^j)$. With the proposed pose similarity D_{pose}, we can split the predicted poses into groups, and then pick the pose with the highest pose confidence, which is defined in the next section (Eq. 2).
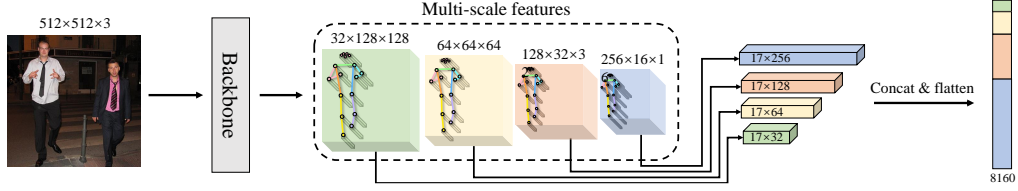
10

Figure 4: Illustration of how to obtain the sampled feature F. Black circles indicate the location to sample the feature maps across different scales based on the estimated human keypoints.

3.3. Pose Confidence Estimation

Existing methods [22, 25] use human bounding box center's confidence or the average of keypoints' confidence [21] to measure the correctness of a predicted human pose in eliminating duplicate poses, which is not a reasonable choice because the confidence of a center/keypoint only reveals its visibility but not the likelihood of the predicted human pose. To this end, we propose to regress the pose confidence, which effectively predicts the confidence of an estimated human pose. We train a Multi-Layer Perceptron (MLP), that takes the concatenation of the sampled intermediate features F (as illustrated in Fig. 4) and predicted heatmaps H according to a predicted pose P as input to regress the Object Keypoint Similarity (OKS) introduced in the COCO evaluation protocol [35] between ground-truth pose $\tilde{P}$ and predicted pose P. The loss of the pose confidence is defined as:

$$L_{pose} = |\mathrm{OKS}(P, \tilde{P}) - \mathrm{MLP}(F \oplus H)|^2. \tag{2}$$

Note that for images with multiple persons, we choose the ground truth $\tilde{P}$ closest to the predicted P to calculate the OKS value. Two examples are provided in Fig. 5 to show the difference between the existing approaches (i.e., bounding box confidence [22] or average of keypoints' confidence [21]) and our pose confidence. We observe that the confidence score of the bounding box center is irrelevant to the correctness of the predicted human pose as shown in the right example (i.e., the predicted pose is wrong), and sensitive to occlusion as shown in the left example (i.e., lower confidence value caused by occlusion). The average of each keypoint's confidence value as an alternative, is affected by occlusion (left) and rare poses (right) as well. In contrast, our pose confidence reflects the correctness of the predicted human pose, at the same time, it is robust to occlusion.
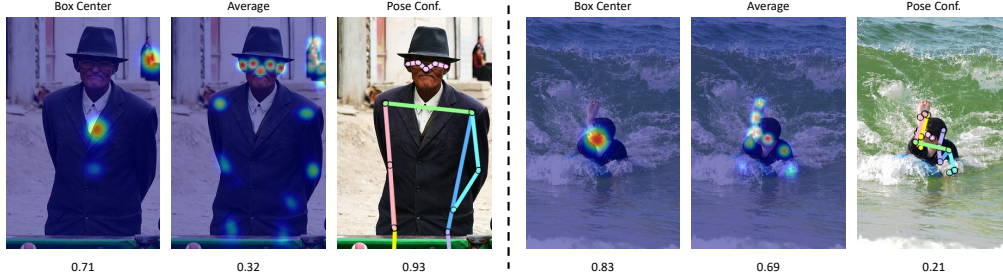
11

Figure 5: Comparisons of the proposed pose confidence with the existing approaches. Two examples are provided, the input image is shown as the background for each example. In each example, the heatmap of the bounding box center, the keypoints' heatmaps, and the predicted human pose are provided. The numbers under each image indicate the confidence value of each approach.

Pose Refinement Unlike most existing methods, we refine our pose estimation further. We use cross-scale information instead of simply merging the regression-based poses and the heatmap-based poses [36], where no additional information is available to fix the errors in the predicted human poses. Based on our proposed pose similarity and pose confidence, we obtain the regression-based poses $\{P^1, ..., P^K\}$ from the estimated centers and offsets. For refinement, we estimate the residual between predicted poses and ground truths, using sampled features across scales to integrate multi-scale information. We use the multi-scale feature, the estimated pose P, and the corresponding heatmap H as input for our refinement network to regress a residual P_δ to its ground-truth pose. The final pose is obtained by $P_{\text{final}} = P + P_\delta$, and L_2 loss between the ground-truth pose and the predicted pose is applied to regress the P_δ for the pose refinement network:

$$L_\delta = |P_\delta - (\tilde{P} - P)|^2. \tag{3}$$

4. Experiments

Datasets and Evaluation Metrics We evaluate our method on COCO [35], OCHuman [37], and CrowdPose [38]. Average precision (AP), AP^{50}, and AP^{75} are used to measure overall human pose estimation accuracy on the COCO dataset. AP is used to measure cross-dataset generalization performance on the OCHuman dataset. However, AP is not an objective performance measurement if a method detects more human poses than the annotated ground truths. In fact, the predicted human poses of a method are
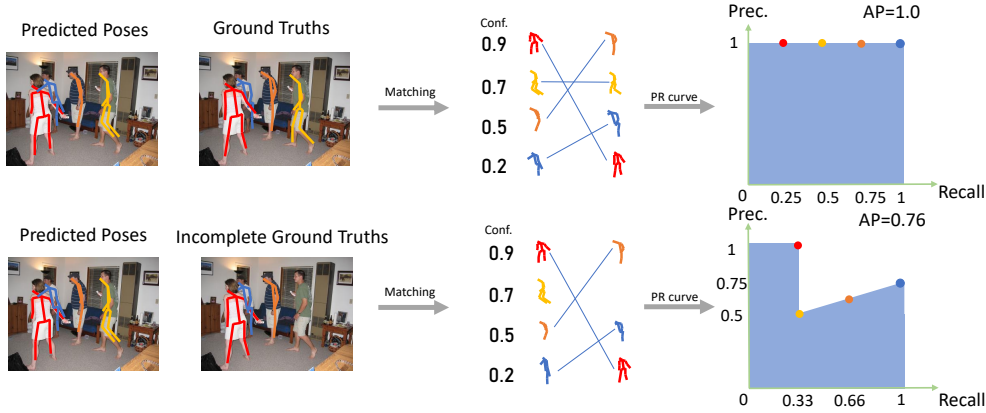
12

Figure 6: An example of the decreasing AP with incomplete ground truths. The first row shows the standard COCO evaluation protocol when every person is annotated in the ground truths and predicted by a pose estimation method. The resulting AP equals to 1.0 (area under the precision-recall curve). However, when one of the ground truths is missing (the yellow pose), the corresponding pose will be treated as a false positive, leading to the lower precision for the yellow pose and the subsequent ones (the yellow, orange, and blue dots in the PR curve). Therefore, the AP decreases to 0.76. Conf. and prec. stand for confidence and precision.

treated as false positives if they are not annotated in the ground truths, resulting in a decrease in the AP value. In particular, an example is provided in Fig. 6 to demonstrate the effect of the incomplete ground truths in AP calculation. In the figure, a standard COCO evaluation protocol is presented, which firstly matches the ground truths with the predicted poses and sorts the poses with their confidence, and then enumerates from the first to the end for the matched poses to calculate the precision and recall. The chart that plots the (precision, recall) pairs forms a precision-recall (PR) curve, where the AP is the area under the PR curve. If a ground-truth pose is missing, the corresponding predicted pose would be considered as a false positive, which decreases every subsequent poses' precision because of the false positive, resulting in the precision to be continuously smaller than one. As a result, the obtained AP (i.e., area under the PR curve) is lower than the one with the complete ground truths. Our method detects a large number of small-scale persons, which are not always annotated in the ground truths, thus resulting in a decrease in our AP values. The situation becomes even worse on the widely used COCO dataset, which contains a great number

of small-scale persons without annotations as shown in Fig. 8. Therefore, simply computing AP cannot properly evaluate our method.

To perform an objective evaluation, bounding box precision (BBP) is used to mitigate the negative influence of the sparse human pose annotations and act as an approximation of AP, where the human bounding box ground truths are mostly available. We also produce the corresponding bounding box recall (BBR) as an indicator of the model's ability to capture persons. In particular, the evaluation of the BBP and BBR follows the standard evaluations of object detection which evaluate the AP and AR of the bounding boxes. Here we set the IoU threshold to be 0.5 which indicates the person is considered as detected if its IoU with the ground truth is larger than 0.5. In other words, the BBP and BBR are AP^{50} and AR^{50} in the standard evaluation of object detection. Furthermore, the evluation process of the BBP can be understood as replacing the human poses with the bounding boxes as shown in the first row of Fig.6 where the matching is decided by the IoU between two bounding boxes in this case. BBP and BBR are used as alternative measures of human pose estimation performance than AP, which are useful when human pose ground truths are missing but human bounding box annotations are available, especially for the small-scale persons, where the ground-truth annotations of the human keypoints are mostly unavailable on the COCO dataset as shown in Fig. 8.

To evaluate the BBP and BBR performance on persons of different scales, COCO validation set is split into small, medium, and large-scale subsets. In particular, a person with bounding box annotation is categorized into one of the three sets based on the bounding box area, where $(0, 64^2]$, $(64^2, 128^2]$, $(128^2, +\infty)$ correspond to small, medium and large, respectively.

Other than splitting COCO and CrowdPose into different subsets for BBP and BBR evaluations of every scale person, we also follow the standard protocol to split COCO test set based on the scale of the persons, where AP^M and AP^L represent the AP values on the subsets of medium-scale and large-scale persons respectively [35]. In addition, we provide the number of parameters and GFLOPs to measure the model size.

Implementation of the Multi-scale Training. For each image in the training, we interpolate it into different scales (256, 512, 1024). We then sequentially input the different scales of images into our model. There will be multiple forward passes with different scales of images in the training stage. This improves a model's ability to handle different scales of the persons. In the end, we gather the losses from the different scales and update the

Method	Center type	COCO validation set				CrowdPose test set			
		S	M	L	All	S	M	L	All
SWAHR [23] (CVPR'21)	-	25.0	73.0	82.2	44.5	38.7*	78.5*	<u>92.2*</u>	<u>71.2*</u>
CGNet [24] (ICCV'21)	single & bounding box	25.3	<u>77.6</u>	**86.4**	<u>46.3</u>	34.8*	<u>79.6*</u>	90.5*	69.4*
DEKR [22] (CVPR'21)	single & bounding box	25.0	71.9	78.3	43.5	<u>41.5</u>	78.0	87.6	69.9
PINet [25] (NeurIPS'21)	multiple & bounding box	<u>27.0</u>	72.6	80.8	45.4	38.2	78.0	89.6	69.7
Ours	dual & anatomical	**35.2**	**80.3**	<u>85.6</u>	**52.9**	**58.4**	**93.2**	**96.4**	**82.9**

Table 1: BBR evaluations on COCO validation set and CrowdPose test set. S, M, L indicate the BBR score for small, medium, large-scale subsets. All means the BBR score for all scales (whole COCO validation set or CrowdPose test set). HRNet-W32 backbone used in all methods for coco validation set. * indicates that the HRNet-W48 backbone is used instead of the HRNet-W32 backbone. Best in **bold**, second best <u>underlined</u>.

Method	Center type	COCO validation set				CrowdPose test set			
		S	M	L	All	S	M	L	All
SWAHR [23] (CVPR'21)	-	<u>22.1</u>	64.6	<u>70.3</u>	<u>38.7</u>	<u>33.6*</u>	67.7*	81.9*	<u>62.1*</u>
CGNet [24] (ICCV'21)	single & bounding box	21.3	<u>64.9</u>	67.9	37.4	30.0*	<u>68.4*</u>	<u>82.4*</u>	61.6*
DEKR [22] (CVPR'21)	single & bounding box	21.0	63.8	68.4	37.6	32.7	66.6	79.7	61.2
PINet [25] (NeurIPS'21)	multiple & bounding box	20.9	65.0	15.9	36.1	30.8	66.0	80.8	60.7
Ours	dual & anatomical	**30.6**	**73.2**	**77.5**	**46.2**	**44.5**	**77.1**	**87.2**	**70.1**

Table 2: BBP evaluations on COCO validation set and CrowdPose test set. S, M, L indicate the BBP score for small, medium, large-scale subsets. All means the BBP score for all scales (whole COCO validation set or CrowdPose test set). HRNet-W32 backbone used in all methods for coco validation set. * indicates that the HRNet-W48 backbone is used instead of the HRNet-W32 backbone. Best in **bold**, second best <u>underlined</u>.

backbone.

4.1. Quantitative Comparisons

Capturing Small-Scale Persons Our method detects more small-scale persons, which are not annotated in the ground truths, thus resulting in a decrease in AP values, which cannot objectively reflect the performance because predicted poses without ground truths are regarded as false positives in the evaluation. Thus, the BBR evaluation is performed on the COCO and CrowdPose datasets to validate the performance of different methods for capturing small-scale persons. The results are shown in Table 1, where our method outperforms all the SOTA methods in small, medium, and all scales.

In particular, columns 3-6 in Table 1 show the BBR values of the SOTA methods and ours on different subsets of the COCO validation set. It is observed that our method achieves a 39.1% improvement (9.9 BBR) on the challenging subset of small-scale persons compared with the SOTA method,

Group	Method	Input	#param.	GFLOPs	AP	AP^{50}	AP^{75}	AP^M	AP^L
Top-down	HRNet-W32 [14]	384×288	-	-	74.9	92.5	82.8	71.3	80.9
	HRNet-W48 [14]	384×288	-	-	75.5	92.5	83.3	71.9	81.5
One-stage	FCPose (ResNet-101) [39]	800	-	-	65.6	87.9	72.6	62.1	72.3
	SPM [40]	-	-	-	66.9	88.5	72.9	62.6	73.1
	Single-scale testing								
Bottom-up	OpenPose [11]	-	-	-	61.8	84.9	67.5	57.1	68.2
	CenterNet-HG [36]	512	-	-	63.0	86.8	69.6	58.9	70.4
	PifPaf [29]	-	-	-	66.7	-	-	62.4	72.9
	PersonLab [20]	1401	-	-	66.5	88.0	72.6	62.4	72.3
	HigherHRNet-W32 [21]	512	28.5	47.9	66.4	87.5	72.8	61.2	74.2
	SWAHR (HRNet-W32) [23]	512	28.6	48.0	<u>67.9</u>	**88.9**	<u>74.5</u>	<u>62.4</u>	75.5
	DEKR (HRNet-W32) [22]	512	29.6	45.4	67.3	87.9	74.1	61.5	**76.1**
	CGNet (HRNet-W32) [24]	512	-	-	67.6	<u>88.6</u>	73.6	62.0	75.6
	PINet (HRNet-W32) [25]	512	-	-	66.7	-	-	-	-
	Ours (HRNet-W32)	512	43.5	46.5	**68.5**	87.8	**75.2**	**63.4**	<u>76.0</u>
	HigherHRNet-W48 [21]	640	63.8	154.3	68.4	88.2	75.1	64.4	74.2
	SWAHR (HRNet-W48) [23]	640	63.8	154.6	<u>70.2</u>	**89.9**	76.9	65.2	<u>77.0</u>
	DEKR (HRNet-W48) [22]	640	65.7	141.5	70.0	89.4	<u>77.3</u>	<u>65.7</u>	76.9
	CGNet (HRNet-W48) [24]	640	-	-	69.5	<u>89.7</u>	76.0	65.0	76.2
	Ours (HRNet-W48)	640	79.8	142.7	**71.0**	89.5	**78.0**	**66.1**	**78.1**
	Multi-scale testing								
Bottom-up	DEKR (HRNet-W32) [22]	512	29.6	45.4	69.8	<u>89.0</u>	76.6	<u>65.2</u>	76.5
	CGNet (HRNet-W32) [24]	512	-	-	**70.3**	**90.0**	<u>76.9</u>	65.4	**77.5**
	Ours (HRNet-W32)	512	43.5	46.5	<u>70.2</u>	88.8	**77.0**	65.4	<u>77.2</u>
	HigherHRNet-W48 [21]	640	63.8	154.3	70.5	89.3	77.2	66.6	75.8
	SWAHR (HRNet-W48) [23]	640	63.8	154.6	**72.0**	**90.7**	**78.8**	**67.8**	<u>77.7</u>
	DEKR (HRNet-W48) [22]	640	65.7	141.5	71.0	89.2	78.0	67.1	76.9
	CGNet (HRNet-W48) [24]	640	-	-	71.4	<u>90.5</u>	78.1	<u>67.2</u>	77.5
	Ours (HRNet-W48)	640	79.8	142.7	<u>71.5</u>	89.1	<u>78.5</u>	67.2	**78.1**

Table 3: Evaluation on COCO test-dev set. Best in **bold**, second best <u>underlined</u> for bottom-up methods. '-' indicates no data available.

CGNet. Moreover, compared with the SOTA regression-based method PINet, we obtain a 30.3% improvement (8.2 BBR) on the COCO small-scale subset. Apart from our superior performance for small-scale persons, clear improvements over the SOTA methods for medium-scale persons and competitive results for large-scale persons are observed as well. All scales combined, our improvement on the COCO validation set is 14.2% (6.6 BBR) against the SOTA, CGNet. The overall increase of BBR is closer to that of the small-scale subset because the COCO dataset contains more small-scale persons, where the ratio of small to medium to large-scale persons is 51:21:29.

Similarly, columns 7-10 in Table 1 show the BBR values on the different subsets of the CrowdPose test set. Again, it is observed that our method demonstrates a large improvement, 50.9% (19.7 BBR), on the challenging subset of small-scale persons compared with the SOTA, SWAHR. Furthermore, a 18.7% (14.7 BBR) and a 4.5% (4.2 BBR) improvements for

Group	Method	Input	#param.	GFLOPs	AP	AP^{50}	AP^{75}	AP^M	AP^L
Top-down	HRNet-W32 [14]	256×192	-	-	74.4	90.5	81.9	70.8	81.0
	HRNet-W32 [14]	384×288	-	-	75.8	90.6	82.5	72.0	82.7
	Single-scale testing								
Bottom-up	CenterNet-HG [36]	512	-	-	64.0	-	-	-	-
	PifPaf [29]	-	-	-	67.4	-	-	-	-
	PersonLab [20]	1401	-	-	66.5	86.2	71.9	62.3	73.2
	HigherHRNet-W32 [23]	512	28.5	47.9	67.1	86.2	73.0	-	-
	SWAHR (HRNet-W32) [23]	512	28.6	48.0	68.9	**87.8**	<u>74.9</u>	<u>63.0</u>	<u>77.4</u>
	DEKR (HRNet-W32) [22]	512	29.6	45.4	68.0	86.7	74.5	62.1	**77.7**
	PINet (HRNet-W32) [25]	512	-	-	67.4	86.8	74.0	62.5	76.3
	CGNet (HRNet-W32) [24]	512	-	-	<u>69.0</u>	87.7	74.4	59.9	75.3
	Ours (HRNet-W32)	512	43.5	46.5	**69.2**	86.4	**75.1**	**63.8**	77.2
	HigherHRNet-W48 [23]	640	63.8	154.3	69.9	87.2	76.1	-	-
	SWAHR (HRNet-W48) [23]	640	63.8	154.6	70.8	<u>88.5</u>	76.8	66.3	77.4
	DEKR (HRNet-W48) [22]	640	65.7	141.5	<u>71.0</u>	88.3	<u>77.4</u>	<u>66.7</u>	<u>78.5</u>
	CGNet (HRNet-W48) [24]	640	-	-	<u>71.0</u>	**88.7**	76.5	63.1	75.2
	Ours (HRNet-W48)	640	79.8	142.7	**72.1**	88.3	**78.2**	**66.9**	**79.6**
	Multi-scale testing								
Bottom-up	HigherHRNet-W32 [23]	512	28.5	47.9	69.9	87.1	76.0	-	-
	SWAHR (HRNet-W32) [23]	512	28.6	48.0	<u>71.4</u>	<u>88.9</u>	<u>77.8</u>	<u>66.3</u>	**78.9**
	DEKR (HRNet-W32) [22]	512	29.6	45.4	70.7	87.7	77.1	66.2	77.8
	CGNet (HRNet-W32) [24]	512	-	-	**71.9**	89.0	**78.0**	63.7	77.4
	Ours (HRNet-W32)	512	43.5	46.5	<u>71.4</u>	87.6	77.5	**66.6**	<u>78.7</u>
	HigherHRNet-W48 [23]	640	63.8	154.3	72.1	<u>88.4</u>	78.2	-	-
	SWAHR (HRNet-W48) [23]	640	63.8	154.6	<u>73.2</u>	**89.8**	<u>79.1</u>	**69.1**	<u>79.3</u>
	CGNet (HRNet-W48) [24]	640	-	-	**73.3**	89.7	**79.2**	66.4	76.7
	DEKR (HRNet-W48) [22]	640	65.7	141.5	72.3	88.3	<u>78.6</u>	<u>68.6</u>	78.6
	Ours (HRNet-W48)	640	79.8	142.7	73.0	88.3	<u>79.1</u>	<u>68.6</u>	**79.8**

Table 4: Evaluation on COCO validation set. Best in **bold**, second best <u>underlined</u> for the bottom-up methods. '-' indicates no data available.

medium-scale and large-scale subset are achieved by our method compared with the SOTA, which indicates our method consistently outperforms the SOTA method at each scale. All scales combined, our improvement over the SOTA on the CrowdPose test set is 16.4% (11.7 BBR). Our improvement on the whole CrowdPose test set includes not only the small-scale persons but also medium and large-scale persons, where the small-scale persons only account for 24% of the whole set. In particular, the ratio of small to medium to large-scale persons in the CrowdPose test set is 24 : 21 : 55.

Anatomical-Center Evaluation We continue to compare the BBP on the COCO and CrowdPose datasets with SOTA methods. BBP exploits exactly the same dataset splits as BBR, while evaluating the precision of the captured human objects, which is an alternative measure to AP when the human pose annotations are largely missing. Similar results are shown in Table 2 , which suggests our anatomical centers outperform the SOTA methods of different center types in every scale of the persons. Column 3-6 in Table 2 shows the

BBP values on the COCO validation set and column 7-10 in Table 2 shows the BBP values on the CrowdPose test set.

To validate the proposed anatomical centers, we further focus on comparing our method with the DEKR [22] and PINet [25], which are the SOTA regression-based methods that exploit different types of bounding box centers to identify human objects. Compared to the single-bounding-box-center method, DEKR, our anatomical centers improve the 22% (8.6 BBP) and 14.5% (8.9 BBP) on the COCO and Crowdpose datasets in all scales. Compared to the multiple-bounding-box-centers method, PINet, our method improves the 27.9% (10.1 BBP) and 15.4% (9.4 BBP) respectively in all scales. We also achieve the best BBP (and also BBR) in small, medium, and large-scale persons when compared with these two methods. In particular, PINet uses three bounding box centers, while our method simply predicts two anatomical centers which intuitively tends to decrease our performance. However, even in this circumstance, we outperform DEKR and PINet at each scale, which clearly demonstrates that our proposed dual anatomical centers are more effective than the single or multiple bounding box centers. A similar result is observed in the AP evaluation in the next section as well.

Apart from DEKR and PINet, we also compared with SWAHR, which is the best-performed existing method in Table 2. It is observed that our method achieves a significant increase of BBP for small-scale persons compared with SWAHR, 38.4% (8.5 BBP) on the COCO validation set and 32.4% (10.9 BBP) on the CrowdPose test set. It is worth noting that the reported numbers of CGNet and SWAHR on the CrowdPose dataset are based on their models with HRNet-W48 backbone, where our results are based on a smaller backbone, HRNet-W32, to be comparable to DEKR and PINet. Even with such a disadvantage, we still outperform both CGNet and SWAHR by a large margin.

Overall Performance Apart from the bounding box recall and precision evaluations at different scales of persons, we also provide the AP comparisons of the overall performance evaluation on the COCO validation and test-dev set as shown in Table 3 and Table 4. Note that the AP values cannot objectively reflect our performance because our method detects a large number of small-scale that are not annotated, which are counted as false positives in AP evaluation. However, even in this circumstance, benefiting from the proposed multi-scale training, our method outperforms existing bottom-up methods consistently in single-scale testing on both COCO validation set and test set. In particular, we achieve 1.1 improvement of AP with HRNet-W48

Metric	AE [12]	HGG [41]	DEKR [22]	SWAHR [23]	CGNet [24]	PINet [25]	Ours
AP	29.5	34.8	36.4	39.5	38.6	37.2	**40.3**

Table 5: Quantitative comparison on OCHuman test set. All results are based on single-scale testing with flipping tests. HRNet-W32 backbone used in all methods.

backbone in Table 4, resulting in 1.5% increase against the SOTA method [22] and 0.2 improvement of AP compared to SOTA with HRNet-W32 backbone. In test-dev set shown in Table 3, our improvement is 0.6 AP over SWAHR's with HRNet-W32 backbone, which is significant by comparing with the latest SOTA method, where PINet's improvement over HigherHR-Net (the previous SOTA) is 0.3 AP with HRNet-W32 backbone. The 0.6 AP increase is significant but not incremental given the fact that 2D pose estimation has been extensively studied for many years.

Specifically, we achieve 71.0 AP with HRNet-W48 backbone in Table 3, which is a 1.1% increase against the SOTA method, SWAHR. The increase should be considered together with the BBR evaluation given the missing annotations in the ground truths, where SWAHR's BBR is 44.5 and ours is 52.9 on the COCO validation set, another 11.6% improvement is achieved by our method compared with SWAHR, the SOTA method focused on small-scale persons handling. Besides single-scale testing, we also conducted the multi-scale testing with different backbones in Table 3 and Table 4. We can also achieve comparable performance compared to SOTA methods even if our method is designed for enhancing the single-scale testing.

Performance of Cross-Dataset Generalization We perform a quantitative comparison on the OCHuman dataset [37], which focuses on the crowded scenes with human interactions. Since OCHuman is a new dataset, where several SOTA methods do not report their performance and do not release their pre-trained models on it [22, 23], we follow the evaluation of PINet [25] to perform testing on OCHuman test set with the models trained on the COCO dataset. Since the OCHuman training set is not used, this comparison evaluates the cross-dataset generalization capability of different methods as shown in Table 5. We observe that our method outperforms all the existing methods and our superior performance shows our method has better generalization capability compared with other SOTA methods in the crowded scenes with occlusions.
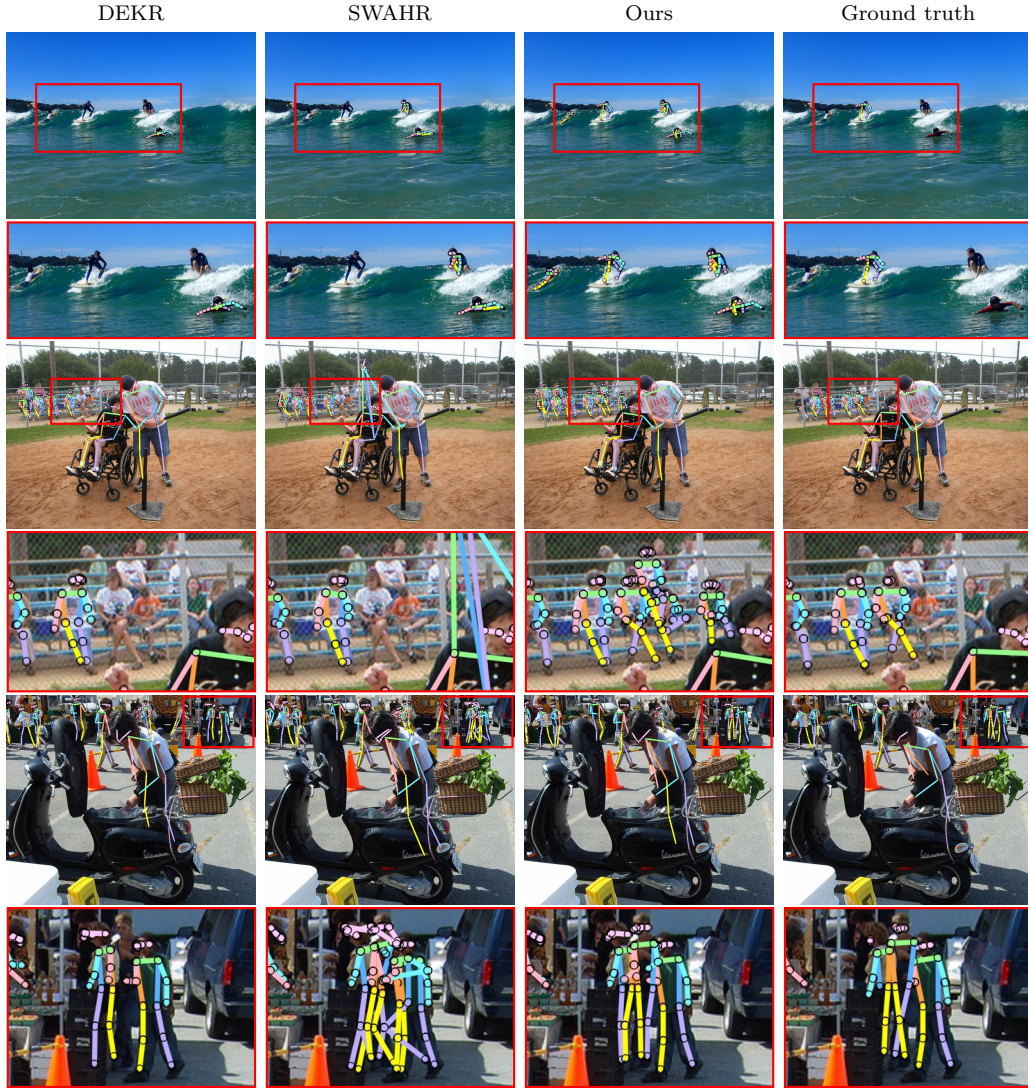
Figure 7: Qualitative comparisons with the SOTA methods (DEKR and SWAHR) on the COCO validation set. Odd rows show the result on the input image, where red bounding box indicates the zoom-in area. Even rows show the enlarged zoom-in area.

4.2. Qualitative Evaluations

Fig. 7 shows the qualitative results of our method compared with DEKR [22] and SWAHR [23]. The first two rows show a surfing image with four small-scale persons, where DEKR predicts one person and SWAHR predicts two. In contrast, our method correctly predicts the poses for all the four
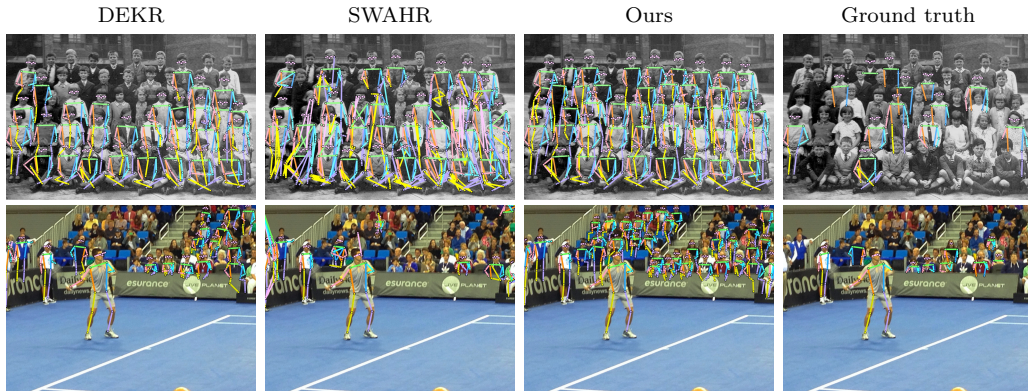
Figure 8: Qualitative comparison on the COCO validation set with the SOTA methods: DEKR [22] and SWAHR [23] in crowd scenes. Each row shows the pose estimation result of each method on the whole image.

persons, while only one person is annotated in the GT. The 3rd and 4th rows show a sports scene with large to small-scale persons. In the zoom-in view, we can see that both DEKR and SWAHR predict two poses, GT has three persons annotated, and our method predicts the three persons correctly and three additional persons. The last two rows show a third example that our method can detect small-scale persons robustly and avoid missing or duplicate predictions.

Fig. 8 shows additional qualitative results in crowd scenes, where many persons are highly overlapped in both examples and most of the persons are small-scale in the second row. Compared to the SOTA methods, our proposed method can produce accurate pose estimation in the crowd scenes. Note that the GT annotation is sparse, where many persons that our method predicts do not have corresponding GT.

4.3. Ablation Studies

The effectiveness of major components of our method is validated in Table 6. We remove all the proposed components from our method as a baseline approach, which uses a single bounding box center for human detection with non-maximum suppression (NMS) to filter out duplicate human poses. The baseline is based on single-scale testing without flipping as stated in Table 6, where flipping is not used to remove its influence on performance (not from proposed components). For example, DEKR's AP drops from 68.0 to 62.3 when flipping is removed.

Method	AP	AP^M	AP^L
baseline	62.4	56.2	<u>74.1</u>
+ two centers	61.3	55.8	71.2
+ multi-scale training	62.8	57.1	73.0
+ pose confidence	63.6	57.4	73.5
+ pose similarity	<u>64.0</u>	<u>57.7</u>	74.0
+ pose refinement	**64.4**	**58.1**	**74.3**

Table 6: Ablation study on COCO validation set. Best in **bold**, second best <u>underlined</u>. Numbers based on single-scale testing without flipping.

On top of the baseline, we add each proposed component at a time and show the resulting performance. It is observed that most of the components help to improve the overall accuracy, where the multi-scale testing brings the largest improvement (1.5 AP). Pose confidence ranks the 2nd in terms of AP improvement, which helps to increase AP by 0.8. Note that in the second row of Table 6, the result of '+ two centers' reduces the AP compared with the baseline, since the average of two centers' confidence is used as the pose confidence to replace our proposed pose confidence for ablation study purposes. Once the pose confidence is added, the AP is improved and better than the baseline.

Validation of the Dual Centers Fig. 9 (left) shows a comparison of the proposed dual centers against the baseline (one bounding box center) used in an existing approach [22], where the AP and recall changes over the baseline are provided. The recall measures the percentage of persons detected. In particular, the statistics are computed based on OKS larger than 0.5, if a pose has an OKS value larger than 0.5 compared to a ground truth, it is counted as a detected person, thus increasing the recall value. The AP and recall of the baseline are 62.4 and 89.5. It is observed that our dual centers approach outperforms the baseline (0.8 recall), which means our approach captures 0.8% more annotated persons that is 51 persons out of 6352 annotated in COCO val set. In fact, our method detects many more persons that are not annotated as shown in Fig. 1.

The effect of using different numbers of the anatomical centers is explored as well in Fig. 9 (left), where we show the results of using either one of the two centers or three centers (the third one is the left shoulder). Using either one of the two centers is worse than the proposed dual-centers approach, and the result of three centers marginally improves recall but hurts the AP. Such evidence supports our choice of using dual anatomical centers.
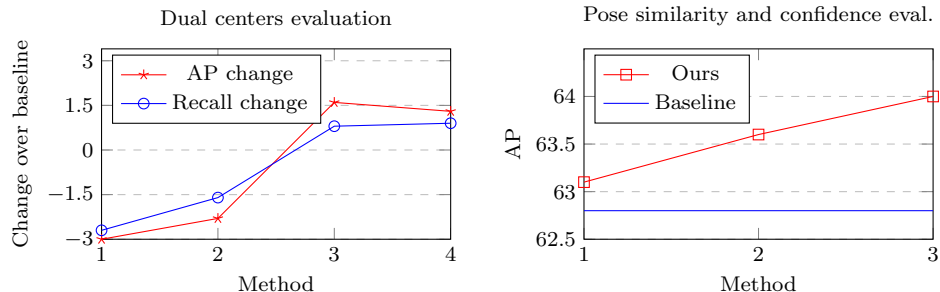
Figure 9: Evaluations of two proposed components on COCO validation set. Left: evaluation of the dual centers approach; Method 1-4: head center only, body center only, the proposed dual centers, and three centers. Right: evaluation of the proposed pose similarity and confidence; Method 1-3: baseline + pose similarity, baseline + pose confidence, baseline + both. Same testing condition as Table 6 in both evaluations.

Pose Similarity and Confidence The effectiveness of the proposed pose similarity and pose confidence is validated in Fig. 9 (right). The baseline's AP is 62.8, where the average of two centers' confidence is used as pose confidence and the poses are grouped and selected based on all joints' average Euclidean distance. On top of the baseline, we use our proposed pose similarity to group the poses, referred as Method 1. Only using the proposed pose confidence without the proposed pose similarity is Method 2, and using both is Method 3, where our pose similarity and confidence bring 1.2 AP improvement against the baseline.

4.4. Limitations and Failure Cases

Our method may have problems processing out-of-boundary large-scale persons. If one or several large-scale persons are present and more than half of the body of them is out of the image boundary, our method could fail. The reason lies in the proposed dual-centers approach that tries to find the head and body centers for a person. A few failure cases are shown in Fig. 10 and we plan to solve this problem in the future.

5. Conclusion

We propose a novel bottom-up multi-person pose estimation method and focus on the performance for small-scale persons. First, we introduce multi-scale training to improve the accuracy of small-scale persons with single-scale testing. Second, our dual anatomical center approach allows the human poses
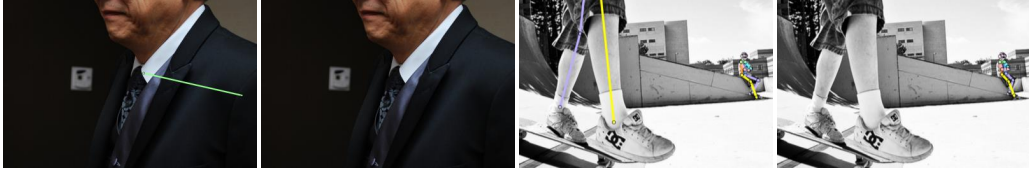
Figure 10: Failure cases of our method on COCO validation set. Column 1, 3: the ground truth, column 2, 4: results of our method.

to be estimated accurately and reliably due to our proposed human centers being anatomically defined and consistent to visual context. Altogether, our method demonstrates its superior performance on the BBP, BBR, and AP metrics across different public datasets, especially in the challenging small-scale-person scenarios.

Acknowledgment

References

[1] S. Yan, Y. Xiong, D. Lin, Spatial temporal graph convolutional networks for skeleton-based action recognition, in: Thirty-second AAAI conference on artificial intelligence, 2018.

[2] L. Ma, X. Jia, Q. Sun, B. Schiele, T. Tuytelaars, L. Van Gool, Pose guided person image generation, in: Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, 4-9 December 2017, Long Beach, CA, USA, 2017, pp. 405–415.

[3] C.-Y. Weng, B. Curless, I. Kemelmacher-Shlizerman, Photo wake-up: 3d character animation from a single photo, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 5908–5917.

[4] H. Dong, X. Liang, X. Shen, B. Wang, H. Lai, J. Zhu, Z. Hu, J. Yin, Towards multi-pose guided virtual try-on network, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019, pp. 9026–9035.

[5] Y. Desmarais, D. Mottet, P. Slangen, P. Montesinos, A review of 3d human pose estimation algorithms for markerless motion capture, Computer Vision and Image Understanding (2021) 103275.

[6] A. Toshev, C. Szegedy, Deeppose: Human pose estimation via deep neural networks, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2014, pp. 1653–1660.

[7] T. Pfister, J. Charles, A. Zisserman, Flowing convnets for human pose estimation in videos, in: Proceedings of the IEEE international conference on computer vision, 2015, pp. 1913–1921.

[8] A. Newell, K. Yang, J. Deng, Stacked hourglass networks for human pose estimation, in: European conference on computer vision, Springer, 2016, pp. 483–499.

[9] L. Pishchulin, E. Insafutdinov, S. Tang, B. Andres, M. Andriluka, P. V. Gehler, 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, 2016, pp. 4929–4937.

[10] E. Insafutdinov, L. Pishchulin, B. Andres, M. Andriluka, B. Schiele, Deepercut: A deeper, stronger, and faster multi-person pose estimation model, in: European Conference on Computer Vision, Springer, 2016, pp. 34–50.

[11] Z. Cao, T. Simon, S.-E. Wei, Y. Sheikh, Realtime multi-person 2d pose estimation using part affinity fields, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 7291–7299.

[12] A. Newell, Z. Huang, J. Deng, Associative embedding: End-to-end learning for joint detection and grouping, Advances in Neural Information Processing Systems 30 (2017).

[13] Z. Cao, G. Hidalgo, T. Simon, S.-E. Wei, Y. Sheikh, Openpose: realtime multi-person 2d pose estimation using part affinity fields, IEEE transactions on pattern analysis and machine intelligence 43 (2019) 172–186.

[14] K. Sun, B. Xiao, D. Liu, J. Wang, Deep high-resolution representation learning for human pose estimation, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 5693–5703.

[15] K. He, G. Gkioxari, P. Dollár, R. Girshick, Mask r-cnn, in: Proceedings of the IEEE international conference on computer vision, 2017, pp. 2961–2969.

[16] Y. Chen, Z. Wang, Y. Peng, Z. Zhang, G. Yu, J. Sun, Cascaded pyramid network for multi-person pose estimation, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2018, pp. 7103–7112.

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

[18] Z. Tian, H. Chen, C. Shen, Directpose: Direct end-to-end multi-person pose estimation, arXiv preprint arXiv:1911.07451 (2019).

[19] M. Kocabas, S. Karagoz, E. Akbas, Multiposenet: Fast multi-person pose estimation using pose residual network, in: Proceedings of the European conference on computer vision (ECCV), 2018, pp. 417–433.

[20] G. Papandreou, T. Zhu, L.-C. Chen, S. Gidaris, J. Tompson, 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), 2018, pp. 269–286.

[21] B. Cheng, B. Xiao, J. Wang, H. Shi, T. S. Huang, 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, 2020, pp. 5386–5395.

[22] Z. Geng, K. Sun, B. Xiao, Z. Zhang, J. Wang, Bottom-up human pose estimation via disentangled keypoint regression, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 14676–14686.

[23] Z. Luo, Z. Wang, Y. Huang, L. Wang, T. Tan, E. Zhou, Rethinking the heatmap regression for bottom-up human pose estimation, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 13264–13273.

[24] G. Brasó, N. Kister, L. Leal-Taixé, The center of attention: Center-keypoint grouping via attention for multi-person pose estimation, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 11853–11863.

[25] D. Wang, S. Zhang, G. Hua, Robust pose estimation in crowded scenes with direct pose-level inference, Advances in Neural Information Processing Systems 34 (2021).

[26] J. J. Tompson, A. Jain, Y. LeCun, C. Bregler, Joint training of a convolutional network and a graphical model for human pose estimation, Advances in neural information processing systems 27 (2014) 1799–1807.

[27] V. Belagiannis, A. Zisserman, Recurrent human pose estimation, in: 2017 12th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017), IEEE, 2017, pp. 468–475.

[28] S.-E. Wei, V. Ramakrishna, T. Kanade, Y. Sheikh, Convolutional pose machines, in: Proceedings of the IEEE conference on Computer Vision and Pattern Recognition, 2016, pp. 4724–4732.

[29] S. Kreiss, L. Bertoni, A. Alahi, Pifpaf: Composite fields for human pose estimation, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019.

[30] R. Girshick, J. Donahue, T. Darrell, J. Malik, Rich feature hierarchies for accurate object detection and semantic segmentation, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2014, pp. 580–587.

[31] R. Girshick, Fast r-cnn, in: Proceedings of the IEEE international conference on computer vision, 2015, pp. 1440–1448.

[32] S. Ren, K. He, R. Girshick, J. Sun, Faster r-cnn: Towards real-time object detection with region proposal networks, Advances in neural information processing systems 28 (2015) 91–99.

[33] X. Nie, J. Feng, J. Xing, S. Yan, Pose partition networks for multi-person pose estimation, in: Proceedings of the european conference on computer vision (eccv), 2018, pp. 684–699.

[34] F. Wei, X. Sun, H. Li, J. Wang, S. Lin, Point-set anchors for object detection, instance segmentation and pose estimation, in: European Conference on Computer Vision, Springer, 2020, pp. 527–544.

[35] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, C. L. Zitnick, Microsoft coco: Common objects in context, in: European conference on computer vision, Springer, 2014, pp. 740–755.

[36] X. Zhou, D. Wang, P. Krähenbühl, Objects as points, arXiv preprint arXiv:1904.07850 (2019).

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

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

[39] W. Mao, Z. Tian, X. Wang, C. Shen, Fcpose: Fully convolutional multi-person pose estimation with dynamic instance-aware convolutions, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 9034–9043.

[40] X. Nie, J. Feng, J. Zhang, S. Yan, Single-stage multi-person pose machines, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019, pp. 6951–6960.

[41] S. Jin, W. Liu, E. Xie, W. Wang, C. Qian, W. Ouyang, P. Luo, Differentiable hierarchical graph grouping for multi-person pose estimation, in: European Conference on Computer Vision, Springer, 2020, pp. 718–734.