SmallPose: A Large-Scale Benchmark for Small Human Pose Estimation

Supplementary Material

8. Comparison to Other Datasets

8.1. Scene Complexity

In this section, we evaluate the scene complexity of SmallPose, CrowdPose and COCO dataset from three key perspectives: scene category entropy, distribution of number of humans and crowding level. The results show that SmallPose has the highest scene complexity, demonstrating greater diversity and challenges.

Scene Category Entropy. To evaluate scene complexity, we analyzed the mean and variance of scene category entropy. The mean value of entropy measures the average uncertainty or randomness in the predicted scene category distributions, while the variance value reflects fluctuations in scene discriminability. Higher values for both metrics indicate greater scene complexity.

Specifically, we used a ResNet model trained on the Place365 [34] dataset to classify images in the SmallPose, COCO, and CrowdPose datasets. The entropy was calculated for each image based on the predicted probability distribution, and the mean and variance value were computed for each dataset. As shown in Table 4, SmallPose exhibits the highest mean and variance value, indicating higher uncertainty in predicted scene distributions and greater fluctuations in scene discriminability compared to other datasets. In contrast, CrowdPose shows the lowest scene complexity, with the lowest mean entropy reflecting minimal uncertainty in predictions, yet significant variance indicating that certain images remain challenging to classify. COCO falls in between, with moderate entropy and variability, suggesting a balanced level of uncertainty in predictions and moderate differences in scene discriminability.

Distribution of Number of Humans. We analyzed the distributions of the number of humans across the SmallPose, COCO, and CrowdPose datasets. Generally, a higher number of humans in an image present greater challenges for human pose estimation models. Thus, by comparing these distributions, we can make a comparison of the scene complexity among these datasets.

Table 4. The mean and variance of the entropy of class probability distributions were calculated for the CrowdPose, COCO, and SmallPose datasets using the ResNet model trained on Place365.

Dataset	Mean of Entropy	Variation of Entropy
CrowdPose [13]	2.42	1.47
COCO [16]	2.61	1.33
SmallPose	2.85	1.48

The distributions of these datasets are depicted in the Figure 7. The SmallPose exhibits the highest complexity, encompassing a full range of the distribution from small to large groups (up to 80 people). The COCO has the lowest complexity, primarily characterized by single-person scenes. It contains a limited number of samples with more than 10 people and is predominantly focused on simpler scenarios The CrowdPose includes a range of multi-person scenarios with 2 to 6 individuals, and relatively few scenarios with more than 10 people, indicating moderate complexity that is intermediate between SmallPose and COCO.

Crowding Level. The crowding level is analyzed using the CrowdIndex metric in Section 3.2 of the main paper. SmallPose shows a moderate level of crowding. It is higher than that of COCO but lower than CrowdPose, which focuses on crowded scenes.

In summary, SmallPose is the dataset with the highest scene complexity, characterized by a moderate level of crowding, a broad distribution of the number of humans in images, and a higher frequency of images with a larger number of humans. Furthermore, the variance of entropy in this dataset indicates that the fluctuation in scene complexity is significantly higher than in other datasets, reflecting greater diversity and challenge.

8.2. The Differences with CrowdPose

To justify the necessity of creating SmallPose, we further conducted analysis on the CrowdPose which is also constructed by integrating multiple datasets and is a highly representative complex scene dataset.

Our analysis revealed several key differences between the CrowdPose and SmallPose datasets. First, as shown in Figure 8(a), large human instances dominate the CrowdPose dataset, while SmallPose focuses primarily on small human instances, making it specifically designed for small target scenarios. Second, as shown in Figure 8(b), the distribution of human body area in the CrowdPose dataset is incomplete, with no samples in the $[0, 32 \times 32]$ area range, whereas SmallPose offers a more comprehensive distribution. Lastly, as depicted in Figure 8(c), our statistical analysis based on keypoint density revealed that the CrowdPose dataset is deficient in dense small human samples, even though such small targets frequently occur in real-world scenarios.

8.3. The Impact of Dataset Scale on Model Performance

The SmallPose dataset is a large-scale benchmark, even larger than the COCO dataset. To assess the impact of

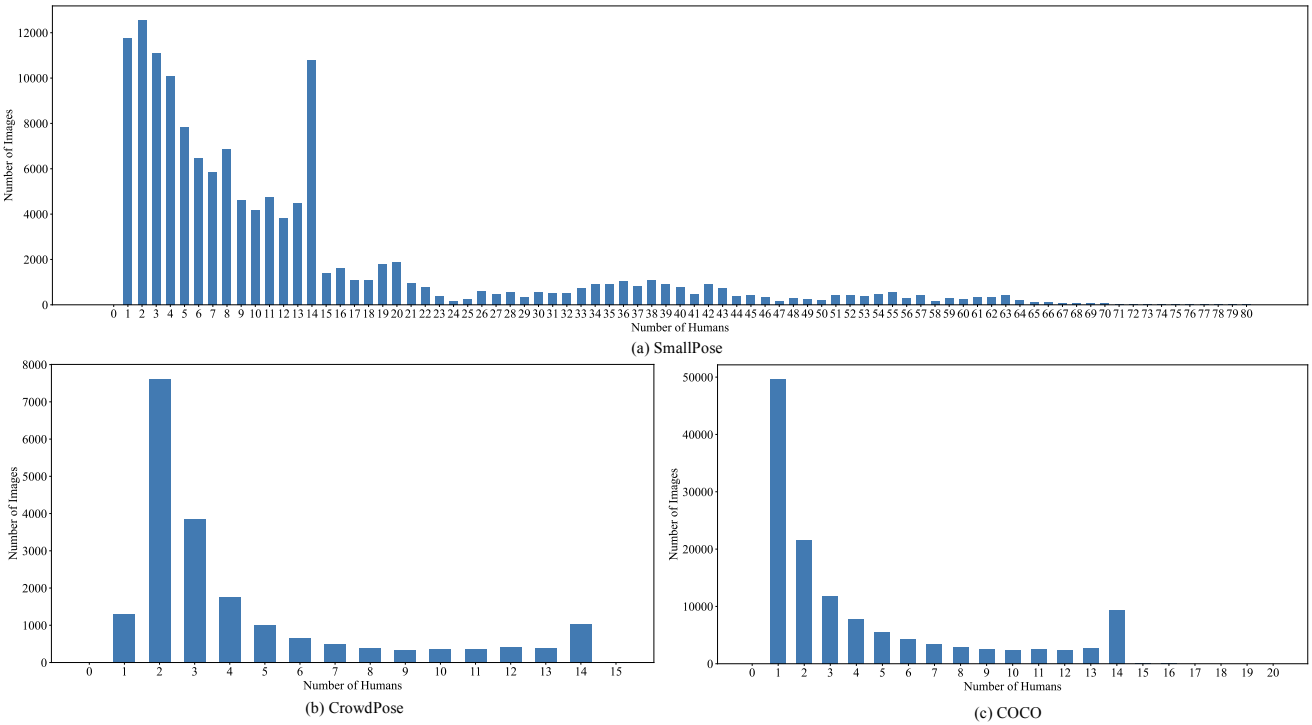

Figure 7. The distribution of the number of humans in (a) SmallPose, (b) CrowdPose and (c) COCO dataset.

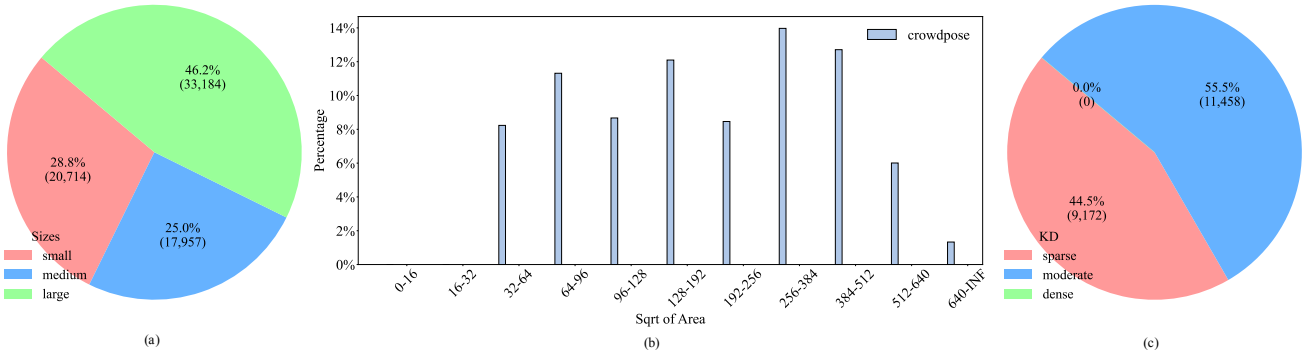

Figure 8. (a) The distribution of the number of large, medium, and small human instances in the CrowdPose dataset. (b) The distribution of human body area in the CrowdPose dataset. (c) The distribution of the number of sparse, moderate and dense small human instances in the CrowdPose dataset.

dataset scale on model performance, we designed ablation experiments focused on dataset scale. We reduced the number of human instances in the training set of SmallPose to match that of the COCO training set. In order to avoid deviation caused by random sampling, we created five sub-training sets from SmallPose and trained the SimpleBaseline model on each one separately. At the same time, we ensured the distribution of human body areas in the sub-training set was consistent with that of the original SmallPose training set All trained models will be tested on the SmallPose testing set, and the experimental results are presented in Table 5.

It is clear that increasing the dataset scale can improve model performance. For example, the SimpleBaseline trained on the original SmallPose training set has 1.2 higher AP and 2.5 higher AP^S compared to the average performance on five sub-training sets, though these improvements are slight. When isolating the effect of dataset scale, we observe that the model trained on the sub-training set, still outperforms the model trained on the COCO dataset. For instance, the mean AP and mean AP^S of the SimpleBaseline models trained on 5 sub-training sets are 16.8 and 31

Table 5. Ablation study on dataset scale. All models are trained on the different training sets (COCO, SmallPose and Subset1-5 from SmallPose), but tested on the SmallPose testing set.

Model	Backbone	Training Set	AP	AP^S	AP^M	AP^L	AP^S_{Sp}	AP^S_{Mo}	AP^S_{De}
SimpleBaseline [30]	ResNet-50	COCO [16]	43.2	32.4	50.5	77.3	8.8	33.9	0.1
SimpleBaseline [30]	ResNet-50	SmallPose	68.2	65.9	69.1	79.7	32.8	69.4	26.0
SimpleBaseline [30]	ResNet-50	Subset1	67.0	63.5	69.0	81.3	32.6	66.6	19.1
SimpleBaseline [30]	ResNet-50	Subset2	67.1	63.4	69.1	81.6	31.9	66.6	18.0
SimpleBaseline [30]	ResNet-50	Subset3	67.1	63.3	69.4	81.6	31.8	66.6	19.4
SimpleBaseline [30]	ResNet-50	Subset4	67.0	63.5	69.1	81.2	32.2	66.5	17.7
SimpleBaseline [30]	ResNet-50	Subset5	67.0	63.4	69.2	81.3	32.8	66.5	19.7

higher than those trained on COCO.

The experimental results demonstrate that while data scale does a certain impact on model performance, the improvement is relatively limited. Instead, the SmallPose, better aligned with the target distribution of small human scenarios, significantly enhances model performance compared to the COCO dataset. This indicates that simply increasing the amount of data is not the most effective way to improve model performance, and the design of the dataset and its alignment with the target scene are more crucial for improving model performance.

9. Dataset Documentation

9.1. Detailed Collection Process

Our main goal for the SmallPose dataset is to include a wide variety of small and medium-sized people across diverse environments. We want to foster the development of robust human pose estimation models that can handle the challenges of various body sizes in real-world scenarios

To achieve this goal, we meticulously selected five popular and challenging datasets: COCO [16], MPII [1], AIC [29], HiEve [17], and PoseTrack [2]. After thorough verification, we confirmed that these five datasets allow for redistribution, modification, and adjustment of the data. For HiEve, since the original training set was composed of video data, we uniformly sampled frames and their corresponding annotations to create the HiEve training set. Additionally, as MPII and HiEve did not have publicly available validation sets, we split their training sets in 8:2 ratio to create validation sets.

During the data collection phase, we devised a novel criterion based on the bounding box area in Section 4, as opposed to the segmentation area used in COCO, to filter small, medium, and large human instances across these selected datasets. To ensure a predominance of small instance samples in the SmallPose dataset, we aggregated all images containing small human instances from the source datasets, and also randomly included some medium and large human instances to ensure comprehensive coverage of diverse body scales. The final SmallPose dataset not only encompasses small human instances across simple to complex scenarios but also includes medium and large human instances, effectively reflecting the various body sizes encountered in real-world situations.

9.2. Annotation Format

The SmallPose dataset follows the COCO dataset format, with the annotations provided in JSON files. This allows researchers to utilize the SmallPose dataset in a similar manner as the widely-adopted COCO dataset.

9.3. Dataset URL

We will provide an anonymous link to access the SmallPose dataset: `http://gofile.me/7aWgk/gra34NfSC`. According to this anonymous link, you will be able to download the training and test sets of the SmallPose dataset, as well as the annotation files used in our experiments.