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

Anonymous CVPR submission

Paper ID 9341

Abstract

Various sizes of human body in the wild pose significant challenges for human pose estimation. A robust human pose estimation model must possess not only cross-scenario generalizability but also the capability for stable predictions across different instance sizes. From the perspective of data-driven model performance, while prior dataset works have mainly focused on scene complexity, we argue that even in simple scenarios, human body size is also a key factor affecting human pose estimation, with small human instances being particularly critical. To fill this gap, we present ***SmallPose****, the first large-scale dataset that specifically targets the human body size with a focus on the challenge of small-size instances. Additionally, to thoroughly assess model performance on small instances, we introduce the concept of* ***keypoint density*** *and establish three novel fine-grained evaluation metrics based on it. By employing a systematic analysis with these metrics on SmallPose, we dissect the efficacy of current pose estimation methodologies across various design paradigms and network architectures. Experimental results demonstrate that models trained on SmallPose not only improve in performance on small instances but also exhibit increased robustness. The SmallPose dataset and the models of previous works optimized on it will be released soon.*

1. Introduction

Recent advancements in human pose estimation have been substantially driven by deep neural networks[21, 24, 25, 28, 30, 31] trained on large-scale datasets[1, 2, 13, 16, 17, 29]. However, the effectiveness of these methods is often constrained by the complexity and diversity inherent in real-world scenarios. In practical applications such as autonomous driving[22, 33] and video surveillance[5, 6, 23], the challenge of accurately estimating human poses is further amplified due to the variability in human body sizes. This is because the human instances of varying sizes makes it challenging for networks to maintain consistent performance. In addition to strategies such as network customization and training paradigm selection, the characteristics of the dataset significantly influence the model's performance and robustness.

However, widely used datasets such as MPII[1] and COCO[16] primarily focus on relatively simple everyday scenarios, while datasets like CrowdPose[13] and HiEve[17] concentrate on crowded or complex event scenes. Although these datasets include small human instances, the number of such instances is relatively low. Moreover, they lack appropriate metrics to evaluate model performance specifically on small-sized human instances, leading to a gap in assessing and improving models for these cases.

In this paper, we have conducted a thorough analysis of existing datasets, leading to the formulation of **SmallPose**, the first large-scale dataset dedicated to the diverse sizes of human instances, with an emphasis on small instances. Figure 1(b) depicts several samples from our dataset that showcase the range of human sizes across diverse real-world settings. Notably, SmallPose also encompasses a broad range of scenarios from simple to complex, but its primary aim is to highlight the ubiquity of small instances across diverse settings. We have also formulated a new taxonomy for small instances, defined by area thresholds, and proposed the performance metric specifically for small human instances. Additionally, we observed that small human instances, such as those where only the human head is visible, pose unique challenges for evaluation due to the incomplete representation of the human body. The coarse-grained metrics are inadequate in these cases, as keypoint density can vary significantly, affecting the estimation difficulty. To address this, we introduced the concept of **keypoint density**. This concept is delineated into three gradations: sparse, moderate, and dense, which categorize the density levels of small human poses. Importantly, the instances with lower keypoint density do not necessarily mean these are easier to estimate. It may contains partial bodies, which may increases the challenge in terms of recognition and localization. We've also established several fine-grained metrics based on it to evaluate model performance on small instances. Employing the SmallPose dataset and novel metrics, we conducted

Figure 1. (a) The detailed breakdown of the SmallPose dataset composition. (b) The visualization of samples in the SmallPose dataset. Our dataset covers a wide range of human sizes from small to large, focusing on the ubiquity of small human bodies. (c) The proposed new human body skeleton. It is more comprehensive than the skeleton of the existing dataset, including 21 joints.

a comparative analysis of diverse models. The experimental results suggest that models trained on SmallPose exhibit enhanced performance on small instances, with improved robustness. Our analysis further reveals the subtle interplay between paradigm choice, network architecture, and the specific demands of human pose estimation, particularly in scenarios involving human instances of different sizes. We summarize the contributions as follows:

1. We introduce **SmallPose**, the first dataset specifically designed to address the challenge of varying human instance sizes, aimed at advancing the robustness of human pose estimation.
2. We propose the novel concept of **keypoint density** and establish three new fine-grained metrics based on it to assess model performance in the real world.
3. We systematically analyzed existing human pose estimation methods using SmallPose and new evaluation metrics, revealing the impact of different network designs and paradigms.
4. The SmallPose dataset and the models of previous works optimized on it will be released soon.

2. Related Work

2.1. 2D Human Pose Estimation Dataset

In the nascent stages of human pose estimation, datasets such as LSP[12], FashionPose[7], PASCAL Person Layout[8], and J-HMDB[11] were pivotal, focusing on single-person pose estimation within controlled environments. These foundational datasets laid the groundwork for understanding and predicting human pose in simplified settings. With the development of this field, the demand for estimating poses in multi-person scenarios has become apparent, leading to the development of datasets like MPII[1], COCO[16], and AI Challenger[29]. These datasets further enhance the real-world applicability of pose estimation models. Afterward, the advent of crowded scenes introduced a new set of challenges, prompting the creation of datasets such as PoseTrack[2], HiEve[17], and CrowdPose[13], designed to accurately identify and track human bodies under challenging conditions. While these datasets have propelled the field forward, a truly robust human pose estimation model demands not only cross-scenario generalizability but also stable predictions across a range of human sizes. To this end, we have introduced SmallPose, the first large-scale dataset to focus on body size. SmallPose encompasses human bodies of various sizes from simple to complex scenes, highlighting the ubiquity of small human instances across diverse settings. Empirical evidence demonstrates that models trained on SmallPose exhibit increased robustness.

2.2. 2D Human Pose Estimation Methods

The field of human pose estimation has progressed through three primary categories: top-down, bottom-up, and single-stage paradigms. The top-down paradigm[9, 24, 30, 31] has been dominant in public benchmarks, utilizing pre-trained detectors to delineate individuals with bounding boxes, simplifying the task by scaling the human body to a uniform size. In contrast, the bottom-up paradigm[3, 4, 9, 19] shifts the focus to keypoint detection, independent of specific instances. By identifying and aggregating keypoints across the entire image, it reconstructs human poses by emphasizing the distribution of keypoints rather than individual bodies. The single-stage paradigm[18, 20, 27, 35] embodies a simplified and efficient methodology, conducting multi-person pose estimation in a single forward pass. This approach directly regresses keypoints from a predefined root,

	Image	Instance			
		Small	Medium	Large	Total
Train	68,311	351,722	225,950	49,453	627,125
Test	10,296	75,841	53,869	12,609	142,319
Total	78,607	427,563	279,819	62,062	769,444

Table 1. The number of images and the numbers for small, medium and large instances in each split.

aiming to enhance efficiency and speed without compromising accuracy. Leveraging the SmallPose dataset, we evaluated multi-person pose estimation methods grounded in these three dominant design paradigms. Our analysis highlights the importance of aligning the chosen paradigm with the network architecture to effectively tackle the unique challenges presented by diverse body sizes in human pose estimation.

3. Dataset

3.1. Dataset Organization and Description

Data Collection. To construct our dataset, we choose the COCO[16], MPII[1], AIC[29], HiEve[17], and Posetrack[2] datasets as the foundational sources. Upon our verification, we confirmed that these five datasets permit distribution, modification, and adjustment of the data. During the phase of data collection, we devised a novel criterion based on the bounding box area, as opposed to the segmentation area used in COCO[16], to identify and filter samples containing small, medium, and large human instances from these datasets. This strategic approach enabled us to amass a substantial collection of images containing small human instances while ensuring comprehensive coverage of the entire spectrum of body sizes, culminating in the creation of the SmallPose dataset. Figure 1(a) provides a detailed breakdown of the dataset composition.

Data Annotation. In the process of annotating human poses for the SmallPose dataset, we encountered the challenge of inconsistent label formats across existing datasets. Specifically, the number of keypoints annotated varied, with MPII[1] having 16 keypoints, AIC[29] and HiEve[17] using 14 keypoints, and PoseTrack[2] and COCO[16] employing 17 keypoints. To address this challenge, we conducted a thorough analysis and proposed a comprehensive human keypoint annotation scheme that integrated the label formats from the existing datasets. This unified scheme encompasses a total of 21 keypoints, with the following numbering and corresponding body parts: 1-nose, 2-head top, 3-left eye, 4-right eye, 5-left ear, 6-right ear, 7-head bottom, 8-chest, 9-left shoulder, 10-right shoulder, 11-left elbow, 12-right elbow, 13-left wrist, 14-right wrist, 15-left hip, 16-right hip, 17-pelvis, 18-left knee, 19-right knee, 20-left ankle, and 21-right ankle. Additionally, we designed a new set of body skeleton connection rules to more accurately capture the human body structure, as illustrated in Figure 1(c).

3.2. Dataset Statistics

Dataset Size. Our SmallPose dataset comprises over 70k images and more than 700k human instances, with a significant portion of over 420k small instances, substantially exceeding the scale of other existing datasets. Specifically, as detailed in Table 1, the training set consists of 68,311 images, with 351,722 small, 255,950 medium, and 49,453 large human instances, a ratio of approximately 56:36:8. The testing set includes 10,296 images, with 75,841 small, 53,869 medium, and 12,609 large human instances, maintaining a similar ratio of 53:39:8. Across the entire dataset, small human instances account for 55.6%, medium human instances for 36.4%, and large human instances for 8%. Due to the abundance of small human samples in the dataset, it is able to demonstrate the prevalence of small instances in various scenarios.

Human Area Distribution. We conducted a comparative analysis for the distributions of human body size across six datasets: MPII[1], AIC[29], PoseTrack[2], COCO[16], HiEve[17], and the proposed SmallPose dataset. The Figure 2 reveals that the MPII[1], AIC[29], and PoseTrack[2] datasets exhibit a more concentrated distribution of samples in the large human size range $[256 \times 192, +\infty)$. This suggests that models trained on these datasets may struggle to handle small human instances, potentially compromising their overall robustness. Additionally, while the COCO[16] and HiEve[17] datasets have a wider distribution of samples in the small human size range $(0, 128 \times 96)$, their coverage and sample scale are not as comprehensive and extensive as those of the SmallPose dataset. the SmallPose, on the

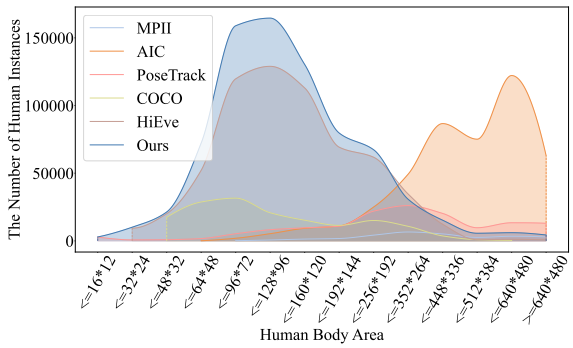


Figure 2. The distribution of human body area in six datasets: MPII[1], AIC[29], PoseTrack[2], COCO[16], HiEve[17], and ours. The proposed SmallPose dataset features a significantly larger sample size of small and medium-sized human instances and a more comprehensive size coverage, showcasing its potential advantages in training robust pose estimation models.

CVPR #9341

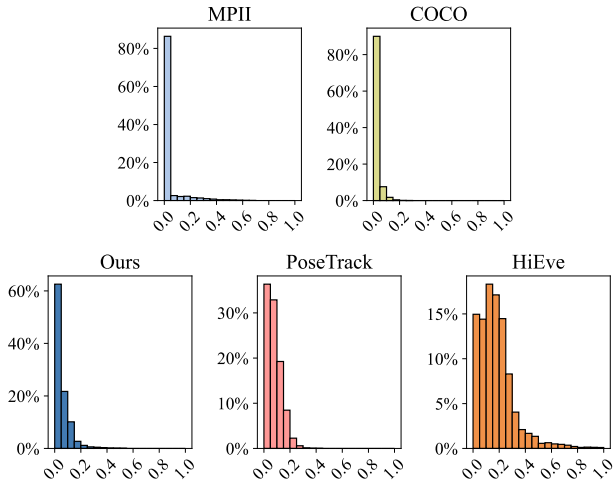


Figure 3. The CrowdIndex distribution of five datasets: MPII[1], COCO[16], PoseTrack[2], HiEve[17], and ours. The proportion of crowded scenes in SmallPose is not significant, indicating that there is no direct correlation between human body size and scene complexity.

other hand, excels with its extensive collection of small and medium human instances, highlighting its benefits for training robust pose estimation models. **CrowdIndex Distribution.** We utilized the CrowdIndex (CI) metric defined in the CrowdPose[13] dataset to quantify the degree of crowding across different datasets. For a given image, the CI is calculated using the formula:

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

where n is the total number of human instances, N_i^a represents the number of keypoints belonging to the i-th instance, and N_i^b denotes the number of keypoints within the i-th instance's bounding box but not belonging to that instance. We evaluated the CI distributions for five datasets: MPII[1], COCO[16], PoseTrack[2], HiEve[17], and the proposed SmallPose. As shown in Figure 3, the results indicate that the MPII[1], and COCO[16] datasets are dominated by simple scenarios, while PoseTrack[2], and HiEve[17] contain a higher proportion of complex scenes. In contrast, the SmallPose dataset focuses on the challenge of small human instances, yet the proportion of crowded scenes is not significantly high, suggesting that human body size and scene complexity are not directly correlated.

4. Evaluation Metrics

4.1. Size-based Metrics

Existing datasets and evaluation metrics lacked a focus on small instances. To address this, we introduced size-based evaluation metrics. Specifically, we defined small, medium, and large human instances based on bounding box areas: small $(0, 128 \times 96)$, medium $[128 \times 96, 256 \times 192)$, and large $[256 \times 192, +\infty)$. We then proposed the AP^S, AP^M, and AP^L metrics to separately evaluate model performance on each size category. This evaluation approach allowed us to more comprehensively assess the models' capabilities in handling instances of varying sizes - a crucial aspect for developing robust and practical human pose estimation systems. For the sake of simplicity of the formula, we will write the defining intervals of small, medium, and large instances as R^S,R^M, and R^L. The detailed formulas are as follows:

$$\begin{aligned} AP^S &= \frac{\sum_p \delta(OKS_p > T)}{\sum_p \delta(S_p \in R^S)}, \\ AP^M &= \frac{\sum_p \delta(OKS_p > T)}{\sum_p \delta(S_p \in R^M)}, \\ AP^L &= \frac{\sum_p \delta(OKS_p > T)}{\sum_p \delta(S_p \in R^L)}, \end{aligned} \tag{2}$$

where S_p represents the ground-truth bounding box area of individual p, T is the threshold, and only an OKS score greater than T is considered accurate.

4.2. Definition of Keypoint Density for Small Human Instance

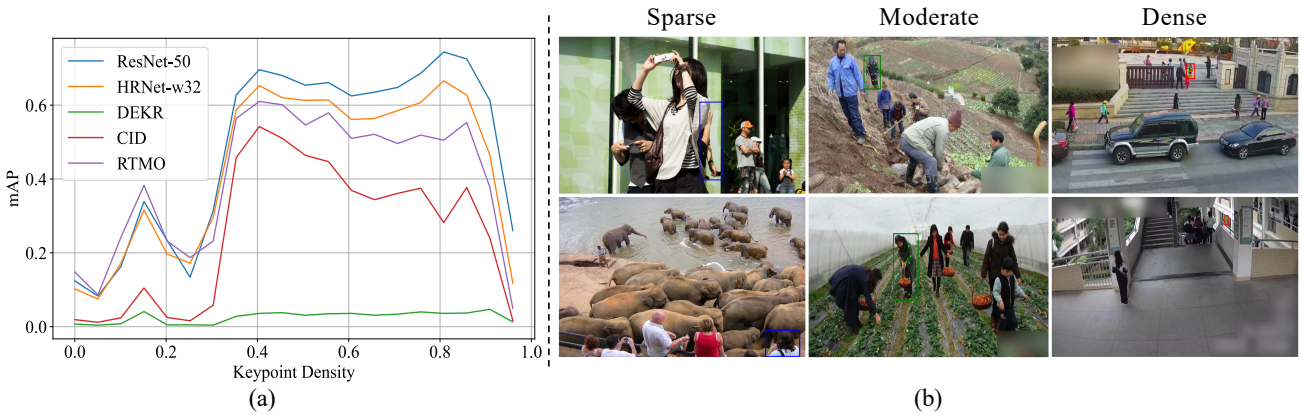
In the SmallPose dataset, we observed that small human instances, such as when only the head of a person is visible, present unique challenges for evaluation due to the incomplete human body. The coarse-grained metrics are not suitable for these cases, as keypoint density can vary significantly, affecting estimation difficulty. To address this issue, we introduce the keypoint density to provide a more fine-grained evaluation of small human instances under different situations. The specific formula is as follows:

$$KD_p = 2\left[1 - \sigma\left(\frac{S_p}{S_{small}} \cdot \frac{K}{n_p}\right)\right] \in (0, 1)\,, \tag{3}$$

where p represents the p-th human instance, S_p denotes the bounding box area of the p-th instance, n_p is the number of keypoint annotations for the p-th instance, S_{small} represents the maximum of the defined area range of small human body, which is 128×96, and K is the number of keypoints in the defined skeleton, such as 17 for COCO[16] and 14 for HiEve[17]. The sigmoid function $\sigma(\cdot)$ is used to restrict the keypoint density value to the range of (0, 1).

4.3. Keypoint-density-based Metrics for Small Human Instance

Based on the definition of keypoint density, we analyzed the relationship between keypoint density and model performance, as shown in Figure 4. To provide further analysis, we defined three categories based on keypoint density: sparse (0-0.35), moderate (0.35-0.95), and dense (0.95-1.0).



(a) (b)

Figure 4. (a) The comparisons of mAP of widely used models, including ResNet[10], HRNet[24], DEKR[9], CID[27] and RTMO[18]. The x-axis represents the keypoint density for small human instances. The results consistently show that the accuracy is lower for both low and high keypoint density, while it is higher for a moderate keypoint density. (b) The visualization of the sparse, moderate, and dense small human instances. The sparse category generally consists of incomplete small instances, while the dense category predominantly contains small instances with dense keypoint annotations. These three types of small instances present varying levels of difficulty.

According to this classification, we introduced the AP^S_{Sp}, AP^S_{Mo}, and AP^S_{De} fine-grained metrics to evaluate model performance on each type of small instances. This fine-grained analysis provided deeper insights into the models' strengths and limitations in handling small human instances with varying keypoint densities. The detailed formulas are as follows:

$$
\begin{aligned}
AP^S_{Sp} &= \frac{\sum_p \delta(OKS_p > T)}{\sum_p^{area_p \in R^S} \delta(KD_p \in (0, 0.35))}, \\
AP^S_{Mo} &= \frac{\sum_p \delta(OKS_p > T)}{\sum_p^{area_p \in R^S} \delta(KD_p \in [0.35, 0.95))}, \\
AP^S_{De} &= \frac{\sum_p \delta(OKS_p > T)}{\sum_p^{area_p \in R^S} \delta(KD_p \in [0.95, 1.0))},
\end{aligned}
\tag{4}
$$

5. Experimental Setup

5.1. Datasets

SmallPose. Our proposed SmallPose dataset comprises a total of 78,607 images and 769,444 human instances, with a remarkable 427,563 small instances. The SmallPose dataset was designed to improve model performance and robustness in handling instances of diverse sizes, particularly small ones, by data samples of various human body sizes.

COCO. The COCO dataset[16], which we used for comparative evaluation, contains a training set of 118,287 images and 156,165 human instances, of which only 98,582 are small instances. Moreover, the COCO samples are predominantly from simple scenes.

5.2. Evaluation Metrics

In our evaluation, we adhere to the metrics defined in Section4. We employ AP, AP^S, AP^M, AP^L, AP^S_{Sp}, AP^S_{Mo}, AP^S_{De}, and the average recall (AR) to evaluate the performance. Notably, in setting A, due to the inconsistency between the number of keypoint types in the COCO dataset[16] and our testing set, we only report the accuracy for the keypoint types present in COCO.

5.3. Experimental Settings

To thoroughly analyze various human pose estimation paradigms and network architectures, we have established two distinct experimental settings:

Setting A: The models were trained on the COCO training set and then tested on our SmallPose testing set, with the keypoint categories aligned.

Setting B: The models were trained on our SmallPose training set and then tested on our SmallPose testing set.

This experimental design allowed us to systematically evaluate the models' performance and generalization capabilities across different training conditions, particularly in relation to handling instances of varying sizes.

5.4. Human Pose Estimation Methods

To comprehensively evaluate the performance of existing models, we employed some representative works from three paradigms (top-down, bottom-up, and single-stage), two types of backbones (single-branch and multi-branch), three prediction architectures (heatmap-based, regression-based, and classification-based), and two different network architectures (CNN-based and Transformer-based) to benchmark the dataset.

SimpleBaseline[30] simply utilizes ResNet[10] as its backbone to regress heatmaps for keypoint localization.

HRNet[24] begins with a high-resolution subnetwork and progressively adds high-to-low resolution subnetworks in later stages, connecting them in parallel. It performs re-

Table 2. The comparative results of various design paradigms and backbones under setting A and setting B.

Method	BackBone	Setting	AP	AP^S	AP^M	AP^L	AP^S_{Sp}	AP^S_{Mo}	AP^S_{De}	AR
Top-down methods										
SimpleBaseline [30]	ResNet-50	A	43.2	32.4	50.5	77.3	8.8	33.9	0.1	45.8
		B	68.2	65.9	69.1	79.7	32.8	69.4	26.0	69.1
HRNet [24]	HRNet-W32	A	45.0	34.3	52.3	79.6	10.3	36.0	0.1	47.7
		B	66.4	62.2	69.3	82.1	32.0	65.3	11.7	67.7
SimCC [15]	ResNet-50	A	43.2	32.9	50.3	76.9	18.7	33.8	0.0	46.2
		B	68.8	68.0	67.3	78.5	38.6	70.5	63.6	69.4
CAL [26]	HRNet-W32	A	45.9	34.4	54.1	81.2	11.0	35.8	0.1	48.4
		B	69.6	66.3	70.8	85.2	35.8	68.9	33.3	70.5
Bottom-up methods										
DEKR [9]	HRNet-W32	A	32.5	15.6	46.5	78.3	3.6	15.9	0.0	36.8
		B	23.2	20.1	32.4	22.8	4.0	20.0	1.3	46.6
Single-stage methods										
CID [27]	HRNet-W32	A	36.6	20.5	47.4	78.4	4.1	21.2	0.0	41.1
		B	60.9	49.4	71.6	83.3	11.8	52.1	1.7	63.0
RTMO [18]	CSPDarknet	A	33.2	17.2	43.7	79.4	7.7	17.3	0.0	37.4
		B	63.1	55.1	70.6	79.9	29.7	58.1	4.9	64.3

peated multi-scale fusion, allowing each resolution to exchange information with others, resulting in rich high-resolution representations for keypoint localization.

DEKR[9] proposes separating regression and adaptive convolution to enhance the quality of regression.

CID[27] utilizes attention mechanisms to separate the instance-level features, and generate instance-level heatmaps for keypoint localization.

RTMO[18] seamlessly integrates coordinate classification by representing keypoints using dual 1-D heatmaps, achieving accuracy comparable to top-down methods while maintaining high speed.

HRFormer[32] leverages the multi-resolution parallel design of HRNet[24] and local-window self-attention to improve memory and computation efficiency.

ViTPose[31] employs plain, non-hierarchical vision transformers as backbones to extract features and a lightweight decoder for pose estimation, demonstrating the strong capabilities of plain vision transformers for pose estimation.

RLE[14] introduces Residual Log-likelihood Estimation (RLE), a regression paradigm that captures the output distribution by learning its change, rather than the unreferenced underlying distribution, to facilitate training.

ViTPose-RLE integrates the ViTPose-S backbone with the RLE paradigm, forming a competitive regression-based method, ViTPose-S-RLE.

SimCC[15] reformulates HPE as two classification tasks for horizontal and vertical coordinates, dividing each pixel into bins to achieve sub-pixel localization and low quantization error. This approach eliminates the need for additional refinement and upsampling layers in certain settings, resulting in a simpler and more efficient HPE pipeline.

CAL[26] introduces Confidence-Aware Learning (CAL) to address two key limitations of offset learning: inconsistent training and testing, decoupled heatmap and offset learning. This method significantly outperforms state-of-the-art approaches for low-resolution human pose estimation.

5.5. Implementation Details

The experiments utilized the PyTorch-based MMPose framework. For setting A, we loaded model weights pretrained on COCO[16] provided by MMPose to test on the SmallPose testing dataset. For setting B, Specifically, the SimpleBaseline[30] model employed the Adam optimizer with a learning rate of 5e-4, a batch size of 128, and was trained for 210 epochs. HRNet[24] used the Adam optimizer with a learning rate of 5e-6, a batch size of 64, and was trained for 210 epochs. DEKR[9] adopted the Adam optimizer with a learning rate of 1e-3, a batch size of 20, and was trained for 140 epochs. CID[27] also used the Adam optimizer with a learning rate of 1e-3, a batch size of 16, and was trained for 140 epochs. RTMO[18] employed the AdamW optimizer with an initial learning rate of 4e-3, a batch size of 32, and was trained for 600 epochs. HRFormer-S[32], ViTPose-S[31] and ViTPose-S-RLE all adopted the AdamW optimizer with a learning rate of 5e-4, a batch size of 64, and was trained for 210 epochs. RLE[14] employed the Adam optimizer with an initial learning rate of 1e-3, a batch size of 128, and was trained for 210 epochs. SimCC[15] also employed the Adam optimizer with an ini-

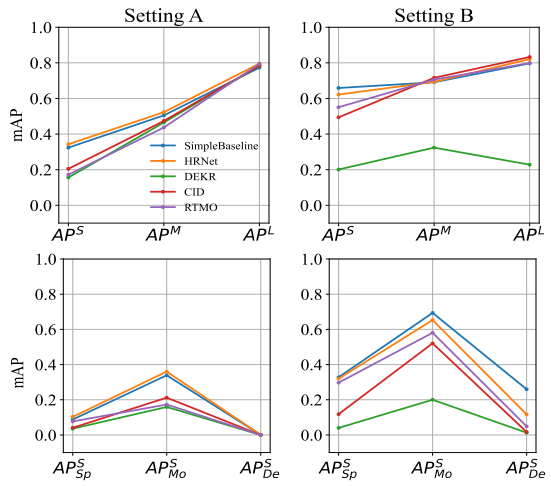


Figure 5. The performance variations across diverse design paradigms and backbones under setting A and setting B.

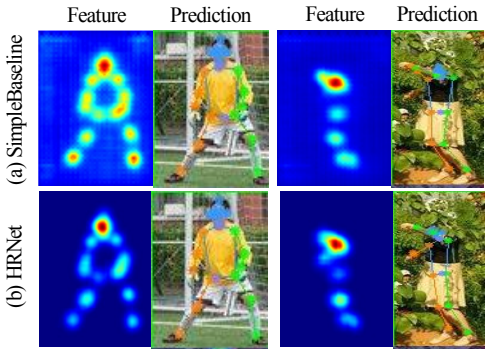


Figure 6. The feature visualization of SimpleBaseline and HRNet for small instances under setting B. It shows that SimpleBaseline offers a more confident and accurate keypoint prediction for small instances, contrasting with HRNet's lower confidence and greater vulnerability to occlusions.

tial learning rate of 1e-3, a batch size of 192, and was trained for 210 epochs. All experiments were conducted on ten 16GB Nvidia V100 GPUs.

6. Results and Analysis

6.1. The Impact on Different Design Paradigms

Our experiments evaluated several popular models across different design paradigms (top-down, single-stage, bottom-up) trained on setting A and B. We present a visual analysis of the experimental results from Table 2, as depicted in Figure 5. Our comparative analysis demonstrated that model performance consistently increased with larger human instances. However, notable performance gaps exist across various instance sizes, indicating the models prioritized larger instances while struggling with smaller ones, limiting robustness. In contrast, top-down and single-stage models trained on SmallPose achieved consistent gains across all sizes, particularly for small instances. The performance gap between different instance sizes was effectively narrowed, enhancing overall robustness. Notably, in the top-down approach, the SimpleBaseline method showed nearly equivalent accuracy on small and medium instances, while the SimCC performed even better on small instances than on medium instances. Conversely, bottom-up models exhibited an inverse trend, seemingly sacrificing the performance for larger instance to compensate for smaller instances. Analysis of keypoint density levels showed the performance of all models trained on SmallPose is improved across sparse, moderate, and dense ranges. In summary, SmallPose enhanced models' ability to handle small instances and reduced performance gaps between different sizes, yielding more robust systems. Top-down models exhibited superior performance for small instances and better robustness compared to single-stage. Bottom-up models face conflicts in handling instances of varying sizes, prioritizing learning one category based on the data distribution.

6.2. The Impact on Different Backbones

Our analysis distinguished between single-branch and multi-branch network architectures based on their handling of feature resolution. Single-branch networks like ResNet and CSPDarknet, which decrease in feature resolution with increasing depth, showed enhanced learning capabilities for small instances. This was evidenced by ResNet's 3.7AP improvement over HRNet on the AP^S metric and the CSPDarknet-based RTMO's 5.7AP lead over the HRNet-w32-based CID on the same metric. Features visualization for ResNet and HRNet in Figure 6 indicated that ResNet offers a more confident and accurate keypoint prediction for small instances, contrasting with HRNet's lower confidence and greater vulnerability to occlusions. These insights emphasize the key role of backbone architectures in advancing model performance for specific tasks, notably the superiority of single-branch networks in detecting smaller human instances.

6.3. The Impact on Different Prediction Paradigms

We conducted a comparison of the performance of heatmap-based, regression-based and classification-based methods for small human pose estimation. As depicted in the Table 3, We highlight the distinct advantages of regression-based methods for the task of small human pose estimation. These methods excel by effectively utilizing global information, which is reflected in their superior performance across sparse and dense keypoint scenarios. Notably, the RLE shows an improvement over the SimpleBaseline,

CVPR #9341

Table 3. The comparison results of various prediction paradigms and network architectures under setting A and setting B.

Method	Backbone	Setting	AP	AP^S	AP^M	AP^L	AP^S_{Sp}	AP^S_{Mo}	AP^S_{De}	AR
Heatmap-based										
SimpleBaseline[30]	ResNet-50	A	43.2	32.4	50.5	77.3	8.8	33.9	0.1	45.8
		B	68.2	65.9	69.1	79.7	32.8	69.4	26.0	69.1
HRNet[24]	HRNet-W32	A	45.0	34.3	52.3	79.6	10.3	36.0	0.1	47.7
		B	66.4	62.2	69.3	82.1	32.0	65.3	11.7	67.7
HRFormer[32]	HRFormer-S	A	44.8	33.9	51.9	76.9	8.9	35.6	0.1	47.3
		B	68.7	66.1	69.1	81.6	33.7	69.1	35.4	69.4
ViTPose[31]	ViT-S	A	45.4	34.6	52.6	79.3	10.6	36.1	0.1	47.8
		B	68.9	66.9	69.1	80.2	33.0	70.3	27.1	69.6
Regression-based										
RLE[14]	ResNet-50	A	44.1	33.4	51.2	77.5	10.9	35.1	0.1	46.4
		B	67.8	65.5	68.3	79.0	36.8	68.7	31.4	68.4
ViTPose-S-RLE	ViT-S	A	29.2	19.8	34.9	63.8	8.9	19.9	0.0	38.9
		B	68.4	66.9	68.2	79.2	37.1	69.9	45.2	68.9
Classification-based										
SimCC[15]	ResNet-50	A	43.2	32.9	50.3	76.9	18.7	33.8	0.0	46.2
		B	68.8	68.0	67.3	78.5	38.6	70.5	63.6	69.4

and a similar trend is observed with ViTPose-S-RLE outperforming ViTPose-S. The integration of regression with transformer-based architectures further amplifies this advantage, as it allows for a more comprehensive capture of features vital for accurate estimation. This synergy underscores the importance of global information in enhancing pose estimation models, especially in the nuanced context of small human instances where other methods may falter. In addition, we found that the classification-based methods, SimCC, which is proposed for quantization error, achieves the best performance in small human instances, especially on dense small human instances, where the problem of quantization error caused becomes particularly serious due to the dense number of joints. This is sufficient to demonstrate that quantization error is an important factor affecting the performance of small human pose estimation. These findings highlight the potential advantages of classification-based methods in addressing the challenges posed by small instances in human pose estimation tasks.

6.4. The Impact on Different Network Architectures

We conducted a comparative analysis of CNN-based and Transformer-based human pose estimation models using the SmallPose dataset, as illustrated in the Table 3. The Transformer-based architecture, trained on the SmallPose dataset, demonstrated superior robustness across various keypoint density levels, particularly excelling at both sparse and dense levels. This indicates that the global attention mechanism utilized by the Transformer-based models provides a more precise detection for small-sized human instances compared to the local receptive fields of CNN-based models.

7. Conclusion

In this paper, we introduce SmallPose, the first dataset dedicated to the challenge of human pose estimation across varying instance sizes. The SmallPose dataset contains a variety of human sizes from simple to complex scenes, focusing on the ubiquity of small instances across diverse settings. With the introduction of the keypoint density concept and corresponding metrics, we have analyzed the human size robustness of different paradigms and different backbones and demonstrated that models trained on SmallPose show improved performance and robustness.

Limitations and Future Plans. While we utilized five benchmark datasets to construct the SmallPose dataset, it is possible that they may not fully represent the distribution of real-world scenarios. Additionally, the concept of keypoint density that we introduced is primarily focused on small human instances, but the notion of keypoint density is applicable to medium and large human bodies as well, albeit with less pronounced challenges. In the future, we can expand the keypoint density concept to encompass the full range of human sizes. This would enable a more comprehensive analysis and potentially provide additional insights for improving human pose estimation models.

References

[1] Mykhaylo Andriluka, Leonid Pishchulin, Peter Gehler, and Bernt Schiele. 2d human pose estimation: New benchmark and state of the art analysis. In *Proceedings of the IEEE Conference on computer Vision and Pattern Recognition*, pages 3686–3693, 2014. 1, 2, 3, 4

[2] Mykhaylo Andriluka, Umar Iqbal, Eldar Insafutdinov, Leonid Pishchulin, Anton Milan, Juergen Gall, and Bernt Schiele. Posetrack: A benchmark for human pose estimation and tracking. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5167–5176, 2018. 1, 2, 3, 4

[3] Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Realtime multi-person 2d pose estimation using part affinity fields. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7291–7299, 2017. 2

[4] Bowen Cheng, Bin Xiao, Jingdong Wang, Honghui Shi, Thomas S Huang, and Lei Zhang. Higherhrnet: Scale-aware representation learning for bottom-up human pose estimation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5386–5395, 2020. 2

[5] Mickael Cormier, Aris Clepe, Andreas Specker, and Jürgen Beyerer. Where are we with human pose estimation in real-world surveillance? In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 591–601. IEEE, 2022. 1

[6] Rita Cucchiara and Matteo Fabbri. Fine-grained human analysis under occlusions and perspective constraints in multimedia surveillance. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, 18(1s): 1–23, 2022. 1

[7] Matthias Dantone, Juergen Gall, Christian Leistner, and Luc Van Gool. Human pose estimation using body parts dependent joint regressors. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3041–3048, 2013. 2

[8] Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *International journal of computer vision*, 88:303–338, 2010. 2

[9] Zigang Geng, Ke Sun, Bin Xiao, Zhaoxiang Zhang, and Jingdong Wang. Bottom-up human pose estimation via disentangled keypoint regression. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14676–14686, 2021. 2, 5, 6

[10] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 5

[11] Hueihan Jhuang, Juergen Gall, Silvia Zuffi, Cordelia Schmid, and Michael J Black. Towards understanding action recognition. In *Proceedings of the IEEE international conference on computer vision*, pages 3192–3199, 2013. 2

[12] Sam Johnson and Mark Everingham. Clustered pose and nonlinear appearance models for human pose estimation. In *bmvc*, page 5. Aberystwyth, UK, 2010. 2

[13] Jiefeng Li, Can Wang, Hao Zhu, Yihuan Mao, Hao-Shu Fang, and Cewu Lu. Crowdpose: Efficient crowded scenes pose estimation and a new benchmark. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10863–10872, 2019. 1, 2, 4

[14] Jiefeng Li, Siyuan Bian, Ailing Zeng, Can Wang, Bo Pang, Wentao Liu, and Cewu Lu. Human pose regression with residual log-likelihood estimation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 11025–11034, 2021. 6, 8

[15] Yanjie Li, Sen Yang, Peidong Liu, Shoukui Zhang, Yunxiao Wang, Zhicheng Wang, Wankou Yang, and Shu-Tao Xia. Simcc: A simple coordinate classification perspective for human pose estimation. In *European Conference on Computer Vision*, pages 89–106. Springer, 2022. 6, 8

[16] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755, 2014. 1, 2, 3, 4, 5, 6

[17] Weiyao Lin, Huabin Liu, Shizhan Liu, Yuxi Li, Rui Qian, Tao Wang, Ning Xu, Hongkai Xiong, Guo-Jun Qi, and Nicu Sebe. Human in events: A large-scale benchmark for human-centric video analysis in complex events. *arXiv preprint arXiv:2005.04490*, 2020. 1, 2, 3, 4

[18] Peng Lu, Tao Jiang, Yining Li, Xiangtai Li, Kai Chen, and Wenming Yang. Rtmo: Towards high-performance one-stage real-time multi-person pose estimation. *arXiv preprint arXiv:2312.07526*, 2023. 2, 5, 6

[19] Zhengxiong Luo, Zhicheng Wang, Yan Huang, Liang Wang, Tieniu Tan, and Erjin Zhou. Rethinking the heatmap regression for bottom-up human pose estimation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 13264–13273, 2021. 2

[20] Debapriya Maji, Soyeb Nagori, Manu Mathew, and Deepak Poddar. Yolo-pose: Enhancing yolo for multi person pose estimation using object keypoint similarity loss. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2637–2646, 2022. 2

[21] Alejandro Newell, Kaiyu Yang, and Jia Deng. Stacked hourglass networks for human pose estimation. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part VIII 14*, pages 483–499. Springer, 2016. 1

[22] Arindam Sengupta, Feng Jin, Renyuan Zhang, and Siyang Cao. mm-pose: Real-time human skeletal posture estimation using mmwave radars and cnns. *IEEE Sensors Journal*, 20 (17):10032–10044, 2020. 1

[23] Pourya Shamsolmoali, Masoumeh Zareapoor, Huiyu Zhou, and Jie Yang. Amil: Adversarial multi-instance learning for human pose estimation. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, 16 (1s):1–23, 2020. 1

[24] Ke Sun, Bin Xiao, Dong Liu, and Jingdong Wang. Deep high-resolution representation learning for human pose estimation. In *Proceedings of the IEEE/CVF conference on*

computer vision and pattern recognition, pages 5693–5703, 2019. 1, 2, 5, 6, 8

[25] Alexander Toshev and Christian Szegedy. Deeppose: Human pose estimation via deep neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1653–1660, 2014. 1

[26] Chen Wang, Feng Zhang, Xiatian Zhu, and Shuzhi Sam Ge. Low-resolution human pose estimation. *Pattern Recognition*, 126:108579, 2022. 6

[27] Dongkai Wang and Shiliang Zhang. Contextual instance decoupling for robust multi-person pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11060–11068, 2022. 2, 5, 6

[28] Shih-En Wei, Varun Ramakrishna, Takeo Kanade, and Yaser Sheikh. Convolutional pose machines. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 4724–4732, 2016. 1

[29] Jiahong Wu, He Zheng, Bo Zhao, Yixin Li, Baoming Yan, Rui Liang, Wenjia Wang, Shipei Zhou, Guosen Lin, Yanwei Fu, et al. Ai challenger: A large-scale dataset for going deeper in image understanding. *arXiv preprint arXiv:1711.06475*, 2017. 1, 2, 3

[30] Bin Xiao, Haiping Wu, and Yichen Wei. Simple baselines for human pose estimation and tracking. In *Proceedings of the European conference on computer vision (ECCV)*, pages 466–481, 2018. 1, 2, 5, 6, 8, 3

[31] Yufei Xu, Jing Zhang, Qiming Zhang, and Dacheng Tao. Vitpose: Simple vision transformer baselines for human pose estimation. *Advances in Neural Information Processing Systems*, 35:38571–38584, 2022. 1, 2, 6, 8

[32] Yuhui Yuan, Rao Fu, Lang Huang, Weihong Lin, Chao Zhang, Xilin Chen, and Jingdong Wang. Hrformer: High-resolution vision transformer for dense predict. *Advances in neural information processing systems*, 34:7281–7293, 2021. 6, 8

[33] Andrei Zanfir, Mihai Zanfir, Alex Gorban, Jingwei Ji, Yin Zhou, Dragomir Anguelov, and Cristian Sminchisescu. Hum3dil: Semi-supervised multi-modal 3d humanpose estimation for autonomous driving. In *Conference on Robot Learning*, pages 1114–1124. PMLR, PMLR, 2023. 1

[34] Bolei Zhou, Agata Lapedriza, Aditya Khosla, Aude Oliva, and Antonio Torralba. Places: A 10 million image database for scene recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2017. 1

[35] Xingyi Zhou, Dequan Wang, and Philipp Krähenbühl. Objects as points. *arXiv preprint arXiv:1904.07850*, 2019. 2