Spatial-Aware Regression for Keypoint Localization

Supplementary Material

Dongkai Wang Shiliang Zhang
National Key Laboratory for Multimedia Information Processing,
School of Computer Science, Peking University
{dongkai.wang, slzhang.jdl}@pku.edu.cn

In the supplemental material, we present the overview of datasets, evaluation metrics and detailed implementation of experiments about each keypoint localization task, experiments on MPII dataset, experiments on ViT backbone, the broader impact of our work, limitation and show some qualitative results on each task.

1. Datasets and Evaluation Metrics

COCO Keypoint [10] is a large scale 2D human pose estimation dataset with 17 annotated keypoints, *e.g.*, nose, left ear, *etc*. It contains more than 200,000 images and 250,000 person instances. COCO Keypoint is divided into three parts: 57k images for training set, 5k images for validation set and 20k images for test-dev set.

MPII [12] is a classical 2D human pose estimation dataset, it contains 25k images with 40k person instances, where 12k instances are used for testing and the remaining instances are used for training. MPII is annotated with 16keypoints.

COCO Whole-Body [7] adopts the images from COCO Keypoint dataset and provides additional keypoint annotation on face, feet and hands for each person. It includes 133 keypoints for each person and follows COCO Keypoint to divide training and validation sets.

CrowdPose [8] is a challenging multi-person pose estimation benchmark that contains both general and crowded scenes. It contains 10k images for training set, 2k images for validation set and 8k images for test set. CrowdPose is annotated with 14 keypoints.

OCHuman [15] focuses on heavily occluded human pose estimation in crowded scenes and is regarded as a validation and test set. It contains 4,731 images and 8,110 persons labeled with 17 keypoints. In OCHuman, on an average 67% of the bounding box area has overlap with other bounding box. OCHuman is divided into validation set (2,500 images, 4,313 persons) and test set (2,231 images, 3,819 persons). Following [6], we train model on COCO Keypoint training set and test model on OCHuman validation set.

Human3.6M [5] is a large scale indoor benchmark for 3D human pose estimation, which consists of 3.6 million poses and corresponding images featuring 11 actors performing 15 daily activities from 4 camera views. Following the standard protocols, the models are trained on subjects 1, 5, 6, 7, 8 and tested on subjects 9, 11.

Evaluation metrics. For 2D human pose estimation, multi-person pose estimation and whole-body pose estimation, we follow previous works to report the OKS-based mAP on COCO Keypoint, COCO Whole-Body, CrowdPose and OCHuman. We also compute the mean error (ℓ_2 norm) of each keypoint to its groundtruth in COCO Keypoint to reflect the accuracy of localization, *i.e.*, Kpt.Error(px) in Table. 1. For MPII, we report official PCKh@0.5/0.1 to evaluate single-person pose estimation results. For 3D human pose estimation task, we report the commonly used Mean Per Joint Position Error (MPJPE) to evaluate the error of each method.

2. Implementation Details

2D Human Pose Estimation. For ablation study in Table 1 and Table 2, we conduct experiments based on SimplePose [13] with ResNet-50 and follow the most setting of [13]. Unless specified separately, the input image is resized to 256×192. The learning rate is set to 1×10^{-3} at first and reduced by a factor of 10 at the 90-th and 120-th epoch. We adopt Adam optimizer to train model for 140 epochs, with a bath size of 320 and 2 NVIDIA RTX 3090 GPUs in total. We adopt single scale test without flip and use groundtruth person bounding box for all methods in Table 1, Table 2 and Fig. 3. When comparing with other methods on COCO Keypoint test-dev set in Table 3, we follow HRNet [11] to train model with 210 epochs and report performance with flip test and with the predicted person bounding box provided by HRNet [11]. During inference, we select the most confident result, *i.e.*, $m = 1$.

Whole-Body Pose Estimation. We adopt MMPose [2] framework to conduct whole-body pose estimation experiments and follow all the setting of existing methods. The input image is resized to 256×192 or 384×288. The learning rate is set to 1×10^{-3} at first and reduced by a factor of 10 at the 170-th and 200-th epoch. We adopt Adam optimizer to train model for 210 epochs, with a bath size of 160 and 2 NVIDIA RTX 3090 GPUs in total. During inference, we select the most confident result, *i.e.*, $m = 1$.

Multi-Person Pose Estimation. We adopt HigherHRNet [1] for bottom-up MPPE experiments and DEKR [3]

for single-stage regression MPPE and follow the most configuration of them. HRNet-W32 [11] is adopted as backbone for all MPPE experiments. The input image is resized to 512×512. The learning rate is set to 1×10^{-3} at first and reduced by a factor of 10 at the 90-th and 120-th epoch. We adopt Adam optimizer to train model for 140 epochs, with a bath size of 40 and 2 NVIDIA RTX 3090 GPUs in total. To evaluate CenterNet and DEKR on OCHuman val set, we directly adopt the model trained on COCO train set. During inference, we follow previous works [1, 3] to select $m = 30$ predictions with confidence score larger than $\gamma = 0.01$.

Monocular 3D Human Pose Estimation. We implement our method based on the codebase provided by RLE [9] to conduct 3D pose estimation experiments. ResNet-50 [4] is adopted as backbone for all experiments. The input image is resized to 256×256. Data augmentation includes random scale (±30%), rotation (±30), color (±20%) and flip. The learning rate is set to 1×10^{-3} at first and reduced by a factor of 10 at the 90th and 120 epochs. We use the Adam solver and train for 140 epochs, with a mini-batch size of 64 per GPU and 4 NVIDIA GTX 1080ti GPUs in total. The 2D and 3D mixed data training strategy (MPII + Human3.6M) is applied. The testing procedure is the same as the previous works [9]. During inference, we select the most confident result, *i.e.*, $m = 1$.

3. 2D Human Pose Estimation on MPII dataset

We conduct 2D human pose estimation experiments on classical MPII dataset and the results are shown in Table 1. We can observe that SAR outperforms regression-based and heatmap-based methods on various backbones under the same setting, especially in PCKh@0.1 metric.

Method	Backbone	Input size	PCKh@0.5	PCKh@0.1
Direct Regression	ResNet-50	256×256	83.8	23.6
RLE [9]	ResNet-50	256×256	85.8	27.1
SimplePose [13]	ResNet-50	256×256	87.1	25.4
HRNet [11]	HRNet-W48	256×256	90.1	33.7
Ours (SAR)	ResNet-50	256×256	87.5	31.7
Ours (SAR)	HRNet-W48	256×256	**90.1**	**38.1**

Table 1. Comparison with other methods on MPII `val` set.

4. Experiments on Vision Transformer Backbone

SAR can be applied to various backbones with different output stride, *e.g.*, ViT [14]. We conduct 2D human pose estimation experiments with other baselines based on ViT-Small in Table 2. We can observe that SAR outperforms regression-based and heatmap-based methods on different output strides in ViT-based backbones, which is consistent with ResNet-based backbones.

Method	Input size	GFLOPs	Deconvs	AP	AP^{50}	AP^{75}
Regression	256×192	4.2		58.0	83.9	64.8
RLE	256×192	4.2		71.6	90.3	79.1
Heatmap	256×192	8.6	✓	72.4	92.1	80.8
Heatmap+Offset	256×192	8.6	✓	72.6	91.3	79.8
SAR	256×192	4.2		73.0	92.0	80.7
SAR	256×192	8.6	✓	**74.0**	**92.3**	**82.1**

Table 2. Comparison with baselines on COCO Keypoint `val` set based on ViTPose [14].

5. Broader Impact

In this work, we propose the spatial-aware regression to improve the performance of keypoint localization, and promote the vision intelligence of machine. The proposed method is simple yet effective and does not introduce a new capability in computer vision. Therefore, our method will not cause misuse of technology that is already available to anyone.

6. Limitation and Future Work

Our work has several limitations. First, SAR adopts Laplacian kernel to measure the quality of regression, which can be viewed as training model with ℓ_1 loss from the perspective of maximum likelihood estimation. As shown in RLE [9], it can be improved by adopting a learnable distribution to model the target keypoint distribution and measure the quality of regression. Second, SAR performs well on localization task where input has spatial structure, *e.g.*, camera images. The effectiveness of SAR on other inputs, *e.g.*, point cloud, need further exploration. We believe these limitations are exciting avenues for future work to explore.

7. Qualitative Results

We show some visualization results on each task in Fig. 1, 2 and 3.

References

[1] 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 *CVPR*, 2020. 1, 2

[2] MMPose Contributors. Openmmlab pose estimation toolbox and benchmark. https://github.com/open-mmlab/mmpose, 2020. 1

[3] Zigang Geng, Ke Sun, Bin Xiao, Zhaoxiang Zhang, and Jingdong Wang. Bottom-up human pose estimation via disentangled keypoint regression. In *CVPR*, 2021. 1, 2

Figure 1. Qualitative results of 2D human pose estimation and multi-person pose estimation on COCO Keypoint.

Figure 2. Qualitative results of whole-body pose estimation on COCO Whole-Body val set.

[4] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 2

[5] Catalin Ionescu, Dragos Papava, Vlad Olaru, and Cristian Sminchisescu. Human3. 6m: Large scale datasets and predictive methods for 3d human sensing in natural environments. *IEEE TPAMI*, 2013. 1

[6] Sheng Jin, Wentao Liu, Enze Xie, Wenhai Wang, Chen Qian, Wanli Ouyang, and Ping Luo. Differentiable hierarchical graph grouping for multi-person pose estimation. In *ECCV*, 2020. 1

[7] Sheng Jin, Lumin Xu, Jin Xu, Can Wang, Wentao Liu, Chen Qian, Wanli Ouyang, and Ping Luo. Whole-body human pose estimation in the wild. In *ECCV*, 2020. 1

[8] 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 *CVPR*, 2019. 1

[9] Jiefeng Li, Siyuan Bian, Ailing Zeng, Can Wang, Bo Pang, Wentao Liu, and Cewu Lu. Human pose regression with residual log-likelihood estimation. In *ICCV*, 2021. 2

[10] 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 *ECCV*, 2014. 1

[11] Ke Sun, Bin Xiao, Dong Liu, and Jingdong Wang. Deep high-resolution representation learning for human pose estimation. In *CVPR*, 2019. 1, 2

[12] Alexander Toshev and Christian Szegedy. Deeppose: Human pose estimation via deep neural networks. In *CVPR*, 2014. 1

[13] Bin Xiao, Haiping Wu, and Yichen Wei. Simple baselines for human pose estimation and tracking. In *ECCV*, 2018. 1, 2

[14] Yufei Xu, Jing Zhang, Qiming Zhang, and Dacheng Tao. Vit-

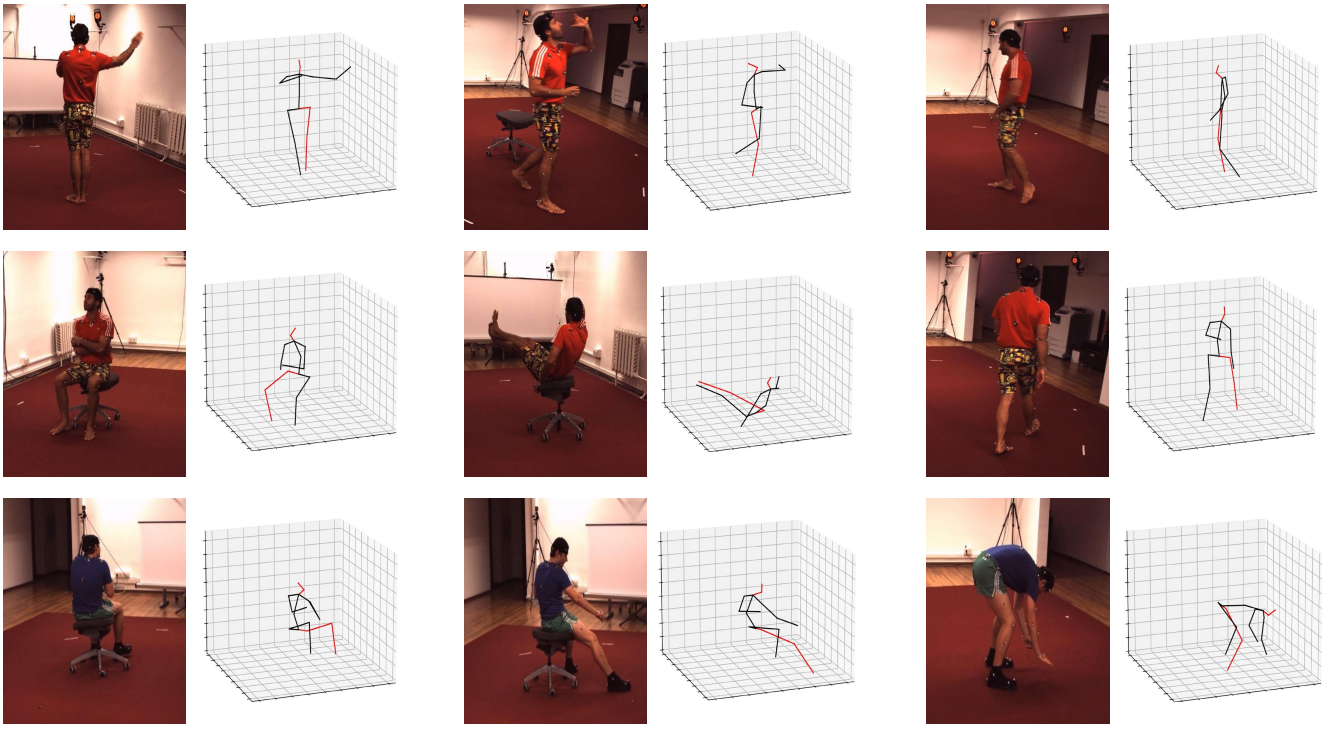

Figure 3. Qualitative results of 3D human pose estimation on Human3.6M.

pose: Simple vision transformer baselines for human pose estimation. *arXiv preprint arXiv:2204.12484*, 2022. 2

[15] Song-Hai Zhang, Ruilong Li, Xin Dong, Paul Rosin, Zixi Cai, Xi Han, Dingcheng Yang, Haozhi Huang, and Shi-Min Hu. Pose2seg: Detection free human instance segmentation. In *CVPR*, 2019. 1