Human Pose Estimation in Monocular Omnidirectional Top-View Images Supplementary Material

Jingrui Yu* Tobias Scheck Roman Seidel Yukti Adya Dipankar Nandi Gangolf Hirtz
Chemnitz University of Technology, Germany
*jingrui.yu@etit.tu-chemnitz.de

1. Example images from THEODORE+

We provide only one example of THEODORE+ in the main paper to illustrate the annotations. Here we provide some more example images without annotations so that the diversity of this dataset can be known to the reader. In Fig. 1 there are nine random example images from THEODORE+ dataset. These examples show the diversity in texture, lighting, objects and human models, as well as the different actions of the persons. This dataset will be publicly available along with PoseFES dataset after the double-blind review process.

2. More Qualitative evaluation results of HRNet

We provide some more qualitative evaluation results of HRNet in this section. In Fig. 2, the person is bending the upper body due to the usage of a walking aid. The legs are hardly visible from the top-view camera. The baseline model is unable to estimate the legs except the ankles. The trained model is able to estimate both legs. In Fig. 3, the same person is sideways from the camera. Although both legs are visible, the baseline model still fails to detect the keypoints on the leg on the farther side (the left leg). The trained model detects the left leg with high accuracy. Note that the left arm is also estimated more accurately than the baseline model. *Record_00791.png* provides two more interesting examples in Fig. 4. The right leg of the first person is partially occluded by the left leg. Our trained model successfully estimates the whole right leg, while the baseline fails to estimate the right knee. The second instance of interest is in the bottom right corner. The left leg of the person further down is occluded, yet it is correctly estimated by our trained model. These examples mainly exhibits the superior performance of our model for occluded body parts.

Our trained model is not perfect, but we find it generally more accurate in harder cases, such as those in the main paper and the above examples. At the same time, it performs as good as the baseline model for easier instances. This is very important for our application in AAL, because our ultimate goal is to recognize human actions. These hard cases are exactly the ones that no popular model can solve. Without the ability to detect the skeletons in these hard cases, it is not possible to perform action recognition reliably.

Additionally, in the situation of monitoring an elderly living on his own, the inference speed of HRNet is sufficiently fast on a capable workstation. Taking the numbers from the main paper, it takes about $20\,\mathrm{ms}$ for person detection and $90\,\mathrm{ms}$ for estimating the keypoints for a single person. This adds up to $110\,\mathrm{ms}$ or $9\,\mathrm{FPS}$, which allows for real-time usage.

3. More qualitative evaluation results of CenterNet

The inference results from CenterNet model with Hourglass backbone (CN-HG) is complicated. Therefore, we want to provide more samples here, so that we can make a more informed analysis.

Fig. 5 shows an extreme case of keypoint estimation with CN-HG where the baseline CenterNet model fails (almost) completely and our trained CenterNet works reasonably well. Interestingly, the baseline model is already able to detect the person in this unusual position of standing directly under the camera, but it cannot detect the keypoints correctly. Our trained model makes one obvious mistake with the right arm.

In *Record_00851.png*, we see more improvement in image areas where persons appear as side-view. In Fig. 6, it is clear that the arms of the person in the left and the legs of the person in the right are estimated better after the finetuning.

There are also many examples where the trained models show no improvement over the baseline model at all. In image *Record_00842.png* (Fig. 7), the first instance shows that the estimation for the right arm is deteriorated. The second instance has no actual improvement because of the falsely detected keypoint *shoulder_right*. In Fig. 8, the keypoint estimation is faulty by both the COCO pretrained model and our finetuned model. The reason is that the person detection is inaccurate, which generated a smaller bounding box than

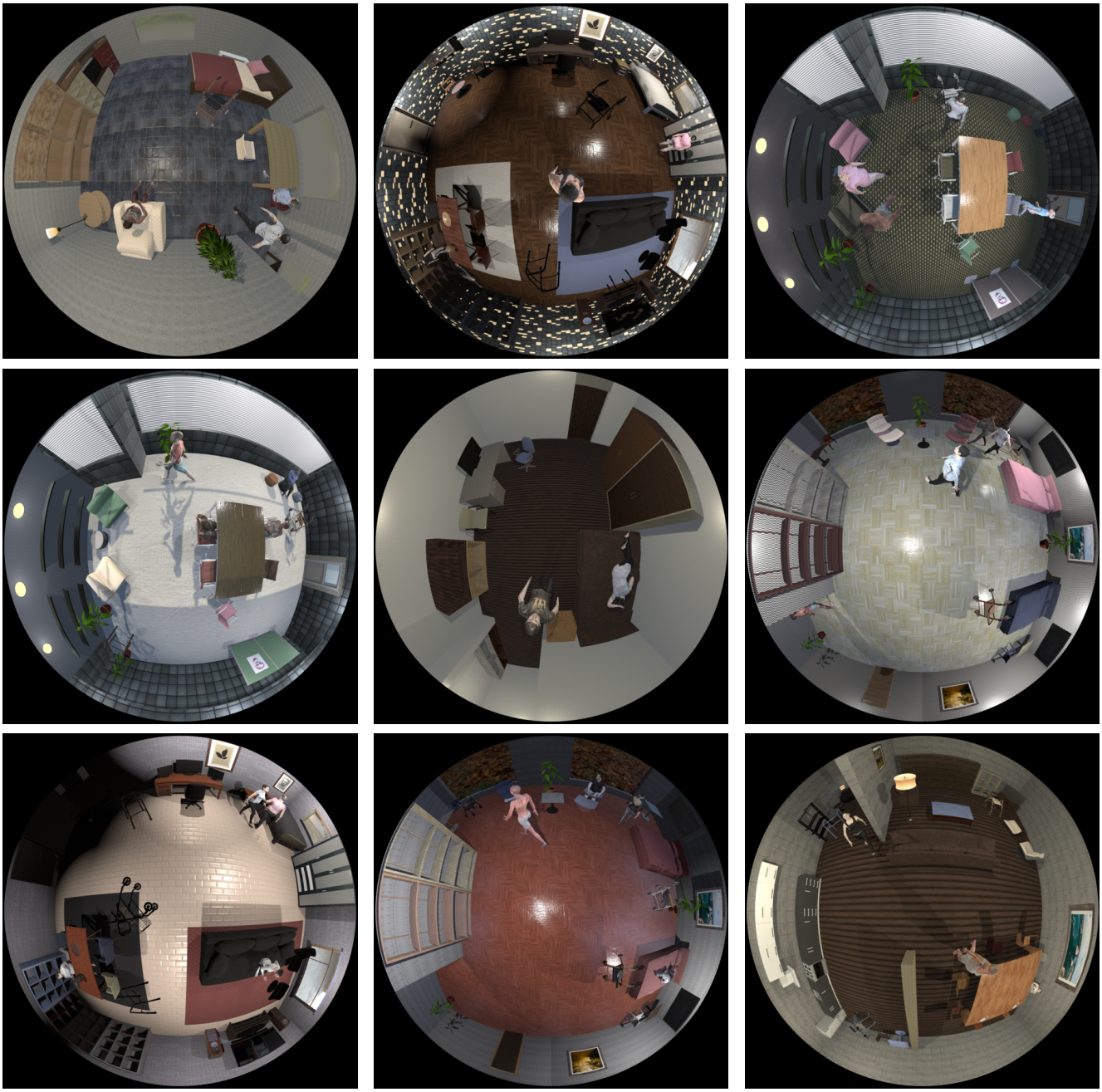

Figure 1. Random example images from THEODORE+ dataset.

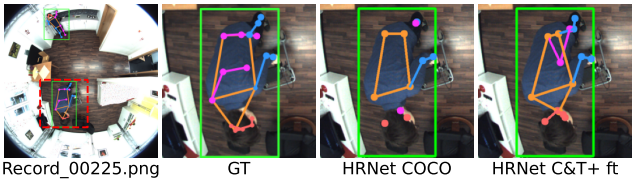

Figure 2. Pose estimation result comparison: Record_ 00225.png.

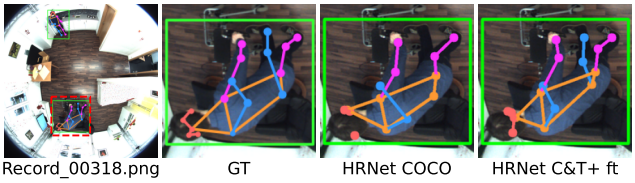

Figure 3. Pose estimation result comparison: Record_ 00318.png.

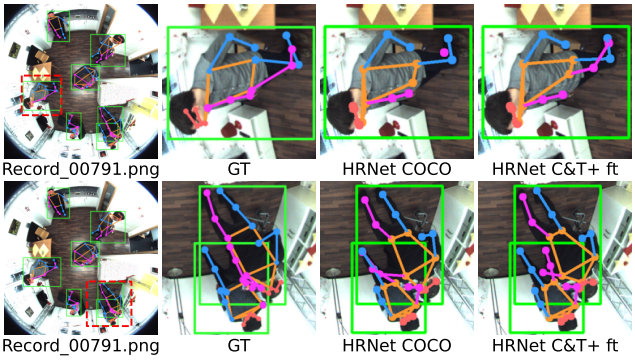

Figure 4. Pose estimation result comparison: Record_ 00791.png.

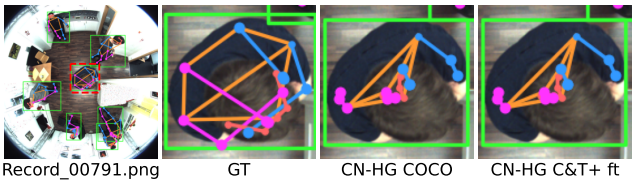

Figure 5. Pose estimation result comparison: Record_ 00791.png.

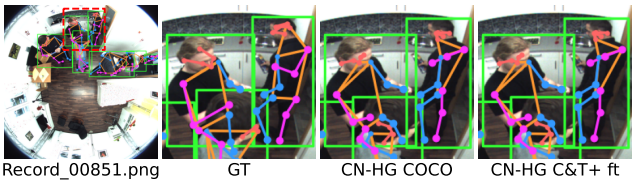

Figure 6. Pose estimation result comparison: Record_ 00851.png.

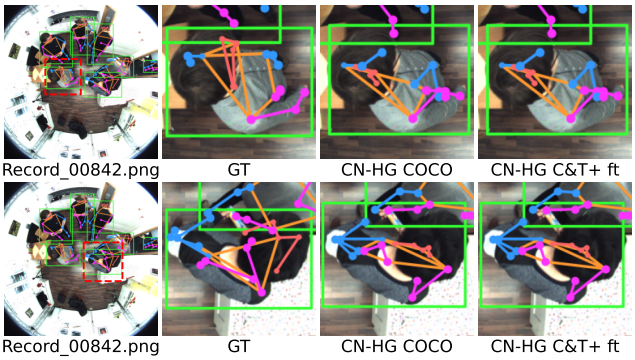

Figure 7. Pose estimation result comparison: Record_ 00842.png.

the actual person.

4. Conclusion

With the research in this paper, we explore the possibility of estimating a detailed skeleton for action recognition in omnidirectional top-view images. It is clear that the top-down model for keypoint estimation performs much better in terms of accuracy, while the CenterNet structure is more

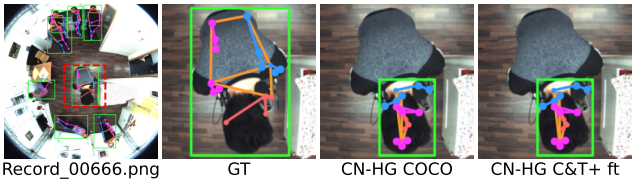

Figure 8. Pose estimation result comparison: Record_ 00666.png.

scalable and better suited for implementation on an embedded system. In our research project, we already deployed a CenterNet model with EfficientNet [1] backbone on a smart sensor that uses the Nvidia TX2 platform. This model has been trained to estimate five keypoints (head, hands, feet). It runs at 3 FPS with acceptable accuracy. Our future plan involves adapting the CenterNet Hourglass model for the embedded platform and / or implementing a top-down keypoint estimation workflow on a local workstation, so that we can have high accuracy detection while protecting the privacy of the potential user of this system.

References

[1] Mingxing Tan and Quoc Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In *International conference on machine learning*, pages 6105–6114. PMLR, 2019. 3