Supplementary Document for Crowd3D: Towards Hundreds of People Reconstruction from a Single Image

Hao Wen[1,†], Jing Huang[1,†], Huili Cui[1], Haozhe Lin[2], Yu-Kun Lai[3], Lu Fang[2], Kun Li[1,*]
[1]Tianjin University, China [2]Tsinghua University, China [3]Cardiff University, United Kingdom
{wenhao, hj00, huilicui_1, lik}@tju.edu.cn, {linhz, fanglu}@tsinghua.edu.cn,
LaiY4@cardiff.ac.uk

In this document, we provide more paper details, including:

- How to automatically set cropping parameters;
- How to obtain the results of other methods for large-scene images;
- Quantitative results on small scenes;
- More ablation studies;
- Qualitative results on *PANDA*.

We also provide a demo video along with this document.

1. Cropping Parameter Settings

Our adaptive human-centric cropping scheme crops a large-scene image into patches with hierarchical sizes which ensures that the height ratio between the person and the corresponding image is as consistent as possible among different cropped images. This is beneficial for the subsequent inference of human poses, shapes and locations. To obtain the cropping parameters automatically, we first crop the images with sliding windows of different scales to ensure that each object is detectable and then use a state-of-the-art pose estimation method [2] to obtain pre-detection results. We obtain the 2D poses of persons at the top and the bottom of the large-scene image by sorting y coordinate of centers of detected poses, and set the heights h_t and h_b of persons at the top and the bottom by respective pose results. We assume that the persons at the top and the bottom are in the center of their respective cropped images, hence the upper and lower bounds of the image area to be processed are set as $b_u = c_{y_t} - h_t$ and $b_l = c_{y_b} + h_b$, where c_{y_t} and c_{y_b} are the y coordinate of pose centers of corresponding persons. We also use some tricks to improve the reliability of the obtained cropping parameters, including detecting multiple images of a scene, requiring the size of the person at the bottom to be larger than the vast majority of the detected people, *etc*. In general, automatically obtaining the cropping parameters is sensitive to false detections and has a high time cost. Therefore, we recommend setting the cropping parameters h_t, h_b, b_u and b_l manually, which is easy and has a low time cost.

2. Running Other Methods on Large-scene Images

Existing methods cannot directly process large-scene images. To obtain the crowd reconstruction results on large-scene images using these methods, we apply our adaptive human-centric cropping to obtain the hierarchical cropped images as their inputs and provide the estimated scene-level focal length f_s of our method.

For CRMH [3], we use its own camera coordinate transformation to convert the camera parameters corresponding to human bounding boxes of cropped images to the global scene-level depths. We restore the bounding boxes of cropped images to the pixel coordinates of large-scene images by upper-left coordinates of the cropped images. We represent the camera parameters corresponding to the bounding box $B_i = [x_{\min}, y_{\min}, x_{\max}, y_{\max}]$ of the i-th person with $\boldsymbol{\pi}_i = \{s_i, x_i, y_i\}$, and define the center and size of B_i as $c_i = [(x_{\min} + x_{\max})/2, (y_{\min} + y_{\max})/2]$ and $\alpha_i = \max(x_{\max} - x_{\min}, y_{\max} - y_{\min})$, respectively. Given these parameters, the global depth of the i-th person is calculated as

$$d_i = \frac{2f_s}{s_i \alpha_i}. \tag{1}$$

Then, the global translation of the i-th person can be obtained by

$$T_{i\_global} = \begin{bmatrix} d_i \left(x_i \alpha_i + c_{i,x} - w_s/2\right)/f_s \\ d_i \left(y_i \alpha_i + c_{i,y} - h_s/2\right)/f_s \\ d_i \end{bmatrix}, \tag{2}$$

where w_s and h_s are the width and height of the large-scene

† Equal contribution.
* Corresponding author.

image, respectively.

For SMAP [11] and BEV [8], we use the method in [1] to implement the local-to-global depth conversion. For the cropped image with width w_c and height h_c, we set the focal length f_c of local camera system for SMAP and BEV by respective model settings. We have $f_c = w_c$ for SMAP and $f_c = f_b \times w_c/s_b$ for BEV, where $f_b = 443.4$ and $s_b = 512$ are the focal length and the input image size of BEV model. We represent the torso center of the i-th person in the local camera coordinate system with $T_i = \{X_i, Y_i, Z_i\}$, and the depth in the global coordinate system is calculated as $Z_{i\_global} = Z_i \times f_s/f_c$. We also restore the 2D pose pixel coordinates of cropped images to large-scene image by the positions of cropped images. With the 2D projection $\{x_i, y_i\}$ of torso center on the large-scene image, the global depth Z_{i_global} and our scene-level camera parameters K, the global position of the i-th person can be calculated as

$$T_{i\_global} = Z_{i\_global} K^{-1} \begin{bmatrix} x_i \\ y_i \\ 1 \end{bmatrix}. \quad (3)$$

3. Quantitative Results on Small Scenes

We evaluate our method on the small-scene datasets *Panoptic* [4] and *MuPoTS* [5], compared with state-of-the-art methods. We directly use small-scene images as inputs to our Crowd3DNet. The ground plane equations that we use are estimated by combining the people of the same scene at different frames since there are only 2-6 people in a small scene image. For *Panoptic*, we use MPJPE (mean per joint position error), root error (RtError) and percentage of correct ordinal depth (PCOD) to evaluate the 3D poses and locations of the reconstructed people. We do not test on *Mafia* because the related 3D annotations cannot be obtained from the official website. We use the SMAP [11] model provided by the authors that is not trained on *Panoptic* for fair comparison. As shown in Table 1, our method achieves the best results, validating the performance of our method on both position inference and pose estimation for small scenes. For *MuPoTS*, we follow the protocol of [6]. The results in Table 2 also demonstrate the effectiveness of our method.

4. Ablation Studies

Besides ablation studies on the HVIP and the cropping score given in the paper, Tab. 3 gives additional ablations of some crucial modules (AC: adaptive cropping, GT-GC: ground-truth ground and camera parameters). The first two lines show the positive impact of adaptive human-centric cropping scheme on crowd reconstruction. The last line shows that better ground estimation can further improve the performance of our method.

Table 1. Comparisons to the state-of-the-art methods on a small-scene dataset *Panoptic*.

	Method	*Haggling*	*Ultim*	*Pizza*	Mean
MPJPE ↓	SMAP [11]	128.5	141.2	236.4	168.7
	CRMH [3]	129.6	153.0	156.7	146.4
	BMP [10]	120.4	140.9	147.5	136.3
	ROMP [7]	110.8	141.6	137.6	130.0
	BEV [8]	100.9	132.4	139.6	124.3
	Crowd3D	**100.1**	**125.7**	**134.1**	**120.0**
RtError ↓	SMAP [11]	432.8	529.8	1297.6	753.4
	CRMH [3]	2384.5	2301.0	2418.7	2368.1
	BEV [8]	786.1	683.0	763.2	744.1
	Crowd3D	**274.8**	**295.3**	**542.1**	**370.7**
PCOD ↑	SMAP [11]	83.5	93.3	76.0	84.3
	CRMH [3]	89.5	93.2	74.8	85.8
	BEV [8]	89.3	98.3	**94.7**	94.1
	Crowd3D	**90.4**	**99.5**	92.8	**94.2**

Table 2. Comparison on *MuPoTS* dataset.

Method	ALL↑	Matched↑
CRMH	69.1	72.2
ROMP	69.9	74.6
BEV	70.2	75.2
Ours	**70.9**	**75.4**

Table 3. Quantitative ablation study on *LargeCrowd* dataset.

Method	PPDS↑	PA-PPDS↑	PCOD↑	OKS↑
w/o AC, w/o GT-GC	81.5	89.1	92.3	69.5
w/ AC, w/o GT-GC	81.5	89.4	92.6	71.7
w/ AC, w/ GT-GC	92.4	92.4	93.5	73.9

5. Qualitative Results on PANDA

In Fig. 1, we present qualitative results on *PANDA* [9]. *PANDA* is a human-centric large-scene video dataset which has gigapixel-level resolutions and contains hundreds of people. Please note that the gigapixel camera used in *PANDA* is different from the camera in *LargeCrowd*, and the image resolutions of the two datasets are also different. The consistent reconstructed results show the generalization ability of our method.

References

[1] Vitor Albiero, Xingyu Chen, Xi Yin, Guan Pang, and Tal Hassner. img2pose: Face alignment and detection via 6DoF, face pose estimation. In *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7617–7627, 2021. 2

[2] Hao-Shu Fang, Shuqin Xie, Yu-Wing Tai, and Cewu Lu. RMPE: Regional multi-person pose estimation. In *Proc.*

Figure 1. Qualitative results on *PANDA*.

IEEE/CVF International Conference on Computer Vision, 2017. 1

[3] Wen Jiang, Nikos Kolotouros, Georgios Pavlakos, Xiaowei Zhou, and Kostas Daniilidis. Coherent reconstruction of multiple humans from a single image. In *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5579–5588, 2020. 1, 2

[4] Hanbyul Joo, Hao Liu, Lei Tan, Lin Gui, Bart Nabbe, Iain Matthews, Takeo Kanade, Shohei Nobuhara, and Yaser Sheikh. Panoptic studio: A massively multiview system for social motion capture. In *Proc. IEEE/CVF International Conference on Computer Vision*, pages 3334–3342, 2015. 2

[5] Dushyant Mehta, Oleksandr Sotnychenko, Franziska Mueller, Weipeng Xu, Srinath Sridhar, Gerard Pons-Moll, and Christian Theobalt. Single-shot multi-person 3D pose estimation from monocular RGB. In *Proc. IEEE International Comference on 3D vision*, 2018. 2

[6] Gyeongsik Moon, Ju Yong Chang, and Kyoung Mu Lee. Camera distance-aware top-down approach for 3D multi-person pose estimation from a single RGB image. In *Proc. IEEE/CVF International Conference on Computer Vision*, pages 10133–10142, 2019. 2

[7] Yu Sun, Qian Bao, Wu Liu, Yili Fu, Michael J Black, and Tao Mei. Monocular, one-stage, regression of multiple 3D people. In *Proc. IEEE/CVF International Conference on Computer Vision*, pages 11179–11188, 2021. 2

[8] Yu Sun, Wu Liu, Qian Bao, Yili Fu, Tao Mei, and Michael J Black. Putting people in their place: Monocular regression of 3D people in depth. In *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022. 2

[9] Xueyang Wang, Xiya Zhang, Yinheng Zhu, Yuchen Guo, Xiaoyun Yuan, Liuyu Xiang, Zerun Wang, Guiguang Ding, David Brady, Qionghai Dai, et al. PANDA: A gigapixel-level human-centric video dataset. In *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3268–3278, 2020. 2

[10] Jianfeng Zhang, Dongdong Yu, Jun Hao Liew, Xuecheng Nie, and Jiashi Feng. Body meshes as points. In *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 546–556, 2021. 2

[11] Jianan Zhen, Qi Fang, Jiaming Sun, Wentao Liu, Wei Jiang, Hujun Bao, and Xiaowei Zhou. SMAP: Single-shot multi-person absolute 3D pose estimation. In *Proc. European Conference on Computer Vision*, pages 550–566, 2020. 2