SEFD: Learning to Distill Complex Pose and Occlusion Supplementary Material

ChangHee Yang[*1] Kyeongbo Kong[*2] SungJun Min[*1,3]
Dongyoon Wee[4] Ho-Deok Jang[4] Geonho Cha[4] SukJu Kang[†1]
Sogang University[1], Pusan National University[2], Samsung Electronics[3], NAVER Cloud Corp[4]
{yangchanghee2251, kkb4723}@gmail.com, sung98.min@samsung.com
{dongyoon.wee, hodeok.jang, geonho.cha}@navercorp.com, sjkang@sogang.ac.kr

Project Page

Our project page can be found in https://yangchanghee.github.io/ICCV2023_SEFD_page/

Appendix

- In section 1, we first describe the implementation details for each of the teacher model and the student model.
- In section 2, we compare State-of-the-Art method in MuPoTs [1].
- In section 3, we provide the qualitative analysis of the proposed method, with the comparison between the baseline method.
- In section 4, we describe the adaptive dilation depending on the size of the object, with how diverse the size of the object is.
- In section 5, we explain the SMPL edge estimator (SEE) and SMPL edge estimator self-supervised deocclusion(SESD). Model parameters and MACs are also compared in this section.
- In section 6, we explain why we used Canny edge detector as the simple edge detector based on the difference between Canny edge [2] and HED [3].

1. Implementation Details

Teacher model. Following the 3DCrowdNet [4], which is chosen as our baseline, the ResNet [5] architecture is used as the encoder, and the ImageNet pretrained weights in [6] are imported for training. The Adam optimizer [7] was used, and training is performed for 22 epochs with a mini-batch size of 64 using a learning rate of 10^{-4}. The learning rate decay of 0.1 was applied at the 14th and 20th epoch. In addition, training was performed by feeding the concatenation of the SMPL edge map and the RGB image as an input. In the case of the structural map, it was trained for a total of 10 epochs, and the learning rate was reduced by 1/10 at the 6th and 8th epoch. Additionally, we used various types of structural maps as an input.

Student model. The student model used the same optimizer and learning rate as the teacher model. We trained for a total of 14 epochs and reduced the learning rate to 1/10 at the 10th epoch. As an input, a structural map with Canny edge [2] with 5×5 dilation, $I^{dilated\_edge}$, which had the best performance, was used. For both teacher and student models, 2D pose estimator [8] was used in the test phase.

As a result of our experiments, Canny edge [2] with 5×5 dilation and HED [3] show the highest accuracy in edge detection. Therefore, with these two edge detection results, we visually compare them to create an SMPL edge.

Method	3DPCK All↑	3DPCK Matched↑	Prior information
SMPLify-X [9]	62.8	68.0	OpenPose [8]
HMR [10]	65.6	68.6	Mask R-CNN [11]
HMR [10]	66.0	70.9	OpenPose [8]
Jiang [12]	69.1	72.2	
3DCrowdNet [4]	70.2	70.9	OpenPose [8]
3DCrowdNet [4]	72.7	73.3	HigherHRNet [13]
SEFD (Ours)	72.7	72.7	OpenPose [8]
SEFD (Ours)	**73.8**	**73.8**	**HigherHRNet** [13]

Table 1. Comparison on the MuPoTs [1] test dataset between SEFD and previous methods. The numbers denote 3DPCK for all annotations (All) and annotations matched to a prediction (Matched).

2. Evaluation of MuPoTs [1]

Unlike the training set, MuPoTs [1] filmed in an outdoor environment. Therefore, since the domain gap exists, I compared it with Baseline and other methods. Prior information in Table 1 means the 2D pose estimator [8, 13] or

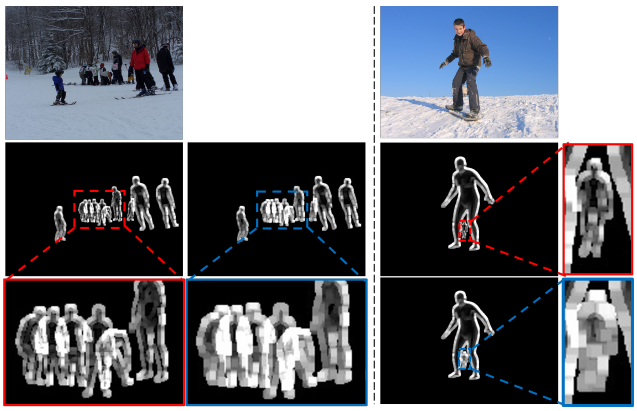


Figure 1. Difference between adaptive dilation and fixed 5×5 dilation kernel; red box shows the results of adaptive dilation and blue box shows the result of fixed 5×5 dilation kernel.

segmentation [11] used for performance evaluation, which is used externally.

3. Visualization on Comparing Proposed Method and Baseline

Fig. 7 shows the results of the 3DPW [14] and the 3DPW-OC benchmark. Figs. 8, 9, and 10 are the results of CrowdPose testset. Fig. 7 shows that SEFD has the better quantitative performance than baseline in all conditions, such as a situation in which a part of the body is occluded due to another object, sitting situation, and occlusion situation between people. Fig. 8 shows the occlusion in the crowded scene and Fig. 9 shows the complex pose case in the crowded scene. In Fig. 10, both the occlusion and complex pose in the crowded scene are shown. In these cases, SEFD shows the better quantitative result than baseline.

4. Adaptive Dilation

As shown in Fig. 1, if a dilation size is used without applying adaptive dilation, the boundary information for small objects cannot be properly obtained. Therefore, it is important to make dilation changeable according to the size of the object.

Fig. 2 is a histogram of the ground-truth person bounding box of MPII [15], MSCOCO [16], MuCo [1], and Human3.6M [17] benchmark datasets used for training the SEFD. The reason for visualizing the ground-truth person bounding box is to check how diverse the size of the object. Since various distribution exist for the bounding box sizes, using only one dilation kernel will cause the boundary information of the small objects to disappear, so various sized dilation kernels should be used.

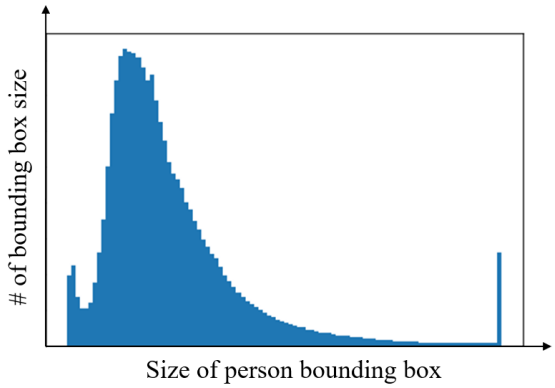


Figure 2. Histogram of bounding box size on MPII [15], MSCOCO [16], MuCo [1], and Human3.6M [17] benchmark dataset. The reason for visualizing the ground-truth person bounding box is to check how diverse the size of the object. The x-axis means the size of the person bounding box, and the y-axis means the number of bounding box sizes.

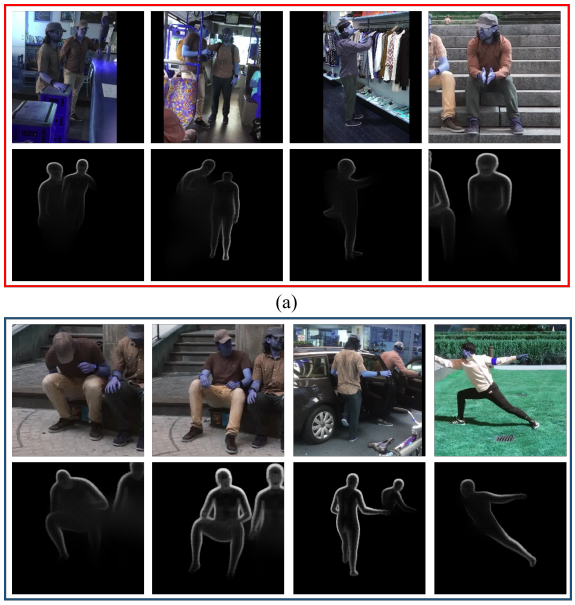
	Baseline	**SEFD**	SEE	SESD
Parameters	30.2M	**30.2M**	61.2M	92.2M
MACs	56.4G	**56.4G**	111.1G	156.8G

Table 2. Comparison of the model parameters and MACs between the baseline, SEFD, SEE, and SESD.

5. SMPL Edge Estimator

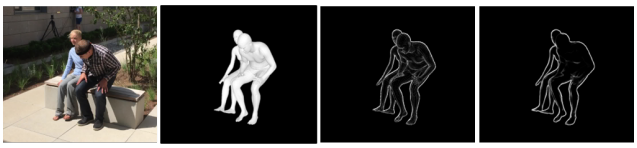
We confirmed that the performance improved when the SMPL overlapping edge $I^{overlap\_edge}$ is used. However, this SMPL overlapping edge could not be applied in the real environment as it was created using ground-truth. Therefore, we devised several approaches to use the $I^{overlap\_edge}$. This section explains the details of the SMPL edge estimator approach to estimate the SMPL edge, prior to the proposed (SEFD) approach.

To directly estimate the SMPL edge, a simple UNet [**?**] structure is utilized. To better predict the SMPL edge map, various combinations of loss are tested between the SMPL edge map and ground-truth edge. $\mathcal{L}_1$ loss is used to consider the general spatial area, and SSIM loss and VGG16 loss are adopted to follow the texture components and the perceptual similarity. $\mathcal{L}_1 \times 0.2 + \mathcal{L}_{VGG16} \times 0.8$ shows the best results. Furthermore, we create an occlusion-resistant SMPL edge estimator self-supervised de-occlusion (SESD) using the idea of [18]. However, as shown in Fig. 3, the occlusion and complex pose are not properly estimated. Also, after checking model parameters and Multiply-ACcumulate (MACs), it is confirmed that both model parameters and MACs show poor results compared to baseline and SEFD, which is shown in Table 2. Therefore, we use the SEFD model.



(a) (b)

Figure 3. Results for the SMPL edge estimator in various complex pose and occlusion: (a) basic UNet structure SMPL Edge Estimator (SEE) and (b) SMPL edge Estimator Self-supervised Deocclusion SESD.



(a) (b) (c) (d)

Figure 4. Comparative visualization of Canny Edge showing human internal information and HED where human internal information disappears. (a) Input image, (b) SMPL map, (c) Canny edge [2], and (d) HED [3].

6. Canny [2] vs HED [3]

For simple edge detection, we used Canny edge detection [2]. The reason Canny edge detector [2] was used is because the results of Canny edge with 5×5 dilation showed the better result than using HED [3]. Fig. 4 also shows why the Canny edge [2] is selected. Canny edge [2] exhibits human internal boundary information, while HED [3] cannot, and the internal boundary information is almost lost. Therefore, we chose the Canny edge detection [2] with 5×5 dilation as simple edge detection.

7. Problems with SMPL Edge Generation

This section explains the problems in generating SMPL edge maps for SEFD and how to solve them. In the case of the MPII [15] and MSCOCO [16] datasets, it is caused by using the pseudo ground truth. For the Human3.6M dataset [17], the multi-view viewpoint becomes a problem, as the SMPL parameter needs to be changed appropriately for each camera extrinsic parameter.

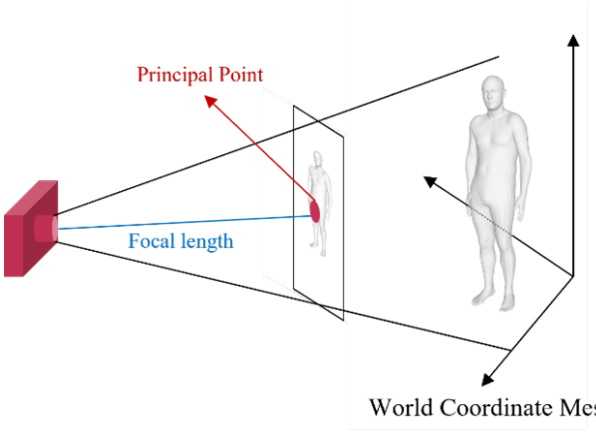


Figure 5. Illustration showing how the 3D mesh is projected onto the image plane according to the local length and principal point.

7.1. Problem with MPII [15] and MSCOCO [16] datasets

Fig. 5 shows how the focal length and the principal point are projected to a human mesh on an image plane. For the MPII [15] and MSCOCO [16] datasets, the focal length and principal point of the people present in an image are different. Therefore, the human mesh is projected to a different image plane depending on the focal length and principal point. Thus, if the coordinates of the mesh are not changed using the focal length and principal point, an inappropriate result will occur. Hence, the following transition must be carried out for generating the correct mesh:

$$\begin{bmatrix} u \\ v \\ 1 \end{bmatrix} = \begin{bmatrix} f_x & 0 & c_x \\ 0 & f_y & c_y \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} r_{11} & r_{12} & r_{13} & t_x \\ r_{21} & r_{22} & r_{23} & t_y \\ r_{31} & r_{32} & r_{33} & t_z \end{bmatrix} \begin{bmatrix} X \\ Y \\ Z \\ 1 \end{bmatrix}, \tag{1}$$

where (u, v) indicates the coordinate of the image plane, f_x, f_y denotes the focal length, c_x, c_y denotes the principal point, $r_{i,j}$ denotes the camera rotation, and t_i denotes the camera translation matrix. We will project the mesh of all the people in one image plane. Therefore, $r_{i,j}$ and t_i should have constant values in one image.

In general, the coordinate of the mesh is defined as (x, y, z). In the case when two people exist in an image, z^1 for the first person can be calculated with focal length for each person, f^1 and f^2.

$$z^2 = \frac{f^2}{f^1} * z^1. \tag{2}$$

As in Fig. 5, z should be changed according to the focal length. Followingly, Eq. (2) is derived. The image projection for each person, (u^i,v^i) can be derived using the following equations.

$$\begin{aligned} u^1 &= \frac{f_x^1}{z^1} * x^1 + c_x^1, \\ u^2 &= \frac{f_x^2}{z^2} * x^2 + c_x^2, \\ v^1 &= \frac{f_y^1}{z^1} * y^1 + c_y^1, \\ v^2 &= \frac{f_y^2}{z^2} * y^2 + c_y^2. \end{aligned} \tag{3}$$

When the second person is projected onto the image plane of the first person, the following equation can be derived:

$$\begin{aligned} u &= \frac{f_x^1}{z^2} * (x^2 + \frac{z^2}{f_x^1}(c_x^2 - c_x^1)) + c_x^1, \\ v &= \frac{f_y^1}{z^2} * (y^2 + \frac{z^2}{f_y^1}(c_y^2 - c_y^1)) + c_y^1. \end{aligned} \tag{4}$$

Following Eq. (4), the image projection to a single image plane is possible.

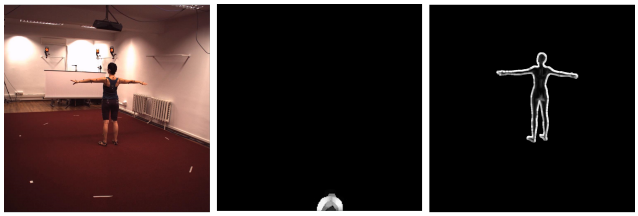
7.2. Problem with Human3.6M [17] dataset

In the case of the multi-view images, the camera extrinsic parameter varies according to each camera view. Therefore, it is necessary to change the SMPL parameter to match the camera extrinsic parameter. The camera extrinsic parameters $\overrightarrow{R} \in \mathbb{R}^{3\times3}$ and $\overrightarrow{t} \in \mathbb{R}^{3\times1}$, the SMPL translation parameter $\overrightarrow{t_{SMPL}} \in \mathbb{R}^{3\times1}$, SMPL 3D poses $\overrightarrow{\theta} \in \mathbb{R}^{23\times3}$, and root pose $\overrightarrow{\theta_{root}} \in \mathbb{R}^{3\times1}$ are included in the SMPL parameter. First of all, to fit $\overrightarrow{t_{SMPL}}$ to the camera extrinsic parameter, the calculated SMPL transition parameter $\overrightarrow{t_{SMPL}}^*$ must be set according to the following equation:

$$\begin{bmatrix} t_{SMPL_1}* \\ t_{SMPL_2}* \\ t_{SMPL_3}* \end{bmatrix} = \begin{bmatrix} r_{11} & r_{12} & r_{13} \\ r_{21} & r_{22} & r_{23} \\ r_{31} & r_{32} & r_{33} \end{bmatrix} \begin{bmatrix} t_{SMPL_1} \\ t_{SMPL_2} \\ t_{SMPL_3} \end{bmatrix} + \frac{1}{1000} \times \begin{bmatrix} t_1 \\ t_2 \\ t_3 \end{bmatrix}, \tag{5}$$

The reason for multiplying 1/1000 is to change from *mm* to *m*. To change $\overrightarrow{\theta_{root}}$, the following equation must be performed:

$$\overrightarrow{\theta_{root}}* = rod(\begin{bmatrix} r_{11} & r_{12} & r_{13} \\ r_{21} & r_{22} & r_{23} \\ r_{31} & r_{32} & r_{33} \end{bmatrix} (rod(\overrightarrow{\theta_{root}}))), \tag{6}$$



(a) (b) (c)

Figure 6. Visualization rendering problems and solutions in Human3.6M [17] datset. The rendering problem is solved by using camera extrinsic parameter to adjust the SMPL translation parameter $\overrightarrow{t_{SMPL}}$ and root pose $\overrightarrow{\theta_{root}}$. (a) an input image, (b) The SMPL edge with the rendering problem, and (c) the SMPL edge with the rendering problem solved.

In the above equation, $rod(coordinate)$ represents the Rodrigues' rotation formula, which is a method of deriving the rotation angle as a result. Through the above process, the problem of the camera extrinsic parameter is solved, and the results are shown in Fig. 6. Fig. 6 (b) shows the result when the rendering problem is not solved, and Fig. 6 (c) shows the result when the rendering problem is solved by using extrinsic camera parameters by adjusting the SMPL translation parameter $\overrightarrow{t_{SMPL}}$ and root pose $\overrightarrow{\theta_{root}}$.

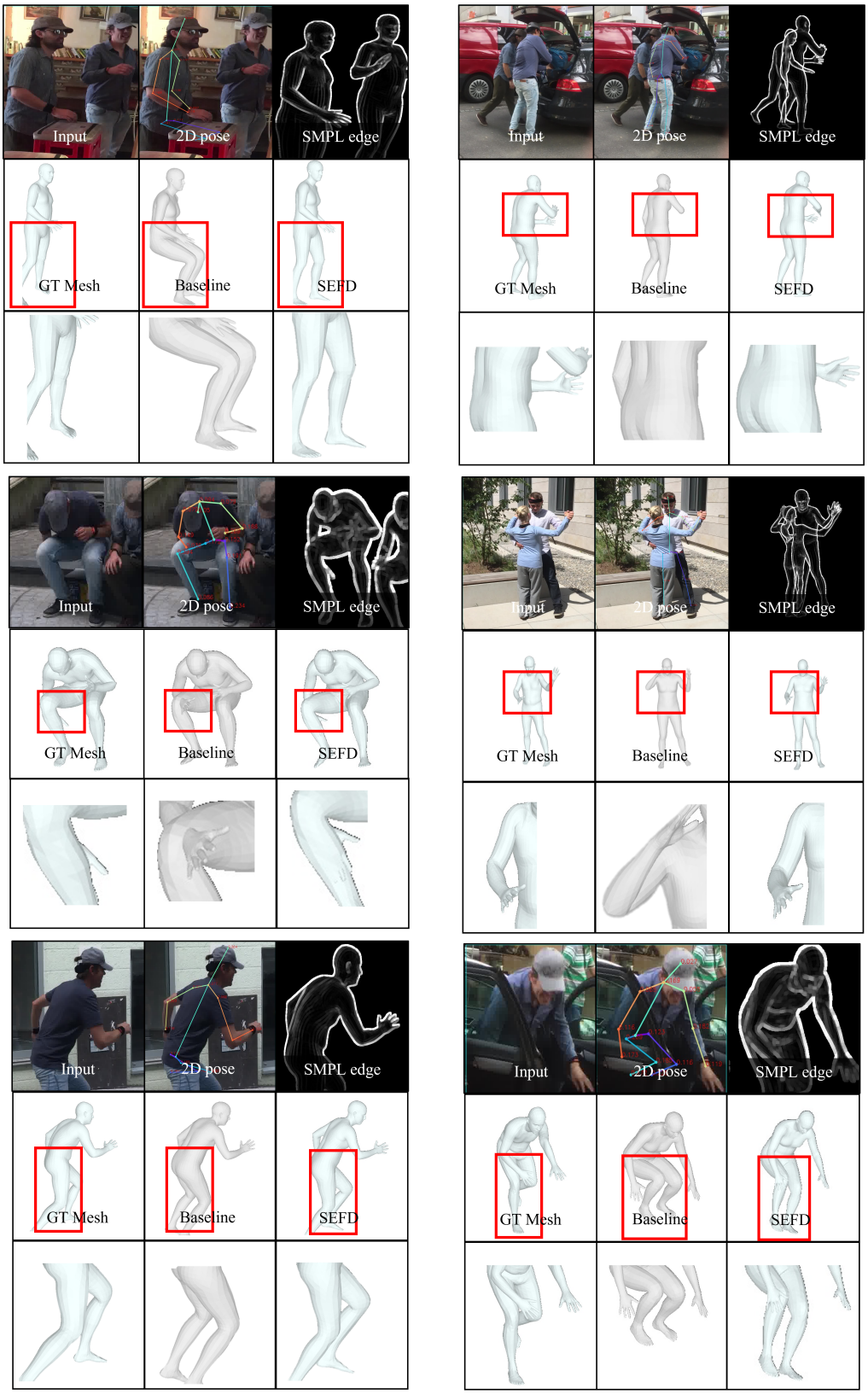


Figure 7. Performance comparison between the proposed and baseline method on the 3DPW test dataset and the 3DPW-OC dataset. The 2D pose could not be found properly, so the occlusion and complex pose could not be estimated properly.

Figure 8. Performance comparison between the proposed and baseline method on the CrowdPose testset. All rows correspond to the occluded cases. Specifically, rows 1, 2, and 4 show results in a low-light condition or blurry environment. Unlike the baseline method, which shows poor performance, the proposed method shows considerable performance robust to these harsh conditions. As in rows 5 and 7, even when more than half of the body part is occluded, the performance tendency between the baseline and the proposed method remains.

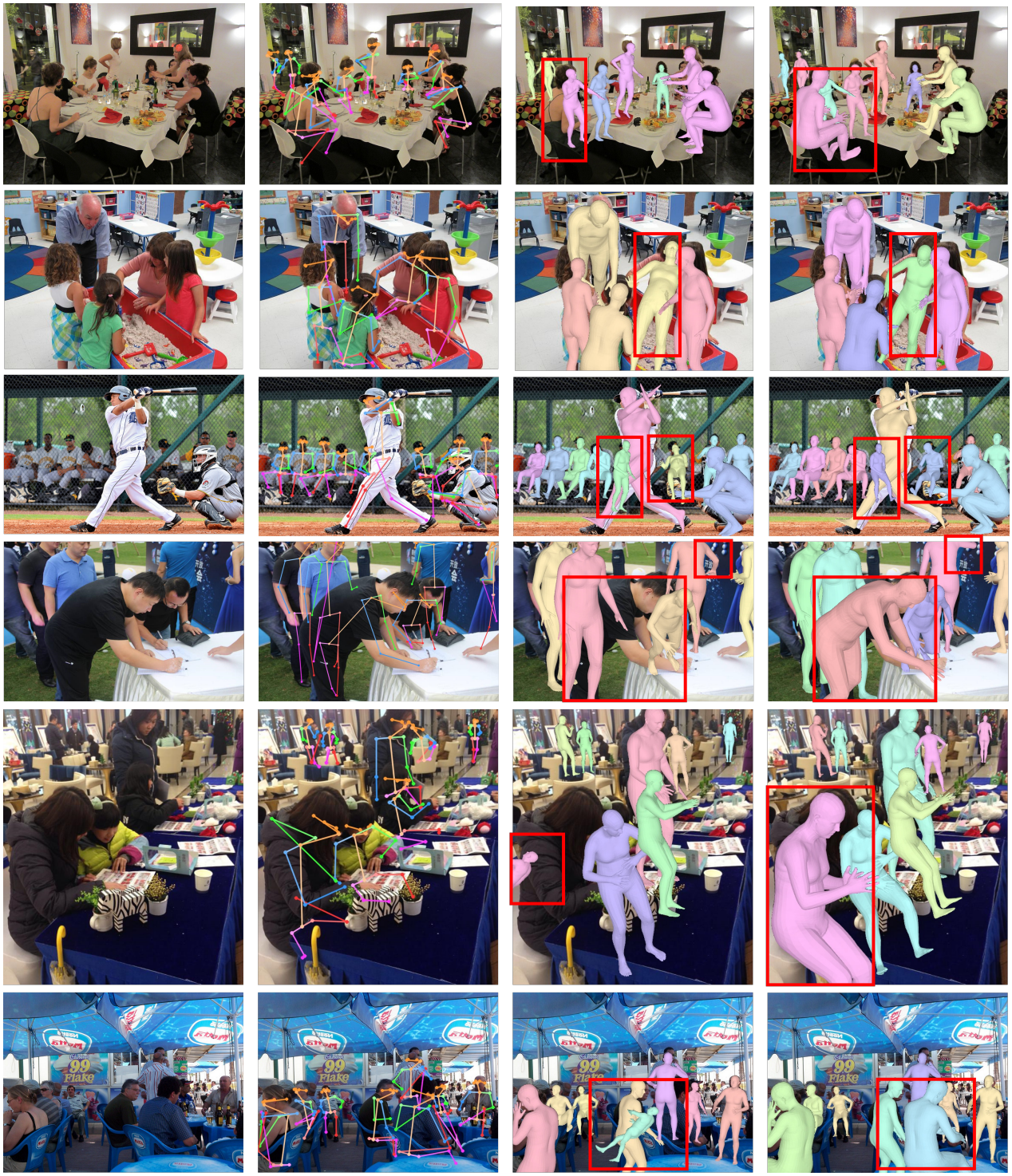


Figure 9. Performance comparison between the proposed and baseline method on the CrowdPose testset. All images correspond to complex pose cases. In particular, as in rows 1, 3, 5, and 6, the baseline method does not properly estimate the complex pose for a sitting person. However, the proposed method estimates properly even in a sedentary situation, showing robustness for the complex pose.

Figure 10. Performance comparison between the proposed and baseline method for CrowdPose testset. All images correspond to occlusion and complex pose cases. Specifically, rows 1 and 3 show results for highly crowded scenes. Unlike the baseline method, which shows poor performance, the proposed method shows considerable performance robust to these harsh conditions. As in rows 2, 4, 6, and 7, even when more than half of the body part is occluded, the performance tendency between the baseline and the proposed method remains.

References

[1] 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 *2018 International Conference on 3D Vision (3DV)*, pages 120–130. IEEE, 2018. 1, 2

[2] John Canny. A computational approach to edge detection. *IEEE TPAMI*, (6):679–698, 1986. 1, 3

[3] Saining Xie and Zhuowen Tu. Holistically-nested edge detection. In *Proceedings of the IEEE international conference on computer vision*, pages 1395–1403, 2015. 1, 3

[4] Hongsuk Choi, Gyeongsik Moon, JoonKyu Park, and Kyoung Mu Lee. Learning to estimate robust 3d human mesh from in-the-wild crowded scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1475–1484, 2022. 1

[5] 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. 1

[6] 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

[7] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 1

[8] 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. 1

[9] Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed AA Osman, Dimitrios Tzionas, and Michael J Black. Expressive body capture: 3d hands, face, and body from a single image. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10975–10985, 2019. 1

[10] Angjoo Kanazawa, Michael J Black, David W Jacobs, and Jitendra Malik. End-to-end recovery of human shape and pose. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7122–7131, 2018. 1

[11] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969, 2017. 1, 2

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

[13] 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. 1

[14] Timo Von Marcard, Roberto Henschel, Michael J Black, Bodo Rosenhahn, and Gerard Pons-Moll. Recovering accurate 3d human pose in the wild using imus and a moving camera. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 601–617, 2018. 2

[15] 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. 2, 3

[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 *European conference on computer vision*, pages 740–755. Springer, 2014. 2, 3

[17] 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 transactions on pattern analysis and machine intelligence*, 36(7):1325–1339, 2013. 2, 3, 4

[18] Xiaohang Zhan, Xingang Pan, Bo Dai, Ziwei Liu, Dahua Lin, and Chen Change Loy. Self-supervised scene deocclusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3784–3792, 2020. 2