Supplementary Material

Shihao Zou[1], Xinxin Zuo[1], Yiming Qian[2], Sen Wang[1], Chi Xu[3], Minglun Gong[4], and Li Cheng[1]

[1] University of Alberta
[2] Simon Fraser University
[3] School of Automation, China University of Geosciences, Wuhan 430074, China
[4] University of Guelph
{szou2,xzuo,sen9,lcheng5}@ualberta.ca,yimingq@sfu.ca,
xuchi@cug.edu.cn,minglun@uoguelph.ca

1 Implementation Details

The encoder-decoder model architecture is used in the stage of normal estimation. In the stage of shape and pose estimation, we use ResNet50 [1] and the average-pooled output is directly regressed to an 85-dimension vector. The polarization image and the predicted normal map are concatenated as the input to estimate the SMPL shape parameters. ResNet50 is trained from scratch.

We synthesize a polarization image (polarizers of 0, 45, 90 and 135 degree) given the rendered depth and color image. In detail, from the depth image, we obtain the normal map and calculate the zenith and azimuth angle, and from the color image, we get the gray image and take it as the polarization image of 0 degree polarizer, denoted by $I(0)$. Assuming diffuse reflection of the human body surface, we can calculate the degree of polarization ρ according to the equation,

$$\rho = \frac{(n - \frac{1}{n})^2 \sin^2\theta}{2 + 2n^2 - (n + \frac{1}{n})^2 \sin^2\theta + 4\cos\theta\sqrt{n^2 - \sin^2\theta}}, \tag{1}$$

with the zenith angle and refractive index known. Then the upper and lower bound of the illumination intensity I_{max} and I_{min} can be solved in closed-form with the constraints,

$$\rho = \frac{I_{max} - I_{min}}{I_{max} + I_{min}}, \tag{2}$$

and

$$I(0) = \frac{I_{max} + I_{min}}{2} + \frac{I_{max} - I_{min}}{2}\cos(2\varphi)). \tag{3}$$

Finally, we can use the equation,

$$I(\phi_{pol}) = \frac{I_{max} + I_{min}}{2} + \frac{I_{max} + I_{min}}{2}\cos(2(\phi_{pol} - \varphi)), \tag{4}$$

to get the image for polarizer ϕ_{pol} of degree 45, 90 and 135. To make it close to the real-world applications, we add Gaussian noise with $\sigma = 1/255$ to each pixel of the synthetic polarization image and then quantize the intensity value to 8

bits. Due to the fact that we only have geometric information for human body, the synthetic polarization images only have values on human body part.

We first train the normal estimation model for 20 epochs by setting λ_c and λ_n to be 2 and 1 respectively. The learning rate starts at 0.001 and decays to 0.0001 after 15 epochs. Then we train the shape estimation model for 30 epochs by setting λ_β, λ_θ, λ_t and λ_J to be 0.2, 0.5, 100 and 3 respectively. The learning rate starts at 0.001 and decays to 0.0001 after 5 epochs. Adam optimizer [2] is used to train our model.

To deform the SMPL based human mesh model towards the predicted normal map, first we integrate a depth map from the predicted normal map with the projected SMPL mesh as the coarse depth. In detail, we define a objective function as

$$E(D) = \lambda_n E_{\mathbf{n}}(D) + \lambda_d E_d(D) + \lambda_s E_s(D), \tag{5}$$

and we get the detailed depth via minimization of this function.

The first term, $E_{\mathbf{n}}(D)$, is used to enforce the predicted normal to be perpendicular to the tangents of the optimized depth surface,

$$E_{\mathbf{n}}(D) = \sum_i T_x^i n_i + T_y^i n_i. \tag{6}$$

The tangents T_x and T_y are defined as below,

$$T_x = [\frac{1}{f_x}(\frac{\partial D}{\partial x}(x - p_x) + D), \frac{1}{f_y}\frac{\partial D}{\partial x}(y - p_y), \frac{\partial D}{\partial x}]^T, \tag{7}$$

$$T_y = [\frac{1}{f_x}\frac{\partial D}{\partial x}(y - p_y), \frac{1}{f_y}(\frac{\partial D}{\partial y}(y - p_y) + D), \frac{\partial D}{\partial y}]^T. \tag{8}$$

In the above function, f_x and f_y are the focal length and p_x and p_y are the camera center of the camera.

For the second term $E_d(D)$, we set the boundary constraints so that the optimized depth to be close to the base depth $\hat{D}_i$,

$$E_d(D) = \sum_i \Big(((\frac{x_i - p_x}{f_x})^2 + (\frac{y_i - p_y}{f_y})^2 + 1)(D_i - \hat{D}_i) \Big)^2. \tag{9}$$

Finally we want to preserve smoothness for the integrated surface and add the smoothness constraints for neighboring pixels in the third term $E_s(D)$,

$$E_s(D) = \sum_{i,j \in N} ||D_i - D_j||^2. \tag{10}$$

We find a linear least squares solution of the objective Eq. (5) and get the detailed depth map.

2 More results

We show the results of normal estimation and shape estimation on SURREAL dataset in Fig. 1 and 2 respectively. We display the corresponding color image in the first column for better demonstration. The results of Linear [3] and Physics [4] are also shown. The synthetic polarization images (displayed the first channel as a gray image) are shown in the second column, which are the input to estimate normal map and also the shape and pose. Due to the fact that we only have geometric information for human body, the synthetic polarization images only have values on human body part.

A video[5] is presented to give a more comphrehensive view of our detailed shape. The video shows the predicted detailed human shape from different view angles. On the one hand, compared with PIFu [5] and Depth Human [6], we can see that our method is more robust to complex poses especially when we change the angle of view. On the other hand, compared with HMD [7], our method can recover more reliable clothing details of human body.

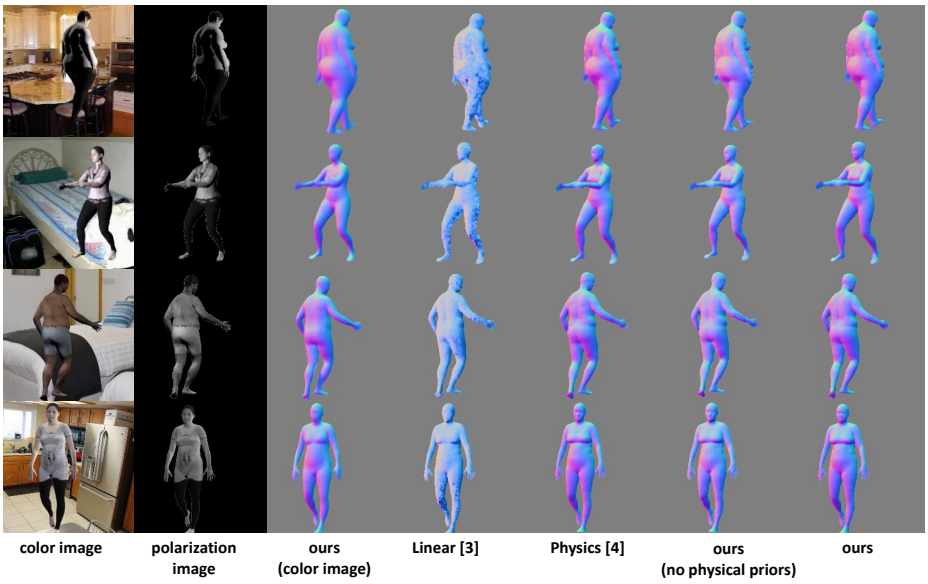

Fig. 1. The figure shows the results of normal estimation on SURREAL dataset. The first column is color image for better visualization. The second column is the synthetic polarization image as the input to estimate normal map. The third column is the result from ours (color image). The fourth and fifth columns are the results from Linear [3] and Physics [4]. The sixth column is ours (no physical priors). Compared with ours (color image), we can see that the better surface normal is predicted by ours.

[5] The video, named *detailed_shape_demo.avi*, is submitted in the supplementary material.

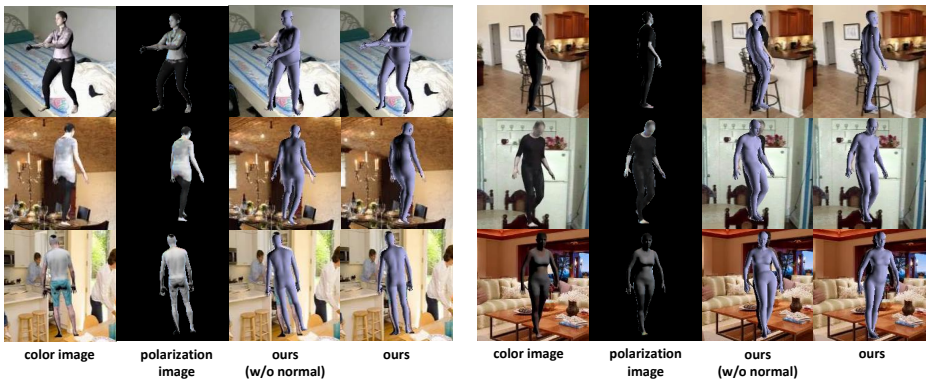

Fig. 2. The figure shows the results of shape and pose estimation on SURREAL dataset. We show the color image in the first column for better visualization. The second column is polarization image, which is the input to estimate the shape. The third column is the result from ours (w/o normal). We can see that normal map is an informative priors to learn to predict better human shape from a polarization image.

3 Polarization Human Shape and Pose Dataset (PHSPD)

More details can be found on the site [6].

3.1 Data Acquisition

Our acquisition system synchronizes four cameras, one polarization camera and three Kinects V2 in three different views (each Kinect v2 has a depth and a color camera). The layout is shown in Fig. 3. The other task is multi-camera synchronization. As one PC can only control one Kinect V2, we develop a soft synchronization method. Specifically, each camera was connected with a desktop (the desktop with the polarization camera is the master and the other three ones with three Kinects are clients). We use socket to send message to each desktop. After receiving certain message, each client will capture the most recent frame from the Kinect into the desktop memory. At the same time, the master desktop sends a software trigger to the polarization camera to capture one frame into the buffer. Fig. 3 shows the synchronization performance of the system that we develop. We let a bag fall down and compare the position of the bag in the same frame from four views. We can find that the positions of the bag captured by four cameras are almost the same in terms of its distance to the ground.

Our dataset has 12 subjects, 9 male and 3 female subjects. Each subject is required to do 3 different groups of actions (18 different actions in total) for 4 times plus one free-style group. Details are shown in Tab. 1. So each subject has

[6] https://jimmyzou.github.io/publication/2020-PHSPDataset

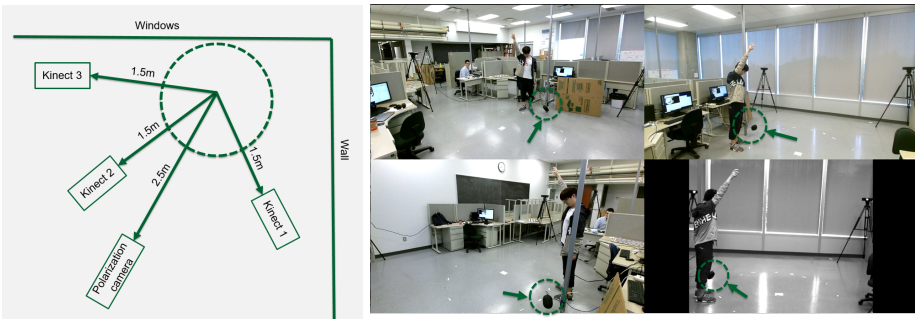

Fig. 3. Left figure: the layout of our multi-camera system. Three Kinects are placed around a circle of motion area with one polarization camera. Right figure: the synchronization result of our multi-camera system. The same frame of the three-view color images and one-view polarization image are displayed. Note that the layout of our multi-camera system has been changed to the left figure, but other settings are the same.

group #	actions
1	warming-up, walking, running, jumping, drinking, lifting dumbbells
2	sitting, eating, driving, reading, phoning, waiting
3	presenting, boxing, posing, throwing, greeting, hugging, shaking hands

Table 1. The table displays the actions in each group. Subjects are required to do each group of actions for four times, but the order of the actions each time is random.

subject #	gender	# of original frames	# of annotated frames	# of discarded frames
1	female	22561	22241	320 (1.4%)
2	male	24325	24186	139 (0.5%)
3	male	23918	23470	448 (1.8%)
4	male	24242	23906	336 (1.4%)
5	male	24823	23430	1393 (5.6%)
6	male	24032	23523	509 (2.1%)
7	female	22598	22362	236 (1.0%)
8	male	23965	23459	506 (2.1%)
9	male	24712	24556	156 (0.6%)
10	female	24040	23581	459 (1.9%)
11	male	24303	23795	508 (2.1%)
12	male	24355	23603	752 (3.1%)
total	-	287874	282112	5762 (2.0%)

Table 2. The table shows the detail number of frames for each subject and also the number of frames that have SMPL shape and 3D joint annotations.

13 short videos and the total number of frames for each subject is around 22K. Overall, our dataset has 287K frames with each frame including one polarization image, three color and three depth images. Quantitative details of our dataset are shown in Tab. 2

3.2 Annotation Process

The reason that we use multi-camera system to acquire image data is that multi-camera system provides much more information than a single-camera system. So the annotation of SMPL human shape and 3D joint position is more reliable using information of three-view Kinects v2.

After camera calibration and plane segmentation of depth images, now we have a point cloud of human fused from three-view depth image and noisy 3D joint position by Kinect SDK at hand. The annotation SMPL human shape and 3D joint position has three main steps. First step is to filter out accurate 3D joint position by three Kinects in three views. For each view, we get the 2D joint estimation by OpenPose [8] and also the 2D Kinect joint by projecting the noisy Kinect 3D joint to the color image.

Then we compare the 2D distance between these two estimated joints. If the distance is larger than 50 pixel distance, we regard the joint estimated by the Kinect as incorrect one. As we have the joint estimation from three views, we simply average the correct joint position of three views and consider it as the initial guess of the position of the joint. If none of the three-view estimated joint is correct, we consider it as a missing joint. In this way, we get the initial guess of 3D joint positions for each frame and we discard the frame with more than 2 joints missing (14 in total). The next step is similar to [9], but instead of fitting to the 2D joints which have inherent depth ambiguity, we fit SMPL model to the initial guess of 3D joints.

Furthermore, as we have the point cloud of a human from three-view depth cameras, our final step is to further iteratively optimize SMPL parameters by minimizing the distance between vertices of SMPL shape to their nearest point. Finally, we have the annotated SMPL shape parameters and 3D joint positions.

Besides, we render the boundary of SMPL shape on the image to get the mask of background, and calculate the target normal using three depth images based on [10]. Although the target normal is noisy, our experiment result shows our model can still learn to predict good and smooth normal maps.

The annotation process is shown in Fig. 4. Starting from the initial guess of 3D pose, we fit SMPL shape to the initial 3D pose and further fit to the point cloud of human mesh. Finally, we get annotated human shape and pose. Besides, we also show our annotated shape on multi-view images (one polarization image and three-view color image) and the human pose in 3D coordinate space in Fig 5. We also have a video [7] to show our annotation results.

[7] The video, named *annotation_demo.avi*, is submitted in the supplementary material.

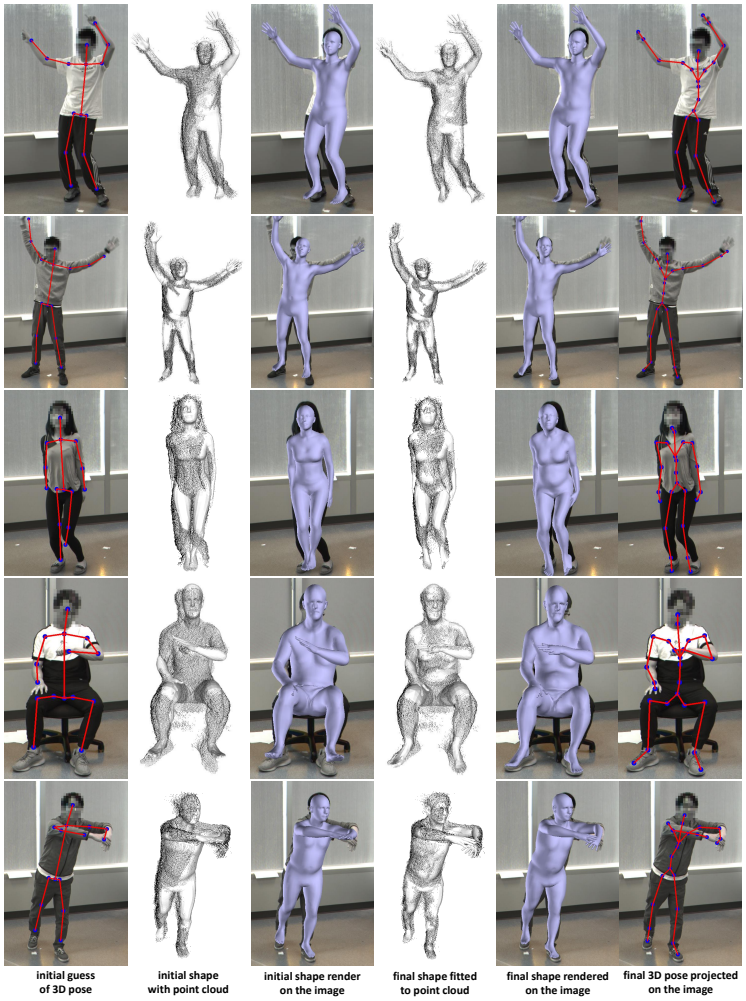

Fig. 4. The figure shows our annotation process. The first column shows the initial guess of 3D pose, which is projected on the polarization image. After fitting the SMPL shape to the initial pose, we show the initial shape with the point cloud of human mesh (black points) in the second column and the rendered shape on the image in the third column. The fourth and fifth columns show the annotated shape after fitting to the point cloud of human mesh. The sixth column shows the corresponding annotated 3D pose.

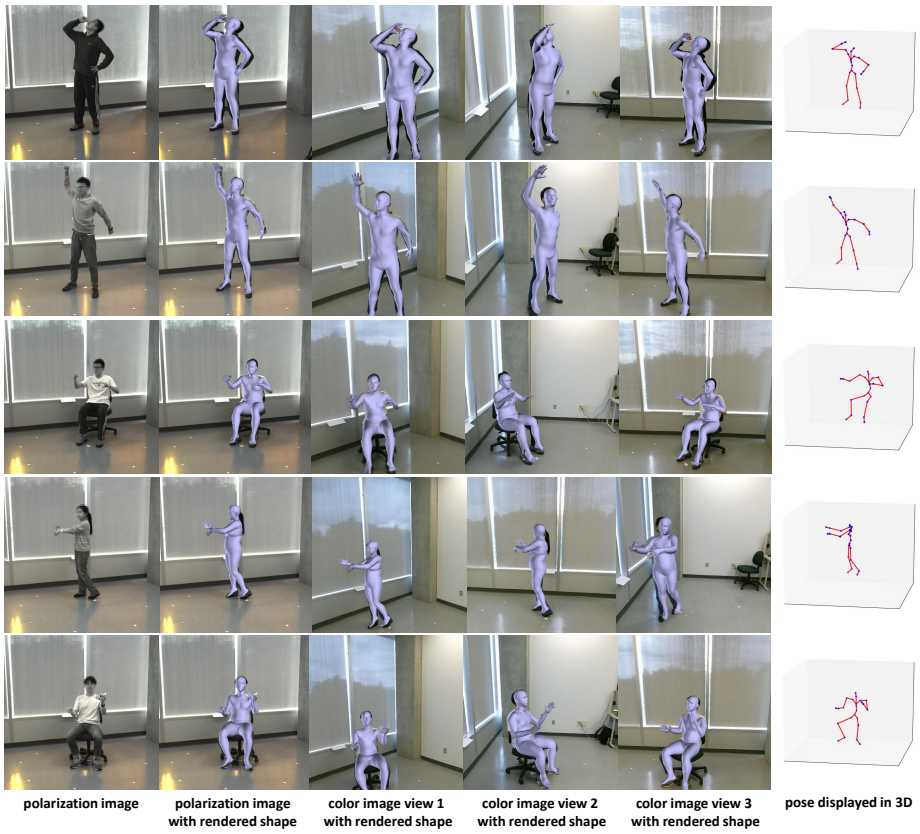

Fig. 5. The figure shows our annotated shapes and poses. The first column is the polarization image for reference. The second to the fifth columns show the annotated shape rendered on the polarization image and three-view color images. The sixth column shows the annotated pose in 3D space.

References

1. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. (2016) 770–778
2. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
3. Smith, W.A., Ramamoorthi, R., Tozza, S.: Linear depth estimation from an uncalibrated, monocular polarisation image. In: European Conference on Computer Vision, Springer (2016) 109–125
4. Ba, Y., Chen, R., Wang, Y., Yan, L., Shi, B., Kadambi, A.: Physics-based neural networks for shape from polarization. arXiv preprint arXiv:1903.10210 (2019)
5. Saito, S., Huang, Z., Natsume, R., Morishima, S., Kanazawa, A., Li, H.: Pifu: pixel-aligned implicit function for high-resolution clothed human digitization. In: Proceedings of the IEEE International Conference on Computer Vision. (2019) 2304–2314
6. Tang, S., Tan, F., Cheng, K., Li, Z., Zhu, S., Tan, P.: A neural network for detailed human depth estimation from a single image. In: Proceedings of the IEEE International Conference on Computer Vision. (2019) 7750–7759
7. Zhu, H., Zuo, X., Wang, S., Cao, X., Yang, R.: Detailed human shape estimation from a single image by hierarchical mesh deformation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (2019) 4491–4500
8. Cao, Z., Martinez, G.H., Simon, T., Wei, S., Sheikh, Y.A.: Openpose: Realtime multi-person 2d pose estimation using part affinity fields. IEEE Transactions on Pattern Analysis and Machine Intelligence (2019)
9. Bogo, F., Kanazawa, A., Lassner, C., Gehler, P., Romero, J., Black, M.J.: Keep it smpl: Automatic estimation of 3d human pose and shape from a single image. In: European Conference on Computer Vision, Springer (2016) 561–578
10. Qi, X., Liao, R., Liu, Z., Urtasun, R., Jia, J.: Geonet: Geometric neural network for joint depth and surface normal estimation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (2018) 283–291