A Neural Network for Detailed Human Depth Estimation from a Single Image
Supplementary Material

Sicong Tang[1,*] Feitong Tan[1,*] Kelvin Cheng[1] Zhaoyang Li[1] Siyu Zhu[2] Ping Tan[1]
[1] Simon Fraser University [2] Alibaba A.I Labs
{sta105, feitongt, kelvinz, zla143, pingtan}@sfu.ca, siting.zsy@alibaba-inc.com

1. Qualitative Results and Failure Cases

We show more qualitative results on the wild Internet data to demonstrate the generality of our method. As shown in the Figure 1, our model can recover faithful details and wrinkles only various challenging cases, even our training data has limited clothes styles, age and gender group (e.g., without women and children). Some failure cases are shown in the Figure 2. Our model have difficulties in dealing with loose clothes, unseen poses, or ambiguous poses.

2. Depth Refinement

Nehab et al. [1] formulate the depth refinement with surface normals as an energy minimization problem and achieve a good result. But its solution requires computing the inverse of a large sparse matrix, which is not suitable for neural networks. We re-formulate this problem with an iterative solution. Here are the details of our derivation.

Depth refinement aims to find a depth map conforming to the initial result and a normal map. Following Nehab et al. [1], we compute the depth map by minimizing the sum of two errors, the *depth error* E_z and the *normal error* E_n:

$$E = \lambda E_z + (1-\lambda)E_n, \tag{1}$$

where λ is a hyper-parameter to balance the two terms. The depth error computes the difference between the final estimated depth and the initial depth, i.e.

$$E_z = \sum_i ||Z_i - Z_i^0||^2, \tag{2}$$

where Z_i and Z_i^0 are the estimated and initial depth at the i-th pixel. The normal term E_n requires normals to be perpendicular to the surface tangent directions:

$$E_n = \sum_i [T_i \cdot N_i]^2 \tag{3}$$

where N_i and T_i are the surface normal and tangent directions of the i-th pixel respectively. In our case, the surface normal direction is a fixed input, the tangent direction is calculated from a set of neighboring pixels. Thus, the normal error is:

$$\begin{aligned} E_n = &\sum_i \sum_{j \in \mathcal{N}_i} \\ &||N_{jx}(X_j - X_i) + N_{jy}(Y_j - Y_i) + N_{jz}(Z_j - Z_i)||^2 + \\ &||N_{ix}(X_i - X_j) + N_{iy}(Y_i - Y_j) + N_{iz}(Z_i - Z_j)||^2 \end{aligned} \tag{4}$$

Here, (N_{ix}, N_{iy}, N_{iz}) and (X_i, Y_i, Z_i) are the normal and position of the i-th pixel. $\mathcal{N}_i$ contains the four neighboring pixels of i. We compute the surface tangents by both forward $(i \to j)$ and backward differentiation $(j \to i)$.

To minimize the energy function E, we iteratively update Z_i by fixing the depths of Z_j for all $j \in \mathcal{N}_i$. Specifically, Z_i can be computed as,

$$Z_i^{n+1} = \lambda Z_i^0 + (1-\lambda)\frac{\sum_{j \in \mathcal{N}_i}\left(Z_{ij}^n + Z_{ji}^n\right)}{8}, \tag{5}$$

where Z_{ij} is the depth of i that makes the edge ij and N_j perpendicular, and Z_{ji} is the depth of i that makes ij and N_i perpendicular. Specifically, these two can be computed as,

$$\begin{aligned} Z_{ij}^n &= \frac{N_{jx}(X_j^n - X_i^n) + N_{jy}(Y_j^n - Y_i^n) + N_{jz}Z_j^n}{N_{jz}}, \\ Z_{ji}^n &= \frac{N_{ix}(X_j^n - X_i^n) + N_{iy}(Y_j^n - Y_i^n) + N_{iz}Z_i^n}{N_{iz}}. \end{aligned} \tag{6}$$

where (X_i^n, Y_i^n, Z_i^n) and $(X_i^{n+1}, Y_i^{n+1}, Z_i^{n+1})$ are the 3D positions of the pixel i in the n-th and $n+1$-th iterations respectively.

3. Normal Evaluation

We compare our Normal-Net with the network in Zhang et al.[2] which is finetuned on our dataset. In Table 1, it shows that our method outperforms the method in Zhang et al. [2] regarding all different metrics. Several qualitative results are also given in the Figure 3.

*These authors contributed equally to this work.

Figure 1. Additional results on Internet data. For each example, from left to right, these images are: the single input RGB image, estimated normal and final depth.

Figure 2. Some failure cases on Internet data. The subfigures are arranged in the same way as in Figure 1.

Method	Accuracy			Error	
	11.25°	22.5°	30°	Mean	Median
Our method	**26.60**	**63.52**	**78.18**	**21.94**	**17.79**
Zhang et al. [2]	22.28	56.23	72.33	24.07	20.20

Table 1. Accuracy and Mean/Median normal errors of different methods.

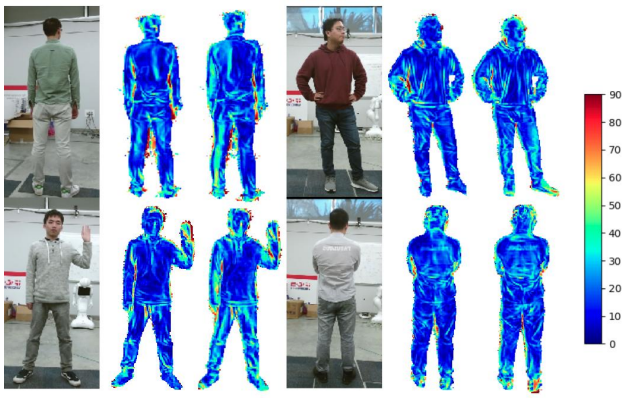

Figure 3. Qualitative comparison with [2]. From left to right, these are input images, error maps of estimated normals by our method and [2] respectively.

References

[1] Diego Nehab, Szymon Rusinkiewicz, James Davis, and Ravi Ramamoorthi. Efficiently combining positions and normals for precise 3d geometry. In *ACM Trans. Gr.*, volume 24, pages 536–543. ACM, 2005. 1

[2] Yinda Zhang, Shuran Song, Ersin Yumer, Manolis Savva, Joon-Young Lee, Hailin Jin, and Thomas Funkhouser. Physically-based rendering for indoor scene understanding using convolutional neural networks. In *CVPR*, 2017. 1, 2