Learning High Fidelity Depths of Dressed Humans by Watching Social Media Dance Videos

Supplementary Material

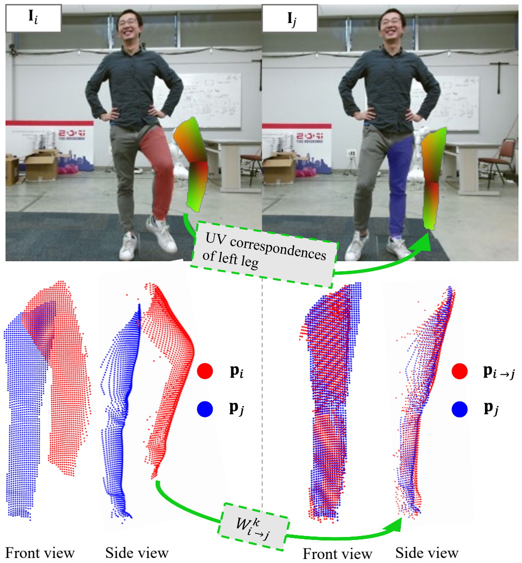

Figure 10: Our method can handle significant depth change.

A. Handling Large Pose Variation

Our method can handle moving body parts, such as arms and legs, that induce significant depth variation across time. Specifically, the 3D translation in $\mathcal{W}^k_{i\to j}$ is designed to account for such changes in depth. Figure 10 shows a large depth and pose change of the left leg between frames where the 3D points ($\mathbf{p}_i$) can be correctly transformed to the other frame ($\mathbf{p}_{i\to j} = \mathcal{W}^k_{i\to j}(\mathbf{p}_i)$).

Pose variant cloth geometry, such as cloth wrinkles, may not be approximated by $SE3$ with UV correspondences. However, our method is agnostic to the choice of transformation, i.e., it can be generalized to affine, perspective, or even non-parametric transformation without loss of generality. Further, our measure of the surface normal consistency enforces the prediction to match the surface of presented image, which can reconstruct the local fine geometry as shown in reconstructed surfaces of Figure 11.

B. Effect of Self-supervision

Our method makes a positive impact on the plausibility of reconstruction. Without it, the trained model is highly over-fitted to the scanned data, which produces unrealistic reconstruction as shown in Figure 8. The head is reconstructed far behind the torso mainly due to the small size of the head. As the mean and median errors are not the best descriptive metrics to capture such qualitative plausibility, we further analyze the error by computing its distribution using error histogram shown in Figure 12. The self-supervision results in majority of pixels remaining in the lower error regions and a smaller number of pixels in outlier regions (shorter tail error distribution). We also report the ablation study of Table 2 for RenderPeople dataset [2] in Table 3. As the training data is similar to the test data, the self-supervision does not make an impact on error reduction.

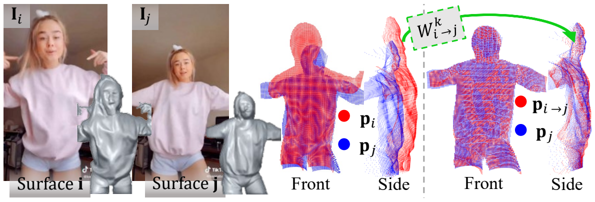

Figure 11: Our method can handle cloth geometry changes in the case of loose clothing.

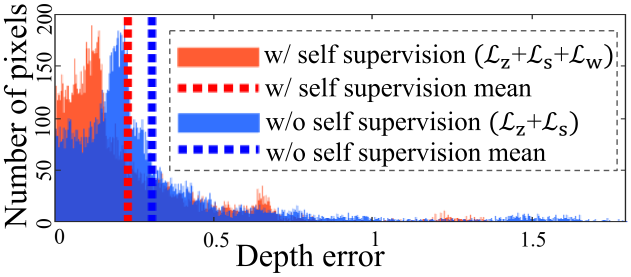

Figure 12: Histogram of error.

Losses	D. error	N. error	R. error
$\mathcal{L}_\mathbf{z}$	1.388±0.860	0.370±0.071	0.066±0.023
$\mathcal{L}_\mathbf{z} + \mathcal{L}_\mathbf{s}$	1.193±0.820	0.282±0.053	**0.059±0.020**
$\mathcal{L}_\mathbf{z} + \mathcal{L}_\mathbf{s} + \mathcal{L}_\mathbf{w}$	**1.115±0.755**	**0.275±0.051**	**0.059±0.021**

Table 3: Ablation study on RenderPeople dataset [2]. D. error (normalized error), N. error (rad) and R. error represent depth error, normal error, and reconstruction error respectively (mean±std).

C. More Qualitative Evaluation

Figure 13 and 14 show more evaluation results on Vlasic et al. dataset [53]. Figure 16 and 17 show more evaluation results on Tang et al. dataset [50] and Figure 15 shows the evaluation results on RenderPeople dataset [2].

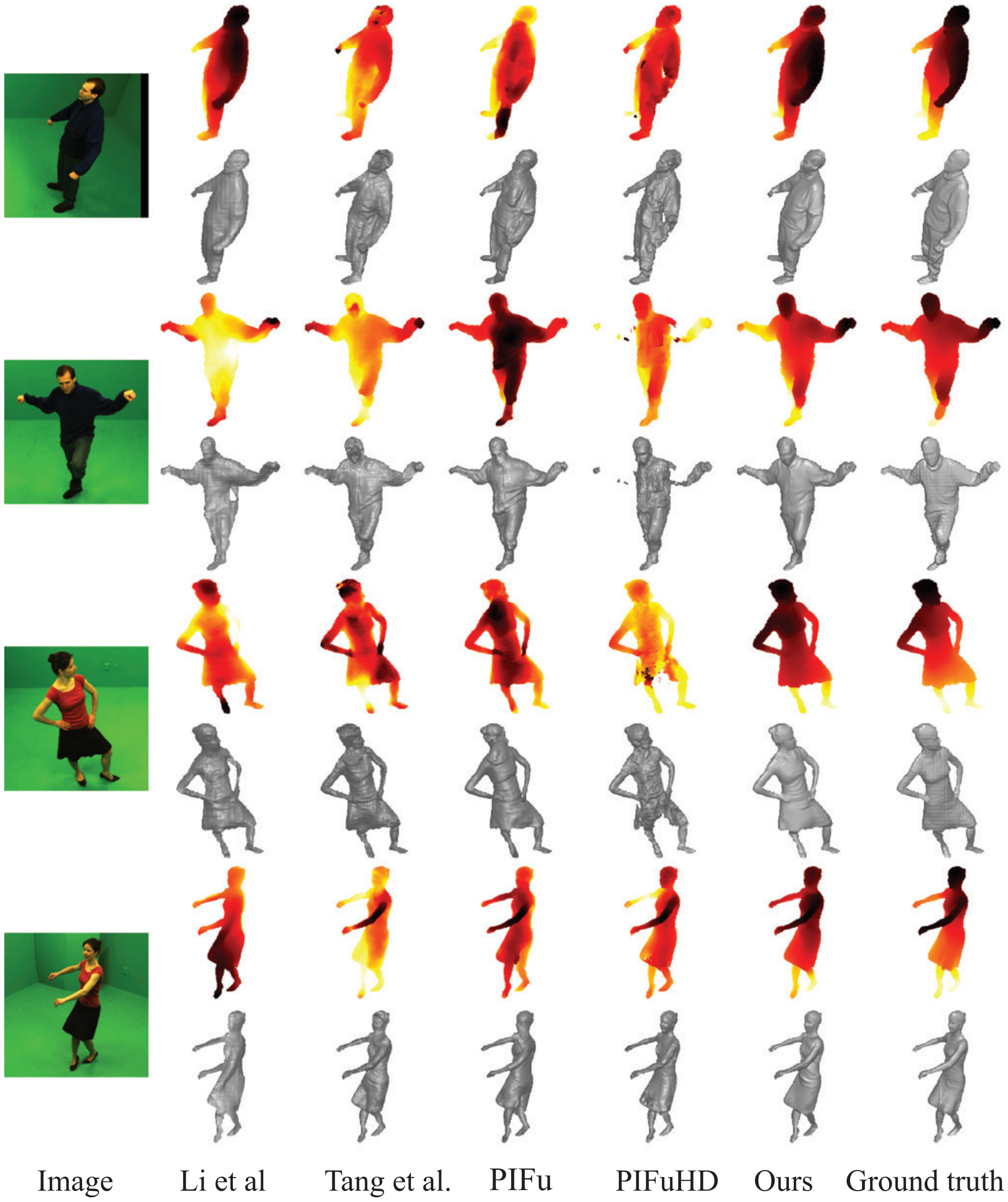

Figure 13: Qualitative comparison on Vlasic et al. dataset

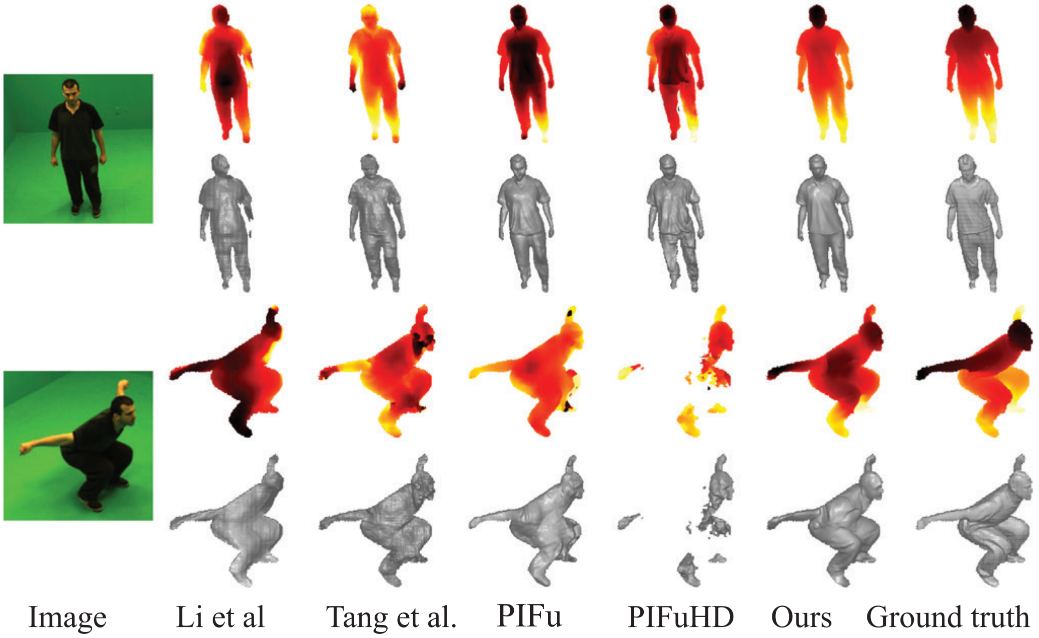

Figure 14: Qualitative comparison on Vlasic et al. dataset

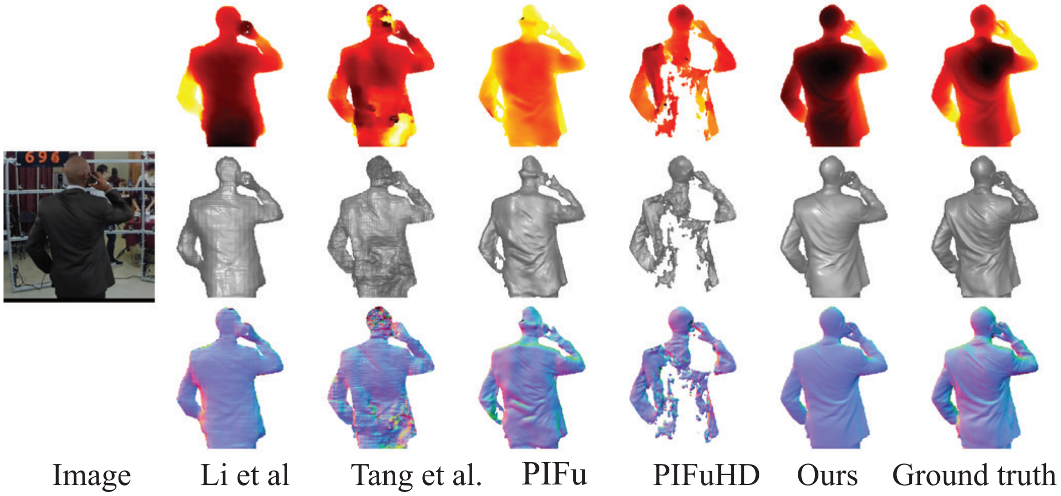

Figure 15: Qualitative comparison on RenderPeople testing dataset

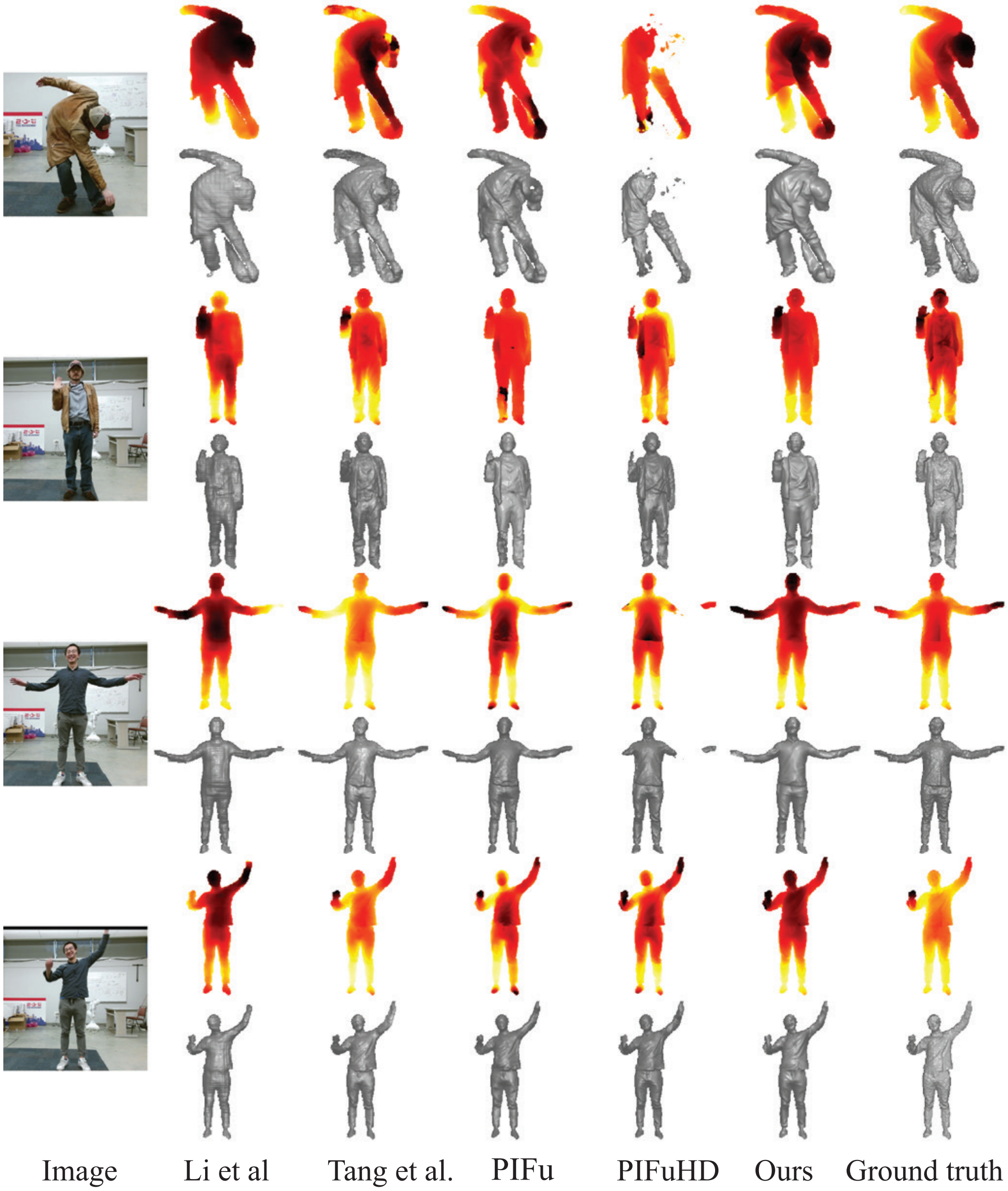

Figure 16: Qualitative comparison on Tang et al. training dataset

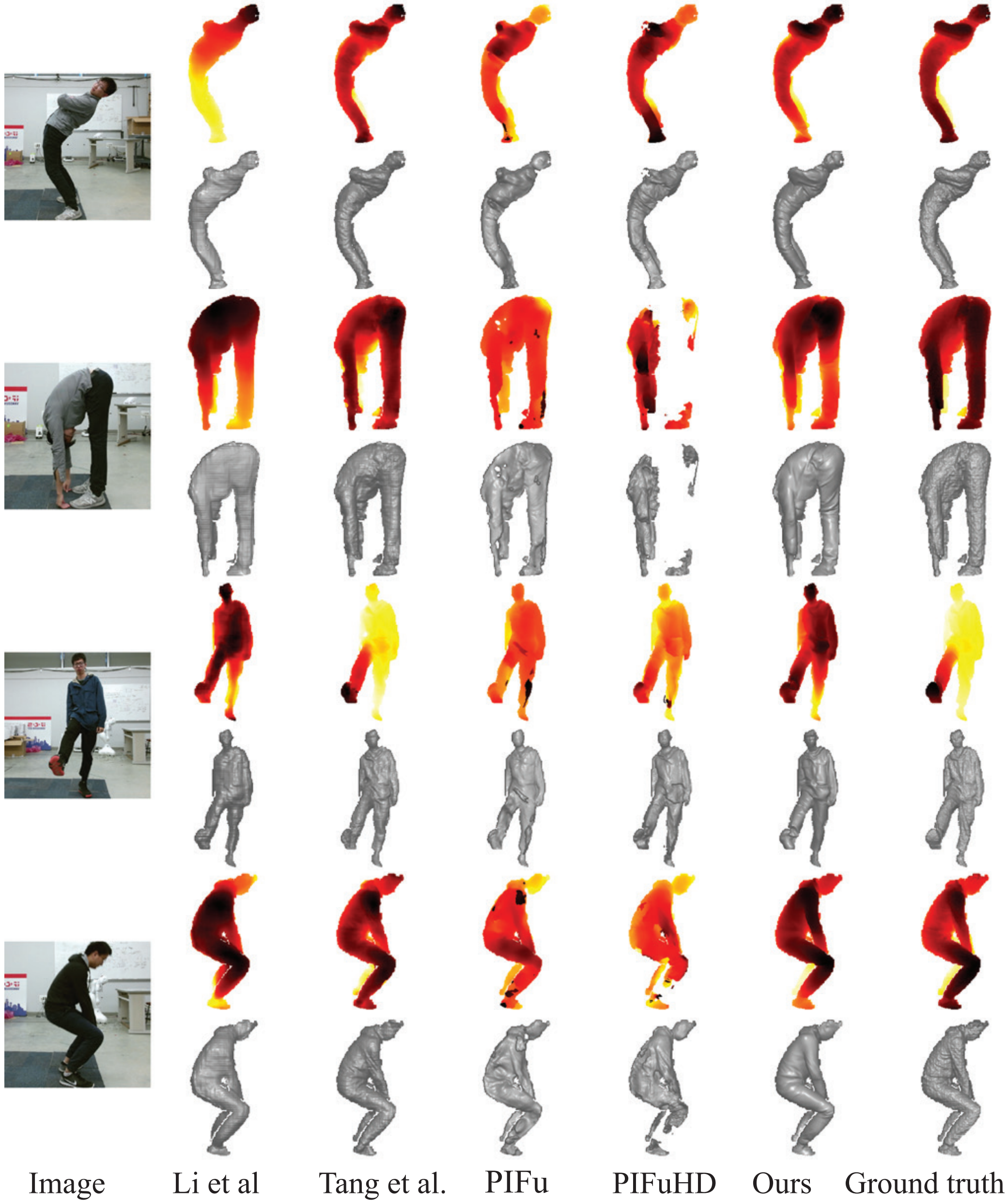

Figure 17: Qualitative comparison on Tang et al. training dataset