Self-Supervised Human Depth Estimation from Monocular Videos

Feitong Tan[1,*] Hao Zhu[2,*] Zhaopeng Cui[3] Siyu Zhu[4] Marc Pollefeys[3] Ping Tan[1]
[1] Simon Fraser University [2] Nanjing University
[3] ETH Zürich [4] Alibaba AI Labs

Abstract

Previous methods on estimating detailed human depth often require supervised training with 'ground truth' depth data. This paper presents a self-supervised method that can be trained on YouTube videos without known depth, which makes training data collection simple and improves the generalization of the learned network. The self-supervised learning is achieved by minimizing a photo-consistency loss, which is evaluated between a video frame and its neighboring frames warped according to the estimated depth and the 3D non-rigid motion of the human body. To solve this non-rigid motion, we first estimate a rough SMPL model at each video frame and compute the non-rigid body motion accordingly, which enables self-supervised learning on estimating the shape details. Experiments demonstrate that our method enjoys better generalization and performs much better on data in the wild.

1. Introduction

Understanding and reconstructing human motion from images and videos is an important problem in computer vision with many applications including surveillance, VR/AR, and tele-presence. Many works focus on estimating a 2D or 3D skeleton model [6, 29, 25, 34, 26, 39]. While a skeleton model could be useful for surveillance, other applications demand a 3D surface model of the undressed human body, which is often represented by the SMPL [24] or SCAPE [3] model. Many works have been proposed to estimate those parametric shape models from images [35, 17, 30, 12, 18]. However, the mid- and high-frequency shape details, *e.g.*, cloth wrinkles and folds, are not captured in the SMPL and SCAPE models, which limits their application in AR/VR and tele-presence applications. Only a handful of recent works [41, 54, 38, 50, 1, 4] can recover those details from a single image, but they all rely on ground-truth 3D data for supervision. This paper aims to develop a self-supervised method for detailed human depth estimation, such that the network can be trained on a much larger dataset, e.g. YouTube videos, for improved performance.

Self-supervised learning has been adopted [10, 9, 21] to train depth estimation from a single image for static scenes by enforcing photo-consistency between the left and right views in a stereo pair. Basically, the left view can be warped to the right view according to the estimated depth, the photo-consistency between the right view and the warped left view can be used to train the network. In principle, if the human motion between two video frames is known, we could adopt the same photo-consistency principle to train the network for human depth estimation. However, the challenge is that human body has non-rigid motion and requires much more complicated motion models such as those in DynamicFusion [28] and VolumeDeform [15], which are difficult to estimate as well.

To address this challenge, we represent the human depth by a SMPL model with an additive residual detail map. This representation is similar to the base and detail shape formulation in [41], but it bears two important advantages. Firstly, the SMPL model estimation has been well studied and is robust even on data in the wild, which makes the base shape estimation more reliable. As we will see in experiments, this helps to reduce large errors in the estimated depth. Secondly, the SMPL model parameters have clear semantic meanings and can be used to induce the non-rigid motion of the human body between two neighboring video frames. In this way, the non-rigid human body motion can be solved, and the photo-consistency for self-supervised learning in [10, 9, 21] can be employed to train the depth estimation network.

Measuring photo-consistency between neighboring video frames is still hard, even when the non-rigid motion to align the two human shapes is known. We design our method to be robust to occlusion, motion inaccuracy, and shading changes to achieve a robust method.

With the proposed self-supervised framework, we can train our network using almost endless online video clips. This vast training data significantly improves the generalization of the trained network on unseen data, making human depth estimation more robust.

*These authors contributed equally to this work.

2. Related Work

Skeleton Pose or Parametric Model Estimation. With the development of deep neural network, the estimation of 2D skeleton joints [6, 29] and 3D skeleton joints [25, 34, 26, 39] has achieved great success with robust performance. Many other works focus on estimating an undressed human body shape from a single image, as the skeleton joints are insufficient to convey shape information. The undressed body shape is often represented by the SCAPE [3], SMPL [24], or SMPL-X [31] model, which encodes the body shape by the pose and shape parameters. These models can be fitted according to estimated skeleton joints [5, 22] or be directly regressed as in [11, 8, 40, 35, 17, 30, 33, 12, 42, 18].

While the estimation of skeleton pose and these parametric body shapes are relatively well studied and robust, they are insufficient for certain applications such as telepresence. In comparison, we strive to recover detailed human shapes, which has broader applications.

Non-parametric shape estimation. As the parametric SMPL or SCAPE captures limited shape details, non-parametric representations have been adopted in human shape estimation. Varol *et al*. [43] and Venkat *et al*. [44] used a 3D volumetric model to represent human shapes for better flexibility. Güler *et al*. [2] recovered a dense 2D-to-3D surface correspondence field for the human body, and the SMPL model can be generated according to the correspondence. Zhu *et al*. [53] and Rematas *et al*. [36] directly predicted the depth map from a single image by training on synthetic data. Li *et al*. [23] exploited motion parallax cues from static scenes to guide the human depth prediction by watching 'frozen people'. Kolotouros *et al*. [20] directly regressed the vertices in the SMPL model while retaining the topology. Natsume*et al*. [27] used 2D silhouettes and 3D joints of a body pose to describe the immense human shape.

All the above methods still cannot capture shape details such as cloth wrinkles. Only a handful of recent works[41, 54, 38, 1, 50, 4] are able to recover those details. Among them, Tang *et al*. [41] proposed a base + detail shape representation, Zheng *et al*. [50] followed the volumetric shape representation, Zhu *et al*. [54] and Alldieck *et al*. [1] improved the undressed SMPL model with hierarchical morphable models and vertex displacements respectively. Saito *et al*. [38] defined a pixel-aligned implicit function to represent the human shape. Bhatnagar [4] designed a neural network to estimate the garment geometry separately, and then dress the SMPL model with the garments.

However, all of them require ground-truth 3D data for supervised training, which is hard to obtain and could lead to serious generalization problem. We also aim to recover human shape details. But we advocate for self-supervised learning to exploit the vast online videos for training. Our approach significantly improves the network performance on in-the-wild data.

Self-supervised Depth Estimation. To alleviate the demands on the expensive ground-truth 3D data, various self-supervised approaches have been proposed to train depth prediction networks. These methods typically train the network by minimizing a photo-consistency loss for some view synthesized according to the inferred depth. The methods in [10, 9, 21] utilize stereo images with known relative motion between the left and right views. Some methods [47, 52, 45, 48] use a monocular moving camera and enforce photo-consistency between neighboring video frames. This setting is more challenging, since these methods need to estimate the camera motion at the same time of estimating scene depth. Khot *et al*. [19] designed a self-supervised method for multi-view stereo with a sophisticated loss function dealing with occlusion and shading changes.

Our method also takes the self-supervised approach to train a depth estimation network. But our method is designed for a moving human, instead of the static scene in all the above methods. Thus, the motion model between our neighboring video frames are much more complicated than those works.

3. Methods

An overview of the proposed framework is shown in Figure 1. It composes of three main components: (1) TrackNet, a neural network to estimate a Skinned Multi-Person Linear (SMPL) model [24] from a single image, which determines the base depth and also the non-rigid motion between consecutive frames. (2) NRMM, a module to compute the non-rigid motion according to the SMPL model to align the 3D human shapes at neighboring frames. (3) ReconNet, a neural network trained in a self-supervised manner to estimate the residual detail shape, which will be added to the base depth to capture shape details.

In the training stage, a short video sequence of a person is fed to the TrackNet to compute a SMPL model for each frame. In the next, a target frame is selected, and the non-rigid motions from the other frames to the target frame are computed through the NRMM module according to their SMPL models. Finally, the ReconNet will be trained in an self-supervised manner given the non-rigid motion by enforcing the photo-consistency loss. In the prediction stage, the TrackNet will estimate the SMPL model to generate the base depth, with which the ReconNet will predict the details. The final result is simply the addition of the base shape and the details.

3.1. Pose Tracking

We estimate a SMPL model at each frame to capture the human pose and rough shape. SMPL is a parameteric undressed human body model with 72 pose parameters θ

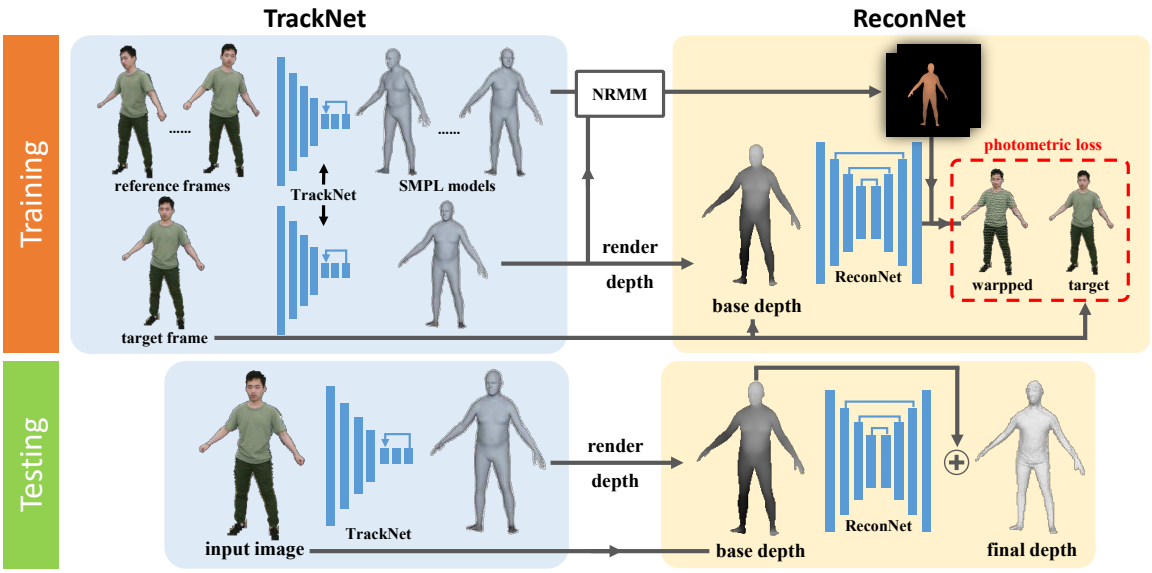

Figure 1: Overview of our system. At training time, our system includes the TrackNet to compute a SMPL model at each input video frame, the NRMM model to compute the non-rigid motion to align two neighboring human shapes, and the ReconNet to estimate shape details. At testing time, our system first compute the SMPL model from the TrackNet and then estimates the details form the ReconNet and combine them as the final result.

and 10 shape parameters β controlling a triangle mesh of 6,890 vertices. The parameters θ define the 3D rotation of each skeleton joint, and the parameters in β describe the height, weight, and other body shape metrics. Compared with a skeleton model, a SMPL model encodes strong human shape prior and provides more information including limbs orientation and the rough shape.

We design a TrackNet module to predict SMPL parameters for each frame of the input video. The undressed body shape defined by SMPL parameters are used as the base shape for further process. Compared to estimating an non-parametric base shape in [41], our approach is more robust since the base shape is constrained into a much smaller parameter space with strong human shape priors.

Our TrackNet adopts the same network architecture as HMR [17], which consists of a ResNet-50 [14] as a feature extractor and an iterative error feedback regressor [7]. The original HMR model is trained on images with annotated 2D joints. In order to produce more accurate results, we captured a small set of videos with 'ground-truth' SMPL coefficients generated by DoubleFusion [49] and further finetuned the TrackNet after pretraining following [17].

The TrackNet outputs a 85-D vector, with 82 parameters as SMPL coefficients and 3 parameters for the weak-perspective camera model. The loss function to finetune TrackNet on our DoubleFusion data is formulated as:

$$L_{tn} = L_{para} + \theta_p L_{J\_pos} + \theta_r L_{J\_rot} \quad (1)$$

where L_{para} is the L_1 loss of SMPL parameters, L_{J_pos} and L_{J_rot} are the L_1 loss of 3D position and rotation of SMPL joints respectively. θ_p and θ_r are the loss weights, and both of them are set to 1 in our experiments. Please note that our method is not limited to the specific HMR [17] model and the TrackNet can be upgraded with other SMPL model estimation networks.

3.1.1 Camera Model Adjustment

State-of-the-art methods [35, 17, 30] for SMPL model estimation employ a weak-perspective camera model to facilitate the computation. However, as it is known, most of videos are captured with a perspective camera. So in order to better utilize the photometric loss, we adjust the camera model from the weak-perspective model to the perspective model. As the focal length of videos in the wild are normally unknown, we use a medium focal length to render SMPL model, and we empirically found that this perspective camera model can align SMPL model to the image well. We take a simple conversion from the weak perspective model to a general perspective model, assuming known camera focal length as the following,

$$V_{tras.} = [t_x, t_y, \frac{f_c}{\frac{1}{2} * img\_size * s}], \quad (2)$$

where vector $t = [t_x, t_y]$ is translation and s is the scale in a weak-perspective camera model. f_c is the focal length of the camera.

3.2. Non-rigid Motion Model

In order to exploit the photometric consistency between different frames for the self-supervision, we need to com-

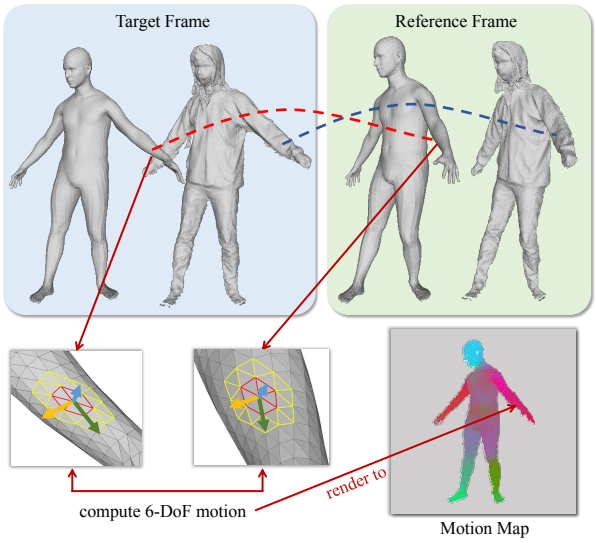

Figure 2: Motion map generation. Given the SMPL models at two neighboring frames, we compute a per-vertex transformation using some nearby vertices (colored in the zoom-in figure). This per-vertex transformation is then rendered to image plane as a motion map. Our method assumes that the non-rigid transformation between the SMPL shapes (red dashed line) is the same as that between the detailed shapes (blue dashed line).

pute the motion fields of the human body between consecutive frames. However, this is non-trivial as the human motion is non-rigid. So we propose a novel approach to compute the non-rigid motion of the human body between consecutive frames based on the estimated SMPL model per video frame. Specifically speaking, for each video, we select the target frame intervally with a gap of 3, then group the consecutive $r = [\pm 4, \pm 5, \pm 7, \pm 8, \pm 9]$ frames as reference frames. After grouping each image tuple, we compute a non-rigid motion field $T_{t \to r}$ to represent the 3D dense spatial transformation of the human body from the target frame to the reference frame. The non-rigid motion is defined as a motion field in 3D space, with a 6-DoF transformation for each vertex of the SMPL model.

To compute the non-rigid motion, we first compute a per-vertex transformation. As shown in Figure 2, since the SMPL models at two neighboring frames share the same topology, we have explicit per-vertex correspondence. Thus, the per-vertex transformation can be computed by registering n two-ring neighboring vertices in the target and reference model. More specifically, the rotation matrix $R \in SO(3)$ and translation vector $t \in \mathbb{R}^3$ can be computed by

$$R = \arg\min \sum_{i=1}^{n} ||R(v_t^i - v_t^c) - (v_r^i - v_r^c)||^2, \qquad t = v_r - R * v_t, \tag{3}$$

where v_r^i and v_t^i represent the ith corresponding vertices in these two-ring neighboring vertices in reference model r and target model t respectively, v^c denotes the center vertex of the two-ring neighbor group.

Then we render the per-vertex transformation to the image plane as a motion map by ray tracing. For each pixel in 2D image, the mean R and t of the 3 vertices in the corresponding triangle is computed as the final transformation. In addition to motion map, we also render the depth of SMPL model in the target frame as base depth D_t.

3.2.1 Occlusion Handling and Baseline Filtering

Exploiting the motion of human shape to measure photo-consistency between neighboring frames faces two technical challenges, namely occlusion and insufficient baseline length. When the motion is too large, part of the body might become invisible in the reference or target view, just like the occlusion problem in wide baseline stereo matching. When the motion is too small, adjusting shape details by maximizing photo-consistency could lead to noisy results. This is similar to the problem when the stereo baseline is too short.

To make the self-supervised training of ReconNet robust, we design careful filtering to deal with this problem. We define a 'baseline length' for each pixel in the motion map, which is the magnitude of its translation t. We then compute the mean baseline over all pixels for each tuple, and remove tuples with mean baseline less than 0.5m. To deal with the occlusion due to large motion, for each $T_{t \to r}$, we compute a validation mask M_r, where we mark a pixel as valid if it is visible in both target and reference view and has a baseline length larger than 5 cm.

3.3. ReconNet Architecture

The ReconNet computes the detail depth layer which will be added to the base depth to compose the final result. We adopt a variant of U-Net [37] using the residual blocks [13] in encoder and decoder with skip connections. The encoder has 6 down-sampling layers, while the decoder has 5 up-sampling layers. We apply a sigmoid function at the end of the last layer to regularize the output from -10 to 10 cm. The input is a concatenation of the 512×512 RGB image and the zero-median rendered depth map from the estimated SMPL model. The output of the network is a 256×256 depth offset map. Finally, we compose the detail depth layer with the base depth layer to obtain our final output.

3.3.1 Self-supervised Learning for Detail

Warping-based view synthesis loss has been proved effective in monocular, stereo, and multi-view stereo depth prediction tasks [10, 52, 19]. We extend it to monocular non-rigid human depth estimation.

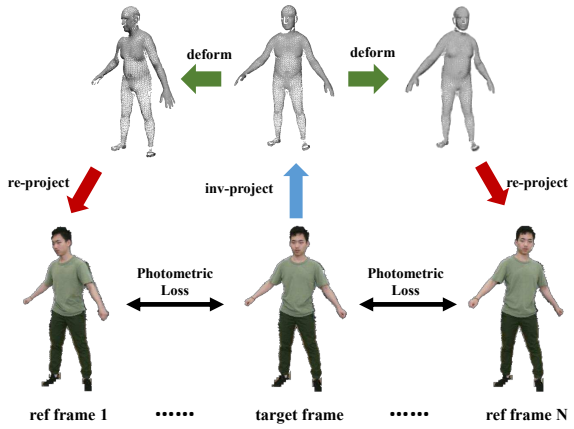

Figure 3: We first inverse-project the composed depth to point clouds, then deform them with non-rigid motion map and finally reproject deformed point clouds back to reference image for bilinear sampling.

Given a clip of temporal continuous frames $\{I_1, ..., I_N\}$ with fixed camera intrinsic parameters K, we select the center frame $\{I_t\}$ as the target frame and the others as reference frames $\{I_r\}(1 \leq r \leq N, r \neq t)$. For each reference frame, a non-rigid motion $\{T_{t\to r}\}_{r=1}^{N-1}$ and the validation map $\{M_{t\to r}\}_{r=1}^{N-1}$ can be pre-computed with the estimated SMPL models. Then our network use these non-rigid motion fields to warp the target frame toward the reference frames with a differentiable bilinear interpolation. This process is illustrated in Figure 3.

Let p_t denote the homogeneous coordinates of a pixel in the target view, and K denote the camera intrinsic matrix. We can obtain p_t's projected coordinates onto the reference view p_r by

$$p_r \sim KT_{t\to r}(p_t)D(p_t)K^{-1}p_t. \tag{4}$$

The inverse-warped images $\{\hat{I}_t^r\}$ from each reference frame can be synthesized according to Equation 4. As a result, we can then formulate a photo-consistency objective function as the following:

$$L_{photo}^r = \left(\alpha\frac{1 - SSIM_{cs}(I_t, \hat{I}_t^r)}{2} + (1-\alpha) \parallel I_t - \hat{I}_t^r \} \parallel\right) \otimes M_r, \tag{5}$$

where SSIM_{cs} denotes the structural similarity index [46] with only the component of contrast and structure ($SSIM_{cs} = \frac{\sigma_{xy}+c}{\sigma_x^2+\sigma_y^2+c}$). We set α to 0.9, because the estimated SMPL is imperfect, which causes misalignment during image warping with the computed non-rigid motion. So the structural similarity measured by SSIM_{cs} is more robust than the intensity difference. Moreover, we also use the validation map to mask out invalid pixels.

Our final photo-consistency function is summed over all reference images for better robustness,

$$L_{photo} = \sum_{r=1, r\neq t}^{N} L_{photo}^r. \tag{6}$$

Following previous self-supervised depth estimation, a smoothness term is also introduced. Since our target is to estimate a human shape, we require the gradient of the composed shape to be close to the base shape, which leads to the following smooth term:

$$L_{smooth} = \sum_{p_t} |\nabla D_{detail}(p_t) - \nabla D_{base}(p_t)|. \tag{7}$$

We also introduce a regularization term to require the final depth to be similar to the base depth:

$$L_{regularizer} = \sum_{p_t} |D_{detail}(p_t) - D_{base}(p_t)|. \tag{8}$$

Finally, our final learning objective function is:

$$L = L_{photo} + \gamma_s L_{smooth} + \gamma_r L_{regularization}, \tag{9}$$

where γ_s and γ_r are the hyperparameters to control the significance of the smooth term L_{smooth} and regularization term $L_{regularizer}$. In all our experiments, γ_s is set as 10^{-5} and γ_r is 10^{-6}.

4. Experiment

4.1. Data

We find the poses tracked by original HMR model trained in the wild images are not accurate enough for the self-supervised learning of ReconNet, so we finetune the TrackNet using our collected data with ground truth SMPL parameters generated by DoubleFusion[49].

Thus, we captured 36 video sequences of different people performing simple action, which contains roughly 48,000 frames in total. Half of the frames have labeled SMPL coefficients recovered from depth streams by DoubleFusion [49], and are used in training of the TrackNet. To augment the background in the video, we randomly use images from the Places Dataset[51] as the background for each sequence. All the image frames are also used to bootstrap our ReconNet in a self-supervised manner, and the SMPL parameters are predicted by our finetuned TrackNet. To train the ReconNet, we grouped the frames from the captured videos. For each clip, we skip every 3 frames to set a target frame, and other consecutive $[\pm4, \pm5, \pm7, \pm8, \pm9]$ frames are set as the reference frames. We sampled in total 12,533 image tuples for training.

To ensure our model can be generalized to in-the-wild data, we select 18 videos from YouTube, and generate about

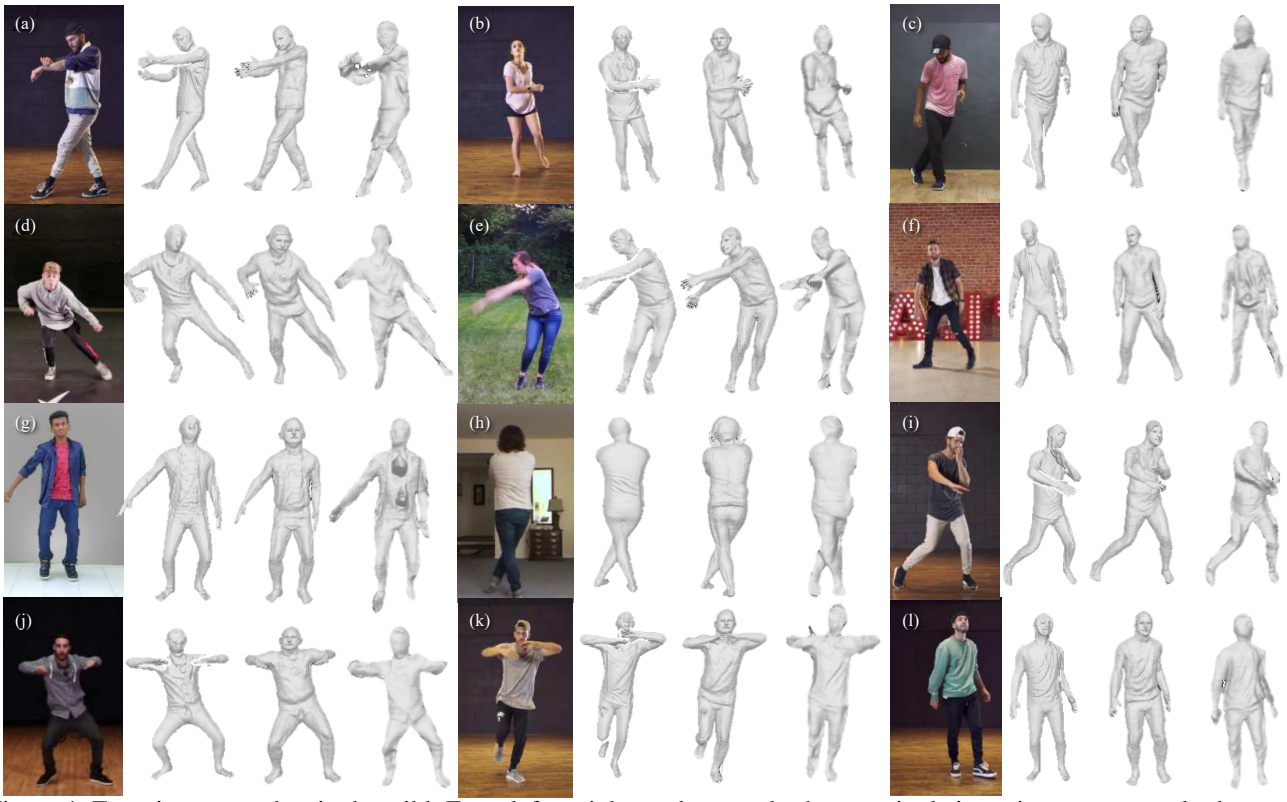

Figure 4: Experiments on data in the wild. From left to right, each example shows a single input image, our result, the result from HMD[54], and the result from Tang *et al*. [41].

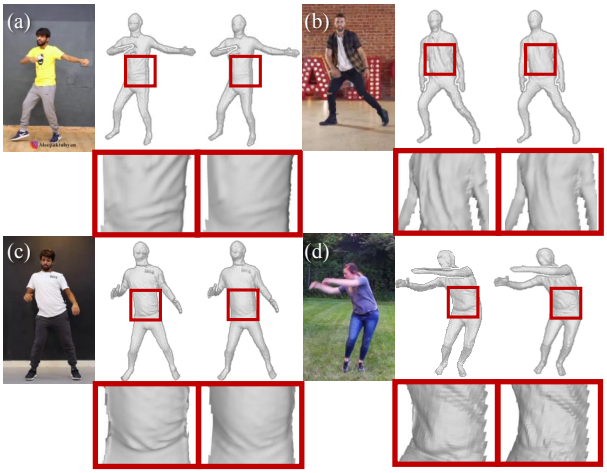

Figure 5: From the left to right, each example shows a single input image, the result after finetuned on YouTube data, the results before finetuning. This finetuning improves both mid- and high- frequency shape details.

3,000 images tuples to finetune our ReconNet. We select YouTube videos with a simple criteria that the video contains a *single* and *complete* person with less occlusion. Note that we can replace TrackNet with other better SMPL model or even SMPL-X [32] model estimation networks for more accurate base shape and motion map generation.

4.2. Training Details

We finetune the TrackNet from the original HMR model with 'Adam' optimizer using our captured data with SMPL parameters from DoubleFusion. The learning rate is set to 1×10^{-6}. We use batch size 20 and train in 20 epochs.

We first bootstrap the ReconNet with our captured videos in self-supervised manner for 2 epochs. The learning rate is set as 4×10^{-4} and the batchsize is 2. We then finetune the ReconNet with YouTube images with the learning rate of 1×10^{-4} for one epoch.

4.3. Experiment on Data in the Wild

We test our method on unseen YouTube videos. We randomly select half of the frames in the video to finetune our ReconNet, and use the other frames for evaluation. Figure 4 shows the comparison with HMD[54] and Tang *et al*. [41] on in-the-wild data. We find our result can capture more fine details compared with other two methods mainly for two reasons: first, our model is finetuned with in-the-wild videos due to the self-supervised learning; second, our photometric loss is more effective in capturing small wrinkles. In comparison, the other two methods are trained only with

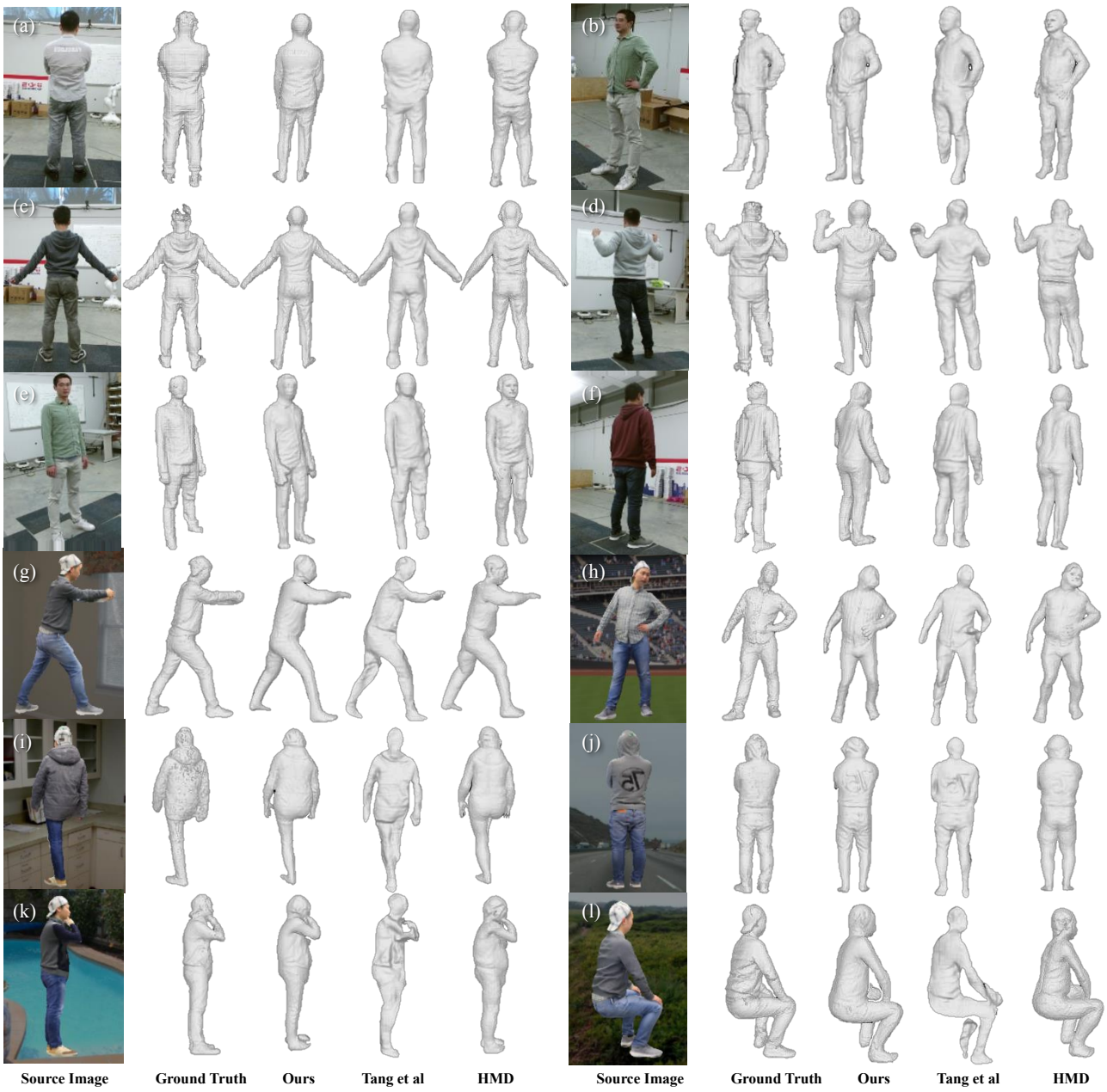

Figure 6: Comparison to Tang *et al*. [41] and HMD[54] on their testing datasets. The source images in top half part are from the testing data of Tang*et al*. [41], and the source images in bottom half part are from the REAL dataset in HMD[54].

limited 'ground truth' depth from consumer depth cameras. The noisy 'ground truth' depth makes it difficult for the network to recover small details. Further more, although both our method and Tang *et al*. [41] separate the human shape to a base shape and a detail shape, we use a SMPL model for the base shape, which is more robust for the in-the-wild data. It is clear that our method generates more details than [41] and [54] in the examples (a), (b), (d), (h), (j). Tang *et al*.'s[41] results have large errors on examples (a), (b), (e), (g), (k). We notice that the complex clothing, faces and hairstyles are not estimated accurately, which is mainly because they are difficult for photo-consistency based reconstruction.

We also show the predicted results with and without the fine-tuning the ReconNet on Youtube data in Figure 5. Our model cannot capture the fine details without finetuning. Finetuning on YouTube data improves the mid- and high-frequency shape details. Please refer to the supplementary material for more results and discussion.

4.4. Quantitative Evaluation

To quantitatively evaluate the accuracy and compare to the previous methods, we evaluate our method on the testing data provided by Tang *et al*. [41] and HMD [54].

Comparison on the Dataset from [41]. Tang *et al*. [41] published a small dataset with ground truth 3D human shapes generated by InfiniTAM [16] for quantitative evaluation. To evaluate our results on their dataset, we first use ICP to register our results to the ground truth, and measure the error at each pixel by the point-to-point nearest neighbor distance. Following [41], we compute the accuracy at different error thresholds, i.e. the percentage of pixels with an error smaller than some threshold. The accuracy of different methods evaluated on this dataset is shown in Table 1. We also compute the Mean Absolute Error (MAE) to evaluate the overall shape accuracy. Our method has smaller MAE both than Tang *et al*. [41] and HMD, and higher accuracy when the error threshold is larger than 4cm. It suggests our method has less large errors than [41] on their published dataset. However, Tang *et al*. can recover better shape details on this dataset. We believe this is because the test data is highly consistent with their training data. The advantages of our self-supervised method is demonstrated on data in the wild in Figure 4. The performance of HMD [54] varies on this dataset, e.g. poor results on example (b), (d) and (f), suggesting its generalization is poorer than our method.

The first three rows in Figure 6 shows some visual results from these methods. Tang *et al*.'s method sometimes generate distorted limbs as shown in example (b) and (d).

Comparing on the Dataset from [54]. HMD provides another small dataset with ground-truth 3D mesh recovered by multi-view reconstruction methods for quantitative evaluation. Here, we focus on comparing our ReconNet with the 'Shading-Net' in HMD, which is also a refinement model on an estimated SMPL model. For fair comparison, we use the same SMPL model for both methods. The accuracy is reported in Table 2. We can see our self-supervised method achieves better accuracy than HMD even on their test data.

The last three rows in Figure 6 are the qualitative comparison between these three methods on this dataset. We can find Tang *et al*.'s performance in this dataset is not good because of its poor generalization and the unusual poses in this dataset. Also, we can find our method can recover more details in example (g), (j), (k).

4.5. Ablation Study

We perform various ablation studies on the dataset of Tang *et al*. [41] in this subsection. We denote the result without using validation mask M_r and SSIM_{cs} as baseline, and compare it with various other settings. All the ablation study are trained with the same hyperparameter.

Validation Mask The baseline method generate poorer accuracy than the proposed method with validation mask as shown in Table 3. This proves the effectiveness of our occlusion handling and baseline filtering.

Table 1: Comparison on the dataset published in [41].

Methods	Accuracy			MAE
	1.0cm	2.0cm	4.0cm	
Tang *et al*. [41]	**33.30**	**59.68**	79.63	2.735
HMD [54]	27.66	54.10	76.31	3.077
Ours	31.47	59.08	**82.13**	**2.609**

Table 2: Comparison on the dataset published in [54]. 'Ours (HMD)' means our method fed with the same undressed SMPL model as HMD.

Methods	Accuracy (%)			MAE (cm)
	1.0cm	2.0cm	4.0cm	
Tang *et al*. [41]	19.07	41.71	73.19	3.125
HMD [54]	21.83	46.10	75.46	3.043
Ours(HMD)	**22.62**	**47.65**	**76.65**	**2.944**

Table 3: Ablation study on Tang *et al*.'s test set. Please see text for more details.

Methods	Accuracy			MAE
	1.0cm	2.0cm	4.0cm	
Ours(Baseline)	28.14	55.57	79.46	2.828
Ours(M)	28.52	56.35	80.67	2.714
Ours(M+SSIM_{cs})	**31.47**	**59.08**	**82.13**	**2.609**

SSIM_{cs} loss. We replace the original SSIM loss with SSIM_{cs} loss in training the ReconNet. SSIM_{cs} is measured on the contrast domain and is more robust to misalignment (due to imperfect non-rigid motion estimation) and shading changes. As shown in Table 3, after replacing SSIM with SSIM_{cs}, our model achieves better performance.

5. Conclusions

We present a self-supervised method to estimate human shapes with fine geometry details such as cloth wrinkles. This self-supervised approach enables the network to be trained on in-the-wild data, such as YouTube videos, which significantly improves the generalization of the network. This result is achieved by introducing the SMPL model for the base shape, and using it to compute the non-rigid human motion at neighboring frames to facilitate photo-consistency evaluation. Extensive evaluation and comparison with state-of-the-art methods proves the effectiveness of our method.

References

[1] Thiemo Alldieck, Gerard Pons-Moll, Christian Theobalt, and Marcus Magnor. Tex2shape: Detailed full human body geometry from a single image. In *Proc. of International Conference on Computer Vision*, 2019. 1, 2

[2] Rıza Alp Güler, Natalia Neverova, and Iasonas Kokkinos. Densepose: Dense human pose estimation in the wild. In *Proc. of Computer Vision and Pattern Recognition*, 2018. 2

[3] Dragomir Anguelov, Praveen Srinivasan, Daphne Koller, Sebastian Thrun, Jim Rodgers, and James Davis. Scape: shape completion and animation of people. In *ACM Trans. on Graphics*, volume 24, pages 408–416, 2005. 1, 2

[4] Bharat Lal Bhatnagar, Garvita Tiwari, Christian Theobalt, and Gerard Pons-Moll. Multi-garment net: Learning to dress 3d people from images. In *Proc. of International Conference on Computer Vision*, 2019. 1, 2

[5] Federica Bogo, Angjoo Kanazawa, Christoph Lassner, Peter Gehler, Javier Romero, and Michael J Black. Keep it smpl: Automatic estimation of 3d human pose and shape from a single image. In *Proc. of European Conference on Computer Vision*, 2016. 2

[6] Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Realtime multi-person 2d pose estimation using part affinity fields. In *Proc. of Computer Vision and Pattern Recognition*, 2017. 1, 2

[7] Joao Carreira, Pulkit Agrawal, Katerina Fragkiadaki, and Jitendra Malik. Human pose estimation with iterative error feedback. In *Proc. of Computer Vision and Pattern Recognition*, 2016. 3

[8] Endri Dibra, Himanshu Jain, Cengiz Öztireli, Remo Ziegler, and Markus Gross. Hs-nets: Estimating human body shape from silhouettes with convolutional neural networks. In *Proc. of International Conference on 3D Vision*, 2016. 2

[9] Ravi Garg, Vijay Kumar BG, Gustavo Carneiro, and Ian Reid. Unsupervised cnn for single view depth estimation: Geometry to the rescue. In *Proc. of European Conference on Computer Vision*, 2016. 1, 2

[10] Clément Godard, Oisin Mac Aodha, and Gabriel J Brostow. Unsupervised monocular depth estimation with left-right consistency. In *Proc. of Computer Vision and Pattern Recognition*, 2017. 1, 2, 4

[11] Peng Guan, Alexander Weiss, Alexandru O Balan, and Michael J Black. Estimating human shape and pose from a single image. In *Proc. of International Conference on Computer Vision*, 2009. 2

[12] Riza Alp Guler and Iasonas Kokkinos. Holopose: Holistic 3d human reconstruction in-the-wild. In *Proc. of Computer Vision and Pattern Recognition*, 2019. 1, 2

[13] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proc. of Computer Vision and Pattern Recognition*, 2016. 4

[14] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Identity mappings in deep residual networks. In *Proc. of European Conference on Computer Vision*, 2016. 3

[15] Matthias Innmann, Michael Zollhöfer, Matthias Nießner, Christian Theobalt, and Marc Stamminger. Volumedeform: Real-time volumetric non-rigid reconstruction. In *Proc. of European Conference on Computer Vision*, 2016. 1

[16] O. Kahler, V. A. Prisacariu, C. Y. Ren, X. Sun, P. H. S Torr, and D. W. Murray. Very High Frame Rate Volumetric Integration of Depth Images on Mobile Device. *IEEE Trans. on Visualization and Computer Graphics*, 22(11), 2015. 8

[17] Angjoo Kanazawa, Michael J Black, David W Jacobs, and Jitendra Malik. End-to-end recovery of human shape and pose. In *Proc. of Computer Vision and Pattern Recognition*, 2018. 1, 2, 3

[18] Angjoo Kanazawa, Jason Y Zhang, Panna Felsen, and Jitendra Malik. Learning 3d human dynamics from video. In *Proc. of Computer Vision and Pattern Recognition*, 2019. 1, 2

[19] Tejas Khot, Shubham Agrawal, Shubham Tulsiani, Christoph Mertz, Simon Lucey, and Martial Hebert. Learning unsupervised multi-view stereopsis via robust photometric consistency. *arXiv preprint arXiv:1905.02706*, 2019. 2, 4

[20] Nikos Kolotouros, Georgios Pavlakos, and Kostas Daniilidis. Convolutional mesh regression for single-image human shape reconstruction. In *Proc. of Computer Vision and Pattern Recognition*, 2019. 2

[21] Yevhen Kuznietsov, Jorg Stuckler, and Bastian Leibe. Semi-supervised deep learning for monocular depth map prediction. In *Proc. of Computer Vision and Pattern Recognition*, 2017. 1, 2

[22] Christoph Lassner, Javier Romero, Martin Kiefel, Federica Bogo, Michael J. Black, and Peter V. Gehler. Unite the people: Closing the loop between 3d and 2d human representations. In *Proc. of Computer Vision and Pattern Recognition*, 2017. 2

[23] Zhengqi Li, Tali Dekel, Forrester Cole, Richard Tucker, Noah Snavely, Ce Liu, and William T Freeman. Learning the depths of moving people by watching frozen people. In *Proc. of Computer Vision and Pattern Recognition*, 2019. 2

[24] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J. Black. SMPL: A skinned multi-person linear model. *ACM Trans. on Graphics*, 34(6):248:1–248:16, 2015. 1, 2

[25] Julieta Martinez, Rayat Hossain, Javier Romero, and James J. Little. A simple yet effective baseline for 3d human pose estimation. In *Proc. of International Conference on Computer Vision*, 2017. 1, 2

[26] Dushyant Mehta, Srinath Sridhar, Oleksandr Sotnychenko, Helge Rhodin, Mohammad Shafiei, Hans-Peter Seidel, Weipeng Xu, Dan Casas, and Christian Theobalt. Vnect: Real-time 3d human pose estimation with a single rgb camera. *ACM Trans. on Graphics*, 36(4):44, 2017. 1, 2

[27] Ryota Natsume, Shunsuke Saito, Zeng Huang, Weikai Chen, Chongyang Ma, Hao Li, and Shigeo Morishima. Siclope: Silhouette-based clothed people. In *Proc. of Computer Vision and Pattern Recognition*, 2019. 2

[28] Richard A Newcombe, Dieter Fox, and Steven M Seitz. Dynamicfusion: Reconstruction and tracking of non-rigid scenes in real-time. In *Proc. of Computer Vision and Pattern Recognition*, 2015. 1

[29] Alejandro Newell, Kaiyu Yang, and Jia Deng. Stacked hourglass networks for human pose estimation. In *Proc. of European Conference on Computer Vision*, 2016. 1, 2

[30] Mohamed Omran, Christoph Lassner, Gerard Pons-Moll, Peter V Gehler, and Bernt Schiele. Neural body fitting: Unifying deep learning and model-based human pose and shape estimation. In *Proc. of International Conference on 3D Vision*, 2018. 1, 2, 3

[31] 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 *Proc. of Computer Vision and Pattern Recognition*, 2019. 2

[32] Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed A. A. Osman, Dimitrios Tzionas, and Michael J. Black. Expressive body capture: 3d hands, face, and body from a single image. In *Proc. of Computer Vision and Pattern Recognition*, 2019. 6

[33] Georgios Pavlakos, Nikos Kolotouros, and Kostas Daniilidis. Texturepose: Supervising human mesh estimation with texture consistency. In *Proc. of International Conference on Computer Vision*, 2019. 2

[34] Georgios Pavlakos, Xiaowei Zhou, Konstantinos G Derpanis, and Kostas Daniilidis. Coarse-to-fine volumetric prediction for single-image 3d human pose. In *Proc. of Computer Vision and Pattern Recognition*, 2017. 1, 2

[35] Georgios Pavlakos, Luyang Zhu, Xiaowei Zhou, and Kostas Daniilidis. Learning to estimate 3d human pose and shape from a single color image. In *Proc. of Computer Vision and Pattern Recognition*, 2018. 1, 2, 3

[36] Konstantinos Rematas, Ira Kemelmacher-Shlizerman, Brian Curless, and Steve Seitz. Soccer on your tabletop. In *Proc. of Computer Vision and Pattern Recognition*, 2018. 2

[37] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Proc. of International Conference on Medical image computing and computer-assisted intervention*, 2015. 4

[38] Shunsuke Saito, Zeng Huang, Ryota Natsume, Shigeo Morishima, Angjoo Kanazawa, and Hao Li. Pifu: Pixel-aligned implicit function for high-resolution clothed human digitization. In *Proc. of International Conference on Computer Vision*, 2019. 1, 2

[39] Xiao Sun, Bin Xiao, Fangyin Wei, Shuang Liang, and Yichen Wei. Integral human pose regression. In *Proc. of European Conference on Computer Vision*, 2018. 1, 2

[40] J Tan, Ignas Budvytis, and Roberto Cipolla. Indirect deep structured learning for 3d human body shape and pose prediction. In *Proc. of British Machine Vision Conference*, 2017. 2

[41] Sicong Tang, Feitong Tan, Kelvin Cheng, Zhaoyang Li, Siyu Zhu, and Ping Tan. A neural network for detailed human depth estimation from a single image. In *Proc. of International Conference on Computer Vision*, 2019. 1, 2, 3, 6, 7, 8

[42] Hsiao-Yu Tung, Hsiao-Wei Tung, Ersin Yumer, and Katerina Fragkiadaki. Self-supervised learning of motion capture. In *Proc. of Advances in Neural Information Processing Systems*, 2017. 2

[43] Gül Varol, Duygu Ceylan, Bryan Russell, Jimei Yang, Ersin Yumer, Ivan Laptev, and Cordelia Schmid. Bodynet: Volumetric inference of 3d human body shapes. In *Proc. of European Conference on Computer Vision*, 2018. 2

[44] Abbhinav Venkat, Sai Sagar Jinka, and Avinash Sharma. Deep textured 3d reconstruction of human bodies. In *Proc. of British Machine Vision Conference*, 2018. 2

[45] Chaoyang Wang, José Miguel Buenaposada, Rui Zhu, and Simon Lucey. Learning depth from monocular videos using direct methods. In *Proc. of Computer Vision and Pattern Recognition*, 2018. 2

[46] Zhou Wang, Alan C Bovik, Hamid R Sheikh, Eero P Simoncelli, et al. Image quality assessment: from error visibility to structural similarity. *IEEE Trans. on Image Processing*, 13(4):600–612, 2004. 5

[47] Junyuan Xie, Ross Girshick, and Ali Farhadi. Deep3d: Fully automatic 2d-to-3d video conversion with deep convolutional neural networks. In *Proc. of European Conference on Computer Vision*, 2016. 2

[48] Zhichao Yin and Jianping Shi. Geonet: Unsupervised learning of dense depth, optical flow and camera pose. In *Proc. of Computer Vision and Pattern Recognition*, 2018. 2

[49] Tao Yu, Zerong Zheng, Kaiwen Guo, Jianhui Zhao, Qionghai Dai, Hao Li, Gerard Pons-Moll, and Yebin Liu. Doublefusion: Real-time capture of human performances with inner body shapes from a single depth sensor. In *Proc. of Computer Vision and Pattern Recognition*, 2018. 3, 5

[50] Zerong Zheng, Tao Yu, Yixuan Wei, Qionghai Dai, and Yebin Liu. Deephuman: 3d human reconstruction from a single image. In *Proc. of International Conference on Computer Vision*, 2019. 1, 2

[51] Bolei Zhou, Agata Lapedriza, Aditya Khosla, Aude Oliva, and Antonio Torralba. Places: A 10 million image database for scene recognition. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 40(6):1452–1464, 2017. 5

[52] Tinghui Zhou, Matthew Brown, Noah Snavely, and David G Lowe. Unsupervised learning of depth and ego-motion from video. In *Proc. of Computer Vision and Pattern Recognition*, 2017. 2, 4

[53] Hao Zhu, Hao Su, Peng Wang, Xun Cao, and Ruigang Yang. View extrapolation of human body from a single image. In *Proc. of Computer Vision and Pattern Recognition*, 2018. 2

[54] Hao Zhu, Xinxin Zuo, Sen Wang, Xun Cao, and Ruigang Yang. Detailed human shape estimation from a single image by hierarchical mesh deformation. In *Proc. of Computer Vision and Pattern Recognition*, 2019. 1, 2, 6, 7, 8