Parametric Human Shape Reconstruction via Bidirectional Silhouette Guidance

Shuang Sun[1] Chen Li [1] Zhenhua Guo[2] Yuwing Tai[1]
[1]Tencent [2]Graduate School of ShenZhen, Tsinghua University
{*shuangsun*, *chaselli*, *yuwingtai*}@tencent.com zhenhua.guo@sz.tsinghua.edu.cn

Figure 1. Results of our human shape reconstruction on challenging examples. The first row shows the input images, the second row shows the reconstruction results, and the third row show the reconstructed body at novel views.

Abstract

We present a method to reconstruct the body geometry of a person by aligning the skinned multi-person linear (SMPL) model to an unconstrained human image. In contrast to previous methods that regress the model parameters from a shared image feature, we decouple the regression of pose and shape parameters in two sub-networks so that we can use different backbone architectures to extract better and more specific features for each regression task while allowing the two sub-networks to work together by our final training loss. We have further proposed a novel bidirectional silhouette constraint to restrict the estimated body geometry. The silhouette constraint is weighted adaptively according to the accuracy of pose estimation in order to handle truncations, occlusions and complex human poses. Experimental results on ***Human3.6M*** *and* ***UP*** *datasets show that our method outperforms state-of-the-art methods and fits the body segmentation better, especially under extreme human pose conditions.*

1. Introduction

Human pose estimation and full body shape reconstruction have been studied for decades. Classical approaches, such as optical flow in monocular image sequences [1, 34, 9], multi-view stereo [44], and depth-range sensors [38, 46, 30, 8, 32], have made great progresses but they usually require more than one input images in order to extract 3D information from 2D images, or require point cloud registration in order to fuse the dense depth point clouds captured/estimated from multiple views. Recent advances, such as [41, 37, 24, 29, 14], have demonstrated the possibility of using deep neural network to directly regress the model parameters of human body shape templates, i.e. SCAPE [3] and SMPL [18], which converts the highly ill-posed optimization problem into a low dimensional parameter regression problem. With the new deep learning framework, single image human body shape reconstruction can be achieved.

Previous deep learning based approaches, however, were mainly focused on the human pose estimation, but less at-

tentions were paid on the full body shape reconstruction. Thus, the resulting body shapes are almost uniform and have less diversities where the estimated body shape usually does not reflect the geometry of the target body.

To address this problem, we present a novel deep neural network architecture to accurately reconstruct human body together with the corresponding 3D joint positions. Our method follows the previous frameworks [14, 29, 24] by aligning the SMPL model [18] to an unconstrained human image. Noticing the deep feature for 3D pose estimation and 3D shape reconstruction might be at different level [29], we decouple the regression of SMPL parameters into two sub-networks and use more appropriate backbone architecture for each of them to better estimate the corresponding parameters respectively. Compared with previous works which first predict semantic human features (*e.g.*, body segmentation [29], 3D joint positions [41, 37, 24], volumetric geometry representation [42]) followed by regressing the SMPL parameters, our method is in a simple end-to-end framework which is easy to train and the converged results are more stable.

In addition, we restrict the estimated body geometry by forcing it to match the projected body shape with a predefined human body segmentation. The geometry-to-silhouette constraint ensures the predicted human body lies within the body segmentation while the opposite silhouette-to-geometry constraint encourages the body segmentation to be covered by its geometry. By considering this bidirectional silhouette constraint, the estimated 3D geometry can better align with the body segmentation which allows the body shape and pose to be more accurately estimated. In order to handle partial truncations or complex human body poses, we segment the body geometry according to its nearest 3D joints, and adaptively weight the silhouette constraint according to the accuracy of estimated 3D joints. Through the adaptive weighting, the silhouette constraint would not be biased by the non-overlapping areas of truncated/self-occluded areas of human segments.

We evaluate the performance of our method on two 3D human datasets, namely **Human3.6M** [13] and **UP**[15]. Our method shows substantial improvements over the previous methods [14, 29, 5, 24, 42], especially for extreme human body pose and shape. The proposed bi-directional silhouette constraint has shown significant contribution and it is general which can be applied to other human reconstruction or 3D pose estimation frameworks. Some of our results on challenging images are shown in Fig. 1, and demonstrate the accurate estimation of human pose as well as a proper body shape which fits the body boundary well.

2. Related works

In this section, we briefly review the recent works in 3D human pose estimation and human shape reconstruction.

3D human pose estimation aims to locate accurate 3D joints from 2D images. Great progress has been made with the establishment of large-scaled 3D human pose datasets, such as **Human3.6M** [13] and **MPI-INF-3DHP** [20]. Because these 3D human datasets are mainly captured in a laboratory setting, 2D human pose datasets with in-the-wild images are always jointly used to enrich the data diversity [48, 40]. Existing approaches can be categorized into the two-stage methods [19, 22, 48, 40, 6, 23, 50] and the direct end-to-end methods [31, 7, 43, 33, 25, 36].

Two-stage methods first predict the 2D projection of 3D joints in image spaces and then estimate the corresponding depth values with various constraints, such as pose prior [6], skeleton prior [23], and geometric prior [50]. End-to-end methods directly estimate the 3D joint positions in a detection or regression framework. Statistical priors, such as kinematic human model [49] and adversarial learning [45], can be used as additional constraints to avoid abnormal poses. Human structural information is difficult to be directly exploited in the regression model, so it can be intentionally enhanced via a joint connection structure [35] or a latent pose representation [39]. The depth information of a 3D joint can be further refined using a coarse-to-fine scheme [28] or supervised by ordinal depth relation label [27].

3D human shape reconstruction estimates the whole 3D body geometry instead of only sparse joint positions. Recent approaches are mainly based on a parametric geometry representation, such as SCAPE [3] and SMPL [18], to estimate body geometry from a single image. The model parameters are optimized with several priors, including 2D joint position [5], edge and shading [11], biological limitation [47], and silhouette constraint [4, 15]. Besides the shape of human body, its clothing and texture can also be modeled simultaneously [2].

Besides optimization based approaches, the model parameters can also be regressed using deep neural networks [41, 37, 24, 29, 14]. These methods are generally supervised by 2D/3D joints and other image features, for example the body segmentation [29, 41, 24]. Tan *et al.* [37] utilizes synthetic data to train an encoder-decoder network, where the decoder network predicts the body masks directly from SMPL parameters and is trained from synthesized SMPL body. Kanazawa *et al.* [14] proposes the first end-to-end model, **HMR**, to directly regress the model parameters by jointly utilizing 2D/3D joints, labeled SMPL parameters, and a pose-adversarial constraint.

3. Parametric human shape reconstruction

In this section, we first introduce the SMPL model and our network architecture, then describe how our network is trained using the SMPL parameter loss, the joint position

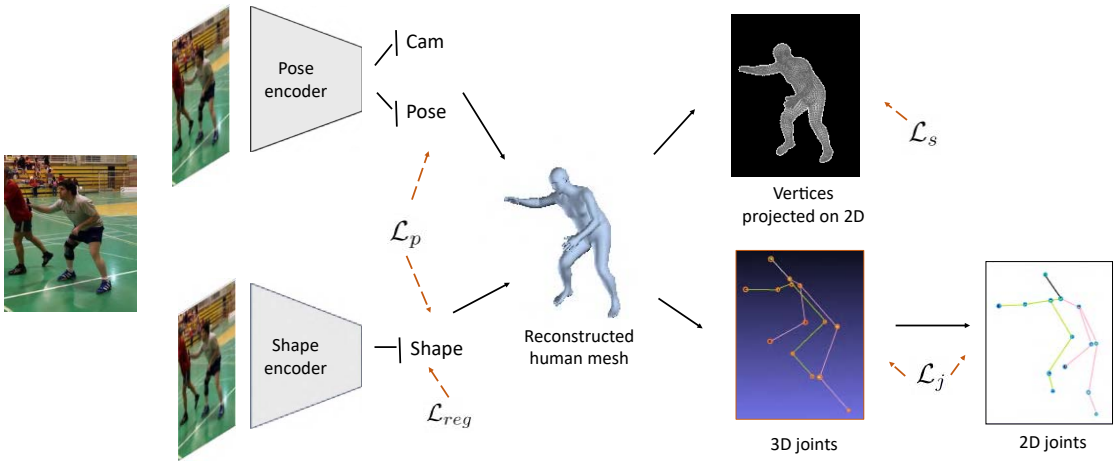


Figure 2. The network architecture of our method. The regression of human pose parameter and shape parameter is decoupled into two sub-networks. The estimated SMPL parameters are restricted by SMPL parameter loss $\mathcal{L}_p$, joint position loss $\mathcal{L}_j$, bidirectional silhouette loss $\mathcal{L}_s$ and regularization loss $\mathcal{L}_{reg}$.

loss, the bidirectional silhouette loss, and the regularization loss. We will also describe the datasets and training details by the end of this section.

3.1. Parametric shape model

We employ the SMPL [18], a skinned multi-person linear model, to represent human 3D geometry. Using a parametric representation has several advantages. First, it can significantly reduce the number of parameters which makes the single image body reconstruction possible. Second, by manipulating a few model parameters, one can create animations of the reconstructed model easily which benefits to many downstream applications.

A SMPL body shape $\mathcal{M}$ has $n = 6890$ vertices and is controlled by a shape parameter $\beta \in \mathbb{R}^{10}$ and a pose parameter $\theta \in \mathbb{R}^{72}$ as:

$$\mathcal{M}(\beta, \theta; \phi) = \mathcal{W}(\bar{T} + B_S(\beta) + B_P(\theta), \mathcal{J}(\beta), \theta), \quad (1)$$

where ϕ is the human model parameters learned from large human 3D scans, $\bar{T} \in \mathbb{R}^{3n}$ is vertices in the mean shape with zero pose, B_S and B_P are the shape-dependent blend function and pose-dependent blend function, $\mathcal{J}$ is a function for computing human joint positions, and $\mathcal{W}$ is a standard blend skinning function.

We use a weak perspective projection $\mathcal{C} = \{\mathcal{C}_s, \mathcal{C}_x, \mathcal{C}_y\}$ to project a 3D vertex $(x, y, z)^\top$ in $\mathcal{M}$ onto the 2D image plane as:

$$proj(x, y, z) = \mathcal{C}_s \cdot (x + \mathcal{C}_x, y + \mathcal{C}_y), \quad (2)$$

where $\mathcal{C}_s$ is a scale factor, and $(\mathcal{C}_x, \mathcal{C}_y)$ is a 2D translation.

Combining Eq. (1) and Eq. (2), recovering a 3D human geometry from one single image is equivalent to regress its SMPL parameter $\{\beta, \theta, \mathcal{C}\}$.

3.2. Network architecture

The architecture of proposed network is shown in Fig. 2. Although previous methods regress reasonable SMPL parameters from a shared generic image feature [24, 14, 29], the projected body shape does not always overlap with human silhouette. We observe that the reconstruction errors caused by the pose parameter errors are significantly larger than the errors caused by the shape parameter errors. Thus, putting these two sets of parameters equally would not solve the inherent bias that a network would focus to regress correct pose parameters before shape parameters. We therefore decouple the regression of shape parameter from pose and projection parameters, so we can use different sub-network architectures to achieve better estimations instead of simply sharing the same feature backbone.

We design our shape sub-network as a simplification to *VGG* network [26] which consists of 5 convolutional layers and 3 fully-connected layers to regress the shape parameter directly. There is an ambiguity between the shape parameter β and projection parameter $\mathcal{C}$ that changing one's body shape can be achieved by either changing the value of scale factor $\mathcal{C}_s$ or tuning the value of β. We avoid such ambiguity by putting the regression of $\mathcal{C}$ in the pose sub-network rather than leaving it in shape sub-network. Our pose sub-network uses *Resnet50* [12] as the backbone network and directly predicts both pose and projection parameters together. An ablation study of our network design is presented in Sec. 4.3

3.3. Training loss

We use four different losses to train our network, including the SMPL parameter loss $\mathcal{L}_p$, the joint position loss $\mathcal{L}_j$, the bidirectional silhouette loss $\mathcal{L}_s$, and the parameter regu-

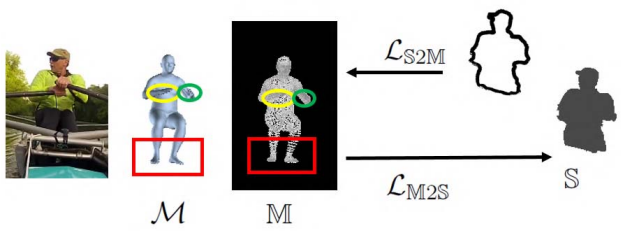


Figure 3. The illustration of proposed bidirectional silhouette loss. The red box indicates the region where $\omega_v = 0$ while the ellipses indicate $\omega_v = 1$. Vertices in green ellipse has larger ω_p than yellow ellipse since the right wrist joint is predicted more accurate than the left wrist.

larization loss $\mathcal{L}_{reg}$. So the entire training loss is:

$$\mathcal{L} = \mathcal{L}_p + \mathcal{L}_j + \mathcal{L}_s + \mathcal{L}_{reg}. \tag{3}$$

SMPL parameter loss. It is possible to get the ground truth of SMPL parameters by aligning its model to sparse landmarks on human shape [14, 15, 17]. So our network can be constrained by measuring the Euclidean distance between the predicted parameters and the corresponding ground truth as:

$$\mathcal{L}_p = \lambda_p(\|\beta - \hat{\beta}\|_2 + \|\theta - \hat{\theta}\|_2), \tag{4}$$

where $\hat{\beta}$, $\hat{\theta}$ are the ground truth and the weight λ_p reflects the importance of this loss.

Joint position loss. With the predicted SMPL parameters, both 3D and 2D positions of selected human joints can be calculated by Eq. (1) and Eq. (2). We therefore define the joint loss as:

$$\mathcal{L}_j = \lambda_{3D} \sum_{j \in \mathcal{F}_\mathcal{M}} \|\mathcal{M}(j) - \hat{j}_{3D}\|_2 + \lambda_{2D} \sum_{j \in \mathcal{F}_\mathcal{M}} \|proj(\mathcal{M}(j)) - \hat{j}_{2D}\|_2,$$

where $\mathcal{F}_\mathcal{M}$ denotes the set of 14 selected human joints, $\hat{j}_{3D}$ and $\hat{j}_{2D}$ is the ground truth position of joint j in 3D and 2D space. We use λ_{3D} and λ_{2D} to balance 3D joints constraint and 2D joints constraint.

Bidirectional silhouette loss. Based on the intuition that an accurate body geometry should align perfectly with the corresponding body segmentation, we propose a bidirectional silhouette loss (shown in Fig. 3) to evaluate how a projected geometry mesh $\mathbb{M} = proj(\mathcal{M})$ aligns with the predefined body segment $\mathbb{S}$ as:

$$\mathcal{L}_s = \lambda_{\mathbb{M}2\mathbb{S}}\mathcal{L}_{\mathbb{M}2\mathbb{S}} + \lambda_{\mathbb{S}2\mathbb{M}}\mathcal{L}_{\mathbb{S}2\mathbb{M}}. \tag{5}$$

The forward geometry-to-silhouette direction $\mathcal{L}_{\mathbb{M}2\mathbb{S}}$ measures the average distance between a mesh point $p \in \mathbb{M}$ to its nearest 2D mask point $\mathbb{S}_p \in \mathbb{S}$ as:

$$\mathcal{L}_{\mathbb{M}2\mathbb{S}} = \sum_{p \in \mathbb{M}} \omega_v \omega_p \|p - \mathbb{S}_p\|_2^2, \tag{6}$$

which ensures the predicted body lies within the segmentation. $\mathcal{L}_{\mathbb{M}2\mathbb{S}}$ might introduce wrong restriction when the human body is truncated or the predicted human joints are incorrect. Thus, we introduce adaptive weights ω_v and ω_p, to reduce the negative effects in these cases. We manually segment the 3D vertices in $\mathcal{M}$ into 14 semantic segments based on their nearby joints and denote the corresponding joint of vertex v as $\mathbb{J}(v)$. We then define ω_v and ω_p as:

$$\omega_v = \begin{cases} 1 & \text{if } \mathbb{J}(v) \text{ is visible} \\ 0 & \text{otherwise} \end{cases}, \tag{7}$$

$$\omega_p = \exp(-\|proj(\mathcal{M}(\mathbb{J}(v))) - \hat{\mathbb{J}}(v)_{2D}\|_2), \tag{8}$$

where $\hat{\mathbb{J}}(v)_{2D}$ is the ground truth of the 2D joint $\mathbb{J}(v)$.

The backward silhouette-to-geometry direction $\mathcal{L}_{\mathbb{S}2\mathbb{M}}$ constraints the estimation of human shape in an inverse way that it should cover the entire body silhouette as:

$$\mathcal{L}_{\mathbb{S}2\mathbb{M}} = \sum_{q \in \mathbb{S}_b} \|q - \mathbb{M}_q\|_2^2, \tag{9}$$

where $\mathbb{M}_q$ identifies the nearest mesh vertex in $\mathbb{M}$ to a 2D point q in the boundary of silhouette $\mathbb{S}_b$.

By jointly enforcing the bidirectional constraints $\mathcal{L}_{\mathbb{M}2\mathbb{S}}$ and $\mathcal{L}_{\mathbb{S}2\mathbb{M}}$, our silhouette loss $\mathcal{L}_s$ ensures that the estimated human shape should align with the body segmentation perfectly. Through the adaptive weights ω_v and ω_p, $\mathcal{L}_s$ influences our network less in the earlier training epochs when the estimation of human pose is not accurate enough. As the training goes on, the influence of $\mathcal{L}_s$ is increased along with the improvement of 3D human pose estimation and finally works together with $\mathcal{L}_p$ and $\mathcal{L}_j$ to achieve an accurate prediction of pose and shape parameters which better fit the body segmentation.

Regularization loss. A regularization [5] or discriminator network [14] of predict parameters is necessary for alleviating abnormal human pose or body shape. In this paper, we incorporate the regularization instead of discriminator network to make our network as simple as possible. Because our network is trained with large human 3D pose datasets [13, 20], we found that ambiguity of human pose can be well relieved even without additional regularization. We therefore employ the $L2$ regularization loss only for shape parameter β as:

$$\mathcal{L}_{reg} = \lambda_{reg}\|\beta\|_2. \tag{10}$$

3.4. Datasets and training details

In order to ensure the proposed method is robust to different imaging conditions, we follow [14] to use 5 datasets together for training. **UP** [15] consists of two subsets, **UP-S1h** containing $26,294$ images with 2D joint positions and body segmentations, **UP-3D** containing $8,515$ images with 2D joints and the ground truth of SMPL parameters. **MS-COCO** [16] provides around $80,000$ images with both 2D joint positions and body segmentations. Besides these three in-the-wild datasets, we also use two large scale 3D human pose datasets, **Human3.6M** [13] and **MPI-INF-3DHP** [23]. **Human3.6M** provides 17 scenarios acted by 11 people from 4 calibrated cameras in a laboratory environment. It consists about 360k images with available annotations of 2D joint positions, 3D joint positions, and SMPL parameters. **MPI-INF-3DHP** is another human 3D pose dataset consisting of 150k images captured in lab scenarios.

Our network takes a $224 * 224$ image as input and the human region is roughly aligned in the center with about 150 pixel height. The network is trained from scratch by using the combination of all datasets for 50 epochs with a learning rate $1e^{-4}$ and then fine-tuned for another 10 epochs with learning $1e^{-5}$. We set the batch-size to 64 and use Adam solver.

4. Experiment and results

We compare the performance of our proposed method with other state-of-the-art methods in this section.

4.1. Evaluation metrics

We employ quantitative evaluation on **Human3.6M** and **UP** datasets with different metrics.

Evaluation on Human3.6M. We follow a common protocol to evaluate the performance of our method on **Human3.6M** [28, 29, 14]. We use the images for subject 1 and $5 \sim 8$ as the training set and test on images for subject 9 and 11. The *Protocol 1* tests on all the four capturing cameras and reports the mean per joint position error(*MPJPE*); the *Protocol 2* only tests on the images captured by the frontal camera (*Cam3*) and reports the mean per joint position error after a rigid alignment via Procrustes Analysis [10](*PA-MPJPE*). *PA-MPJPE* is a variation of *MPJPE* that eliminates the global misalignments by Procrustes Analysis and exactly measures the accuracy of 3D pose estimation.

Evaluation on UP. The evaluation on **UP** dataset also consists of two parts: 1) an evaluation on **UP-S1h** focus on the binary prediction accuracy between the 2D mask of estimated human geometry and the ground truth of body segmentation ,and we use two metrics, *Accuracy* and *F1-score*, as the measurement; 2) an evaluation on **UP-3D** focuses the accuracy of reconstructed body shape and measures the reconstruction error as mean per vertex position error after Procrustes Analysis(*PA-MPVPE*).

Method	*PA-MPJPE*(mm)
*Tome *et al.* [40]	70.7
*Martinez *et al.* [19]	47.7
*Pavlakos *et al.* [27]	41.8
*Yang *et al.* [45]	37.7
SMPLify [5]	82.3
Lassner *et al.* [15](UP-P91)	80.7
Lassner *et al.* [15](Direct predict)	93.9
Pavlakos *et al.* [29]	75.9
NBF [24]	59.9
HMR [14]	**56.8**
Ours	57.3

Table 1. Evaluation for 3D pose estimation on **Human3.6M** *Protocol 2* with images only from the frontal camera. * indicates methods only estimate 3D joint positions.

4.2. Evaluation with the state-of-the-art methods

We compare our results with both the state-of-the-art 3D pose estimation methods [40, 19, 27, 45] and 3D shape reconstruction methods [14, 29, 15, 5, 24, 42]. The evaluations for previous methods are obtained from their original papers.

3D pose estimation. The evaluation on 3D pose estimation with **Human3.6M** [13] *Protocol 1* and *Protocol 2* are shown in Tab. 2 and Tab. 1, respectively. Tab. 2 lists the *MPJPE* results on all test images and Tab. 1 lists the *PA-MPJPE* results on images only from the frontal camera. The methods denoted by * only estimate the positions of sparse 3D joints and the rests, including our method, reconstruct the whole body geometry.

Since the 3D human shape reconstruction problem is generally a more complex problem than the 3D human pose estimation, methods directly predicting 3D joint positions show better results than methods regressing SMPL parameters. As shown in Tab. 1 and Tab. 2,two state-of-the-art 3D human shape reconstruction methods, NBF [24] and HMR [14] and our method have very competitive results and are significantly better than others [5, 15, 29] in *Protocol 2*. Comparing to *Protocol 2* which only uses images from the frontal camera, the *Protocol 1* results shown in Tab. 2 demonstrate our method outperforms HMR [14] in a more general evaluation that the test images are from all the four cameras in **Human3.6M** [13]. The result of our method shows more accuracy and robust estimations of 3D joint positions than previous human reconstruction approaches.

Method	*MPJPE*(mm)
*Vnect [21]	80.5
*Pavlakos *et al.* [28]	71.9
*Martinez *et al.* [19]	62.9
*Yang *et al.* [45]	58.6
HMR [14]	88.0
Ours	**84.4**

Table 2. Evaluation for 3D pose estimation on **Human3.6M** *Protocol 1* with images from all the four cameras. * indicates methods only estimate 3D joint positions.

Method	*Acc.*	*F1.*	*Time*	*Platform*
SMPLify [5]	0.919	0.88	60s	Desktop
Lassner *et al.* [15](DP)	0.867	0.80	0.13s	GTX970
Bodynet [42]	**0.928**	0.84	0.28s	Modern GPU
SMPLify-anchor [29]	0.922	0.88	60s	Desktop
HMR [14]	0.917	0.87	0.040s	Titan 1080ti
Our	0.919	**0.88**	0.012s	Tesla P40
3D ground truth [15]	0.93	0.88	\	\

Table 3. Evaluation for human body segmentation on **UP**. The running time and platform for each method is obtained from their original paper.

Human body segmentation. Besides the evaluation on 3D pose estimation task, we further analyze the performance of our method focusing more on the quality of reconstructed body shape via evaluating the body segmentation results on **UP** [15] dataset with segmentation *accuracy* and *F1-score* as done in [14, 15].

A result with 3D ground truth SMPL parameters is provided by [15] and sets the upper bound for this evaluation. Although Bodynet [42] achieved the highest performance in *accuracy*, but its *F1-score* is relative lower than other methods. Furthermore, it outputs a 3D body geometry in the volume space and is generally time consuming than others which directly regress the SMPL parameters instead. We believe the *F1-score* is a more appropriate metric than *accuracy* in human body segmentation task because it balances the precision and recall, and a significant improvement over human pose or shape may only increase *F1-score* a little. Three methods, SMPLify [5], SMPLify-anchor [29] and our approach arrive the upper bound of *F1-score* while our method is extremely faster than others. Although our method and HMR [14] are very competitive in this evaluation, our method is slightly better in a higher *F1-score* which demonstrates the robustness of our method. The running time and platform of each method is obtained from relevant papers and also included in Tab. 3. Because the running platforms are different, we only regard it as a reference for evaluating the overall computational cost.

Qualitative evaluation. Besides the aforementioned quantitative evaluations on 3D pose estimation and human body segmentation, we further present some qualitative comparisons with HMR in Fig. 4. Thanks to the proposed bidirectional silhouette constraint, our results are visually better in matching with human boundaries which means a more accurate pose and body shape are estimated.

In order to validate our method on more general images rather than only selecting images from the datasets, we additionally show some results of our method on collected Internet images in Fig. 5 and it demonstrates our method is accurate and robust enough for handling various scenarios.

4.3. Ablation study

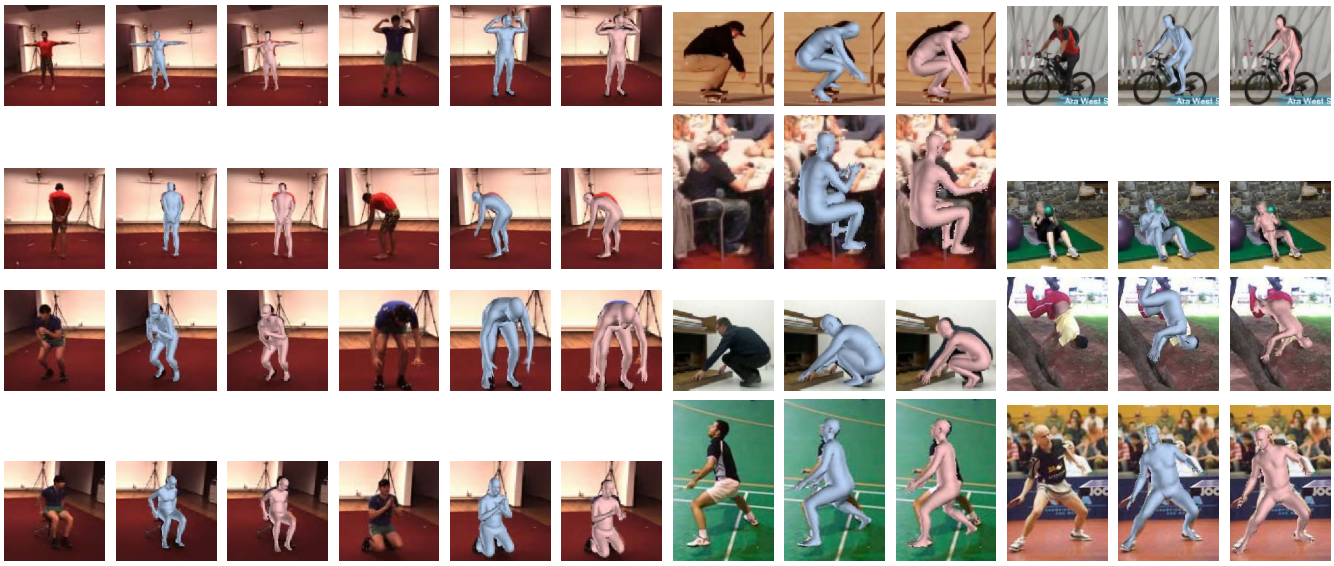
To better analyze the network architecture and understand the benefit by incorporating the bidirectional silhouette constraint, we compare several different network structures with/without our silhouette constraint in Tab. 4. All the results presented in this subsection are trained and evaluated on **UP** dataset to eliminate the possible improvement achieved by using multiple training datasets.

Network architecture. As discussed in Sec. 3.2, decoupling the regression of shape and pose parameters achieves better estimation than simply sharing the same feature backbone. So we design our-base1 network using a simple architecture that one *ResNet*-50 [12] is used to extract a generic image feature and then directly regresses all the SMPL parameters from it. We further separate the regression of pose and shape into two sub-networks to form our-base2 network and two *ResNet*-50 [12] are used individually. Considering the freedom of human shape is much less than human poses, we propose our final design by using a simplification of *VGG* network [26] as the backbone for shape sub-network.

According to the evaluation listed in Tab. 4, separating the regression of pose and shape performs better than sharing the same feature network. Although the *accuracy* and *PA-MPVPE* decline a little bit after replacing the shape backbone by simplified *VGG*, it achieves a higher *F1-score* which is more balance for evaluating a segmentation problem. Taking the performance and efficiency into account, we choose to present a decoupled network architecture with simplified *VGG* as the shape backbone in Sec. 3.2.

Bidirectional silhouette constraint. One of our major contribution is the proposed adaptive bidirectional silhouette constraint which improves the estimation of both human pose and shape via encouraging the reconstructed body geometry aligns well with the body segmentation/boundary. We validate its capability by employing it with four different network architectures, including our-base1, our-base2, our-proposed network, and the state-of-the-art HMR [14].

Based on the evaluation results listed in Tab. 4, all the four networks perform better after training with the pro-



Results on **Human3.6M** [14] Results on **UP** [15]

Figure 4. Qualitative evaluation of our method on images in **Human3.6M** and **UP**. The images in **Human3.6M** are captured in the laboratory environment and the images in **UP** are collected in-the-wild. Our results are rendered as blue and the results of HMR [14] are rendered as pink.

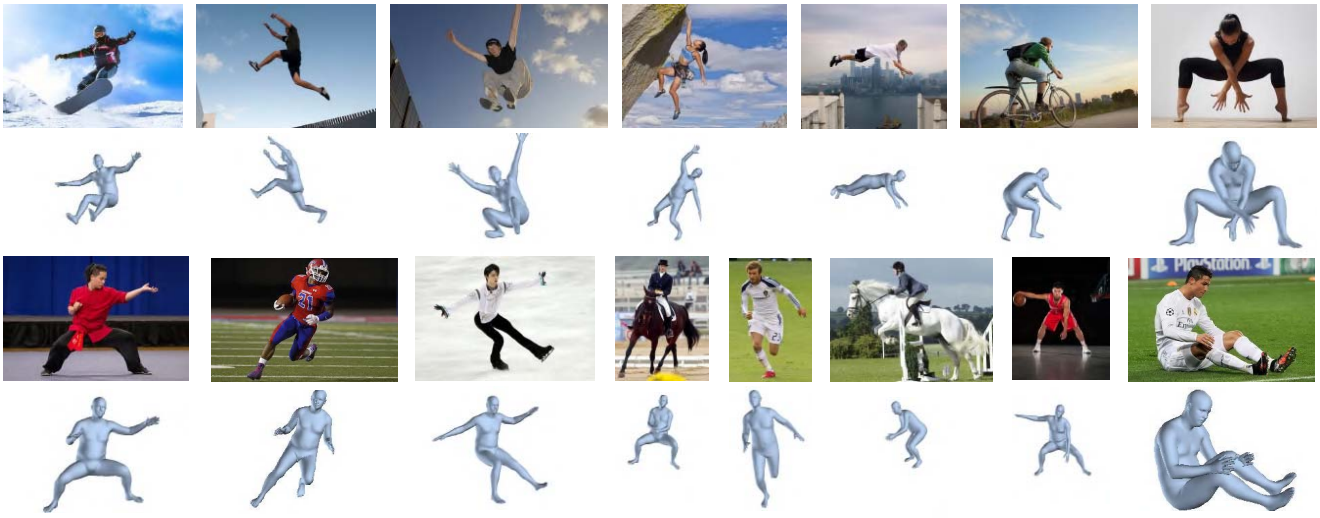


Figure 5. Qualitative results of our method on challenging Internet images which are totally outside of the datasets.

posed bidirectional silhouette constraint, especially achieving a significant improvement in *F1-score*. It is because these networks achieve a better alignment between the estimated body and human mask via the incorporation of silhouette constraint and result in a rising of true positive and decrease of false negative. Results shown in Fig. 6 visually demonstrate the benefit of incorporating silhouette constraint in some specific cases. We highlight the regions where silhouette constraint shows substantial improvements over the networks without using silhouette constraint in a better body boundary fitting. The silhouette constraint not only help the estimated body geometry fit the body segmentation well, but also correct inaccurate 3D pose estimations, especially for the results of HMR [14] in the second and third case.

4.4. Failure cases

Although our method outperforms other state-of-the-art single image human 3D reconstruction approaches, it still suffers from failure cases caused by extreme human pose or shape as shown in Fig. 7. Even incorporating with the silhouette constraint, our method still cannot handle such ex-

Method	Network			$\mathcal{L}_s$	*Accuracy*	*F1-score*	*PA-MPVPE*(mm)
	Pose-Net	Shape-Net	Regressor				
HMR [14]	*ResNet*-50		IEF-FC	w/o	0.890	0.811	96.3
				w/	0.893 ↑	0.845 ↑	92.7 ↓
Our-base1	*ResNet*-50		Straight-FC	w/o	0.900	0.827	95.7
				w/	0.905 ↑	0.840 ↑	90.9 ↓
Our-base2	*ResNet*-50	*ResNet*-50	Straight-FC	w/o	0.907	0.844	89.9
				w/	**0.908** ↑	0.847 ↑	88.5 ↓
Our-proposed	*ResNet*-50	*Simple VGG*	Straight-FC	w/o	0.898	0.849	91.1
				w/	0.903 ↑	**0.856** ↑	**86.3** ↓

Table 4. Ablation study for the network architecture and validation for the bidirectional silhouette constraint. IEF-3 identifies the iterative error feedback structure with three fully-connected layers used in HMR and Straight-FC identifies straight fully-connected layers used in our method. w/o means without while w/ means with. ↑ and ↓ denote the metric changes caused by the utilization of silhouette constraint.

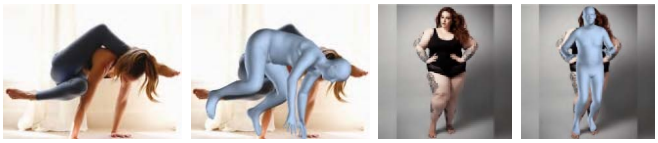
(a) (b) (c) (d) (e)

Figure 6. Visual benefits of incorporating silhouette constraint.(a) The input image. (b) Results of HMR w/o $\mathcal{L}_s$. (c) Results of HMR w/ $\mathcal{L}_s$. (d) Results of our-proposed network w/o $\mathcal{L}_s$. (e) Results of our-proposed network w/ $\mathcal{L}_s$. Substantial improved regions are masked using red boxes.

treme cases because these images are very rare in all public human 3D datasets. This problem might be addressed by extending the current datasets or training the network using synthetic images and real images jointly [42].

5. Conclusion

We propose a novel end-to-end deep neural network architecture to accurately reconstruct 3D human shape from a single image. We decouple the regression of pose and shape parameters into two sub-networks and employ different backbone architectures to extract more specific features for each individual regression. The network is constrained by a bidirectional silhouette constraint which forces the estimated geometry to match the predefined human body segmentation. Our method shows substantial improvement over the previous methods in its accurate estimation on both 3D human pose and shape. The proposed bi-directional silhouette constraint is also general and significantly contributes to other human reconstruction or 3D pose estimation frameworks.



(a) (b)

Figure 7. Failure case study of our method. (a) Failure case on image with extreme pose. (b) Failure case on image with extreme body shape.

References

[1] T. Alldieck, M. Kassubeck, B. Wandt, B. Rosenhahn, and M. Magnor. Optical flow-based 3d human motion estimation from monocular video. In *Proc. German Conference on Pattern Recognition*, 2017. 1

[2] T. Alldieck, M. A. Magnor, W. Xu, C. Theobalt, and G. Pons-Moll. Video based reconstruction of 3d people models. In *Proc. of Computer Vision and Pattern Recognition*, 2018. 2

[3] D. Anguelov, P. Srinivasan, D. Koller, S. Thrun, J. Rodgers, and J. Davis. Scape: shape completion and animation of people. *ACM Transactions on Graphics*, 24(3):408–416, 2005. 1, 2

[4] A. O. Balan, L. Sigal, M. J. Black, J. E. Davis, and H. W. Haussecker. Detailed human shape and pose from images. In *Proc. of Computer Vision and Pattern Recognition*, 2007. 2

[5] F. Bogo, A. Kanazawa, C. Lassner, P. Gehler, J. Romero, and M. 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, 4, 5, 6

[6] C.-H. Chen and D. Ramanan. 3d human pose estimation= 2d pose estimation+ matching. In *Proc. of Computer Vision and Pattern Recognition*, 2017. 2

[7] R. Dabral, A. Mundhada, U. Kusupati, S. Afaque, A. Sharma, and A. Jain. Learning 3d human pose from structure and motion. In *Proc. of European Conference on Computer Vision*, 2018. 2

[8] S. R. Fanello, C. Keskin, S. Izadi, P. Kohli, D. Kim, D. Sweeney, A. Criminisi, J. Shotton, S. B. Kang, and T. Paek. Learning to be a depth camera for close-range human capture and interaction. *ACM Transactions on Graphics*, 33(4):86:1–86:11, 2014. 1

[9] K. Fragkiadaki, H. Hu, and J. Shi. Pose from flow and flow from pose. In *Proc. of Computer Vision and Pattern Recognition*, 2013. 1

[10] J. C. Gower. Generalized procrustes analysis. *Psychometrika*, 40(1):33–51, 1975. 5

[11] P. Guan, A. Weiss, A. O. Balan, and M. J. Black. Estimating human shape and pose from a single image. In *Proc. of Internatoinal Conference on Computer Vision*, 2009. 2

[12] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proc. of Computer Vision and Pattern Recognition*, 2016. 3, 6

[13] C. Ionescu, D. Papava, V. Olaru, and C. 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, 2014. 2, 4, 5

[14] A. Kanazawa, M. J. Black, D. W. Jacobs, and J. Malik. End-to-end recovery of human shape and pose. In *Proc. of Computer Vision and Pattern Recognition*, 2018. 1, 2, 3, 4, 5, 6, 7, 8

[15] C. Lassner, J. Romero, M. Kiefel, F. Bogo, M. J. Black, and P. V. Gehler. Unite the people: Closing the loop between 3d and 2d human representations. In *Proc. of Computer Vision and Pattern Recognition*, 2017. 2, 4, 5, 6, 7

[16] T. Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick. Microsoft coco: Common objects in context. In *Proc. of European Conference on Computer Vision*, 2014. 5

[17] M. Loper, N. Mahmood, and M. J. Black. Mosh:motion and shape capture from sparse markers. *ACM Transactions on Graphics*, 33(6):1–13, 2014. 4

[18] M. Loper, N. Mahmood, J. Romero, G. Pons-Moll, and M. J. Black. Smpl: A skinned multi-person linear model. *ACM Transactions on Graphics*, 34(6):248:1–248:16, 2015. 1, 2, 3

[19] J. Martinez, R. Hossain, J. Romero, and J. J. Little. A simple yet effective baseline for 3d human pose estimation. In *Proc. of Internatoinal Conference on Computer Vision*, 2017. 2, 5, 6

[20] D. Mehta, H. Rhodin, D. Casas, P. Fua, O. Sotnychenko, W. Xu, and C. Theobalt. Monocular 3d human pose estimation in the wild using improved cnn supervision. In *Proc. of International Conference on 3D*, 2017. 2, 4

[21] D. Mehta, S. Sridhar, O. Sotnychenko, H. Rhodin, M. Shafiei, H.-P. Seidel, W. Xu, D. Casas, and C. Theobalt. Vnect: Real-time 3d human pose estimation with a single rgb camera. *ACM Transactions on Graphics*, 36(4):44:1–44:14, 2017. 6

[22] F. Moreno-Noguer. 3d human pose estimation from a single image via distance matrix regression. In *Proc. of Computer Vision and Pattern Recognition*, 2017. 2

[23] B. X. Nie, P. Wei, and S.-C. Zhu. Monocular 3d human pose estimation by predicting depth on joints. In *Proc. of Internatoinal Conference on Computer Vision*, 2017. 2, 5

[24] M. Omran, C. Lassner, G. Pons-Moll, P. V. Gehler, and B. Schiele. Neural body fitting: Unifying deep learning and model-based human pose and shape estimation. In *Proc. of International Conference on 3D*, 2018. 1, 2, 3, 5

[25] S. Park, J. Hwang, and N. Kwak. 3d human pose estimation using convolutional neural networks with 2d pose information. In *Proc. of European Conference on Computer Vision*, 2016. 2

[26] O. M. Parkhi, A. Vedaldi, and A. Zisserman. Deep face recognition. In *Proc. of British Machine Vision Conference*, 2015. 3, 6

[27] G. Pavlakos, X. Zhou, and K. Daniilidis. Ordinal depth supervision for 3d human pose estimation. In *Proc. of Computer Vision and Pattern Recognition*, 2018. 2, 5

[28] G. Pavlakos, X. Zhou, K. G. Derpanis, and K. Daniilidis. Coarse-to-fine volumetric prediction for single-image 3d human pose. In *Proc. of Computer Vision and Pattern Recognition*, 2017. 2, 5, 6

[29] G. Pavlakos, L. Zhu, X. Zhou, and K. 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, 5, 6

[30] T. Probst, A. Fossati, and L. Van Gool. Combining human body shape and pose estimation for robust upper body tracking using a depth sensor. In *Computer Vision – ECCV 2016 Workshops*, 2016. 1

[31] H. Rhodin, M. Salzmann, and P. Fua. Unsupervised geometry-aware representation for 3d human pose estimation. In *Proc. of European Conference on Computer Vision*, 2018. 2

[32] K. M. Robinette, H. Daanen, and E. Paquet. The caesar project: a 3-d surface anthropometry survey. In *Proc. International Conference on 3-D Digital Imaging and Modeling*, 1999. 1

[33] G. Rogez and C. Schmid. Mocap-guided data augmentation for 3d pose estimation in the wild. In *Proc. of Advances in Neural Information Processing Systems*, 2016. 2

[34] J. Romero, M. Loper, and M. J. Black. Flowcap: 2d human pose from optical flow. In *Proc. German Conference on Pattern Recognition*, 2015. 1

[35] X. Sun, J. Shang, S. Liang, and Y. Wei. Compositional human pose regression. In *Proc. of Internatoinal Conference on Computer Vision*, 2017. 2

[36] X. Sun, B. Xiao, F. Wei, S. Liang, and Y. Wei. Integral human pose regression. In *Proc. of European Conference on Computer Vision*, 2018. 2

[37] J. Tan, I. Budvytis, and R. Cipolla. Indirect deep structured learning for 3d human body shape and pose prediction. In *Proc. of British Machine Vision Conference*, 2017. 1, 2

[38] Y. Tao, Z. Zheng, K. Guo, J. Zhao, D. Quionhai, H. Li, G. Pons-Moll, and Y. Liu. Doublefusion: Real-time capture of human performance with inner body shape from a depth sensor. In *Proc. of Computer Vision and Pattern Recognition*, 2018. 1

[39] B. Tekin, I. Katircioglu, M. Salzmann, V. Lepetit, and P. Fua. Structured prediction of 3d human pose with deep neural networks. In *Proc. of British Machine Vision Conference*, 2016. 2

[40] D. Tome, C. Russell, and L. Agapito. Lifting from the deep: Convolutional 3d pose estimation from a single image. In *Proc. of Computer Vision and Pattern Recognition*, 2017. 2, 5

[41] H.-Y. Tung, H.-W. Tung, E. Yumer, and K. Fragkiadaki. Self-supervised learning of motion capture. In *Proc. of Advances in Neural Information Processing Systems*, 2017. 1, 2

[42] G. Varol, D. Ceylan, B. Russell, J. Yang, E. Yumer, I. Laptev, and C. Schmid. Bodynet: Volumetric inference of 3d human body shapes. In *Proc. of European Conference on Computer Vision*, 2018. 2, 5, 6, 8

[43] G. Varol, J. Romero, X. Martin, N. Mahmood, M. J. Black, I. Laptev, and C. Schmid. Learning from synthetic humans. In *Proc. of Computer Vision and Pattern Recognition*, 2017. 2

[44] D. Vlasic, P. Peers, I. Baran, P. Debevec, J. Popović, S. Rusinkiewicz, and W. Matusik. Dynamic shape capture using multi-view photometric stereo. *ACM Transactions on Graphics*, 28(5):174:1–174:11, 2009. 1

[45] W. Yang, W. Ouyang, X. Wang, J. Ren, H. Li, and X. Wang. 3d human pose estimation in the wild by adversarial learning. In *Proc. of Computer Vision and Pattern Recognition*, 2018. 2, 5, 6

[46] Z. Zheng, T. Yu, H. Li, K. Guo, Q. Dai, L. Fang, and Y. Liu. Hybridfusion: Real-time performance capture using a single depth sensor and sparse imus. In *Proc. of European Conference on Computer Vision*, 2018. 1

[47] S. Zhou, H. Fu, L. Liu, D. Cohen-Or, and X. Han. Parametric reshaping of human bodies in images. *ACM Transactions on Graphics*, 29(4):126:1–126:10, 2010. 2

[48] X. Zhou, Q. Huang, X. Sun, X. Xue, and Y. Wei. Towards 3d human pose estimation in the wild: A weakly-supervised approach. In *Proc. of Internatoinal Conference on Computer Vision*, 2017. 2

[49] X. Zhou, X. Sun, W. Zhang, S. Liang, and Y. Wei. Deep kinematic pose regression. In *Proc. of European Conference on Computer Vision*, 2016. 2

[50] X. Zhou, M. Zhu, S. Leonardos, K. G. Derpanis, and K. Daniilidis. Sparseness meets deepness: 3d human pose estimation from monocular video. In *Proc. of Computer Vision and Pattern Recognition*, 2016. 2