3D Human Mesh Regression with Dense Correspondence

Wang Zeng[1], Wanli Ouyang[2], Ping Luo[3], Wentao Liu[4], and Xiaogang Wang[1,4]

[1]The Chinese University of Hong Kong [2]The University of Sydney [3]The University of Hong Kong [4]SenseTime Research
{zengwang@link, xgwang@ee}.cuhk.edu.hk, wanli.ouyang@sydney.edu.au, pluo@cs.hku.hk,
liuwentao@sensetime.com

Abstract

*Estimating 3D mesh of the human body from a single 2D image is an important task with many applications such as augmented reality and Human-Robot interaction. However, prior works reconstructed 3D mesh from global image feature extracted by using convolutional neural network (CNN), where the dense correspondences between the mesh surface and the image pixels are missing, leading to suboptimal solution. This paper proposes a model-free 3D human mesh estimation framework, named DecoMR, which explicitly establishes the dense correspondence between the mesh and the local image features in the UV space (*i.e. *a 2D space used for texture mapping of 3D mesh). DecoMR first predicts pixel-to-surface dense correspondence map (*i.e.*, IUV image), with which we transfer local features from the image space to the UV space. Then the transferred local image features are processed in the UV space to regress a location map, which is well aligned with transferred features. Finally we reconstruct 3D human mesh from the regressed location map with a predefined mapping function. We also observe that the existing discontinuous UV map are unfriendly to the learning of network. Therefore, we propose a novel UV map that maintains most of the neighboring relations on the original mesh surface. Experiments demonstrate that our proposed local feature alignment and continuous UV map outperforms existing 3D mesh based methods on multiple public benchmarks. Code will be made available at* `https://github.com/zengwang430521/DecoMR`.

1. Introduction

Estimation of the full human body pose and shape from a monocular image is a fundamental task for various applications such as human action recognition [12, 35], VR/AR [11] and video editing [10]. It is challenging mostly due to the inherent depth ambiguity and the difficulty to obtain the ground-truth 3D human body data. There are several popular representations for 3D objects in literature, *e.g.*, point clouds, 3D voxels and 3D meshes. Because of its compatibility with existing computer graphic engines and the efficiency to represent object surface in details with reasonable storage, 3D mesh representation has been widely adopted for 3D human body reconstruction [18, 4, 20, 8, 27, 38, 11, 26, 25, 37, 21, 39].

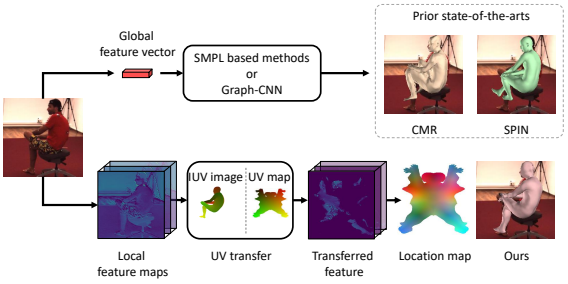


Figure 1. Prior methods (*e.g.*, SPIN [20] and CMR [21]) usually reconstruct 3D meshes of human body from the global image feature vector extracted by neural networks, where the dense correspondences between the mesh surface and the image pixels are missing, leading to suboptimal results (top). Our DecoMR framework explicitly establishes such correspondence in the feature space with the aid of a novel continuous UV map, which results in better results in mesh details (bottom).

However, unlike 3D voxel representation, the dense correspondence between the template human mesh surface and the image pixels is missing, while this dense correspondence between the input and the output has been proven crucial for various tasks [24, 39]. Due to this limitation, most existing 3D mesh based methods, either model-based [18, 26, 25, 20] or model-free [21], have to ignore the correspondence between the mesh representation and pixel representation. And they have to estimate the human meshes based on either global image feature [18, 21, 20], or hierarchical projection and refinement [39], which is time consuming and sensitive to initial estimation.

To utilize the 3D mesh representation without losing

the correspondence between the mesh space and the image space, we propose a 3D human mesh estimation framework that explicitly establishes the dense correspondence between the output 3D mesh and the input image in the UV space.

Representing output mesh by a new UV map: Every point on the mesh surface is represented by its coordinates on the continuous UV map. Therefore, the 3D mesh can be presented as a location map in the UV space, of which the pixel values are the 3D coordinates of its corresponding point on the mesh surface, as shown in Figure 1. Instead of using SMPL default UV map, we construct a new continuous UV map that maintains more neighboring relations of the original mesh surface, by parameterizing the whole mesh surface into a single part on the UV plane, as shown in Figure 1.

Mapping image features to the UV space: To map the image features to the continuous UV map space, we first use a network that takes a monocular image as input for predicting an IUV image [2], which assign each pixel to a specific body part location. Then the local image features from the decoder are transferred to the UV space with the guidance of predicted IUV image to construct the transferred feature maps that are well aligned with the corresponding mesh area.

Given the transferred local features, we use both the local features and the global feature to estimate the location map in the UV space, which is further used to reconstruct the 3D human body mesh with the predefined UV mapping function. Since our UV map is continuous and maintains the neighboring relationships among body parts, details between body parts can be well preserved when the local features are transferred.

In summary, our contributions are twofold:

- We propose a novel UV map that maintains most of the neighboring relations on the original mesh surface.
- We explicitly establish the dense correspondence between the output 3D mesh and the input image by the transferred local image features.

We extensively evaluate our methods on multiple widely used benchmarks for 3D human body reconstruction. Our method achieves state-of-the-art performance on both 3D human body mesh reconstruction and 3D human body pose estimation.

2. Related Work

2.1. Optimization-based methods

Pioneer works solve the 3D human body reconstruction by optimizing parameters of an predefined 3D human mesh models, *e.g.*, SCAPE [3] and SMPL [23], with respect to the ground-truth body landmark locations [8], or employing a 2D keypoints estimation network [4]. To improve the precision, extra landmarks are used in [22]. Recent work [38] enables multiple persons body reconstruction by incorporating human semantic part segmentation clues, scene and temporal constrains.

2.2. Learning-based methods

Model-based methods: Directly reconstruction of the 3D human body from a single image is a relatively hard problem. Therefore, many methods incorporate a parameterized 3D human model and change the problem into the model parameter regression. For example, HMR [18] regresses the SMPL parameters directly from RGB image. In order to mitigate the lack of robustness caused by the inadequacy of in-the-wild training data, some approaches employ intermediate representations, such as 2D joint heatmaps and silhouette [26], semantic segmentation map [25] or IUV image [36]. Recently, SPIN [20] incorporates 3D human model parameter optimization into network training process by supervising network with optimization result, and achieves the state-of-the-art results among model-based 3D human body estimation approaches.

Compared with optimization-based methods, model parameter regression methods are more computationally efficient. While these methods can make use of the prior knowledge embedded in 3D human model, and tend to reconstruct more biologically plausible human bodies compared with model-free methods, the representation capability is also limited by the parameter space with these predefined human models. In addition, as stated in [21], 3D human model parameter space might not be so friendly to the learning of network. On the contrary, our framework does not regress model parameters. Instead, it directly outputs 3D coordinates of each mesh vertex.

Model-free methods: Some methods do not rely on human models and regress 3D human body representation directly from image. BodyNet [33] estimates volumetric representation of 3D human with a Voxel-CNN. A recent work [6] estimates visible and hidden depth maps, and combines them to form a point cloud of human. Voxel and point cloud based representations are flexible and can represent objects with different topology. However, the capability of reconstructing surface details is limited by the storage cost.

CMR [21] uses a Graph-CNN to directly regress 3D coordinates of vertices from image features. Densebody [37] estimates vertex location in the form of UV-position map. A recent work [28] represents the 3D shapes using 2D geometry images, which can be regarded as a special kind of UV-position map. These methods do not use any human model. However, they still lack correspondence between human mesh and image and estimate the whole surface only relying on global image feature. On the contrary, our method can employ local feature for the reconstruction

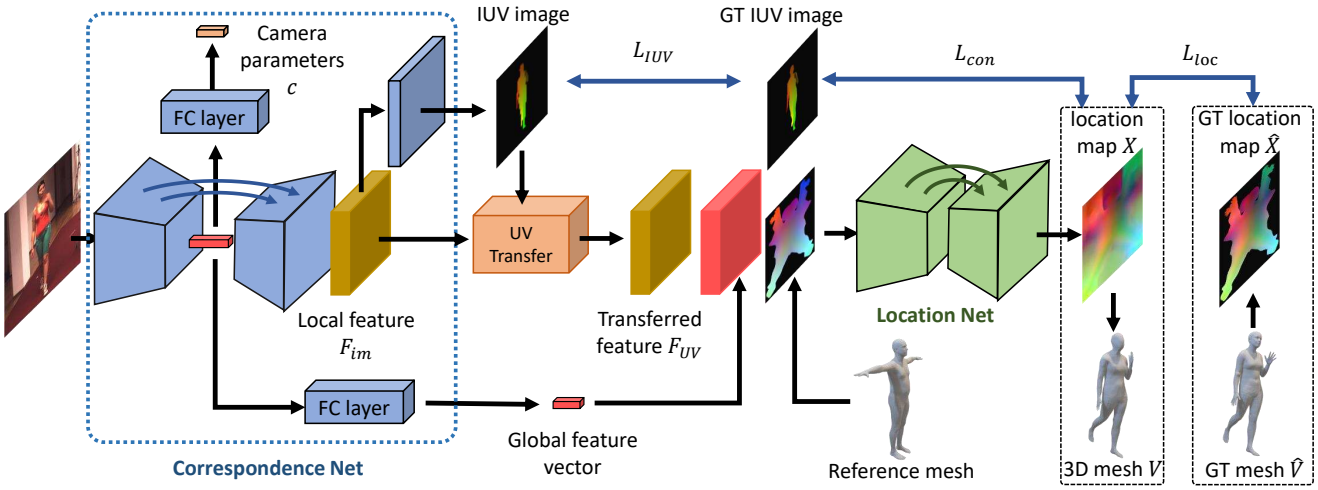


Figure 2. Overview of our framework. Given an input image, an IUV map is first predicted by the correspondence net. Then local image features are transferred to the UV space. Location net takes transferred local features, expanded global feature and reference location map as input, and regresses a location map. Finally, 3D mesh is reconstructed from the location map.

of corresponding surface area.

The efficacy of the UV space representation has been demonstrated in recent work Tex2Shape [1], where the 3D human shape is estimated from the texture map which is obtained by transferring images pixels according to the IUV image estimated by DensePose [2]. We also use the IUV image to guide the human mesh estimation. However, in [1], the UV transfer is used to preprocess the raw image and is independent from the model learning, while we incorporate the UV transfer into our network to enable the end-to-end learning. We observe the efficacy of learning the transferred features end-to-end, which has also been proved by prior works, *e.g.*, Spatial Transformer Networks [15] and Deformable ConvNets [5].

Very recently, HMD [39] refines initial estimated human mesh by hierarchical projection and mesh deformation. PIFu [30] reconstructs 3D human as implicit function. HMD and PIFu are able to utilize local image features to achieve impressive details in the reconstruction results. However, HMD is computationally intensive and sensitive to the initial estimation, while implicit function lacks the semantic information of human body. In contrast, we estimate the pixel-to-surface dense correspondence from images directly, which is computationally efficient and more robust, and the location map maintains the semantic information of human body.

3. Our Method

Overview. As shown in Figure 2, our framework DecoMR consists of two components, including a dense correspondence estimation network (CNet), which preforms in the image space, as well as a localization network (LNet), which performs on a new continuous UV map space. The CNet has an encoder-decoder architecture to estimate an IUV image. It also extracts local image features $\mathcal{F}_{im}$, and then uses the the estimated IUV image for transferring the image features $\mathcal{F}_{im}$ to the transferred local features $\mathcal{F}_{UV}$ in the UV space. LNet takes the above transferred local features $\mathcal{F}_{UV}$ as input, and regresses a location map X, whose pixel value is the 3D coordinates of the corresponding points on the mesh surface. Finally, the 3D human mesh V is reconstructed from the above location map by using a predefined UV mapping function. As a result, the location map and the transferred feature map are well aligned in the UV space, thus leading to dense correspondence between the output 3D mesh and the input image.

Although the SMPL UV map [23] is widely used in the literature [37, 1, 7], it loses the neighboring relationships between different body parts as shown in Figure 3 (a), which is crucial for network learning as stated in [21]. Therefore, we design a new UV map that is able to maintain more neighboring relationships on the original mesh surface as shown in Figure 3 (b).

The overall objective function of DecoMR is

$$\mathcal{L} = \mathcal{L}_{IUV} + \mathcal{L}_{Loc} + \lambda_{con}\mathcal{L}_{con}. \quad (1)$$

It has three loss functions of different purposes. The first loss denoted as $\mathcal{L}_{IUV}$ minimizes the distance between the predicted IUV image and the ground-truth IUV image. The second loss function denoted as $\mathcal{L}_{Loc}$ minimizes the dissimilarity between the regressed human mesh (*e.g.* location map) and the ground-truth human mesh. In order to encourage the output mesh to be aligned with the input image, we add an extra loss function, denoted as $\mathcal{L}_{con}$, which is a consistent loss to increase the consistency between the regressed location map and the ground-truth IUV image. The λ_{con} in Equation 1 is a constant coefficient to balance the

consistent loss $\mathcal{L}_{con}$. We first define the new UV map below and then introduce different loss functions in details.

3.1. The Continuous UV map

First we define a new continuous UV map that preserves more neighboring relationships of the original mesh than the ordinary UV map of SMPL. As shown in Figure 3 (a), multiple mesh surface parts are placed separately on the SMPL default UV map, which loses the neighboring relationships of the original mesh surface. Instead of utilizing SMPL UV map as [1, 7, 37], we design a new continuous UV map. We first carefully split the template mesh into an open mesh, while keeping the entire mesh surface as a whole. Then we utilize an algorithm of area-preserving 3D mesh planar parameterization [14, 16], to minimize the area distortion between the UV map and the original mesh surface, in order to obtain an initial UV map. To maintain symmetry for every pair of symmetric vertices on the UV map, we further refine the initial UV map by first aligning the fitted symmetric axis with v axis and then averaging the UV coordinates with the symmetric vertex flipped by v axis.

Comparisons. Here we quantitatively show that our continuous UV map outperforms the SMPL UV map in terms of preserving connection relationships between vertices on the mesh. To do so, we compute the distance matrix, where each element is the distance between every vertex pair. We also compute the distance matrix on the UV map. Figure 4 shows such distance matrices. This distance matrix can be computed by using different types of data. For the mesh surface, the distance between two vertices is defined as the length of the minimal path between them on the graph built from the mesh. For the UV map, the distance between two vertices is directly calculated by the the distance between their UV coordinates.

Now we quantitatively evaluate the similarity between the distance matrices of UV map and original mesh in two aspects as shown in Table 1. In the first aspect, we calculate the 2D correlation coefficient denoted as S_1. We have

$$S_1 = \frac{\sum_m \sum_n \left(A_{mn} - \bar{A}\right)\left(B_{mn} - \bar{B}\right)}{\sqrt{\left(\sum_m \sum_n \left(A_{mn} - \bar{A}\right)^2\right)\left(\sum_m \sum_n \left(B_{mn} - \bar{B}\right)^2\right)}}, \tag{2}$$

where A and B are the distance matrices of original mesh and UV map, respectively. $\bar{A}$ and $\bar{B}$ are the mean value of A and B respectively. m and n are the indices of mesh vertices.

In the second aspect, we calculate the normalized cosine similarity between the distance matrices of UV map and original mesh, denoted as S_2. From Table 1, we see that our continuous UV map outperforms SMPL UV map by large margins on both metric values, showing that our UV map preserves more neighboring relationships than the SMPL UV map.

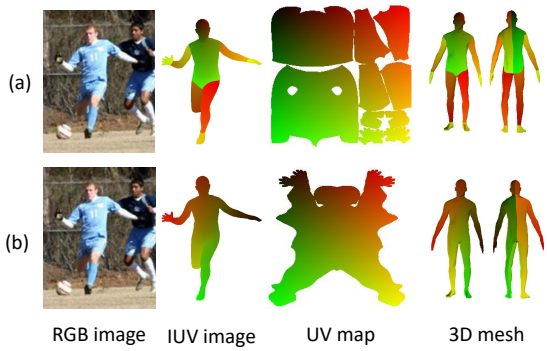


Figure 3. Comparisons of UV maps. Row (a) shows SMPL default UV map and row (b) shows our continuous UV map.

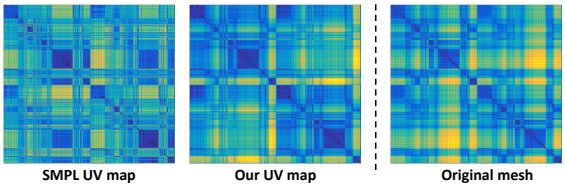


Figure 4. Comparisons of distance matrices between vertices calculated on SMPL UV map , the proposed UV map, and the original mesh surface. Compared to SMPL UV map, the distance matrix of the proposed UV map is more similar to that of the original mesh.

UV map	2D correlation (S_1)	cosine similarity (S_2)
SMPL [23]	0.2132	0.8306
Ours	0.7758	0.9458

Table 1. Comparisons of the similarity between the vertices' distance matrices of the original mesh surface and different types of UV maps. S_1 is the 2D correlation coefficient and S_2 is the normalized cosine similarity. We see that the proposed UV map outperforms SMPL default UV map on both metrics.

Pixel-to-Mesh Correspondence. With the proposed UV map, every point on the mesh surface can be expressed by its coordinates on the UV map (*i.e.* UV coordinates). Therefore, we can predict the pixel-to-surface correspondence by estimating the UV coordinates for each pixel belonging to human body, leading to an IUV image as shown in Figure 3. More importantly, we can also represent a 3D mesh with a location map in the UV space, where the pixel values are 3D coordinates of the corresponding points on the mesh surface. Thus it is easy to reconstruct 3D mesh from a location map with the following formula,

$$V_i = X(u_i, v_i), \tag{3}$$

where V_i denotes 3D coordinates of vertex, X is the location map, u_i and v_i are UV coordinates of the vertex.

3.2. Dense Correspondence Network (CNet)

CNet establishes the dense correspondence between pixels of the input image and areas of 3D mesh surface. As

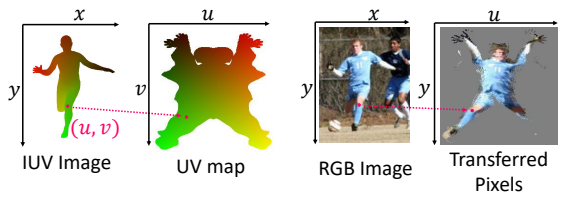


Figure 5. Illustration of the UV transferring of raw image pixels. Elements in the image space can be transferred to the UV space with the guidance of IUV image.

illustrated in Figure 2, CNet has an encoder-decoder architecture, where the encoder employs ResNet50 [9] as backbone, and the decoder consists of several upsampling and convolutional layers with skip connection with encoder. In particular, the encoder encodes the image as a local feature map and a global feature vector, as well as regresses the camera parameters, which are used to project the 3D mesh into the image plane. The decoder first generates a mask of the human body, which distinguishes fore pixels (*i.e.* human body) from those at the back. Then, the decoder outputs the exact UV coordinates for the fore pixels, constituting an IUV image as shown in Figure 3. With the predicted IUV image, the corresponding point on the mesh surface for every image pixel can be determined. The loss function for the CNet contains two terms,

$$\mathcal{L}_{IUV} = \lambda_c \mathcal{L}_c + \lambda_r \mathcal{L}_r, \tag{4}$$

where $\mathcal{L}_c$ is a dense binary cross-entropy loss for classifying each pixel as 'fore' or 'back', $\mathcal{L}_r$ is an l_1 dense regression loss for predicting the exact UV coordinates, and λ_c and λ_r are two constant coefficients.

3.3. Vertex coordinates regression

The location net (LNet) aims to regress 3D coordinates of mesh vertices by outputting a location map, from which the 3D mesh can be reconstructed easily. As shown in Figure 2, the LNet first transfers image features from the image space to the UV space with the guidance of predicted IUV image:

$$\mathcal{F}_{UV}(u, v) = \mathcal{F}_{im}(x, y), \tag{5}$$

where (x, y) are the coordinates in image space of the pixels classified as fore, and (u, v) are the predicted coordinates in UV space of these pixels. $\mathcal{F}_{im}$ is the feature map in image space and $\mathcal{F}_{UV}$ is the transferred feature map in UV space.

The feature map $\mathcal{F}_{UV}$ is well aligned with the output location map. So the LNet can predict location map utilizing corresponding local image features. In this way, the dense correspondence between image pixels and mesh surface areas is established explicitly. An example of raw image pixels transferred to UV space is shown in Figure 5. Note that our framework transfers features instead of pixel values.

The LNet is a light CNN with skip connections taking the transferred local image features, expanded global image feature and a reference location map as input. Intuitively, we apply an weighted l_1 loss between the predicted location map X and ground-truth location map $\hat{X}$, *i.e.*,

$$\mathcal{L}_{map} = \sum_u \sum_v W(u, v) \cdot \left\| X(u, v) - \hat{X}(u, v) \right\|_1. \tag{6}$$

W is a weight map used to balance the contribution of different mesh areas, where areas away from torso are assigned higher weights.

We also reconstruct a 3D human mesh from the predicted location map and get 3D joints from human mesh employing joint regressor as previous works [18, 21, 20]. Then we add supervision on the 3D coordinates and projected 2D coordinates in the image space of the joints, *i.e.*,

$$\mathcal{L}_J^{3D} = \sum_i^k \left\| Z_i - \hat{Z}_i \right\|_1, \tag{7}$$

$$\mathcal{L}_J^{2D} = \sum_i^k \| v_i (z_i - \hat{z}_i) \|_2^2, \tag{8}$$

where Z_i and z_i are the regressed 3D and 2D coordinates of joints, while $\hat{Z}_i$ and $\hat{z}_i$ refer to the coordinates of the ground-truth joints, and v_i denotes the visibility of joints.

Finally, the full loss for LNet is

$$\mathcal{L}_{loc} = \mathcal{L}_{map} + \mathcal{L}_J^{3D} + \mathcal{L}_J^{2D}. \tag{9}$$

Consistent Loss: Besides the above widely used supervision, we add an extra supervision between regressed location map and ground-truth IUV image to improve the alignment between 3D mesh and image.

As shown in Figure 6, with an IUV image, we can also transfer location map from the UV space back to the image space and get 3D coordinates for every foreground pixel. The 3D coordinates are then projected to image plane to get 2D coordinates, which should be consistent with the coordinates of the pixels in the image space. Then the consistent

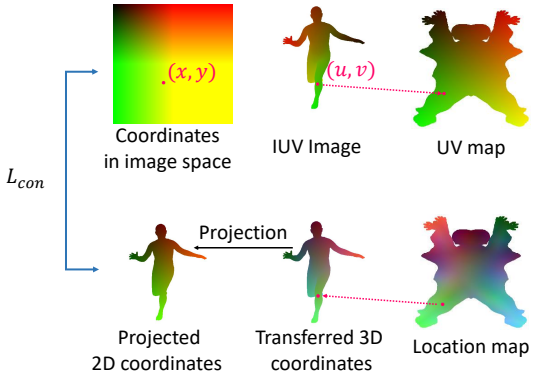


Figure 6. Illustration of our consistent loss between the location map and the IUV image. 3D coordinates in the location map are transferred back to the image space using IUV image, and then projected to the image plane. The projected 2D coordinates are supervised by the coordinates of image pixels in the image space.

loss is constructed as follows:

$$\mathcal{L}_{con} = \sum_{(x,y)} \|(x,y) - \pi(X(u,v),c))\|_2^2, \qquad (10)$$

where X is the predicted location map, $\pi(X,c)$ denotes the projection function with predicted camera parameters c, and x, y, u, v are the same as that in Equation 5. This consistent loss is similar to the loss item $\mathcal{L}_{dense}$ in recent work of Rong *et al*. [29]. However, in our framework there is no need to calculate the corresponding point on mesh surface as in [29], because the correspondence between mesh surface and image pixel is already established.

3.4. Implementation details

We set λ_c, λ_r and λ_{cons} to 0.2, 1 and 1 respectively and optimize the framework with an Adam optimizer [19], with batch size 128 and learning rate 2.5e-4. The training data is augmented with randomly scaling, rotation, flipping and RGB channel noise. We first train the CNet for 5 epochs and then train the full framework end-to-end for 30 epochs.

4. Experiments

4.1. Datasets

In the experiment, we train our model on the Human3.6M [13], UP-3D [22] and SURREAL [34] dataset, while we provide evaluations on the test set of Human3.6M, SURREAL and LSP dataset [17].

Human3.6M: Human3.6M [13] is a large scale indoor dataset for 3D human pose estimation, including multiple subjects performing typical actions like walking, sitting and eating. Following the common setting [18], we use subjects S1, S5, S6, S7 and S8 as training data and use subjects S9 and S11 for evaluation. For evaluation, results are reported using two widely used metrics (MPJPE and MPJPE-PA) under two popular protocols: P1 and P2, as defined in [18],

UP-3D: UP-3D [22] is an outdoor 3D human pose estimation dataset. It provides 3D human body ground truth by fitting SMPL model on images from 2D human pose benchmarks. We utilize the images of training and validation set for training.

SURREAL: SURREAL dataset [34] is a large dataset providing synthetic images with ground-truth SMPL model parameters. We use the standard split setting [34] but remove all images with incomplete human body and evaluate on the same sampled test set as BodyNet [33].

LSP: LSP [17] dataset is a 2D human pose estimation benchmark. In our work, we evaluate the segmentation accuracy of each model on the segmentation annotation [22].

4.2. Comparison with the state-of-the-art

In this section, we present comparison of our method with other state-of-the-art mesh-based methods.

Methods	MPJPE-PA
Lassner *etc*. [22]	93.9
SMPLify [4]	82.3
Pavlakos *etc*. [26]	75.9
HMR[18]	56.8
NBF[25]	59.9
CMR[21]	50.1
DenseRaC[36]	48.0
SPIN[20]	41.1
Ours	**39.3**

Table 2. Comparison with the state-of-the-art mesh-based 3D human estimation methods on Human3.6M test set. The numbers are joint errors in mm with Procrustes alignment under P2, and lower is better. Our approach achieves the state-of-the-art performance.

Methods	Surface Error
SMPLify++ [22]	75.3
Tung*et al*. [32]	74.5
BodyNet[33]	73.6
Ours	**56.5**

Table 3. Comparison with the state-of-the-art methods on SURREAL dataset. The numbers are the mean vertex errors in mm, and lower is better. Our methods outperform baselines with a large margin.

	FB Seg.		Part Seg	
	acc.	f1	acc.	f1
SMPLify *oracle* [4]	**92.17**	**0.88**	88.82	0.67
SMPLify [4]	91.89	**0.88**	87.71	0.67
SMPLify on [26]	92.17	**0.88**	88.24	0.64
HMR [18]	91.67	0.87	87.12	0.60
CMR [21]	91.46	0.87	88.69	0.66
SPIN [20]	91.83	0.87	89.41	0.68
Ours	92.10	**0.88**	**89.45**	**0.69**

Table 4. Comparison with the state-of-the-art methods on LSP test set. The numbers are accuracy and f1 scores, and higher is better. SMPLify [4] is optimization based, while HMR [18], CMR [21], SPIN [20] and our method are regression based. Our framework achieves the state-of-the-art result among regression based methods and is competitive with optimization based methods.

Table 2 shows the results on Human3.6M test set. We train our model following the setting of CMR [21] and utilize Human3.6M and UP-3D as the training set. Our method achieves the state-of-the-art performance among the mesh-based methods. It's worth notice that SPIN [20] and our method focus on different aspect and are compatible. SPIN [31] focus on the training using data with scarce 3D ground truth and the network is trained with extra data from 2D human pose benchmarks. While we focus on the dense correspondence between mesh and image, and do not include data from 2D human pose benchmarks.

Similarly, we show the results on SURREAL dataset in

UV map	$\mathcal{F}_G$	$\mathcal{F}_L$	raw pixel	MPJPE P1	MPJPE P2	MPJPE-PA P1	MPJPE-PA P2
SMPL	✓			72.1	68.9	51.9	49.1
		✓		71.9	69.6	47.4	44.8
	✓	✓		65.0	61.7	45.1	42.6
	✓		✓	65.0	63.2	46.5	44.7
Ours	✓			69.5	67.7	49.4	47.1
		✓		69.8	68.4	44.6	42.3
	✓	✓		**62.7**	**60.6**	**42.2**	**39.3**
	✓		✓	63.2	61.0	45.5	42.6

Table 5. Comparison on Human3.6M test set with different UV map and input of location net. The numbers are 3D joint errors in mm. $\mathcal{F}_G$ and $\mathcal{F}_L$ refer to global feature vector and local feature map, respectively. With both UV maps, the framework use local feature outperforms the baseline using global feature with a large margin. Combining global feature and local feature further improves the performance. However, transferring raw image pixels brings a gain much smaller. With the same input, the frameworks using our UV map outperform these using SMPL default UV map.

Table 3. Our model is trained only with training data of SURREAL dataset and outperforms the previous methods by a large margin. The human shape in SURREAL dataset is of great variety, and this verifies the human shape reconstruction capability of our method.

We also investigate human shape estimation accuracy by evaluating the foreground-background and part-segmentation performance on the LSP test set. During the evaluation, we use the projection of the 3D mesh as segmentation result. The predicted IUV image is not used in evaluation for fair comparison. The results are shown in Table 4. Our regression based method outperforms the state-of-the-art regression based methods and is competitive with the optimization based methods, which tend to outperform the regression based methods on this metric but are with much lower inference speed.

4.3. Ablative studies

In this section, we provide the ablation studies of the proposed method. We train all networks with training data from Human3.6M and UP-3D dataset, and evaluate the models on Human3.6M test set.

Dense correspondence: We first investigate the effectiveness of the dense correspondence between 3D mesh and image features. We train networks that only use global feature or transferred local feature as the input of LNet. The comparison is shown in Table 5. With both UV maps, the framework utilizing transferred local feature outperforms the baseline using global feature with a large margin, which proves the effectiveness of the established dense correspondence. Combining global feature with local feature further improves the performance.

We also train frameworks that transfer raw image pixels rather than image features and observe much less improvement than transferring local features. We attribute this phenomenon to the lack of human pose information in transferred raw pixels. For images with the same person in different poses, the pixels of a certain body part will be transferred to the same position in the UV space, which generates similar inputs for the LNet. So the LNet can only use transferred pixels to refine the estimation of human shape, and predict human pose only based on global feature.

On the contrary, the CNet is able to embed human pose information into image features. Then the LNet can resort to transferred features to refine both human shape and pose estimation.

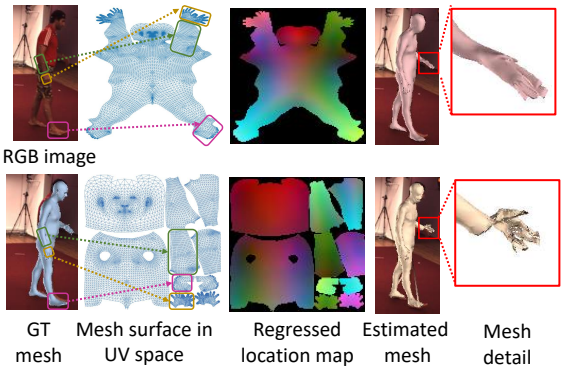


Figure 7. An example of mesh reconstructed using our new UV map (top) and SMPL default UV map (bottom). SMPL default UV map may cause discontinuity between different parts as well as erroneous estimation of some vertices near part edges. While our new UV map mitigates these problems.

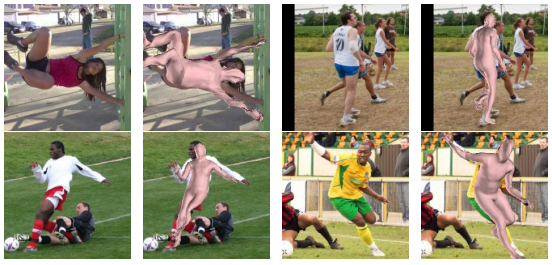
UV map: For the second ablative study, we investigate the influence of different UV maps. We compare the performance of frameworks using SMPL default UV map [23], and our continuous UV map.

As shown in Table 5, with the same input of LNet, the frameworks using our continuous UV map outperforms these frameworks using SMPL default UV map with a large margin. We attribute the gain to the continuity of the new UV map. As shown in Figure 7, some neighboring parts on mesh surface are distant on SMPL default UV map, such as arms and hands. This may lead to discontinuity of these parts on the final 3D mesh. Additionally, some faraway surface parts are very close on the UV plane, such as hands and foots, which might cause erroneous estimation of vertices on edges of these parts. These phenomenons are both shown in Figure 7. On the contrary, our UV map preserves more neighboring relations of the original mesh surface, so these problems are mitigated.

4.4. Qualitative result

Some qualitative results are presented in Figure 8, and Figure 9 includes some failure cases. Typical failure cases can be attributed to challenging poses, viewpoints rare seen

Figure 8. Qualitative results of our approach. Rows 1-3: LSP [17]. Rows 4-5: Human3.6M [13].



(a) Image (b) Result (c) Image (d) Result

Figure 9. Examples of erroneous reconstruction of our methods. Typical failures can be attributed to challenging poses, viewpoints rare seen in training set, severe self-osculation, as well as confusion caused by interaction among multiple people.

in training set, severe self-osculation, as well as confusion caused by interaction among multiple people.

5. Conclusion

This work aims to solve the problem of lacking dense correspondence between the image feature and output 3D mesh in mesh-based monocular 3D human body estimation. The correspondence is explicitly established by IUV image estimation and image feature transferring. Instead of reconstructing human mesh from global feature, our framework is able to make use of extra dense local features transferred to the UV space. To facilitate the learning of frame work, we propose a new UV map that maintains more neighboring relations of the original mesh surface. Our framework achieves state-of-the-art performance among 3D mesh-based methods on several public benchmarks. Future work can focus on extending the framework to the reconstruction of surface details beyond existing human models, such as cloth wrinkles and hair styles.

Acknowledgement

We thank reviewers for helpful discussions and comments. Wanli Ouyang is supported by the Australian Research Council Grant DP200103223.

References

[1] T. Alldieck, G. Pons-Moll, C. Theobalt, and M. Magnor. Tex2shape: Detailed full human body geometry from a single image. *arXiv preprint arXiv:1904.08645*, 2019.

[2] R. Alp Güler, N. Neverova, and I. Kokkinos. Densepose: Dense human pose estimation in the wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7297–7306, 2018.

[3] D. Anguelov, P. Srinivasan, D. Koller, S. Thrun, J. Rodgers, and J. Davis. Scape: shape completion and animation of people. In *ACM transactions on graphics (TOG)*, volume 24, pages 408–416. ACM, 2005.

[4] 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 *European Conference on Computer Vision*, pages 561–578. Springer, 2016.

[5] J. Dai, H. Qi, Y. Xiong, Y. Li, G. Zhang, H. Hu, and Y. Wei. Deformable convolutional networks. In *ICCV*, 2017.

[6] V. Gabeur, J.-S. Franco, X. Martin, C. Schmid, and G. Rogez. Moulding humans: Non-parametric 3d human shape estimation from single images. *arXiv preprint arXiv:1908.00439*, 2019.

[7] A. Grigorev, A. Sevastopolsky, A. Vakhitov, and V. Lempitsky. Coordinate-based texture inpainting for pose-guided human image generation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 12135–12144, 2019.

[8] P. Guan, A. Weiss, A. O. Balan, and M. J. Black. Estimating human shape and pose from a single image. In *2009 IEEE 12th International Conference on Computer Vision*, pages 1381–1388. IEEE, 2009.

[9] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[10] P. Huang, M. Tejera, J. Collomosse, and A. Hilton. Hybrid skeletal-surface motion graphs for character animation from 4d performance capture. *ACM Transactions on Graphics (ToG)*, 34(2):17, 2015.

[11] Y. Huang, F. Bogo, C. Lassner, A. Kanazawa, P. V. Gehler, J. Romero, I. Akhter, and M. J. Black. Towards accurate marker-less human shape and pose estimation over time. In *2017 International Conference on 3D Vision (3DV)*, pages 421–430. IEEE, 2017.

[12] M. E. Hussein, M. Torki, M. A. Gowayyed, and M. El-Saban. Human action recognition using a temporal hierarchy of covariance descriptors on 3d joint locations. In *Twenty-Third International Joint Conference on Artificial Intelligence*, 2013.

[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):1325–1339, 2013.

[14] A. Jacobson and D. Panozzo. libigl: prototyping geometry processing research in c++. In *SIGGRAPH Asia 2017 courses*, page 11. ACM, 2017.

[15] M. Jaderberg, K. Simonyan, A. Zisserman, et al. Spatial transformer networks. In *Advances in neural information processing systems*, pages 2017–2025, 2015.

[16] Z. Jiang, S. Schaefer, and D. Panozzo. Simplicial complex augmentation framework for bijective maps. *ACM Transactions on Graphics*, 36(6), 2017.

[17] S. Johnson and M. Everingham. Clustered pose and nonlinear appearance models for human pose estimation. In *Proceedings of the British Machine Vision Conference*, 2010. doi:10.5244/C.24.12.

[18] A. Kanazawa, M. J. Black, D. W. Jacobs, and J. Malik. End-to-end recovery of human shape and pose. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7122–7131, 2018.

[19] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[20] N. Kolotouros, G. Pavlakos, M. J. Black, and K. Daniilidis. Learning to reconstruct 3d human pose and shape via model-fitting in the loop. *arXiv preprint arXiv:1909.12828*, 2019.

[21] N. Kolotouros, G. Pavlakos, and K. Daniilidis. Convolutional mesh regression for single-image human shape reconstruction. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4501–4510, 2019.

[22] 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 *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6050–6059, 2017.

[23] M. Loper, N. Mahmood, J. Romero, G. Pons-Moll, and M. J. Black. Smpl: A skinned multi-person linear model. *ACM transactions on graphics (TOG)*, 34(6):248, 2015.

[24] A. Newell, K. Yang, and J. Deng. Stacked hourglass networks for human pose estimation. In *European conference on computer vision*, pages 483–499. Springer, 2016.

[25] M. Omran, C. Lassner, G. Pons-Moll, P. Gehler, and B. Schiele. Neural body fitting: Unifying deep learning and model based human pose and shape estimation. In *2018 International Conference on 3D Vision (3DV)*, pages 484–494. IEEE, 2018.

[26] G. Pavlakos, L. Zhu, X. Zhou, and K. Daniilidis. Learning to estimate 3d human pose and shape from a single color image. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 459–468, 2018.

[27] L. Pishchulin, E. Insafutdinov, S. Tang, B. Andres, M. Andriluka, P. V. Gehler, and B. Schiele. Deepcut: Joint subset partition and labeling for multi person pose estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4929–4937, 2016.

[28] A. Pumarola, J. Sanchez-Riera, G. Choi, A. Sanfeliu, and F. Moreno-Noguer. 3dpeople: Modeling the geometry of dressed humans. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2242–2251, 2019.

[29] Y. Rong, Z. Liu, C. Li, K. Cao, and C. C. Loy. Delving deep into hybrid annotations for 3d human recovery in the wild. In *The IEEE International Conference on Computer Vision (ICCV)*, October 2019.

[30] S. Saito, Z. Huang, R. Natsume, S. Morishima, A. Kanazawa, and H. Li. Pifu: Pixel-aligned implicit function for high-resolution clothed human digitization. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2304–2314, 2019.

[31] B. Tekin, P. Márquez-Neila, M. Salzmann, and P. Fua. Learning to fuse 2d and 3d image cues for monocular body pose estimation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3941–3950, 2017.

[32] H.-Y. Tung, H.-W. Tung, E. Yumer, and K. Fragkiadaki. Self-supervised learning of motion capture. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 5236–5246. Curran Associates, Inc., 2017.

[33] G. Varol, D. Ceylan, B. Russell, J. Yang, E. Yumer, I. Laptev, and C. Schmid. Bodynet: Volumetric inference of 3d human body shapes. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 20–36, 2018.

[34] G. Varol, J. Romero, X. Martin, N. Mahmood, M. J. Black, I. Laptev, and C. Schmid. Learning from synthetic humans. In *CVPR*, 2017.

[35] L. Xia, C.-C. Chen, and J. K. Aggarwal. View invariant human action recognition using histograms of 3d joints. In *2012 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, pages 20–27. IEEE, 2012.

[36] Y. Xu, S.-C. Zhu, and T. Tung. Denserac: Joint 3d pose and shape estimation by dense render-and-compare. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 7760–7770, 2019.

[37] P. Yao, Z. Fang, F. Wu, Y. Feng, and J. Li. Densebody: Directly regressing dense 3d human pose and shape from a single color image. *arXiv preprint arXiv:1903.10153*, 2019.

[38] A. Zanfir, E. Marinoiu, and C. Sminchisescu. Monocular 3d pose and shape estimation of multiple people in natural scenes-the importance of multiple scene constraints. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2148–2157, 2018.

[39] H. Zhu, X. Zuo, S. Wang, X. Cao, and R. Yang. Detailed human shape estimation from a single image by hierarchical mesh deformation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4491–4500, 2019.