Learning Attention Map for 3D Human Recovery from a Single RGB Image

Peng Xu
2191002048@cnu.edu.cn
Na Jiang†
jiangna@cnu.edu.cn
Jun Li
junmuzi@gmail.com
Zhiping Shi
shizp@cnu.edu.cn

Capital Normal University
Beijing, China
†Corresponding author

Abstract

3D human recovery from a single RGB image is a promising topic in computer vision, virtual reality, and image processing. It focuses on estimating 3D pose and shape of human from a 2D image. Due to the lack of depth and local information, the task remains challenging. Targeting to solve these problems, this work proposes a LAMNet with three branches that learning attention map from depth and parsing features for 3D human recovery. The first branch explicitly leverages the depth and pose cues to learn an adjusted depth map, which alleviates the recovery error between 3D space and 2D plane. The second branch explicitly leverages human parsing to provide local information, which supplements the shape or edge details of 3D recovery. The last branch is the main branch which is responsible for learning 2D global features to estimate 3D pose and shape of human. Inspired by attention mechanism, an attention aware fusion is designed to integrating three branches. It can achieve an attention map containing depth and local cues, which effectively improves the precision of 3D recovery, especially in details and different perspectives. Extensive experimental results demonstrate that our proposed approach significantly outperforms most state-of-the-art methods on the popular Human3.6m, UP-3D, and 3DPW datasets.

1 Introduction

Recovering a 3D human body from monocular images is an appealing and challenging task [4, 7, 42]. Early traditional 3D human recovery works rely on manual annotations, which require a lot of time and cost. With the rapid development of deep learning [18], deep learning is widely used in the field of computer vision. 3D recovery algorithm based on deep learning has become the focus of research, therefore, and greatly fostered the development of 3D human recovery. Such methods can be divided into two categories. Some do not require any prior information and statistical models to estimate the geometric shape of the body, which are called a model-free method [4, 16, 40]. The others which utilize parametric body model to realize 3D human recovery are called model-based method [14, 15, 30].

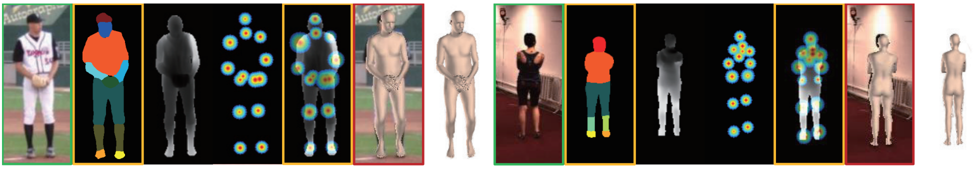

Figure 1: An illustration of all the inputs and outputs involved in LAMNet. For the two examples, these images shown from left to right are the raw image, human parsing, depth map, 2D keypoints, depth with adjusted keypoints, recovery result of LAMNet, and groundtruth, respectively. The green box denotes original input, the yellow box indicates auxiliary input generated from original input, and the red box marks the output of LAMNet.

In these existing methods, some representative works are worthy of analysis and discussion. For instance, HS-Nets [5] leveraging silhouette and NBF [22] using body semantic segmentation. Both of them considered introducing additional information as input to enhance feature learning for 3D recovery. Meanwhile, the weak perspective camera model is used into 3D recovery [15], which can improve the pose accuracy of recovery model by estimating camera parameters. However, they are easy to obtain unpredictable errors in the observation perspective different from the inputs. And when the image resolution is too low or the image content is incomplete, the camera position is difficult to be estimated, which will bring serious projection deviation to 3D recovery. It can be seen from these problems that 2D information has obvious limitations. Therefore, the cue that can provide 3D spatial information began to be considered. Shotton *et al.* [28] directly exploited depth maps to make up for the lack of three-dimensional information. They greatly promoted the development of 3D human recovery. However, due to the lack of effective complementary combination of 2D and 3D information, there are still errors in the shape and pose of 3D recovery model.

Targeting to solve the problem mentioned above, we propose a multi-branch network named LAMNet as shown in Figure 1. The raw image displayed in the first column is an indispensable input for recovering 3D shape and pose of human model. Based on it, human parsing is introduced as an auxiliary input. It can provide shape and local information to LAMNet, which helps to reduce the interference of fuzzy edges between the foreground and the background. If only considering parsing which can provide 2D information in network, however, the reconstructed body model are prone to pose distortion or keypoint deviation. Therefore, a novel depth with adjusted keypoints, which can provide 3D spatial information, is designed as another auxiliary input for LAMNet. It combines the keypoint position and body depth in a complementary way, which guides the key learning in the region of interest. Facing different inputs, LAMNet adopts three branches with different blocks to learn features. To fully integrate the features from depth, parsing and the raw image, an attention aware fusion module is proposed. Extensive experimental results demonstrate that LAMNet achieves significantly improvement on multiple popular datasets [12, 17, 35]. As shown in Figure 1, our final outputs very close to groundtruth.

In summary, this paper's contributions are threefold:

1) Propose a multi-branch LAMNet with attention aware fusion module to estimate 3D pose and shape of human, which explicitly utilizes depth and parsing cues to improve 3D human recovery.

2) Design a depth map with adjusted keypoints as auxiliary input, which can provided

meaningful attention guidance by learning the effect of depth on 3D pose and shape. Its introduction effectively reduces the 3D recovery error from different perspectives.

3) The proposed LAMNet achieves the state-of-the-art performance on popular Human3.6M, UP-3D, and 3DPW datasets.

2 Related Work

Traditional 3D human recovery tasks rely on manual data annotation. For example, [9, 29] uses annotated human silhouette to make the SCAPE [31] closer to the ground truth in the generation process. Since the emergence of the human body frame SMPL [19] model, Bogo *et al*. [2] designed an optimization method SMPLify on the basis of SMPL, and first automatically fitted the SMPL model to the keypoints of the 2D human body. Now 3D human recovery methods are divided into two categories. One is to use a model similar to the SMPL to estimate the human pose and shape, which is called a model-based method. The other method does not use statistical models and other methods to directly estimate the geometric body shape from 2D images, which is called a model-free method.

Model-free method. The commonly used methods for model-free are voxels, 3D mesh fitting, and so on. For example, BodyNet [34] shared Voxel-CNN to estimate the volumetric representation of the human body, which represents the occupancy rate of the 3D human shape on the voxel grid. Other recent work has shared many effective methods for estimating the geometric shape of the human body. Densebody [39] divides the human body into 24 pieces, and uses the UV position map to store the human 3D coordinate information. Zeng *et al*. [40] designed a new UV location map based on the dense map, which contains more accurate body texture locations. Moon *et al*. [4] designed the Pose2Mesh convolutional network, in which PoseNet uses the [3, 20] network structure to convert the standardized 2D input pose to the 3D output pose. Input the PoseNet result into MeshNet to estimate the 3D mesh. HMD [44] divides the grid structure into a three-layer structure of joint handles, anchor handles, and vertex handles, and continuously refines the human body surface information from coarse to fine. The model-free method focuses on global information and pays attention to the human surface, which leads to high storage and computing costs. And it is easy to lose detailed or local information.

Model-based method. From the early SCAPE model to the popular SMPL model, the parametric body model based methods have become the essential research direction. The classic HMR [14] model is proposed to achieve end-to-end 3D human recovery. On this basis, SPIN [15] with keypoint inputs is proposed. It added a 2D pose constraint by introducing the weak perspective camera model. To further enhance the 2D constraints, Rhodin *et al*. [25] introduced body contour to estimate the human pose and shape. Some other works tried to exploit 2D keypoint heatmaps [23], silhouette [30]. These supplementary information effectively improve 3D human recovery by providing the 2D edges and location constraints. Only considering 2D constraints is insufficient for current 3D human recovery. In recent years, many works have tried to introduce UV texture maps [42], depth maps [8] to provide estimated 3D spatial information. For example, Wang *et al*. [36] estimated depth and silhouette from a raw image to provide high dimensional spatial details, and then used the MANO [26] model to recovery 3D hand. Compared with the model-free methods mentioned above, these model-based methods generally have higher efficiency and better accuracy. However, the combination of 2D and 3D information that can further improve 3D human recovery is still worth exploring, which is the focus of this paper.

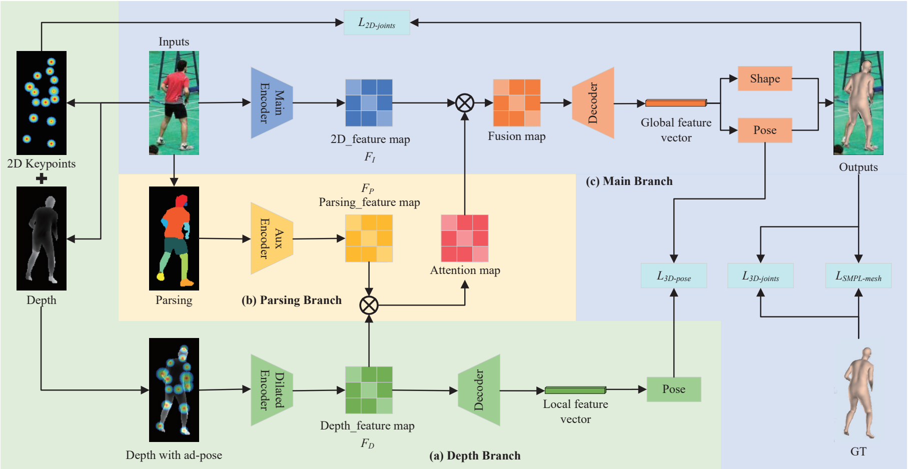

Figure 2: Outline of the proposed LAMNet. **(a)** is the depth branch, which is responsible for encodering depth feature map. Its input is the depth map with adjust keypoints. **(b)** illustrates the parsing branch with human parsing as input, which is responsible for learning the parsing feature map to provide shape and local cues. **(c)** demonstrates the main branch with the raw image as input. It combines the features from three branches through attention aware fusion module, and then recovery the shape and pose of 3D human model.

3 Methodology

In this section, a novel 3D human recovery algorithm named LAMNet is presented. As shown in Figure 2, the proposed LAMNet is a multi-branch framework. It consists of depth branch (in green), parsing branch (in yellow), and main branch (in blue). For any raw input, its depth, pose and parsing are first estimated to provide two auxiliary inputs. The depth branch adopts depth with adjusted keypoints as input, which learns the effect of keypoint area with different depth on 3D human recovery. The parsing branch directly learns the effect of different body part on 3D recovery from the human parsing. The two effects together form a valuable attention map, which guides LAMNet to realize the complementary combination of 2D and 3D cues. The main branch subtly blend the achieve attention map with global feature map by attention aware fusion, and then recover the shape and pose of 3D human model. More implementation details are explained below.

3.1 Depth Branch with Dilated Convolution

The task of this branch is to use the depth with adjusted pose (ad-pose) to learns the effect of keypoint area with different depth on 3D human recovery. To achieve this goal, we first use the work [13] to generate depth maps. But the depth map alone cannot accurately mark the coordinate depth of the keypoints. Therefore, an novel depth with ad-pose is designed as auxiliary input. It is synthesized by the keypoint heatmap and the depth map. The abstract illustration is shown in Figure 3. During synthesis, the depth map is responsible for adjusting the radius r_i of each keypoint confidence region. The calculation refer to Eq. 1. The keypoint heatmap is responsible for providing the normalized heat value C_j in keypoint confidence region according to Eq. 2.

$$r_i = d_r \sqrt{\exp\left(1 - \frac{H_i - H_{MIN}}{H_{MAX} - H_{MIN}}\right)} \tag{1}$$

where H_i represents the gray value of the i-th keypoints, and H_{MIN} and H_{MAX} represent the minimum and maximum gray values, respectively. r_i represents the radius of the i-th keypoints region, and d_r is initialized to 10. According to the calculated radius and the estimated keypoint positions, 14 confidence regions can be determined. In these regions of the adjusted map, the original depth and normalized heat value will be superimposed according to $min\{H_i{+}C_j, 255\}$.

$$C_j = \begin{cases} 0 & (x_j, y_j) \ \textit{not in region} \\ N\left(\frac{M_j - M_{MIN}}{M_{MAX} - M_{MIN}}\right) * 255 & (x_j, y_j) \ \textit{in region} \end{cases} \tag{2}$$

In Eq.2, M_j represents the confidence of the heatmap at the j-th pixel position. M_{MAX} and M_{MIN} represent the maximum and minimum heatmap confidence in all heatmaps, respectively. (x_j, y_j) represents the coordinate of the j-th pixel position in heatmap. If (x_j, y_j) not in any keypoint confidence region, the C_j is directly set to 0. If (x_j, y_j) in one region, the C_j is normalized $N(\cdot)$ and multiplied by 255 for unified expression. Through the above calculations, the depth with ad-pose can be obtained and highlight the relative spatial relationship between keypoints with different depths.

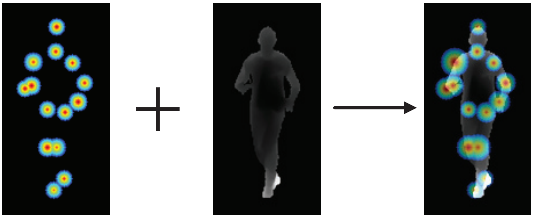

Figure 3: The depth with ad-pose is synthesized by the keypoint heatmap and the depth map.

The achieved depth with ad-pose contains sparse features. It is easy to lose the neighboring position relationship in the normal convolutions. To alleviate this problem, we use an encoder with dilated convolution. The purpose is to increase the receptive field and retain adjacent position information. Meanwhile, the dilated convolution can obtain multi-scale information. In addition, we add a self-attention block [6] at the end of down-sampling under this branch, which can enhance the feature perception for the keypoint confidence region. That is, for any adjusted depth map, a conventional feature map A will be extracted from down-sampling module (D-Sampling). And feed A into the three convolutional layers in the attention module to generate three feature maps X, Y and Z. We perform matrix multiplication of X and Y according to Eq. 3, and apply the softmax layer to calculate the feature map s. Then the matrix multiplication of Z and s is performed again and superimposed on the feature map A to obtain the depth feature F_D displayed in Figure 2. This calculation is defined in Eq. 4.

$$s_{ji} = softmax(X_i \cdot Y_j) \tag{3}$$

$$F_{D\_j} = \alpha \sum\nolimits_{i=1}^{N} (s_{ji} Z_i) + A_j \tag{4}$$

where s_{ji} represents the impact of the i-th position on the j-th position. F_{D_j} denotes the j-th value in F_D. α is automatically changed during continuous learning, and is initially set to 0. According to Eq. 4, the F_D can be inferred and each position can represent the weighted sum of the features across all positions and original features.

3.2 Human Parsing Branch

This branch uses ResNet18 [11] as the backbone network, extracts 2D planar features, and obtains the parsing feature map F_P. The human parsing isolates the background and the body, provides certain 2D shape information. Meanwhile, it uses different colors to represent different body parts, which can provide further sufficient local details for 3D human recovery.

We use the CorrPM model [43] to generate the human parsing, but there is a problem of incomplete body parts in the parsing image, which directly reduces the feature extraction ability. In response to this problem, we analyze the relationship between the pixel dimensions of each part and propose a parsing prior constraint using parsing pixels statistics. We set the color type $c \geqslant 3$, count the number of pixels k_c of each color, and arrange them in descending order to ensure that the number of pixels of the third color $k_c \geqslant 400$. If the parsing of a image does not satisfy the parsing prior constraint, the image is considered to be unable to provide effective 2D shape and local part cues. It will be filtered out and thus can not participate in any training. Parsing that meets the parsing prior constraint will become an auxiliary input, which is encoded into F_P by ResNet18. F_P can reflect the effects of different body part on 3D human recovery, and can provide relatively accurate 2D shape and local details. This will effectively help improve 3D human recovery.

3.3 Main Branch with Attention Aware Fusion

The main branch is modified with reference to the single branch in Rong *et al.* [27] to estimate the pose and shape of the human body. It uses down-sampling resnet50 network for feature extraction to obtain 2D feature map F_I.

Then we designed an attention aware fusion module so that the main branch concentrates the features from different branches. The calculation of attention aware fusion is defined in Eq. 5. According to it, F_P from the parsing branch and F_D from the depth branch can be integrated in a complementary way. As shown in Figure 2, the parsing branch extracts the parsing feature map F_P, and the depth map branch passes through the self-attention module to obtain the depth feature map F_D. Multiply F_D and F_P, and then pass through the softmax layer to get attention map $F_{Weighted}$. The attention map contains detailed information about the local details of human joints, as well as the spatial depth information between different keypoints. Finally, multiply it with the 2D feature map F_I of the main branch to obtain the final fusion map fusion map F_{Fusion}.

$$F_{Fusion} = softmax(F_D \odot F_P) \odot F_I \tag{5}$$

where $\odot$ is element-wise product. F_{Fusion} refers to the final feature map of multi-branch network fusion.

The specific network structure is shown in Table 1. The main branch and the parsing branch use residual networks of different depths, and the depth map branch lists the main down-sampling and up-sampling structures.

Main Branch	Parsing Branch	Depth Branch	
Resnet50	Resnet18	D-Sampling	K3-S2-C64-D2
			(K3-S1-C64)×3
			(K3-S2-C128)×4
			K3-S2-C256-D2
		U-Sampling	K3-S2-C128-D2
			K3-S2-C64-D2

Table 1: Configurations of The Multi-branch Estimatior. ’K’ = kernel size, ’S’ = stride, ’C’ = channel, ’D’ = dilation rate.

3.4 Loss Function

The main branch generates a global feature vector from F_{Fusion}, and obtains 3D pose θ and shape β from this vector. In order to reduce the 3D pose error in the training process, we design 3D pose loss L_{3D_pose} between depth branch and main branch. It is defined in Eq. 7, which consists of one 3D rotation matrix (from Rodrigues formula) error and three 2D pose projection errors. Take *xoy* in Cartesian coordinate system as an example, we first need to calculate the projection $P(\theta)$ of 3D pose vector θ on 2D plane through Eq. 6. Similarly, the projection of the other two planes is also calculated in the same way. The 3D pose loss can then be obtained using Eq. 7.

$$P_{xoy}\left(\theta_i\right) = \theta_i - \frac{\theta_i \cdot \overrightarrow{u_{xoy}}}{\left\|\overrightarrow{u_{xoy}}\right\|^2}\overrightarrow{u_{xoy}} \tag{6}$$

$$\begin{aligned} L_{3D\_pose} = & \sum\nolimits_{i=1}^{O}\left(\left\|R\left(\theta_i\right) - R\left(\theta_{H\_i}\right)\right\|_2 + \left\|P_{xoy}\left(\theta_i\right) - P\left(\theta_{H\_i}\right)\right\|_2\right. \\ & \left. + \left\|P_{xoz}\left(\theta_i\right) - P\left(\theta_{H\_i}\right)\right\|_2 + \left\|P_{yoz}\left(\theta_i\right) - P\left(\theta_{H\_i}\right)\right\|_2\right) \end{aligned} \tag{7}$$

where θ_i in Eq. 6 represents the *i*-th joint vectors of the 3D pose. $\overrightarrow{u_{xoy}}$ is the normal vector of the plane. $R(\cdot)$ means rodrigues formula calculation method. In Eq. 7, O is the number of SMPL parameters. θ_{H_i} represents the human 3D pose estimated from the depth branch. The first term in the formula calculates the 3D rotation matrix loss, and the last three terms calculate the 2D pose projection errors. They can enhance the pose consistency between 3D recovery results and estimated pose, which are conducive to improving final human model.

The 3D human body is restored by SMPL model $E(\theta, \beta)$, the method is the same as [24]. In order to ensure that the generated 3D model more closely fits the human pose of the image I, the SMPL mesh loss constraint L_{SMPL_mesh} is added to the main branch.

$$L_{SMPL\_mesh} = \sum\nolimits_{i=1}^{O}\left(\left\|R\left(\theta_i\right) - R\left(\hat{\theta}_i\right)\right\|_2 + \left\|\beta_i - \hat{\beta}_i\right\|_2\right) \tag{8}$$

where $\hat{\theta}_i$ represents ground truth pose. $\hat{\beta}_i$ represents ground truth shape.

Overall, the full loss function of LAMNet is defined in Eq. 9.

$$L_{total} = \lambda_1\left(L_{3D\_joints} + L_{SMPL\_mesh}\right) + \lambda_2 L_{3D\_pose} + \lambda_3 L_{2D\_joints} \tag{9}$$

where L_{3D_joints} and L_{2D_joints} calculate the loss functions of 3D keypoints and 2D keypoints, respectively. Their mathematical definitions are the same as the existing work [41]. λ_1, λ_2 and λ_3 are set to 1, 1, 0.5, respectively.

4 Experiments

4.1 Implementation Details and Datasets

Due to the huge results of the multi-branch network and the large amount of training data each time, we set the image resize to 224*224. Set the batch size to decrease from 128 to 32, and increase the training epoch from 20 to about 100. In addition, set the initial learning rate 1e-5. The update strategy adopts the Adam with 0.5.

In the experiment, we train and test on multiple datasets. The main datasets used are Human3.6M [12], UP-3D [17], and 3DPW [35] datasets. Human3.6M is a large indoor human body dataset. The human body is unobstructed and contains rich actions. It is one of the commonly used datasets in 3D human recovery.UP-3D is a collection of existing 2D human pose datasets, including LSP, LSP-Extended, MPII HumanPose and FashionPose datasets. 3DPW is a type of outdoor dataset. The dataset contains 61 video sequences mainly shot under outdoor conditions. The evaluation indicators mainly use the average per-vertex error (PVE), the vertex rigidity is transformed into the vertex error in the 'T' pose (T-PVE), mean keypoints error (MPJPE) and MPJPE-PA. MPJPE-PA means that the output is rigidly transformed to ground truth alignment and then MPJPE is calculated.

4.2 Ablation Study

To evaluate the effectiveness of the key components proposed in our method, we conduct ablation experiments on 3DPW under various settings. The baseline comes from the advanced work [27]. For a convincing and clear comparative experiment, we remove its branches related to UV and dense correspondence. Only keep the single branch with the raw image as input. For this branch, we perform a retraining and achieve better MPJPE results than their published results. Therefore, the retraining baseline is used in the ablation study. We add improved parts in turn on the baseline to evaluate their effectiveness.

As shown in Table 2, the parsing, ad-pose, L_{SMPL_mesh} and L_{3D_pose} can improve the indicators of 3D human recovery. When all of them are added to the baseline to form our proposed method, the best result can be achieved. This significant promotion is largely due to four reasons: (1) the human parsing can provide more detailed 2D shape information for 3D human recovery; (2) the ad-pose can provide more accurate relative depth and spatial relationship of keypoints than the single depth or pose auxiliary input; (3) L_{SMPL_mesh} and L_{3D_pose} can enhance 3D spatial constraints and reduce the position deviation of 3D model in different perspectives; (4) the designed attention aware fusion can give full play to the role of these auxiliary inputs and loss functions in 3D human recovery.

Baseline	Parsing	Pose	Depth	Ad-pose	L_{SMPL_mesh}	L_{3D_pose}	PVE	MPJPE	T-PVE
√							154.7	134.4	68.3
√	√						122.8	105.4	30.2
√	√	√					116.7	100.3	30.1
√	√		√				116.0	99.5	30.6
√	√			√			111.5	95.7	29.4
√	√			√	√		106.9	91.1	28.3
√	√			√	√	√	**101.3**	**86.2**	**24.9**

Table 2: Ablation experiments on the 3DPW dataset.

4.3 Effectiveness of Our Approach

In this section, we will compare our method with other advanced methods. Use different evaluation indicators to compare in different datesets.

Method	3DPW	
	MPJPE-PA	MPJPE
Pose2Mesh [4]	58.3	88.9
I2L-MeshNet [21]	60.8	95.4
DSD+SATN [33]	69.5	-
RSC-Net [37]	58.9	96.3
ETC-Net [1]	72.2	-
Ours	**57.6**	**86.2**

Table 3: Quantitative comparison with state-of-the-art methods on the 3DPW dataset.

Method	Human3.6M	
	MPJPE-PA	MPJPE
Pose2Mesh [4]	46.2	64.9
HoloPose [10]	46.52	60.2
CMR [16]	-	74.2
DecoMr [40]	42.2	62.7
ETC-Net [1]	54.3	77.8
Ours	**41.2**	**59.3**

Table 4: Quantitative comparison with state-of-the-art methods on the Human3.6M dataset.

Method	UP-3D	
	Accuracy	F1 score
DenseRaC [38]	92.4	0.88
BSG-Net [32]	91.9	0.88
BodyNet [34]	92.8	0.84
HMR [14]	91.7	0.87
DecoMr [40]	92.1	0.88
Ours	**93.2**	**0.89**

Table 5: Quantitative comparison with state-of-the-art methods on the UP-3D dataset.

Since other methods use different indicators on different datasets, in order to be fair, we adjust the test indicators and compare them with different methods. MPJPE and MPJPE-PA are used for testing on Human3.6M and 3DPW datasets, while Accuracy and F1 score test indicators are used on UP-3D dataset.

Table 3 is tested on the Human3.6M dataset. Compared with Pose2Mesh [4], MPJPE-PA increased by 4.0%, and MPJPE increased by 5.6%. It is verified that the recovery effect of LAMNet is more accurate on the rich indoor human pose dataset. Table 4 verifies that the recovery ability on the 3DPW dataset is better than the previous method. Compared with I2L-MeshNet [21], MPJPE-PA increased by 3.2%, and MPJPE increased by 9.2%. We also investigate human shape estimation accuracy by evaluating the foreground-background performance on the UP-3D. It can be seen from Table 5 that compared with DecoMr [40], our Accuracy is improved by 1.1%, reaching the best accuracy. In general, the performance of our multi-branch network on multiple datasets is better than the current state-of-the-art methods.

4.4 Visualization Analysis

In addition to quantitative analysis, we perform qualitative visual analysis on different datasets. We visually compare with SPIN [15] and HMR [14] under the same datasets. As shown in

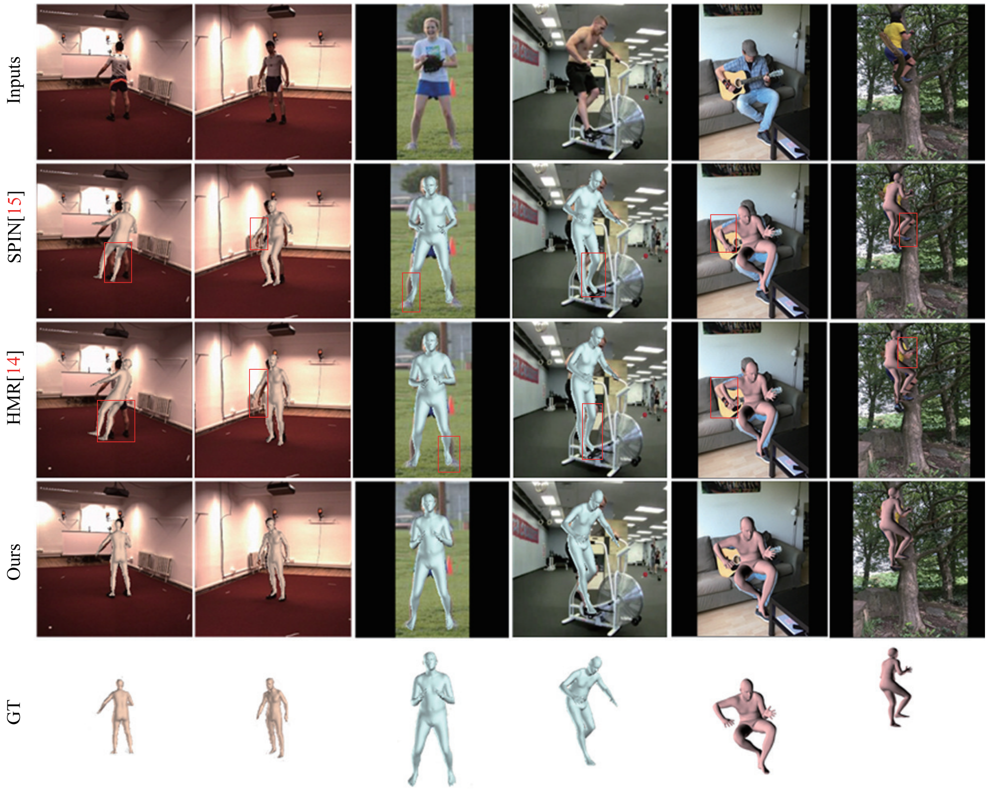

Figure 4: Qualitative comparison of recovery results on the Human36m, UP-3D and 3DPW datasets.

Figure 4, some inaccurate 3D recovery details from different methods are marked in the red box. Compared with SPIN and HMR, our 3D recovery effect is closer to ground truth. LAMNet processes some joints in special positions more finely, which benefits from the learning of relative depth and human parsing.

5 Conclusion

In this paper, we propose a multi-branch network named LAMNet to realize 3D human recovery from a single image. The LAMNet consist of depth branch, parsing branch, and main branch, The depth branch adopts the novel depth with adjusted pose to learn the impact of different keypoint regions on 3D human recovery. The parsing branch exploits the human parsing to achieve the shape and local details that are meaningful to recovery the 3D model. On this basis, the main branch enhances the complementary combination of 2D and 3D information by attention aware fusion, and thus improves 3D recovery accuracy by valuable attention map. As a result, LAMNet significantly outperforms the state-of-the-art methods on three popular datasets. In future work, we will explore more efficient 3D human recovery, and pay attention to its promising applications.

6 Acknowledgments

This work was supported by the National Natural Science Foundation of China under Grant No.62002247, the Open Project Program of State Key Laboratory of Virtual Reality Technology and Systems, Beihang University (No. VRLAB2020A03), and the general project numbered KM202110028009 of Beijing Municipal Education Commission.

References

[1] Anurag Arnab, Carl Doersch, and Andrew Zisserman. Exploiting temporal context for 3d human pose estimation in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3395–3404, 2019.

[2] 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 *European conference on computer vision*, pages 561–578. Springer, 2016.

[3] Ju Yong Chang, Gyeongsik Moon, and Kyoung Mu Lee. Poselifter: Absolute 3d human pose lifting network from a single noisy 2d human pose. *arXiv preprint arXiv:1910.12029*, 2019.

[4] Hongsuk Choi, Gyeongsik Moon, and Kyoung Mu Lee. Pose2mesh: Graph convolutional network for 3d human pose and mesh recovery from a 2d human pose. In *European Conference on Computer Vision*, pages 769–787. Springer, 2020.

[5] Endri Dibra, Himanshu Jain, Cengiz Öztireli, Remo Ziegler, and Markus Gross. Hs-nets: Estimating human body shape from silhouettes with convolutional neural networks. In *2016 fourth international conference on 3D vision (3DV)*, pages 108–117. IEEE, 2016.

[6] Jun Fu, Jing Liu, Haijie Tian, Yong Li, Yongjun Bao, Zhiwei Fang, and Hanqing Lu. Dual attention network for scene segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3146–3154, 2019.

[7] Valentin Gabeur, Jean-Sébastien Franco, Xavier Martin, Cordelia Schmid, and Gregory Rogez. Moulding humans: Non-parametric 3d human shape estimation from single images. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2232–2241, 2019.

[8] Liuhao Ge, Zhou Ren, Yuncheng Li, Zehao Xue, Yingying Wang, Jianfei Cai, and Junsong Yuan. 3d hand shape and pose estimation from a single rgb image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10833–10842, 2019.

[9] Peng Guan, Alexander Weiss, Alexandru O Balan, and Michael 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.

[10] Riza Alp Guler and Iasonas Kokkinos. Holopose: Holistic 3d human reconstruction in-the-wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10884–10894, 2019.

[11] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[12] Catalin Ionescu, Dragos Papava, Vlad Olaru, and Cristian 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.

[13] Yasamin Jafarian and Hyun Soo Park. Learning high fidelity depths of dressed humans by watching social media dance videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12753–12762, 2021.

[14] Angjoo Kanazawa, Michael J Black, David W Jacobs, and Jitendra 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.

[15] Nikos Kolotouros, Georgios Pavlakos, Michael J Black, and Kostas Daniilidis. Learning to reconstruct 3d human pose and shape via model-fitting in the loop. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2252–2261, 2019.

[16] Nikos Kolotouros, Georgios Pavlakos, and Kostas Daniilidis. Convolutional mesh regression for single-image human shape reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4501–4510, 2019.

[17] 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 *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6050–6059, 2017.

[18] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

[19] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J Black. Smpl: A skinned multi-person linear model. *ACM transactions on graphics (TOG)*, 34(6):1–16, 2015.

[20] Julieta Martinez, Rayat Hossain, Javier Romero, and James J Little. A simple yet effective baseline for 3d human pose estimation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2640–2649, 2017.

[21] Gyeongsik Moon and Kyoung Mu Lee. I2l-meshnet: Image-to-lixel prediction network for accurate 3d human pose and mesh estimation from a single rgb image. *arXiv preprint arXiv:2008.03713*, 2020.

[22] Mohamed Omran, Christoph Lassner, Gerard Pons-Moll, Peter Gehler, and Bernt 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.

[23] Georgios Pavlakos, Luyang Zhu, Xiaowei Zhou, and Kostas 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.

[24] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10975–10985, 2019.

[25] Helge Rhodin, Nadia Robertini, Dan Casas, Christian Richardt, Hans-Peter Seidel, and Christian Theobalt. General automatic human shape and motion capture using volumetric contour cues. In *European conference on computer vision*, pages 509–526. Springer, 2016.

[26] Javier Romero, Dimitrios Tzionas, and Michael J Black. Embodied hands: Modeling and capturing hands and bodies together. *ACM Transactions on Graphics (ToG)*, 36(6): 1–17, 2017.

[27] Yu Rong, Ziwei Liu, Cheng Li, Kaidi Cao, and Chen Change Loy. Delving deep into hybrid annotations for 3d human recovery in the wild. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5340–5348, 2019.

[28] Jamie Shotton, Ross Girshick, Andrew Fitzgibbon, Toby Sharp, Mat Cook, Mark Finocchio, Richard Moore, Pushmeet Kohli, Antonio Criminisi, Alex Kipman, et al. Efficient human pose estimation from single depth images. *IEEE transactions on pattern analysis and machine intelligence*, 35(12):2821–2840, 2012.

[29] Leonid Sigal, Alexandru Balan, and Michael Black. Combined discriminative and generative articulated pose and non-rigid shape estimation. *Advances in neural information processing systems*, 20:1337–1344, 2007.

[30] Brandon M Smith, Visesh Chari, Amit Agrawal, James M Rehg, and Ram Sever. Towards accurate 3d human body reconstruction from silhouettes. In *2019 International Conference on 3D Vision (3DV)*, pages 279–288. IEEE, 2019.

[31] Praveen Srinivasan, Dragomir Anguelov, Daphne Koller, Sebastian Thrun, Jim Rodgers, and James Davis. Scape: shape completion and animation of people. In *ACM SIGGRAPH 2005 Papers*, pages 408–416. 2005.

[32] Shuang Sun, Chen Li, Zhenhua Guo, and Yuwing Tai. Parametric human shape reconstruction via bidirectional silhouette guidance. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, 2019.

[33] Yu Sun, Yun Ye, Wu Liu, Wenpeng Gao, Yili Fu, and Tao Mei. Human mesh recovery from monocular images via a skeleton-disentangled representation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5349–5358, 2019.

[34] Gul Varol, Duygu Ceylan, Bryan Russell, Jimei Yang, Ersin Yumer, Ivan Laptev, and Cordelia Schmid. Bodynet: Volumetric inference of 3d human body shapes. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 20–36, 2018.

[35] Timo von Marcard, Roberto Henschel, Michael J Black, Bodo Rosenhahn, and Gerard Pons-Moll. Recovering accurate 3d human pose in the wild using imus and a moving camera. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 601–617, 2018.

[36] Jiayi Wang, Franziska Mueller, Florian Bernard, Suzanne Sorli, Oleksandr Sotnychenko, Neng Qian, Miguel A Otaduy, Dan Casas, and Christian Theobalt. Rgb2hands: real-time tracking of 3d hand interactions from monocular rgb video. *ACM Transactions on Graphics (TOG)*, 39(6):1–16, 2020.

[37] Xiangyu Xu, Hao Chen, Francesc Moreno-Noguer, László A Jeni, and Fernando De la Torre. 3d human shape and pose from a single low-resolution image with self-supervised learning. In *European Conference on Computer Vision*, pages 284–300. Springer, 2020.

[38] Yuanlu Xu, Song-Chun Zhu, and Tony Tung. Denserac: Joint 3d pose and shape estimation by dense render-and-compare. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7760–7770, 2019.

[39] Pengfei Yao, Zheng Fang, Fan Wu, Yao Feng, and Jiwei Li. Densebody: Directly regressing dense 3d human pose and shape from a single color image. *arXiv preprint arXiv:1903.10153*, 2019.

[40] Wang Zeng, Wanli Ouyang, Ping Luo, Wentao Liu, and Xiaogang Wang. 3d human mesh regression with dense correspondence. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7054–7063, 2020.

[41] Hongwen Zhang, Jie Cao, Guo Lu, Wanli Ouyang, and Zhenan Sun. Learning 3d human shape and pose from dense body parts. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020.

[42] Tianshu Zhang, Buzhen Huang, and Yangang Wang. Object-occluded human shape and pose estimation from a single color image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7376–7385, 2020.

[43] Ziwei Zhang, Chi Su, Liang Zheng, and Xiaodong Xie. Correlating edge, pose with parsing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8900–8909, 2020.

[44] Hao Zhu, Xinxin Zuo, Sen Wang, Xun Cao, and Ruigang Yang. Detailed human shape estimation from a single image by hierarchical mesh deformation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4491–4500, 2019.