Object-Occluded Human Shape and Pose Estimation from a Single Color Image

Tianshu Zhang* Buzhen Huang* Yangang Wang†
{tszhang,hbz,yangangwang}@seu.edu.cn
Southeast University, China

Abstract

Occlusions between human and objects, especially for the activities of human-object interactions, are very common in practical applications. However, most of the existing approaches for 3D human shape and pose estimation require human bodies are well captured without occlusions or with minor self-occlusions. In this paper, we focus on the problem of directly estimating the object-occluded human shape and pose from single color images. Our key idea is to utilize a partial UV map to represent an object-occluded human body, and the full 3D human shape estimation is ultimately converted as an image inpainting problem. We propose a novel two-branch network architecture to train an end-to-end regressor via the latent feature supervision, which also includes a novel saliency map subnet to extract the human information from object-occluded color images. To supervise the network training, we further build a novel dataset named as ***3DOH50K****. Several experiments are conducted to reveal the effectiveness of the proposed method. Experimental results demonstrate that the proposed method achieves the state-of-the-art comparing with previous methods. The dataset, codes are publicly available at* https://www.yangangwang.com.

1. Introduction

3D human shape and pose estimation from color images attracts lots of research interests in the area of computer vision. It may promote several promising virtual reality applications, such as body shape animation, shape retargeting, motion mimic and *etc*. Conventionally, the full 3D human body shape estimation has experienced from complex hardware (*e.g*., multi-view cameras, IMU sensors) to single devices (*e.g*., color camera, Kinect). In recent years, deep learning based techniques [17, 33, 51] have witnessed the rapid progress of recovering the full body human shape from single color images, though most of the existing approaches are addressed for the scenarios that human bodies are well captured without occlusions or with minor self-occlusions. However, occlusions between human and objects, especially for the activities of human-object interactions, are very common in practical applications. It is hard to achieve good performance for recovering the full 3D human body shape from object-occluded color images when the occlusions are not explicitly taken into account. In this paper, we mainly focus on the problem of estimating the object-occluded full 3D human shape and pose from single color images.

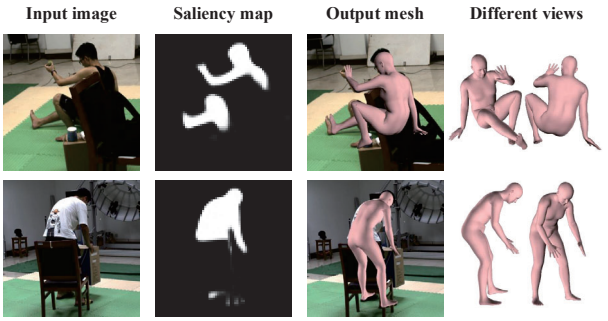

Figure 1. Our method performs well for the human shape and pose estimation from single object-occluded color images.

Historically, 3D human shape and pose estimation from single color images is flourished after the Skinned Multi-Person Linear model (SMPL) [27] has been presented. It goes through several stages, including SMPL parameters optimization via fitting to 2D visual cues [3, 21], directly regressing the SMPL parameters with Convolutional Neural Network (CNN) [33, 30], volumetric representation for 3D human shape [15, 43], and 2D UV map representation for the human body geometry surface [51, 54]. Although deep learning has become the mainstream of full 3D human body estimation due to its accuracy and runtime efficiency, it cannot be directly shifted to handle the object-occluded human body estimation without explicitly considering occlusions. There are two main challenges. The first one is that there is a lack of sufficient data for the network training. Existing datasets are not originally designed for the occluded human shape estimation. The other one is object-occlusions would introduce severe ambiguities into the network training, and

†Corresponding author.
*Contribute equally. The authorship was determined by a coin toss.

thus confuse the full 3D human body shape estimation.

To tackle the obstacles, we investigate different human body representations and take the up-to-date 2D UV map [6, 51] to describe the 3D human shape. Even so, it is still hard to directly regress the full 2D UV map via CNN due to the ambiguities caused by occlusions. Our key idea is to utilize a partial UV map to describe an object-occluded human body and **convert the full 3D human shape estimation as a UV map inpainting problem**, as shown in Fig.2. We propose a novel two-branch encoder-decoder network architecture, where the first branch is a UV map inpainting and the second branch keeps an input color image to be consistent with its partial UV map in their latent feature spaces. Both of the two branches share the same decoder, and they are trained separately with different datasets. Typically, the UV map inpainting branch, which could be regarded as a prior for body shapes, can be trained without color images. It is also worth noting that the pixels in color images that are not part of bodies may fool the color image encoder, we then introduce a saliency map estimation subnetwork to emphasize the importance of human pixels in color images as shown in Fig.2 (b). Although the proposed network focuses on the object-occluded human shape estimation, it does not affect the performance of non-occluded human shape estimation and also achieves the state-of-the-art, as shown in Sec.5.

Our network can be run efficiently both for training and inference. At the training stage, a two-step training strategy is adopted to optimize the network parameters. We first train the UV map inpainting branch, and then its parameters are fixed to supervise the training of the color image encoder (Fig.2 (c) and (d)). At the inference stage, a single color image is passed through the saliency map subnetwork, the color image encoder, and the decoder of UV map inpainting branch. However, to train the proposed network, we found that human-object occlusion datasets are far from sufficient. We first added virtual objects into existing datasets (*e.g.*, Human3.6M [13]) to synthesize occlusions. To further facilitate the network training, we build a new dataset named as **3DOH50K**. The new dataset contains more than 51600 images, where all images were captured from real scenes with 6 viewpoints, and we used the modified SMPLify-X[32] to fit the SMPL model. Finally, each instance has an accurate 2D pose, a 3D pose, SMPL parameters, and a binary mask. To the best of our knowledge, **3DOH50K** is the first real dataset for the problem of human-object occlusion. The proposed dataset could provide a new challenge benchmark for human reconstruction and pose estimation in occlusion scenarios.

The main contributions of this work are summarized as follows.

- We take a partial UV map representation for an object-occluded 3D human body, and describe the full 3D human shape and pose estimation as an image inpainting problem.
- We propose a novel two-branch network architecture to train an end-to-end human shape regressor for estimating the full 3D human shape and pose from single color images.
- We build a novel object-occluded human dataset, which is named as **3DOH50K**, to ease the network training. The dataset, codes are publicly available at https://www.yangangwang.com.

2. Related Work

Human pose and shape estimation. Traditional human pose and shape estimation methods mostly use the complex hardware to obtain the human body's cues, and estimate the full 3D human body pose and shape through iterative optimization[12, 45, 23] or deep learning[41, 2, 25]. Due to various limitations of hardware, they cannot be easily applied to real-world scenarios. In order to accurately estimate the pose and shape of the human body from single RGB camera, [17, 30, 33, 55] parameterized the mesh in terms of 3D joint angles and a low dimensional linear shape space. Unlike the previous methods[48, 3], they directly infer 3D mesh parameters from image features which avoided two stage training and also avoided throwing away lots of image information. In order to avoid complex nonlinear mapping of parameter prediction methods, Venkat *et al*. [44] proposed HumanMeshNet that regressed a template mesh's vertices. Some recent works[51, 1, 22] turned a hard 3D inference problem into an image-to-image translation which is amenable to CNNs by encoding appearance and geometry layout on a common SMPL UV-space.

Occlusion. Huang *et al*. [11] presented a method capable of recovering 3D human pose when a person is partially or heavily occluded in the scene from monocular images. However, the occlusions are limited to two rectangles. [36] presented a systematic study of various types of synthetic occlusions in 3D human pose estimation from a single RGB image. Since synthetic data can not fully depict the real occlusion, Girshick *et al*. [8] learned from real data and used grammar models with explicit occluding templates to reason about occluded people. To avoid specific design for occlusion patterns, [7] presented a method for modeling occlusion that was aimed at explicitly learning the appearance and statistics of occlusion patterns. [34] integrated depth information about occluded objects into 3D pose estimation framework. In the scope of face de-occlusion, [42] tried to address the problem of detailed face reconstruction from occluded images. [53] proposed a novel deep face de-occlusion framework, which can handle face images under challenging conditions. In [46], a very effective occluded face recognition algorithm, GD-HASLR, was proposed. It has strong robustness to the shape and the size of

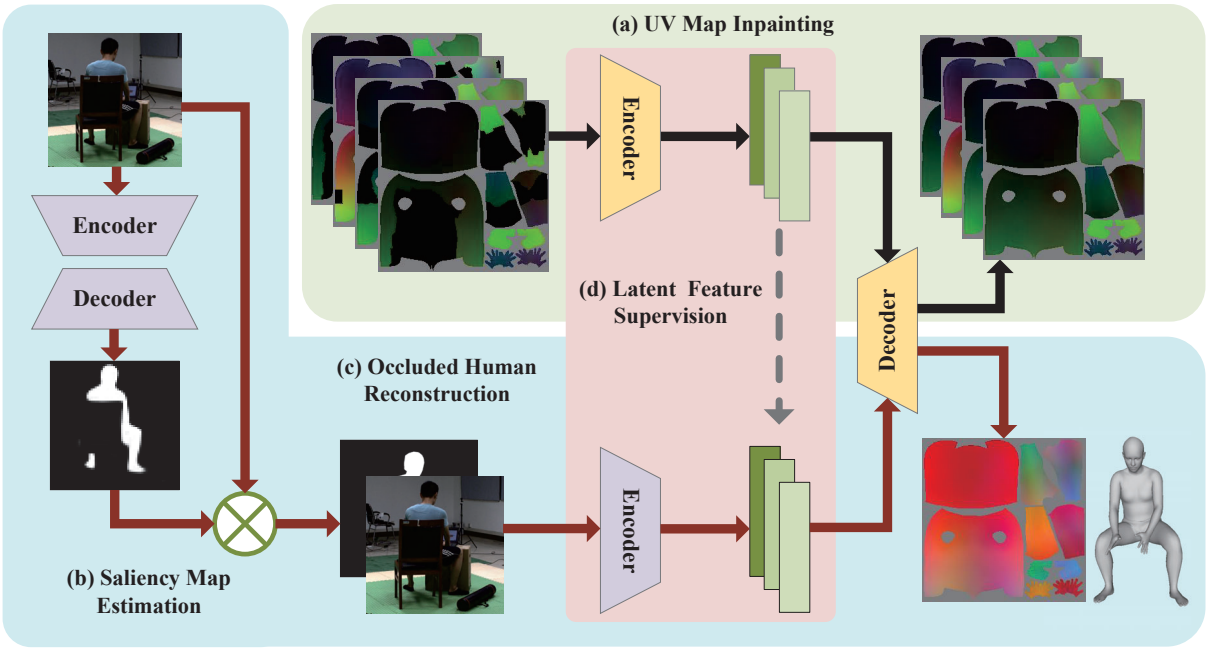

Figure 2. **Overview of the proposed two-branch network.** At the training stage, UV map inpainting branch (a) is trained at first. Then, the occluded color image is concatenated with its saliency map (b) and fed to color image encoder (c). The corresponding partial UV map is encoded by fixed inpainting network and used for supervising the color image encoder in latent space (d). At the inference stage, a single color image is passed through the saliency map sub-net (b) and the occluded human reconstruction sub-net (c). The output mesh is directly re-sampled from the UV position map.

the occlusion object. Due to the fact that a human visual system explicitly ignores occlusions and only focuses on non-occluded areas, [39] proposed an occlusion robust face recognition approach with the pairwise differential siamese network (PDSN) that explicitly build the correspondence between occluded facial blocks and corrupted feature elements.

Image Inpainting. Pathak *et al*. [31] proposed Context Encoders – the first work applies deep neural networks for image inpainting. It consists of an encoder capturing the context of an image into a compact latent feature representation and a decoder which uses that representation to produce the missing image content. [40] promoted this task with dividing it into inference and translation as two separate steps and each step with a deep neural network. Xiong *et al*. [47] learned to predict the foreground contour first, and then inpainted the missing region using the predicted contour as guidance. [52] proposed a new deep generative model-based approach which can not only synthesize novel image structures but also explicitly utilized surrounding image features as references during network training to make better predictions. [50] proposed a multi-scale neural patch synthesis approach based on joint optimization of image content and texture constraints, which preserves contextual structures and produces high-frequency details by matching and adapting patches with the most similar mid-layer feature correlations of a deep classification network. [49] proposed Shift-Net, which inherits the advantages of exemplar-based and CNN-based methods, and can produce inpainting result with both plausible semantics and fine detailed textures.

3. Method

An overview of the proposed method is shown in Fig.2. We use a partial UV map to represent the object-occluded human body, and human shape estimation is finally formulated as a UV map inpainting problem.

3.1. Object-Occluded Human Representation

We use the representation of 3-channel UV position map [6] to describe a human body for network training. The RGB values in a UV position map record 3D positions of a body mesh, where the map encodes the geometry topology of body surface. Based on the UV position map, we further promote a representation of 3D object-occluded human shapes. In our method, UV coordinates of all mesh vertices are provided by SMPL [27].

Fig.3 shows how we generate a partial UV position map from a body mesh and a segmentation mask. We project the body mesh into the image plane by a weak perspective projection. The projected points that are outside the segmentation mask are regarded as occluded vertices. Otherwise,

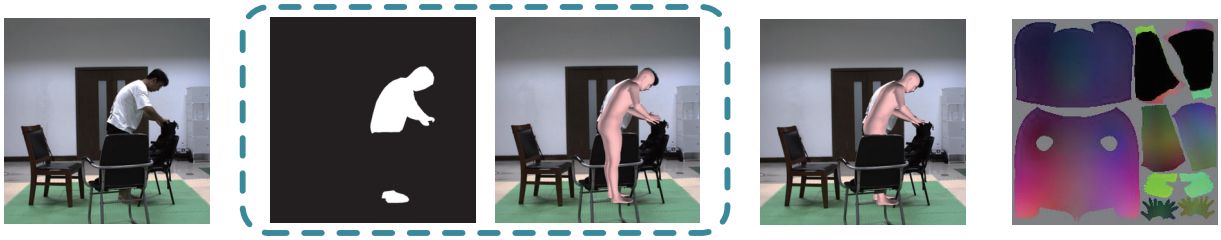

Figure 3. **The Representation of Object-Occluded Human.** Given an occluded human image (a) and the corresponding occlusion segmentation (b), we render the fitted human body model onto the 2D image plane (c). For the visible part, we store the normalized x, y, z coordinates of the vertices as r, g, b color values in the UV map. For the occlusion part, we set the r, g, b values of the UV map to -0.5 (e).

positions of visible vertices are normalized into a range of -0.5 to 0.5, and their x, y, z coordinates are stored as 3-channel (RGB) values in the UV map. For occluded parts, we set the values of the UV map as $[-0.5, -0.5, -0.5]$. Partial UV position map can accurately represent the object-occluded mesh. It should be noted that our partial UV map only considers the object-occlusion. For self-occlusion, [51, 1, 44] have proved that it can be easily estimated through the visible part of the body and the part of self-occlusion is encoded as the supervision in latent space. Notice that the output mesh can be re-sampled from a complete UV position map.

3.2. UV Map Inpainting Sub-Network

Estimating a full UV map from a partial UV map is an image-to-image translation problem [14]. As shown in Fig.2 (b), the process of partial-to-full is not affected by the background of the occluded color images. Thus, we can synthesize occlusions to train an inpainting network, which is robust to various types of occlusions. We follow the work [36] to perform the synthetic occlusion data synthesis on the Human3.6M dataset.

We use an encoder-decoder structure to train the UV map inpainting sub-network and our loss function has three terms

$$L = L_1 + \lambda L_{tv} + \mu L_p, \tag{1}$$

which is a little different from [51].

The first term L_1 performs the supervision between predicted UV maps and ground-truth UV maps, which is,

$$L_1 = \sum_{j=1}^{H} \sum_{i=1}^{W} \beta_{i,j} \left(\left| P_{i,j} - P_{i,j}^{gt} \right| \right), \tag{2}$$

where $\beta_{i,j}$ is a weight mask and the weight is inversely proportional to the part area. W and H are the width and height respectively. P is the pixel RGB value.

The second term L_{tv} ensures the smoothness among each body parts, which has the form,

$$L_{tv} = \sum_{k} \sum_{(i,j) \in R_k} \left(|P_{i+1,j} - P_{i,j}| + |P_{i,j+1} - P_{i,j}| \right), \tag{3}$$

where R_k is defined as the k^{th} body part.

Since L_{tv} only guarantees the smoothness in the same part, it cannot guarantee the body connection part smooth. We then propose a third term named as part loss, which is

$$L_p = \sum_{v_i \in V_b} \left| \overline{P}\left(v_i\right) - p_i^{gt} \right|, \tag{4}$$

where V_b is a set of vertices that have multiple UV coordinates. $\bar{P}(v_i)$ means the average RGB value of UV coordinates corresponding to vertex v_i. p_i^{gt} is the ground truth position of vertex v_i.

3.3. Saliency Map Estimation Sub-Network

To reduce the influences of invalid information such as background and occlusion for human shape and pose recovery, we introduce a sub-network to estimate human saliency map as shown in Fig.4. We use different scales of masks as intermediate supervision. The proposed saliency map, which can be treated as a representation of visual attention, is not an accurate segmentation. Even the state-of-the-art instance segmentation methods[9, 24] are hard to give the correct segmentation in the case of occlusion. However, the imperfect saliency map is good enough to reduce the influence of background and avoid extra cropping operation required by previous methods [17, 19, 20]. In Sec 5, we compare the results with and without saliency map to demonstrate the effectiveness of saliency map.

3.4. Latent Feature Supervision

For the UV map inpainting task, we assume that high dimensional features in the encoder part have a certain degree of prior knowledge for human body shapes. This drives us to utilize the high dimensional features extracted from the

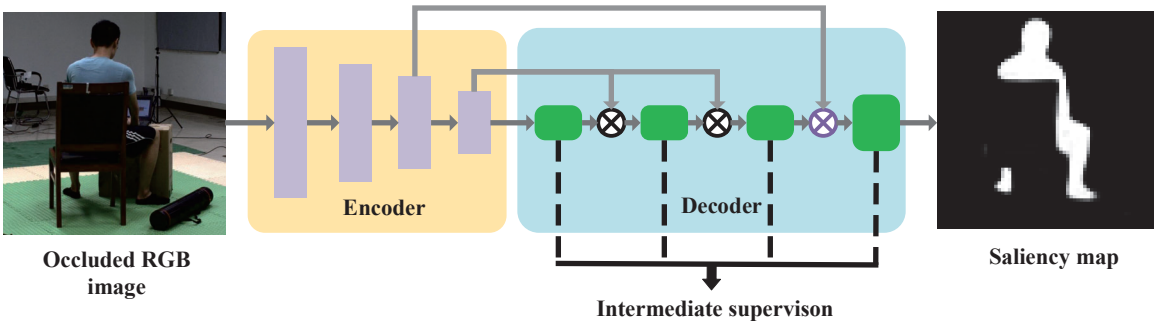

Figure 4. **Overview of saliency map estimation Sub-Network.** We propose a sub-network to estimate human saliency map which aims to reduce invalid information such as background and occlusion. We use different scales of masks as intermediate supervision.

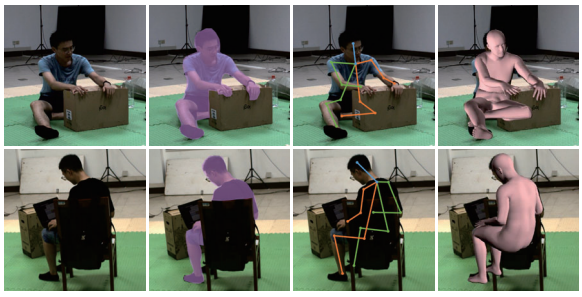

Figure 5. **Samples of the 3DOH50K.** Each image in our dataset includes segmentation, 2D and 3D skeleton keypoints annotation and SMPL parameters.

UV map inpainting branch to supervise the training of color image encoder (Fig.2 (d)). We found that using UV map high dimensional features as constraint could accelerate the convergence speed of training and is more accurate than directly predicting full UV position maps from color images as shown in the experiment section.

4. 3D Occlusion Human Dataset

Most of existing 3D human datasets *e.g.*, [13, 37] focus on the complexity and diversity of poses. Nevertheless, they often overlook the occlusions generated by the interactions between the human and objects, which are commonplaces in the real world. Therefore, human pose and shape estimation methods trained on such datasets are sensitive to occlusions. To solve this challenging problem, we propose our dataset **3D Occlusion Human 50K(3DOH50K)**. It contains 51600 images, most of which are human activities in occlusion scenarios. Fig.5 shows some examples. All images are captured from real scenes with six views. **3DOH50K** is the first real 3D human dataset for the problem of occlusion. Our dataset could provide a new challenge benchmark for human reconstruction and pose estimation in occlusion scenarios.

4.1. Annotations

Obtaining accurate annotations in occluded scenes is extremely difficult. We tried a variety of the state-of-the-art instance segmentation and pose estimation methods[9, 24, 5, 4]. It turns out that all of them do not achieve the desired results. Therefore, for each image, we first use Mask-RCNN[9] and Alphapose[5] to automatically segment the mask and estimate the 2D keypoints. For the inaccurate parts, we manually corrected the mask and keypoints. Then, we fit the SMPL model with annotated keypoints by using SMPLify-X[32] in a multi-view strategy:

$$E(\beta,\theta,T) = E_J + \lambda_{\theta_b} E_{\theta_b} + \lambda_\alpha E_\alpha + \lambda_\beta E_\beta \tag{5}$$

where E_{θ_b}, E_α and E_β are same as [32]. For the data term E_J we use 6-views re-projections to constrain the SMPL model:

$$E_J\left(\beta,\theta,T;K,J_{\text{est}}\right) = \sum_{\text{view}j}\sum_{\text{joint}i} w_{j,i}\rho\left(\Pi_{K_j}\left(R_\theta\left(J(\boldsymbol{\beta})_i\right)+T\right) - J_{\text{est},j,i}\right) \tag{6}$$

Finally, 3DOH50K has camera parameters of 6 views. Each image has an accurate 2D pose and 3D pose, SMPL parameters, and a binary mask.

4.2. Dataset Statistics

We compare different public datasets related to 3D pose estimation in Tab.1. Although existing datasets have high-quality annotations and a large amount of data, they hardly contain examples with occlusions. There are some occlusion sequences in CMU Panoptic[16] and 3DPW[45], but they have a similar pose for the entire occlusion sequence.(*e.g.*, in the subset of Musical Instruments in CMU Panoptic Dataset, the instrument produces a large proportion of occlusion, but the pose and occlusion in the entire sequence are similar.) There is also a small amount of occluded samples in UP-3D dataset. However, since it is fitted through a monocular method [3], the depth information of ground truth is not processed well. [29] provides accurate

Dataset	Occlusion Data	Real Data	2D Pose	3D Pose	Occlusion Seg.	Mesh	Camera Param.
CMU Panoptic[16]	++	✓	✓	✓	–	✓	✓
3DPW[45]	++	✓	✓	✓	–	✓	✓
Human3.6M[13]	–	✓	✓	✓	–	–	✓
UP-3D[21]	+	✓	✓	✓	–	✓	–
MPI-INF-3DHP[29]	+	✓	✓	✓	✓	–	✓
3DOH50K (ours)	++++	✓	✓	✓	✓	✓	✓

Table 1. Comparison among different public datasets related to 3D pose estimation. Occlusion data refers to the object-occluded data and + denotes the amount of occluded samples.

occlusion segmentation, however it only contains very few types of occlusion. These shortcomings of existing datasets lead to their inability to perform 3D pose and shape estimation in occlusion situations.

5. Experiments

5.1. Datasets

Human3.6M [13] is one of the most widely used 3D human datasets. It has 11 subjects, 15 kinds of action sequences and 1.5 million training images with accurate 3D annotations. Since Human3.6M dataset has no object occlusion, we adopt the method of [36] to add a **synthetic occlusion** on the image, an example is shown in Fig. 6 (row 3 right). Similar to [17], we use MoSH[26] to process the marker data in the original dataset, and obtain the ground truth SMPL parameters. For a fair comparison, we use 300K data in S1, S5, S6, S7, S8 for network training, and test in S9, S11.

3DOH50K is the first 3D human occlusion dataset proposed by us. It contains 50310 training images and 1290 test images. It provides 2D, 3D annotations and SMPL parameters for generating meshes. Detailed information has been provided in Sec 4.2.

3DPW [45] is captured via IMUs and contains both indoor and outdoor scenes. It provides accurate SMPL parameters and calibrated camera parameters. However, the occluded samples in the dataset are few and not representative. In order to demonstrate the effectiveness of our approach, we selected the occluded sequences from the entire dataset as a new testset. The names of these sequences are provided in the Supplementary Material.

5.2. Implementation Details

The U-Net structure [35] is adopted for Saliency Map Estimation and the model is supervised using the segmentation of human. In order to reduce the redundant latent features and make it easier to be consistent, a modified ResNet-18 [10] and VGG-19 [38] are respectively used as encoder for partial UV maps and color images. The decoder part is simply composed of 6 consequent up-sampling and convolutional layers. The size of UV maps and color images in this work are all scaled to 256×256. To generate ground-truth UV position maps, we transform all meshes into the same normalized camera coordinate frame by weak perspective projection. For testing, SMPL fitting is performed to estimate SMPL parameters for quantitative comparison among different algorithms. Non-optimized L-BFGS algorithm is adopted for fitting, which takes about 30s. Since most of UV maps and saliency maps have zero values regions, we use leaky-ReLu[28] instead of ReLu. We use the Adam optimizer[18] with a batch-size of 10 and the initial learning rate set to 1e-3. Running time of our work with a 2080Ti GPU is 13ms per image, which is effectively real-time.

5.3. Quantitative Evaluation and Comparison

To demonstrate the effectiveness of our method, we performed quantitative evaluations on Human3.6M, 3DPW and 3DOH50K. Numerous comparisons are conducted with the state-of-the-art methods. It is noted that previous works do not specifically target the object-occluded problem, we retested their released model on the occlusion dataset. Fig.6 and Fig.7 present some results and more detailed comparisons are described in the following.

We first tested our method on S9 and S11 of the original Human3.6M, verifying that our method can also achieve the state-of-art performance without occlusions. The 3rd column in Tab.2 shows the performance of our approach on the origin Human3.6M dataset. Our method can obtain the very similar results as the recent work [19].

Then, we compared our method on those occlusion datasets. On Synthetic Occlusion Human3.6M, we compared ours with the methods generating human meshes (based on SMPL model). Detailed results are shown in the 4th column of Tab.2, and our method outperforms all the other methods. As 3DPW is not designed for occlusions, we selected only the sequences with occlusions for evaluation. For a fair comparison, the evaluation is performed similar as [19]. The 5th column in Tab.2 shows that our method can also obtain better performance than other methods. It is noted that 3DPW contains numerous outdoor scenes, the errors on this dataset are higher than others. We then performed the comparison on the proposed 3DOH50K. The last 3 columns in Tab.2 show the comparison results. From these columns, we can find that our method performs better than all previous methods for about 10mm.

Furthermore, we studied the effect of occlusion ratios on

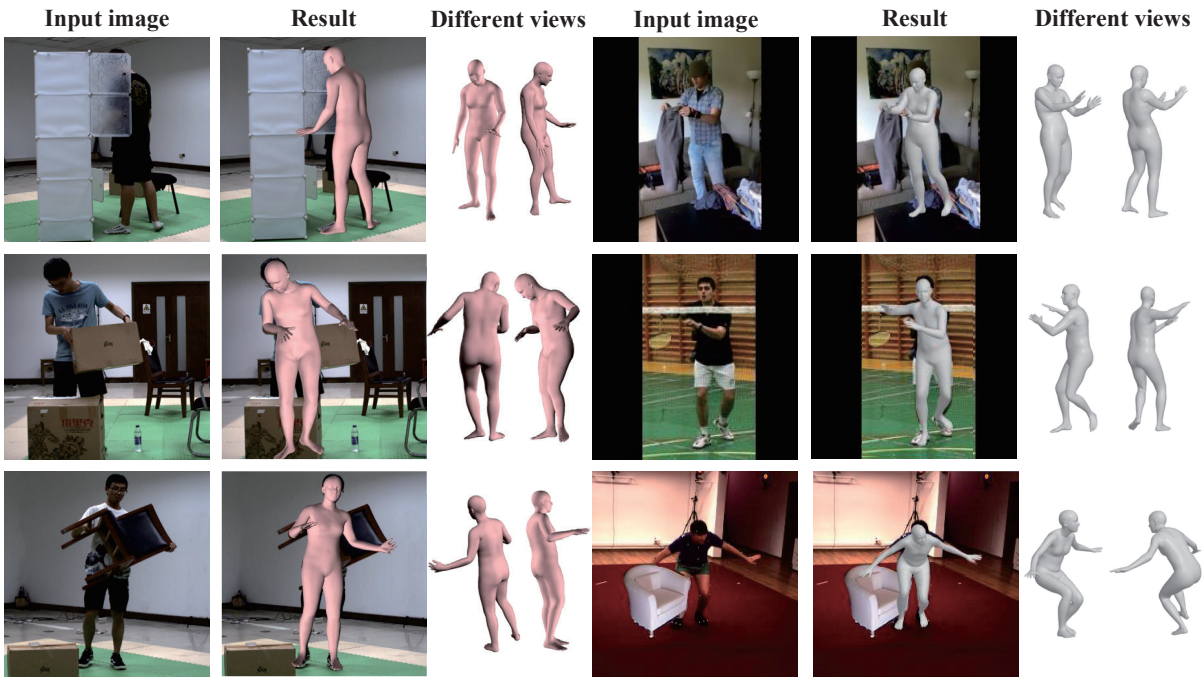

Figure 6. Qualitative results of our approach. The left 3 columns show the results on our 3DOH50K. The right 3 columns are the examples on 3DPW[45], UP-3D[21] and Synthetic Occlusion Human3.6M[13] respectively. **More results without SMPL fitting are shown in the Sup. Mat.**

Method	Runtime	*Human3.6M [13]* ***(Protocol#2)*** PA-MPJPE	*Human3.6M [13]* ***(Synthetic Occlusion)*** PA-MPJPE	*3DPW [45]* PA-MPJPE	*Our 3DOH50K* PA-MPJPE	Surface Error
SMPLify[3]	100 sec	82.3	159.4	114.0	156.4	177.3
SMPLify-X[32]	30 sec	–	145.6	151.3	117.2	132.4
HMR[17]	0.420 sec	56.8	82.2	103.8	83.2	92.9
GraphCMR[20]	0.033 sec	50.1	74.4	104.8	76.3	84.0
SPIN[19]	0.016 sec	**41.1**	64.9	95.4	67.5	73.6
Ours	**0.013 sec**	41.7	**56.4**	**72.2**	**58.5**	**63.3**

Table 2. Comparisons with the state-of-the-art methods on Human3.6M, 3DPW and our 3DOH50K. Synthetic Occlusion Human3.6M means that we randomly render the synthetic occlusion on images, the minimum is 30% of the bounding-box pixels occluded, details can be found in 5.1. Numbers are 3D joint errors and Surface errors in mm.

the reconstruction accuracy by synthesizing different ratios of occlusion on the Synthetic Occlusion Human3.6M. The results are presented in Fig.8. The curves prove that our method can maintain good performance even with an occlusion ratio of more than 50%. In addition, due to the effectiveness of the proposed UV inpainting branch, our method is relatively insensitive to the increase of occlusion ratios.

5.4. Ablation Study

Importance of the UV inpainting branch. In Tab.3, we tested different model structures to demonstrate the importance of our UV inpainting branch. Results show that it is difficult to directly predict object-occluded human shape and pose from color images. We also tried to estimate partial UV maps from occluded color images and then inpaint partial maps in a cascade manner. It turns out that latent space supervision performs better in our method.

Importance of the saliency map estimation. Since color images contain a lot of invalid information, we performed a salient detection on color images to obtain valid human features. In order to verify the importance of saliency map estimation network, we compared the results of occluded color images input and color images combined with saliency map as input. As shown in Tab.3, extra saliency map input improves the performance. Furthermore, it makes our method also have a good performance in outdoor scenes.

Importance of the proposed part loss. In UV maps, the whole body is divided into several parts which means that resampled meshes may have a coarse connection between different parts. Therefore, we put an extra constraint on

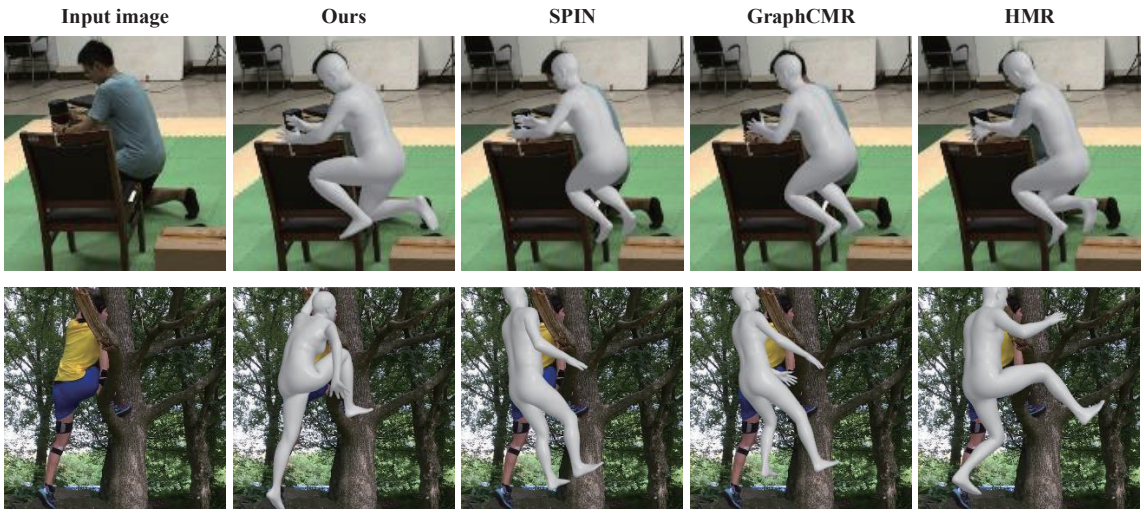

Figure 7. Comparison with different methods. Our method could obtain more visually appealing results.

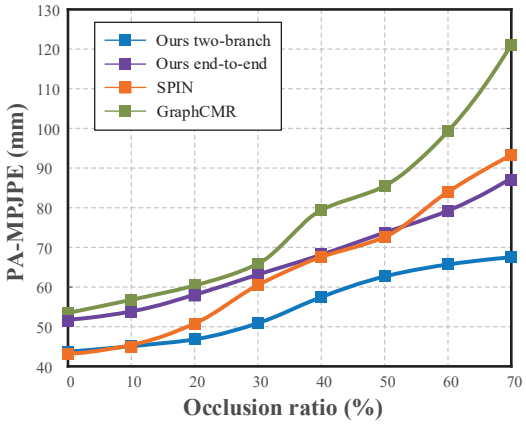

Figure 8. Relationship between the reconstruction accuracy and the occlusion ratio.

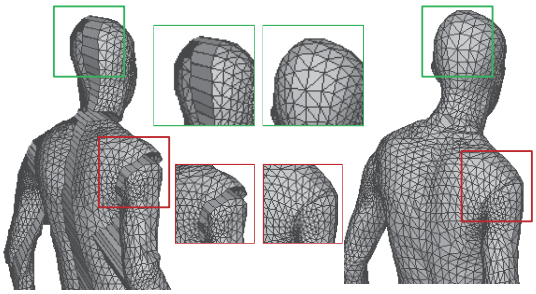

Figure 9. Importance of the part loss. The right sub-figure shows a smoother result with the proposed part loss.

the connection vertices on UV maps. Fig. 9 shows the result where we can find that part loss is essential to improve smoothness.

6. Conclusion

In this paper, we propose a novel method for the object-occluded human shape and pose estimation from single color images. Our main contribution is to utilize a partial UV map representation to describe the human body occlusion, and convert the occluded human shape estimation as a UV map inpainting problem. A novel two-branch network architecture is proposed to train an efficient regressor via the latent feature matching. We also introduce a saliency map sub-net to extract the human information from object-occluded color images. For fertilizing the network training, we further build a new dataset named as **3DOH50K**. We hope the dataset would promote the future research on object-occluded human shape and pose estimation.

Method	MPJPE	PA-MPJPE
end-to-end	73.1	67.3
cascade	62.9	61.9
(w/o) saliency map	60.8	57.9
two-branch	**58.2**	**56.4**

Table 3. Comparison of different network structures on the testing dataset. **end-to-end**: without UV map inpainting network. **cascade**: cascade the UV map inpaiting network with the color image encoder. **(w/o) saliency map**: without saliency map estimation network. **two branch**: the proposed two-branch network. **More details are shown in the Sup. Mat.**

Acknowledgments

This work was supported in part by National Natural Science Foundation of China (No. 61806054), in part by Natural Science Foundation of Jiangsu Province (No. BK20180355), in part by National Key R&D Program of China (No. 2018YFB1403900), in part by Shenzhen Science and Technology Innovation Committee (STIC) (No. JCYJ20180306174459972) and "Zhishan Young Scholar" Program of Southeast University.

References

[1] T. Alldieck, G. Pons-Moll, C. Theobalt, and M. Magnor. Tex2shape: Detailed full human body geometry from a single image. In *ICCV*, 2019. 2, 4

[2] F. Bogo, M. J. Black, M. Loper, and J. Romero. Detailed full-body reconstructions of moving people from monocular rgb-d sequences. In *ICCV*, 2015. 2

[3] 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 *ECCV*, 2016. 1, 2, 5, 7

[4] Z. Cao, T. Simon, S.-E. Wei, and Y. Sheikh. Realtime multi-person 2d pose estimation using part affinity fields. In *CVPR*, 2017. 5

[5] H.-S. Fang, S. Xie, Y.-W. Tai, and C. Lu. RMPE: Regional multi-person pose estimation. In *ICCV*, 2017. 5

[6] Y. Feng, F. Wu, X. Shao, Y. Wang, and X. Zhou. Joint 3d face reconstruction and dense alignment with position map regression network. In *ECCV*, 2018. 2, 3

[7] G. Ghiasi, Y. Yang, D. Ramanan, and C. C. Fowlkes. Parsing occluded people. In *CVPR*, 2014. 2

[8] R. Girshick, P. Felzenszwalb, and D. Mcallester. Object detection with grammar models. *IEEE TPAMI*, 33, 11 2010. 2

[9] K. He, G. Gkioxari, P. Dollar, and R. Girshick. Mask r-cnn. In *ICCV*, 2017. 4, 5

[10] K. He, X. Zhang, S. Ren, and S. Jian. Deep residual learning for image recognition. In *CVPR*, 2016. 6

[11] J.-B. Huang and M.-H. Yang. Estimating human pose from occluded images. In *ACCV*, 2009. 2

[12] Y. Huang, F. Bogo, C. Classner, A. Kanazawa, P. Gehler, I. Akhter, and M. Black. Towards accurate markerless human shape and pose estimation over time. In *3DV*, 2017. 2

[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, jul 2014. 2, 5, 6, 7

[14] P. Isola, J. Zhu, T. Zhou, and A. A. Efros. Image-to-image translation with conditional adversarial networks. In *CVPR*, 2016. 4

[15] A. S. Jackson, C. Manafas, and G. Tzimiropoulos. 3d human body reconstruction from a single image via volumetric regression. In *ECCV*, 2018. 1

[16] H. Joo, T. Simon, X. Li, H. Liu, L. Tan, L. Gui, S. Banerjee, T. S. Godisart, B. Nabbe, I. Matthews, T. Kanade, S. Nobuhara, and Y. Sheikh. Panoptic studio: A massively multiview system for social interaction capture. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2017. 5, 6

[17] A. Kanazawa, M. J. Black, D. W. Jacobs, and J. Malik. End-to-end recovery of human shape and pose. In *CVPR*, 2018. 1, 2, 4, 6, 7

[18] D. Kingma and J. Ba. Adam: A method for stochastic optimization. *International Conference on Learning Representations*, 12 2014. 6

[19] N. Kolotouros, G. Pavlakos, M. J. Black, and K. Daniilidis. Learning to reconstruct 3d human pose and shape via model-fitting in the loop. In *ICCV*, 2019. 4, 6, 7

[20] N. Kolotouros, G. Pavlakos, and K. Daniilidis. Convolutional mesh regression for single-image human shape reconstruction. In *CVPR*, 2019. 4, 7

[21] C. Lassner, J. Romero, M. Kiefel, F. Bogo, M. Black, and P. Gehler. Unite the people: Closing the loop between 3d and 2d human representations. In *CVPR*, 2017. 1, 6, 7

[22] V. Lazova, E. Insafutdinov, and G. Pons-Moll. 360-degree textures of people in clothing from a single image. In *3DV*, 2019. 2

[23] K. Li, N. Jiao, Y. Liu, Y. Wang, and J. Yang. Shape and Pose Estimation for Closely Interacting Persons Using Multi-view Images. *Computer Graphics Forum*, 2018. 2

[24] R. Li, X. Dong, Z. Cai, D. Yang, H. Huang, S. Zhang, P. L. Rosin, and S. Hu. Pose2seg: Human instance segmentation without detection. In *CVPR*, 2019. 4, 5

[25] J. Liang and M. C. Lin. Shape-aware human pose and shape reconstruction using multi-view images. In *ICCV*, 2019. 2

[26] M. Loper, N. Mahmood, and M. Black. Mosh: Motion and shape capture from sparse markers. volume 33, 12 2014. 6

[27] M. Loper, N. Mahmood, J. Romero, G. Pons-Moll, and M. J. Black. SMPL: A skinned multi-person linear model. *ACM Trans. Graphics (Proc. SIGGRAPH Asia)*, 34(6):248:1–248:16, oct 2015. 1, 3

[28] A. L. Maas. Rectifier nonlinearities improve neural network acoustic models. In *ICMLW*, 2013. 6

[29] 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 *3DV*, 2017. 5, 6

[30] 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 *3DV*, 2018. 1, 2

[31] D. Pathak, P. Krahenbuhl, J. Donahue, T. Darrell, and A. Efros. Context encoders: Feature learning by inpainting. In *CVPR*, 2016. 3

[32] G. Pavlakos, V. Choutas, N. Ghorbani, T. Bolkart, A. Osman, D. Tzionas, and M. Black. Expressive body capture: 3d hands, face, and body from a single image. In *CVPR*, 2019. 2, 5, 7

[33] G. Pavlakos, L. Zhu, X. Zhou, and K. Daniilidis. Learning to estimate 3d human pose and shape from a single color image. In *CVPR*, 2018. 1, 2

[34] U. Rafi, J. Gall, and B. Leibe. A semantic occlusion model for human pose estimation from a single depth image. 2015. 2

[35] O. Ronneberger, P. Fischer, and T. Brox. U-net: Convolutional networks for biomedical image segmentation. In *MICCAI*, 2015. 6

[36] I. Sárándi, T. Linder, K. O. Arras, and B. Leibe. How robust is 3d human pose estimation to occlusion? *CoRR*, abs/1808.09316, 2018. 2, 4, 6

[37] L. Sigal, A. Balan, and M. Black. Humaneva: Synchronized video and motion capture dataset and baseline algorithm for

evaluation of articulated human motion. *International Journal of Computer Vision*, 87:4–27, 03 2010. 5

[38] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. In *ICLR*, 2015. 6

[39] L. Song, D. Gong, Z. Li, C. Liu, and W. Liu. Occlusion robust face recognition based on mask learning with pairwisedifferential siamese network. In *ICCV*, 2019. 3

[40] Y. Song, C. Yang, Z. Lin, X. Liu, Q. Huang, H. Li, and C. C. J. Kuo. Contextual-based image inpainting: Infer, match, and translate. In *ECCV*, 2018. 3

[41] Y. Tao, Z. Zheng, K. Guo, J. Zhao, and Y. Liu. Doublefusion: Real-time capture of human performances with inner body shapes from a single depth sensor. In *CVPR*, 2018. 2

[42] A. Tran, T. Hassner, I. Masi, E. Paz, Y. Nirkin, and G. Medioni. Extreme 3d face reconstruction: Seeing through occlusions. In *CVPR*, 2018. 2

[43] G. Varol, D. Ceylan, B. Russell, J. Yang, E. Yumer, I. Laptev, and C. Schmid. Bodynet: Volumetric inference of 3d human body shapes. In *ECCV*, 2018. 1

[44] A. Venkat, C. Patel, Y. Agrawal, and A. Sharma. Humanmeshnet: Polygonal mesh recovery of humans. In *ICCV*, 2019. 2, 4

[45] T. von Marcard, R. Henschel, M. J. Black, B. Rosenhahn, and G. Pons-Moll. Recovering accurate 3d human pose in the wild using imus and a moving camera. In *ECCV*, 2018. 2, 5, 6, 7

[46] C. Wu and J. Ding. Occluded face recognition using low-rank regression with generalized gradient direction. *Pattern Recognition*, 80, 03 2018. 2

[47] W. Xiong, J. Yu, Z. Lin, J. Yang, X. Lu, C. Barnes, and J. Luo. Foreground-aware image inpainting. In *CVPR*, 2019. 3

[48] W. Xu, A. Chatterjee, M. Zollhöfer, H. Rhodin, D. Mehta, H.-P. Seidel, and C. Theobalt. Monoperfcap: Human performance capture from monocular video. *ACM Transactions on Graphics*, 37, 08 2017. 2

[49] Z. Yan, X. Li, M. Li, W. Zuo, and S. Shan. Shift-net: Image inpainting via deep feature rearrangement. In *ECCV*, 2018. 3

[50] C. Yang, X. Lu, Z. Lin, E. Shechtman, O. Wang, and H. Li. High-resolution image inpainting using multi-scale neural patch synthesis. In *CVPR*, 2017. 3

[51] 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. 1, 2, 4

[52] J. Yu, Z. Lin, J. Yang, X. Shen, X. Lu, and T. Huang. Generative image inpainting with contextual attention. In *CVPR*, 01 2018. 3

[53] X. Yuan and I. Kyu Park. Face de-occlusion using 3d morphable model and generative adversarial network. In *ICCV*, 2019. 2

[54] H. Zhang, J. Cao, G. Lu, W. Ouyang, and Z. Sun. Danet: Decompose-and-aggregate network for 3d human shape and pose estimation. In *ACM MM*, 2019. 1

[55] H. Zhu, X. Zuo, S. Wang, X. Cao, and R. Yang. Detailed human shape estimation from a single image by hierarchical mesh deformation. In *CVPR*, 2019. 2