SEE THROUGH OCCLUSIONS: DETAILED HUMAN SHAPE ESTIMATION FROM A SINGLE IMAGE WITH OCCLUSIONS

Tianyi Zhang[⋆] *Jin Wang*[⋆] *Qing Zhu*[⋆] *Baocai Yin*[†]

[⋆] Beijing University of Technology, Beijing 100124, China
[†]Dalian University of Technology, Dalian 116024, China

ABSTRACT

3D human body shape and pose reconstructing from a single RGB image is a challenging task in the field of computer vision and computer graphics. Since occlusions are prevalent in real application scenarios, it's important to develop 3D human body reconstruction algorithms with occlusions. However, existing methods didn't take this problem into account. In this paper, we present a novel depth estimation Neural Network, named Detailed Human Depth Network(DHDNet), which aims to reconstruct the detailed and completed depth map from a single RGB image contains occlusions of human body. Inspired by the previous works [1, 2], we propose an end-to-end method to obtain the fine detailed 3D human mesh. The proposed method follows a coarse-to-fine refinement scheme. Using the depth information generated from DHDNet, the coarse 3D mesh can recover detailed spatial structure, even the part behind occlusions. We also construct DepthHuman, a 2D in-the-wild human dataset containing over 18000 synthetic human depth maps and corresponding RGB images. Extensive experimental results demonstrate that our approach has significant improvement in 3D mesh reconstruction accuracy on the occluded parts.

Index Terms— 3D reconstruction, deep learning, monocular camera, depth estimation, human body

1. INTRODUCTION

3D human shape reconstruction from a single image is a challenging task and a hot topic in recent years. It has a wide range of applications in the field of VR/AR such as virtual dressing, and in video/image editing. Currently available approaches can be divided into two categories: template adaption and feature matching. Methods based on template adaption rely on pre-trained generative human models, such as the SMPL [3] or SCAPE [4] model. Bogo et al. [5] presented SMPLify which manages to minimize the difference between detected 2D joints which comes from a CNN-based method and the projected 3D model joints. Kanazawa et al. [1] present HMR, an end-to-end framework to reconstruct human body pose and shape based on SMPL model using an adversarial loss to constrain the pose effectively with only 2D joints annotations. Using the HMR results as standard model, Zhu et al. [2] deform the SMPL mesh in three levels to minimize the objective functions. Making the mesh have cloth detail, both the shape and pose are more accurate. Latest work from Alldieck et al. [6] use UV-mapping to unfold the body surface into a 2D image. Instead of regressing details on the mesh, they propose to regress shape as UV-space normal map and displacement. Methods based on feature matching directly regress 3D geometry from single image, instead of optimizing a standard human body shape. DoubleFusion [7] and HybridFusion [8] use monocular depth camera to capture human motions in real time. The later one uses sparse inertial measurement units (IMUs) as auxiliary means. DeepHuman [9] fuses image features into 3D model through volumetric feature transformation. Then, using a normal refinement network to optimize the visible surface details.

Although there is significant progress to solve this challenging problem, existing methods hardly consider the occlusion in input images, which is prevailing in practical application scenarios such as large crowd photos, surveillance videos and TV shows. Most existing methods have poor performance on reconstructing detailed and complete 3D human shapes when the input image contains occlusions or truncation of human bodies.

To solve this problem, in this paper we propose an end-to-end 3D human reconstruction framework, which takes a single RGB image with occlusions as input and outputs detailed and complete human shape. According to our investigation, the detailed depth map is of crucial importance in the 3D human mesh reconstruction. To exploit the shading information in the input image to add surface details on the reconstructed human models, we design an effective depth reconstruction network detailed human depth network(DHDNet) to obtain a detailed and complete human depth map from a single RGB image with occlusions. The generator of the DHDNet aims to generate depth map from RGB image supervised by multiple losses including a photometric loss, while the discriminator focus on the occluded part of generated depth map. We apply a coarse-to-fine optimization strategy, using the robust parametric models as standard human shape. Then the standard shape is deformed in three levels, using the depth information to recover occluded surface details.

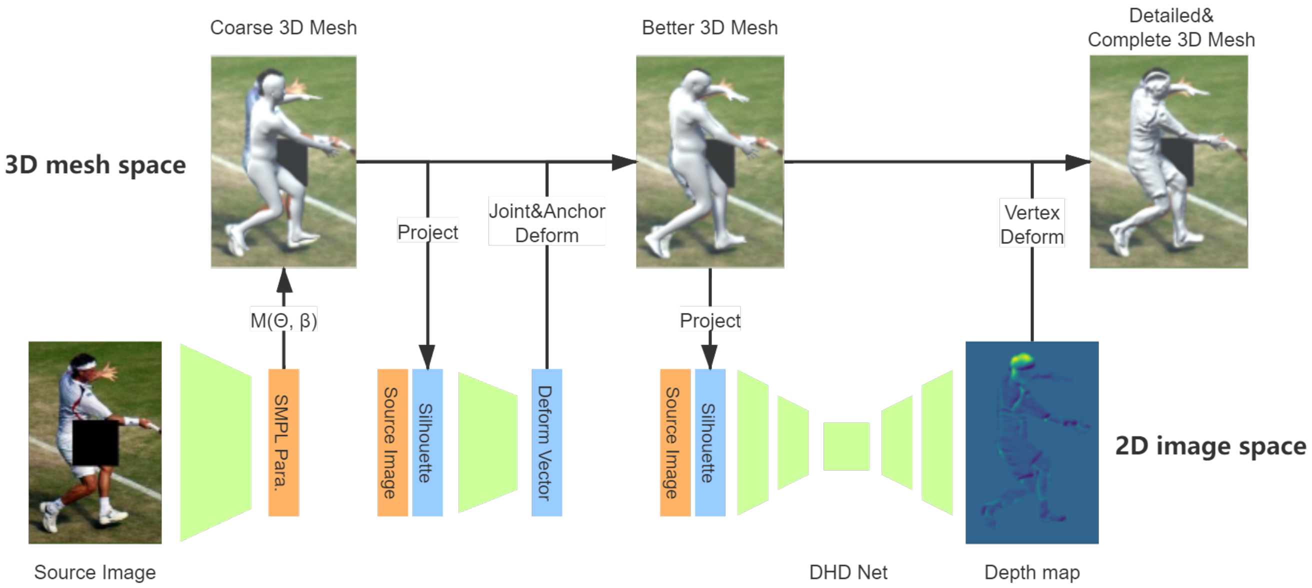


Fig. 1: The end-to-end frame we proposed to obtain detailed and complete 3D human model from a single RGB image with occlusions. The coarse model is generated by HMR. The joint and anchor refinement stage follows HMD's method.

The rest of this paper is organized as follows. Section 2 elaborates the proposed end-to-end framework and detailed human depth network. Section 3 presents experimental results and analysis, and Section 4 concludes this paper.

2. PROPOSED METHOD

In this section, we present the architecture and training scheme of our DHDnet. The details of our end-to-end framework, including standard model construction stage and optimization stage.

2.1. End-to-end framework for 3D reconstruction

We establish an end-to-end framework for detailed and complete 3D human reconstruction from a single occluded RGB image. The overall framework is as shown in Fig.1. We take a coarse-to-fine strategy to finish this task. First, a standard model is constructed using HMR [1]. In this stage, any method that constructs a SMPL model can be applied. Here the HMR model is applied for its state-of-the-art performance on human shape recovery from a single image. Then, 3D space points on the standard mesh are deformed in joint and anchor level to obtain a better human model with more accurate joint position and smooth shape in the same way as HMD[2]. Vertices around 20 body joints are selected for joint level deformation, while 200 vertices evenly over the standard model are selected for anchor level deformation. After the preparation, the mesh is deformed in vertex level to adding details to the human model surface, using the reconstructed depth information from DHDNet. After a quadruple subdivision, all the vertices of SMPL mesh are selected for the vertex level deformation. In these two stages, a challenging task is how to deform the 3D space mesh from vertices defined in 2D space. We address this problem by Laplacian mesh deformation. Specifically, the motion vector for each group of selected vertices is predicted from the network which takes the joints and silhouettes of the 2D image as input. Using Laplacian deformation, the human mesh will maintain the local 3D geometry as much as possible while deforming the vertices according to the depth information, which helps to recover reasonable 3D surface details according to the 2D features. This deformation method has been used in many multi-view 3D reconstruction problems [10, 11, 12].

2.2. Detailed human depth network

Since detailed depth map is of crucial importance in the 3D human reconstruction, we propose a detailed human depth network(DHDNet) to reconstruct an accurate and complete depth map from a single occluded RGB image. To learn the mapping relations between input incomplete RGB image and corresponding depth map, we propose a training model of DHDNet based on generative adversarial networks(GAN)[13]. The generator is based on U-Net and the discriminator is simply an encoder net which inputs the generated depth map and outputs a binary classification result. To train the proposed model, we also construct a large human depth map dataset using synthesized images. More details of this dataset are in Section 3.2. Using this dataset, we first try to infer the depth information from a complete

RGB image without occlusions. We modify the original U-Net output layer to one channel and use the MSE loss to supervise the network weights for minimizing the difference between the output results and ground truth depth map.

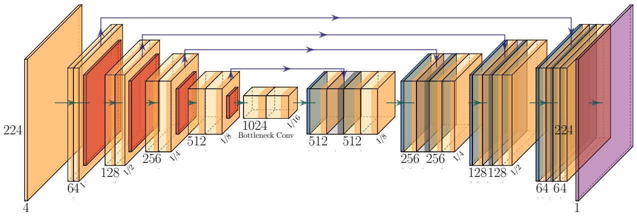


Fig. 2: The generator part of our DHDNet. The input is 3 channel RGB image plus one channel occlusion mask. After an encoder-decoder progress, the network ouputs one channel depth map of the original image.

Next, according to the occlusions in the image, we make a mask layer which represents the invisible regions. Adding this mask to the modified U-Net inputs, we get the generator part of our DHDNet as shown in Fig.2. This part is supervised by multiply loss functions, including the depth loss, a GAN loss and a photometric loss. The GAN loss is defined as:

$$\begin{aligned} L_{gan} =& \underset{G}{min}\,\underset{D}{max} \mathbb{E}_{x \in X}[log(D(x))] \\ &+ \mathbb{E}_{z \in Z}[log(1 - D(G(z)))] \end{aligned} \tag{1}$$

where G and D is the generator and the discriminator, X and Z denote the distribution of ground-truth images and input images respectively. To reconstruct more details of the depth map, inspired by DDRNet[14], the photometric loss is introduced and defined as:

$$L_{photo} = \left\| \rho \sum_{k=1}^{9} l_k H_k(n) - I \right\|_2 \tag{2}$$

where ρ is the albedo computed by the same way as in [15]. Under the Lambertian surface assumption, we use the second spherical harmonics(SH) for illumination representation. H_k is the basis of SH, and l_n represents the SH coefficients. To fill the occluded part more properly, we set up a encoder network as adversarial net. It focus on the occluded part, supervised by the GAN loss and tries to distinguish the generator results from ground truth. In order to recover high-frequency details, a VGG feature extractor is utilized on the predicted depth map and the ground truth, with a style loss and a content loss as in Shift-Net [16]. The content loss is defined as:

$$L_c(p, x, l) = \frac{1}{2} \sum_{i,j} (F^l_{i,j} - P^l_{i,j})^2 \tag{3}$$

where p and x are the original image and the generated image respectively, P^l and F^l are the feature representations in layer l. The style loss is defined as:

$$L_s(a, x) = \sum_{l=0}^{L} w_l E_l \tag{4}$$

where $E_l = \frac{1}{4N_l^2 M_l^2} \sum_{i,j} (G^l_{i,j} - A^l_{i,j})^2$, $G^l_{i,j}$ is a Gram matrix and w_l are weighting factors of the contribution of each layer, a and x are the original image and the generated image, A_l and G_l are the feature representation in layer l. Sum these losses up, we got our final objective function:

$$\begin{aligned} L_{final} =& L_{depth} + \lambda_{gan} L_{gan} + \lambda_{photo} L_{photo} \\ &+ \lambda_c L_c + \lambda_s L_s \end{aligned} \tag{5}$$

where λ_{gan}, λ_{photo}, λ_c, λ_s denote the tradeoff parameters for the corresponding losses respectively.

3. EXPERIMENTAL RESULTS

3.1. Experimental settings

The proposed model is implemented in PyTorch. The network is trained using images with center mask, which means the mask layer mentioned in Section 2.1 consists of a black square in the center of the complete image. The batch size is set to 1. Adam optimizer is applied with a learning rate of 1×10^{-4} and $\beta = 0.9$. It takes about 30 hours to train DHDNet on a single QUADRO M5000 graphic card. Three datasets are used to validate the efficiency of our proposed method: RECON, SYN [2] and our DepthHuman dataset. RECON dataset contains 150 3D ground truth meshes, which are reconstructed by the traditional multi-view 3D reconstruction method. SYN dataset contains 300 3D ground truth meshes, which are rendered from PVHM dataset [17].

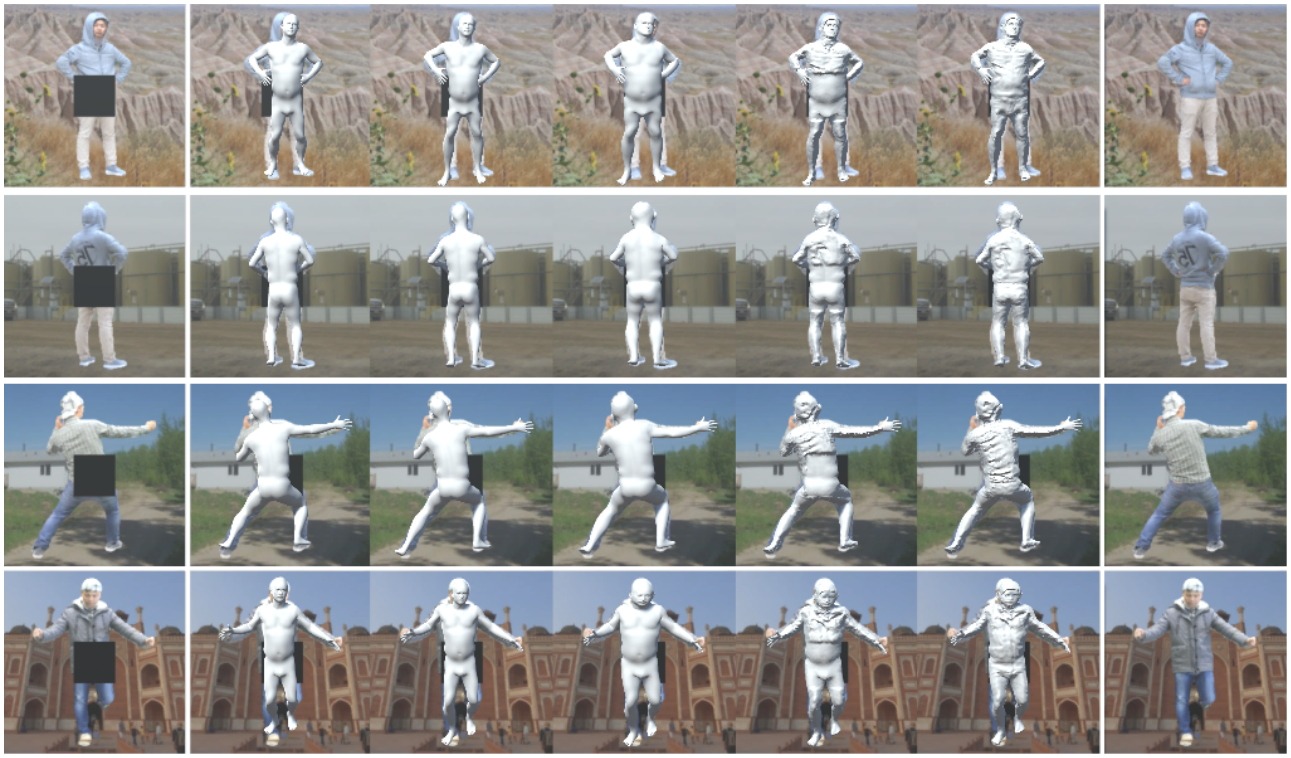
Table 1: Depth dataset components.

Data Source	LSP	LSPET	MPII	COCO	TOTAL
training num	987	5376	8035	4004	18402
testing num	703	0	1996	606	3305

By collating the public available human datasets, including Leeds Sports Pose dataset(LSP) [18] and its extension dataset(LSPET) [19], MPII human pose database (MPII) [20], Common Objects in Context dataset (COCO) [21], we build up DepthHuman, a large dataset containing over 18000 joints annotated images of human in diverse poses. The data is divided into training and testing parts according to the rules of each dataset. We use the pretrained model of HMD to synthesize depth maps as ground truth. Therefore, every single RGB image has its corresponding depth ground truth. The details are shown in Table 1.

3.2. Performance comparison with SOTA methods

To evaluate the performance of our proposed model, we compared our method with state-of-the-art methods with different datasets. Subjective results comparisons are shown in Fig.3. As it shows, HMR, HMD-j and HMD-a barely recover the surface details of human body. HMD-s creates noticeable



(a)Input (b)HMR (c)HMD-j (d)HMD-a (e)HMD-s (f)Proposed (g)Original

Fig. 3: Quantitative comparison results: from right to left, there are input images, HMR results, HMD-j results, HMD-a results, HMD-s results and ours results. On the far right, there are original images as reference.

Table 2: Quantitative comparison results.

	SYN dataset			RECON dataset		
Methods	**3d_err**	**3d_err_visi**	**iou**	**3d_err**	**3d_err_visi**	**iou**
SMPLify	77.310	75.670	66.200	61.840	60.690	69.900
BodyNet	69.410	61.550	65.200	52.750	51.050	68.500
HMR	67.993	62.467	66.901	60.047	51.501	70.358
HMD-j	63.876	59.591	69.977	58.918	51.047	73.231
HMD-a	61.224	58.373	**73.799**	53.215	49.814	**79.114**
HMD-s	58.584	55.191	-	50.859	**50.387**	-
Ours	**58.503**	**54.317**	-	**50.592**	50.617	-

potholes in the surface of the model. Models reconstructed by our method are significantly better than the other results. Body surface details are more complete, meanwhile the integrity of human body is promised. With the 3D groundtruth meshes, we also get quantitative comparison results. We measure the accuracy of the reconstructed shape with three metrics. As Table 2 shows, the first column of result of each dataset shows the average distance of vertices between the groundtruth mesh and the predicted mesh as the 3D error.

$$D_{3d\_err} = \frac{1}{N}\sum_{i=1}^{N} \|v_i - \hat{v}_i\|_2 \qquad (6)$$

where N denotes the vertex number in SMPL model, v_i is vertex on the predicted mesh, $\hat{v}_i$ is the vertex on the groundtruth mesh. The second column shows the same distance but only considers the visible side of human body. The third column is the silhouette Intersection over Union (IoU), which represents the matching rate of the groundtruth silhouette and the projected silhouette of reconstructed 3D mesh. Due to the fact that depth information only transforms point in the vertical direction of camera plane, the silhouette IoU of HMD-s and ours stay the same after vertex deformation as HMD-a result. Further details are as shown in Table 2.

4. CONCLUSION

In this paper, we proposed an effective method to reconstruct detailed 3D human body from a single RGB image with occlusions. The key to filling missing areas is recovering a complete depth map from an occluded RGB image. We address this problem by proposing DHDNet, a CNN based generative adversarial network. The DHDNet supervised by multiple loss functions and trained by our established dataset. Extensive experiments shows excellent results on filling the missing regions. Qualitative and quantitative results are both significantly better than other state-of-the-art methods.

5. REFERENCES

[1] Angjoo Kanazawa, Michael J. Black, David W. Jacobs, and Jitendra Malik, "End-to-End Recovery of Human Shape and Pose," *CVPR*, pp. 7122–7131, 2018.

[2] Hao Zhu, Xinxin Zuo, Sen Wang, Xun Cao, and Ruigang Yang, "Detailed Human Shape Estimation from a Single Image by Hierarchical Mesh Deformation," *CVPR*, 2019.

[3] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J. Black, "SMPL: A skinned multi-person linear model," *ACM Transactions on Graphics*, vol. 34, no. 6, 2015.

[4] Dragomir Anguelov, Praveen Srinivasan, Daphne Koller, Sebastian Thrun, Jim Rodgers, and James Davis, "SCAPE: Shape Completion and Animation of People," *ACM Transactions on Graphics*, vol. 24, no. 3, pp. 408–416, 2005.

[5] Federica Bogo, Angjoo Kanazawa, Christoph Lassner, Peter Gehler, Javier Romero, and Michael J. Black, "Keep it SMPL: Automatic estimation of 3D human pose and shape from a single image," in *European Conference on Computer Vision*, 2016, vol. 9909 LNCS, pp. 561–578.

[6] Thiemo Alldieck, Gerard Pons-Moll, Christian Theobalt, and Marcus Magnor, "Tex2Shape: Detailed Full Human Body Geometry From a Single Image," *CVPR*, pp. 1–10, 2019.

[7] Tao Yu, Zerong Zheng, Kaiwen Guo, Jianhui Zhao, Qionghai Dai, Hao Li, Gerard Pons-Moll, and Yebin Liu, "DoubleFusion: Real-Time Capture of Human Performances with Inner Body Shapes from a Single Depth Sensor," *CVPR*, pp. 7287–7296, 2018.

[8] Zerong Zheng, Tao Yu, Hao Li, Kaiwen Guo, Qionghai Dai, Lu Fang, and Yebin Liu, "HybridFusion: Real-time performance capture using a single depth sensor and sparse IMUs," in *European Conference on Computer Vision*, 2018, vol. 11213 LNCS, pp. 389–406.

[9] Zerong Zheng, Tao Yu, Yixuan Wei, Qionghai Dai, and Yebin Liu, "DeepHuman: 3D Human Reconstruction from a Single Image," in *Proceedings of the IEEE International Conference on Computer Vision*, 2019.

[10] Thiemo Alldieck, Marcus Magnor, Weipeng Xu, Christian Theobalt, and Gerard Pons-Moll, "Video Based Reconstruction of 3D People Models," *CVPR*, pp. 8387–8397, 2018.

[11] Hao Zhu, Yebin Liu, Jingtao Fan, Qionghai Dai, and Xun Cao, "Video-Based Outdoor Human Reconstruction," *IEEE Transactions on Circuits and Systems for Video Technology*, 2017.

[12] Miao Liao, Qing Zhang, Huamin Wang, Ruigang Yang, and Minglun Gong, "Modeling deformable objects from a single depth camera," in *ICCV*. IEEE, 2009, pp. 167–174.

[13] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio, "Generative adversarial nets," in *Advances in neural information processing systems*, 2014, pp. 2672–2680.

[14] Shi Yan, Chenglei Wu, Lizhen Wang, Feng Xu, Liang An, Kaiwen Guo, and Yebin Liu, "DDRNet: Depth map denoising and refinement for consumer depth cameras using cascaded CNNs," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, vol. 11214 LNCS, pp. 155–171.

[15] Sean Bell, Kavita Bala, and Noah Snavely, "Intrinsic images in the wild," *ACM Transactions on Graphics*, vol. 33, no. 4, pp. 159, 2014.

[16] Zhaoyi Yan, Xiaoming Li, Mu Li, Wangmeng Zuo, and Shiguang Shan, "Shift-net: Image inpainting via deep feature rearrangement," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, vol. 11218 LNCS.

[17] Hao Zhu, Hao Su, Peng Wang, Xun Cao, and Ruigang Yang, "View extrapolation of human body from a single image," in *CVPR*, June 2018.

[18] Sam Johnson and Mark Everingham, "Clustered pose and nonlinear appearance models for human pose estimation.," in *BMVC*, 2010, vol. 2, p. 5.

[19] Sam Johnson and Mark Everingham, "Learning effective human pose estimation from inaccurate annotation," in *CVPR*. IEEE, 2011, pp. 1465–1472.

[20] Mykhaylo Andriluka, Leonid Pishchulin, Peter Gehler, and Bernt Schiele, "2d human pose estimation: New benchmark and state of the art analysis," in *CVPR*, 2014, pp. 3686–3693.

[21] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick, "Microsoft coco: Common objects in context," in *ECCV*. Springer, 2014, pp. 740–755.