LEARNING MONOCULAR MESH RECOVERY OF MULTIPLE BODY PARTS VIA SYNTHESIS

Yu Sun[1†] *Tianyu Huang*[3] *Qian Bao*[2] *Wu Liu*[2⊠] *Wenpeng Gao*[1] *Yili Fu*[1⊠]
[1] Harbin Institute of Technology [2] JD AI Research [3] University of Maryland

ABSTRACT

In this paper, we focus on simultaneously recovering the 3D mesh of multiple body parts from a single RGB image. One of the main challenges is that available datasets with full-body 3D annotations are very limited. This results in poor generalization ability of existing learning-based methods. Existing optimization-based methods iteratively fit the 3D mesh to the 2d pose, which is very time-consuming. To address these limitations, we propose to integrate multiple 3D single-body-part datasets to create a highly diverse whole-body 3D motion space for learning from controllable synthetics. Compared with the learning-based approaches, the proposed method greatly alleviates the reliance on training data. Compared with the optimization-based approaches, the proposed method is a hundred times faster. Our proposed method also outperforms previous state-of-the-art methods on CMU Panoptic dataset.

Index Terms— 3D mesh recovery, synthetic training, multi-part estimation

1. INTRODUCTION

Human 3D mesh recovery (reconstruction/modeling) is a hot and challenging research topic in computer vision, which has drawn more and more attention recently. Different from the general pose estimation that detects several 2D or 3D keypoints, the 3D human mesh contains thousands of vertices that can provide subtle cues for understanding human posture, behavior, and interaction.

In this paper, we investigate simultaneously estimating the 3D mesh of multiple body parts from a single RGB image, as shown in Fig 2. Some previous works [1, 2] adopt an optimization-based framework, which is very time-consuming. They estimate the 3D keypoints of each body part separately using individual modules, and then iteratively fit the statistical body model to the 2D keypoint detections. The entire process may take tens of seconds. The main challenge of training a non-iterative learning-based model is lacking sufficient datasets containing the images with full-body 3D annotations. Most existing datasets are captured with 3D annotations of either body or hand only. As shown in Fig. 1, commonly used 3D hand datasets contain high-resolution 2D hand images. While in the whole-body datasets, the hands are usually small, blurred, and easily occluded by objects. A few datasets (e.g. CMU Panoptic [1]) containing both 3D body and hand annotations. While they are all captured in constrained indoor environments and the diversity of the motion space is very limited. Consequently, methods trained on these datasets are hard to generalize well to the in-the-wild cases. Therefore, we propose to break the reliance on training data via learning from controllable synthetics.

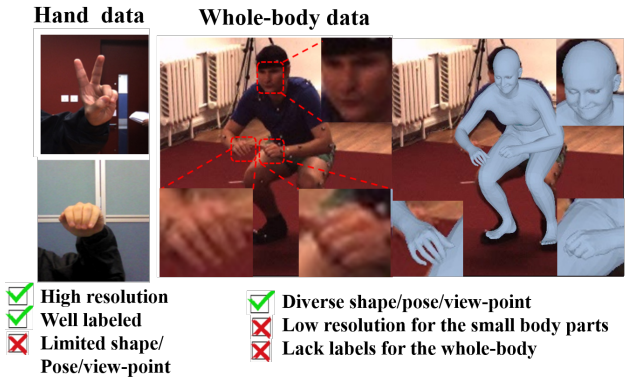

Fig. 1. Challenges in the monocular mesh recovery of multiple body parts.

To achieve this, we design a disentangled framework and establish a whole-body 3D motion space with high diversity for synthetic learning. To deal with the large-scale differences between different body parts, we are supposed to provide individual receptive field in model for each of them. Therefore, we designed a disentangled multi-branch architecture to facilitate separate learning of different body properties. The architecture consists of two parts, 2D pose estimation of multiple body parts and 3D mesh recovery with 2D pose assistance. We focus on enriching the 3D motion space and learning the mapping from 2D poses to 3D meshes. In this way, we can avoid synthesizing complex textures while being able to learn from richer poses. To ensure the rationality and diversity of the synthetic data, we combine the 3D poses of different body parts from the existing datasets [3, 4, 5] to generate the whole-body poses. However, directly combing the 3D pose of individual body parts from multiple datasets would cause weird poses since they are 1) usually in different scales; 2) captured in diverse camera configurations; 3) in diverse shape/pose space. A synthetic training scheme (STS) is developed to integrate them naturally and generate the train-

[†] This work was done when Yu Sun was an intern at JD AI Research.
[⊠] Corresponding author.

Fig. 2. Qualitative comparisons to previous methods, MTC [1] and SMPLify-X [2].

ing data for directly supervised training via sampling the 3D motion space. In this way, we can generate data with controllable settings (e.g. full shooting angles, shape/pose space).

Our method is qualitatively and quantitatively evaluated and compared with previous methods. As shown in Fig. 2, compared with MTC [1] and SMPLify-X [2], the model trained with the proposed STS shows an obvious advance in the estimation of body and hand pose. Compared with optimization-based methods [2, 1], our single-step method trained with proposed STS can be hundreds of times faster. Compared with learning-based method, DetNet [6], the proposed synthetic-based method alleviates the reliance on the training data and achieves comparable results. We evaluate on three benchmarks. State-of-the-art results are achieved on CMU Panoptic [1]. The main contributions are: 1) a controllable synthetic method is proposed to create a high-diversity 3D motion space for training, which significantly alleviates the data reliance; 2) a decoupled framework is designed to facilitate individual learning and estimation of body parts at diverse scales; 3) promising pose accuracy and computational efficiency are achieved on relevant benchmarks.

2. RELATED WORK

3D Human Mesh Recovery. Many approaches regard this problem as recovering parameters of the statistical 3D human model, like SMPL [7], to reduce the complexity. Most recently, one-stage solutions [8, 9] are explored for more efficient estimation. More and more researchers [10] realize the importance of 2D poses for better generalization. In this work, we employ the 2D pose as an intermediate representation to connect the 3D space and 2D image plane.

3D Hands Mesh Recovery. Many methods for 3D hands mesh recovery also employ a statistical model, like MANO [4], to reduce complexity. Zhou et al. [11] develop a multi-stage framework to estimate MANO pose and shape parameters from the 3D pose position. However, most existing methods for 3D hands mesh recovery focus on estimating from the 2D high-resolution cropped images of hands, which is different from the blurred whole-body image we used.

Joint learning of multiple body parts recovery. Xiang et al. [1] use separate CNN networks for the body, hand, and face, and then jointly fits the Adam [12] to the outputs of all body parts using the optimization-based algorithm. SMPLify-X [2] iteratively fits the SMPL-X [2] to 2D keypoints of face, hands, and body. Those previous works are all based on the iterative optimization algorithm, which is time-consuming. Recent ConvNet-based method [6] first detect all body parts and then crop them out for separate mesh regression. Their two-stage framework is to estimate the mesh of each body part from the cropped image and suffered from limited paired 3D training data. We explore an effective synthetic training scheme to learn a robust model without data reliance.

3. METHOD

3.1. Background

SMPL-X [2] is a unified parametric model of full-body 3D human mesh. Statistical parameters of pose $\boldsymbol{\theta}$, shape $\boldsymbol{\beta}$, and expression $\boldsymbol{\psi}$ are used to control the human 3D mesh variations. An efficient mapping $M(\boldsymbol{\theta}, \boldsymbol{\beta}, \boldsymbol{\psi}; \Phi) : \mathbb{R}^{|\boldsymbol{\theta}|\times|\boldsymbol{\beta}|\times|\boldsymbol{\psi}|} \mapsto \mathbb{R}^{3\times N}$ is established to recover the 3D human mesh with $N = 10,475$ vertices, where Φ represents the statistical prior of human body mesh. The pose parameter $\boldsymbol{\theta} \in \mathbb{R}^{3\times(K+1)}$ represents the relative 3D rotation of $K = 54$ keypoints, including poses of body $\boldsymbol{\theta_b} \in \mathbb{R}^{3\times 21}$, each hand $\boldsymbol{\theta_h} \in \mathbb{R}^{3\times 15}$, jaw $\boldsymbol{\theta_j} \in \mathbb{R}^{3\times 1}$, each eye $\boldsymbol{\theta_e} \in \mathbb{R}^{3\times 1}$ and global rotation $\boldsymbol{\theta_{gr}} \in \mathbb{R}^{3\times 1}$. The shape parameter $\boldsymbol{\beta} \in \mathbb{R}^{10}$ represents the joint shape of body, face and hands. Besides, a linear regressor $\boldsymbol{P_{3D}}$ is employed to regress the 3D keypoints $\boldsymbol{J_{3D}} \in \mathbb{R}^{3\times K}$ from vertices of the human 3D mesh by

$$\boldsymbol{J_{3D}} = M(\boldsymbol{\theta}, \boldsymbol{\beta}, \boldsymbol{\psi}; \Phi)\boldsymbol{P_{3D}}. \quad (1)$$

Weak-perspective projection. To learn from in-the-wild 2D pose datasets, we need a weak-perspective camera model to project J_{3D} to 2D coordinates J^{p2D} on image plane. Given an un-calibrated image, we estimate the scale $s \in \mathbb{R}$ and 2D location $\boldsymbol{t} \in \mathbb{R}^2$ of each person on image. The 2D projection $\boldsymbol{J_i^{p2D}}$ of i-th 3D keypoints $\boldsymbol{J_i^{3D}}$ is

$$\boldsymbol{J_i^{p2D}} = s\Pi(\boldsymbol{R}\boldsymbol{J_i^{3D}}) + \boldsymbol{t}, \quad (2)$$

where Π is an orthographic projection operation.

3.2. Overview

As shown in Fig. 3, a two-branch framework is designed to predict the SMPL-X and camera parameters from a 2D image. Given images, we use a 2D pose detector (like OpenPose [13]) and the shape encoder to predict the whole-body

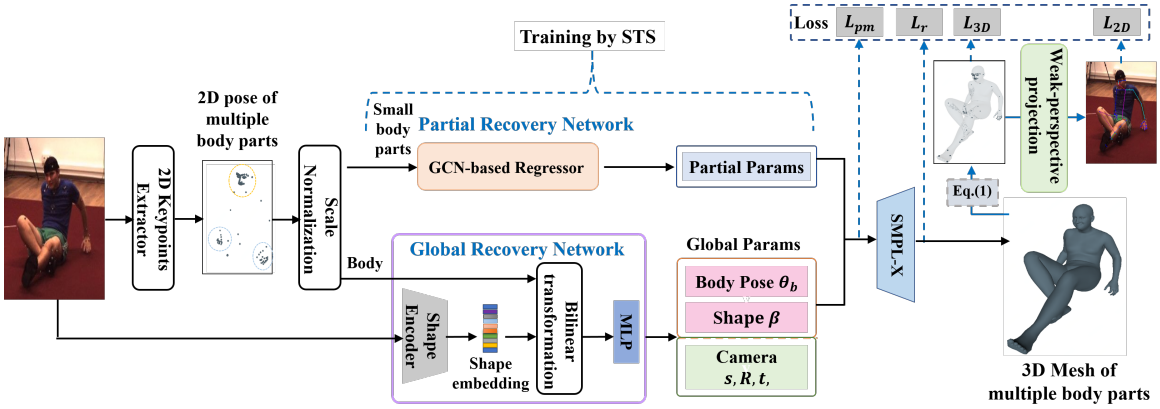

Fig. 3. Overview of the proposed single-step whole-body 3D mesh recovery framework. The network contains two branches for retrieving the global and partial parameters separately. The supervision is carried out on the estimated parameters and 2D/3D keypoints of the 3D mesh. We develop STS to train the GCN-based regressor for partial parameter regression.

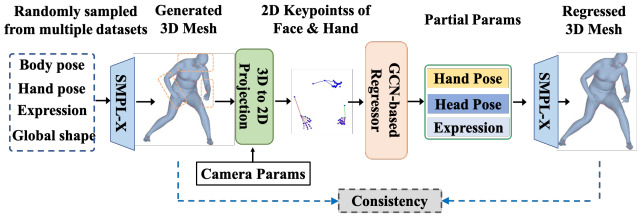

Fig. 4. The pipeline of the synthetic training scheme. Without full annotations, the small body part recovery sub-network can be trained by supervising the consistency between the first generated 3D mesh and the final regressed 3D mesh.

2D keypoints and shape embedding. To balance the scale differences between different body parts, we develop a scale normalization layer to transform the 2D keypoints of each body part into a standardized space. Specifically, keypoints of different parts are first translated to their center by subtracting the mean value and then scaled by dividing the half size of their bounding boxes. The re-scaled 2D keypoints are used to retrieve the global and partial parameters.

In global branch, with normalized 2D body keypoints and extracted shape embedding, we develop a global recovery sub-network to estimate global parameters, including the body shape $\boldsymbol{\beta}$, pose parameters $\boldsymbol{\theta_b}$, and the camera parameters $(s, \boldsymbol{t})$. In partial branch, since small body parts have the structure prior (e.g., symmetry, connection), we leverage the GCNs to explore their structural and spatial relations. With these estimated parameters $\boldsymbol{\Theta} = \{\boldsymbol{\theta}, \boldsymbol{\beta}, \boldsymbol{t}\}$ from the global and partial branches, we directly connect all these parameters and put into the SMPL-X model described in Sec.3.1 to obtain the final 3D human mesh.

3.3. STS: Synthetic Training Scheme

As shown in Fig. 4, STS is designed to generate reasonable training samples from unpaired 3D data and learn the small body parts recovery in a self-supervised manner. Firstly, we collect the 3D pose $\boldsymbol{\theta}$ and shape $\boldsymbol{\beta}$ parameters from these unpaired datasets by fitting the statistical models to establish a parameter bank. We combine multiple 3D datasets to collect real whole-body parameters. 3D body poses are obtained from the AMASS dataset [3]. It contains 300 subjects, more than 11000 motions. 3D hand poses are obtained from the MANO [4], Freihand [5], and AMASS [3]. Then, as shown in Fig. 4, we randomly sample these parameters from the bank to form the complete parameter vector $\boldsymbol{\Theta}$ and put it into the SMPL-X model to generate the whole-body 3D mesh. Next, we randomly sample the parameters in a pre-defined perspective camera space and use it to project the $\boldsymbol{J_{3D}}$ of 3D body mesh back to the 2D image plane. In this way, 2D pose along with its complete 3D annotations are generated. With these synthetic paired data, in (c) and (d), we exploit the GCN-based regression network for estimating small body parts from their 2D landmarks. Finally, the small body parts recovery sub-network can be trained by self-supervising their consistency.

Loss functions. The total loss $L_{\boldsymbol{\Theta}}$ is the weighted sum of the parameter loss L_{pm}, 3D keypoints loss L_{3D}, 2D projected keypoints loss L_{2D} and the rationality loss L_r where $w_{pm}, w_{3D}, w_{2D}, w_r$ are the weights of those losses. L_{pm} and L_{3D} are L_2 loss to supervise the estimated parameters and $\boldsymbol{J_{3D}}$ of all body parts. L_{2D} is L_1 loss to supervise the projected 2D pose $\boldsymbol{J^{p2D}}$. L_r is a multivariate Gaussian prior to supervise the rationality of the keypoint angles to prevent physically impossible 3D human mesh.

4. EXPERIMENTS

4.1. Implementation Details

For training, we employ four 3D datasets (Human3.6M, MPI-INF-3DHP [14], CMU Panoptic, and STB Hand Dataset [15]) and two in-the-wild 2D pose datasets (AI-CH [16] and MPII [17]). We evaluate on three benchmarks, CMU Panoptic, Human3.6M, and STB Hand Dataset. Following [1], on CMU Panoptic, we compute Mean Per-Joint Position Error (MPJPE) and MPJPE after Procrustes Alignment (PA-MPJPE). On STB Hand Dataset, we follow the evaluation protocol of [18].

Table 1. 3D body pose evaluation on Panoptic Studio Dataset.

Method	MPJPE(↓)	PA-MPJPE(↓)
MTC [1]	63.0	–
DetNet [6]	66.8	61.5
Ours	**57.5**	**41.3**

Table 2. 3D body pose evaluation on Human3.6M dataset.

Method	MPJPE(↓)	PA-MPJPE(↓)
HMR [21]	87.9	58.1
GCMR [22]	74.7	51.9
MTC [1]	58.3	-
SMPLify-X [2]	-	75.9
Ours	67.4	52.0

Considering that only a part of the data has 3D hand pose annotations, the STS is used one time in every 5 iterations. The Adam [19] optimizer is adopted with betas=(0.9,0.999), momentum=0.9. The learning rate and batch size are set to 1e-4 and 16 respectively. The weights of loss items are set as $w_{pm} = 20, w_r = 0.5, w_{2D} = 6, w_{3D} = 60$.

4.2. Comparisons to the State-of-the-arts

CMU Panoptic. MTC deals with the same task as ours, while they independently estimate each human part via combining CNNs and optimization algorithms. As shown in Tab. 1, our proposed method outperforms MTC by 8.7% in terms of MPJPE. It indicates the superiority of our single-step learning-based framework over the multi-step optimization-based method. Our proposed method outperforms learning-based DetNet [6] by 13.9% and 32.8% in terms of MPJPE and PA-MPJPE, which demonstrates the superiority of our disentangled synthetic-based framework.

Human3.6M. As shown in Tab. 2, our method achieves competitive performance. It demonstrates that we preserve promising performance of estimating 3D body pose, meanwhile, accomplish 3D mesh recovery of multiple body parts.

STB Hand Dataset. As shown in Tab. 3, our method can achieve comparable performance with GraphHand [18] which is the state-of-the-art method specially designed for hand recovery only. It demonstrates that our framework can well deal with the partial part recovery with the whole-body setting.

Efficiency analysis. In Tab.4, we compare the computational efficiency with SMPLify-X [2] and MTC [1]. For a fair comparison, we use the released official code and test on the same dataset [20] on the same device (Tesla P40 GPU). Considering that both SMPLify-X and our method employ the OpenPose to estimate the whole-body 2D keypoints, we only compare the time consuming of 3D mesh recovery. As shown in Tab. 4, the proposed method is hundreds of times faster than SMPLify-X and MTC. Our single-step design enables us to enjoy parallel computation, which greatly accelerates the calculation.

Table 3. Evaluation on the **STB hand dataset**. ⋆ stands for using the ground truth 2D hand keypoints as input.

Method	GraphHand [18]	GR	GR+STS	GR+STS⋆
MPJPE (↓)	11.3	13.1	12.1	9.1

Table 4. Comparisons to state-of-the-art methods on computational efficiency.

Method	SMPLify-X [2]	MTC [1]	**Ours**
Avg. runtime (s)	12	82	**0.09**

Fig. 5. Qualitative results on internet images, Freihand [5], and STB hand dataset [15].

4.3. Ablation Study of STS

When directly training the GCN-based regressor (GR) on the STB dataset in a pure supervised manner, the performance (GR in Tab. 3) is unsatisfactory. With the help of the proposed STS, 3D hand pose accuracy (GR+STS in Tab. 3) gets improved by 7.6%. This improvement indicates that exploring 3D hand motion space with the proposed STS is helpful for purely supervised training. Besides, when using ground truth 2D pose as input, the performance gets greatly improved, which demonstrates that we can achieve better performance with a better 2D pose estimator.

5. CONCLUSION

In this paper, we present a synthesis-based framework for learning a single-step regression model that jointly recovers the mesh of multiple 3D body parts from a single RGB image. Compared with the previous optimization-based method, the proposed method shows an obvious advantage in computational efficiency. The proposed synthetic training scheme greatly alleviate the reliance on training data for learning-based frameworks. To deal with the large-scale difference among different body parts, the network architecture is designed in a disentangled manner. Experiments with ablation analysis show that our proposed method achieves promising performance on relevant benchmarks.

6. REFERENCES

[1] Donglai Xiang, Hanbyul Joo, and Yaser Sheikh, "Monocular total capture: Posing face, body, and hands in the wild," in *CVPR*, 2019.

[2] 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 *CVPR*, 2019.

[3] Naureen Mahmood, Nima Ghorbani, Nikolaus F. Troje, Gerard Pons-Moll, and Michael J. Black, "AMASS: Archive of motion capture as surface shapes," in *ICCV*, 2019.

[4] Javier Romero, Dimitrios Tzionas, and Michael J. Black, "Embodied hands: Modeling and capturing hands and bodies together," *ACM Transactions on Graphics*, vol. 36, pp. 1–17, 2017.

[5] Zimmermann Christian, Ceylan Duygu, Yang Jimei, Russel Bryan, Argus Max, and Brox Thomas, "Freihand: A dataset for markerless capture of hand pose and shape from single rgb images," in *ICCV*, 2019.

[6] Yuxiao Zhou, Marc Habermann, Ikhsanul Habibie, Ayush Tewari, Christian Theobalt, and Feng Xu, "Monocular real-time full body capture with inter-part correlations," in *CVPR*, 2021.

[7] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J. Black, "SMPL: A skinned multi-person linear model," *ACM Transactions on Graphics*, 2015.

[8] Yu Sun, Qian Bao, Wu Liu, Yili Fu, Michael J. Black, and Tao Mei, "Monocular, one-stage, regression of multiple 3d people," in *ICCV*, 2021.

[9] Yu Sun, Wu Liu, Qian Bao, Yili Fu, Tao Mei, and Michael J Black, "Putting people in their place: Monocular regression of 3d people in depth," *arXiv preprint arXiv:2112.08274*, 2021.

[10] 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 *ICCV*, 2019.

[11] Yuxiao Zhou, Marc Habermann, Weipeng Xu, Ikhsanul Habibie, Christian Theobalt, and Feng Xu, "Monocular real-time hand shape and motion capture using multi-modal data," in *CVPR*, 2020.

[12] Hanbyul Joo, Tomas Simon, and Yaser Sheikh, "Total capture: A 3d deformation model for tracking faces, hands, and bodies," in *CVPR*, 2018.

[13] Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh, "Realtime multi-person 2d pose estimation using part affinity fields," in *CVPR*, 2017.

[14] Dushyant Mehta, Helge Rhodin, Dan Casas, Pascal Fua, Oleksandr Sotnychenko, Weipeng Xu, and Christian Theobalt, "Monocular 3d human pose estimation in the wild using improved cnn supervision," in *3DV*, 2017.

[15] Jiawei Zhang, Jianbo Jiao, Mingliang Chen, Liangqiong Qu, Xiaobin Xu, and Qingxiong Yang, "3d hand pose tracking and estimation using stereo matching," *arXiv:1610.07214*, 2016.

[16] Jiahong Wu, He Zheng, Bo Zhao, Yixin Li, Baoming Yan, Rui Liang, Wenjia Wang, Shipei Zhou, Guosen Lin, Yanwei Fu, et al., "Ai challenger: A large-scale dataset for going deeper in image understanding," *arXiv*, 2017.

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

[18] 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 *CVPR*, 2019.

[19] Diederik P. Kingma and Jimmy Ba, "Adam: A method for stochastic optimization," in *ICLR*, 2015.

[20] Raviteja Vemulapalli and Aseem Agarwala, "A compact embedding for facial expression similarity," in *CVPR*, 2019.

[21] Angjoo Kanazawa, Michael J. Black, David W. Jacobs, and Jitendra Malik, "End-to-end recovery of human shape and pose," in *CVPR*, 2018.

[22] Nikos Kolotouros, Georgios Pavlakos, and Kostas Daniilidis, "Convolutional mesh regression for single-image human shape reconstruction," in *CVPR*, 2019.