Generating Multiple Hypotheses for 3D Human Mesh and Pose using Conditional Generative Adversarial Nets

Xu Zheng, Yali Zheng[0000−0002−2906−7984], Shubing Yang

School of Automation Engineering,
University of Electronic and Scientific Technology of China, Chengdu, China
zhengyl@uestc.edu.cn

Abstract. Despite recent successes in 3D human mesh/pose recovery, the human mesh/pose reconstruction ambiguity is a challenging problem that can not be avoided as lighting, occlusion or self-occlusion in scenes happens. We argue that there could be multiple 3D human meshes corresponding a single image from a view point, because we really do not know what happens in extreme lighting or behind occlusion/self occlusion. In this paper, we address the problem using Conditional Generative Adversarial Nets (CGANs) to generate multiple hypotheses for 3D human mesh and pose from a single image under the condition of 2D joints and relative depth of adjacent joints. The initial estimation of 2D human skeletons, relative depth and features is taken as input of CGANs to train the generator and discriminator in the first stage. Then the generator of CGANs is used to generate multiple human meshes via different conditions which are consistent with human silhouette and 2D joint points in the second stage. Selecting and clustering are utilized to eliminate abnormal and redundant human meshes. The number of hypothesis is not unified for each single image, and it is dependent on 2D pose ambiguity. Unlike the existing end-to-end 3D human mesh recovery methods, our approach consists of three task-specific deep networks trained separately to mitigate the training burden in terms of time and datasets. Our approach has been evaluated not only on the datasets of laboratory and real scenes but also on Internet images qualitatively and quantitatively, and experimental results demonstrate the effectiveness of our approach.

Keywords: Human mesh · CGAN · Multiple Hypotheses.

1 Introduction

Recovering 3D human mesh and pose from images is a fundamental problem in the field of computer vision. It is challenging to reconstruct 3D human mesh accurately and robustly from a single image, which has drawn a lot of attention from researchers. Essentially this is an ill-posed problem whether using traditional 3D geometric methods or using deep learning methods to restore. The number of geometric constraints is less than the number of unknown variables, which causes the uncertainty of 3D human mesh and pose reconstruction. Further, as extreme lighting, occlusion or self-occlusion in real scenes happens, it exacerbates the uncertainty [51, 10]. Most of the existing methods are only able to recover one plausible human mesh from a single image, however, there

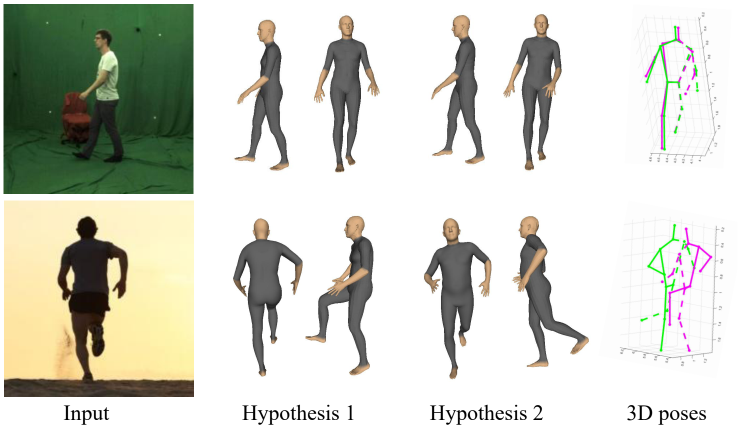

Fig. 1. Our approach takes a single image as input, and recovers multiple diverse hypotheses of 3D human mesh and pose. Here two hypotheses for human mesh are shown in the picture in two different views. (Best view in color)

exists multiple possible reconstruction for a number of images, see the first column in Fig. 1, that is one image leads to multiple reconstructions.

Generative Adversarial Nets are the popular frameworks to train generative model in an alternative manner, however, they suffer from the *mode collapse* easily, and they are hard to train. So Mirza and Osindero proposed CGANs, and conditioning could be class labels, some part of data or even data from different modality [29]. CGANs have many benefits. They can generate models under the control of condition, and speed the convergence of model training. This property of CGANs is quite suitable to solve the problem of multiple hypotheses reconstruction for human mesh and pose. In this paper, we try to generate multiple hypotheses for reasonable 3D human mesh via CGANs which are consistent with human silhouette and 2D joint points detected from images, see Fig. 1.

Unlike the existing end-to-end methods for 3D human mesh reconstruction, we propose a two-stage approach to generate multiple hypotheses of 3D human mesh under the CGANs framework. In the first stage, human silhouette, 2D joints and features are detected automatically, and CGANs are trained to generate human mesh bases in an adversarial manner. Since 2D joints and relative depth may be estimated incorrectly due to the extreme light or occlusion/self-occlusion, so we expand the conditions to control the generation of 3D human meshes and poses. So in the second stage, limited hypotheses for 3D human mesh and pose are produced by varying the conditions of CGANs due to being tolerant to detection ambiguity, then clustered in a cascade pipeline. Here, the conditions include 2D joints and the relative depth of adjacent joints. One of the benefits of our system is that the background influence is removed since only human silhouette, 2D joints and relative depth of 2D joints are taken as input for CGANs, it is easily generalized to different scenarios. Our contributions include:

1) We propose a two-stage weakly-supervised approach to address the problem of multiple hypotheses for human mesh and pose by CGANs, which only takes human silhouette as input under the constraints of 2D joints and relative depth without any 3D supervision.

2) We generate multiple 3D human body meshes by importing limited and discrete hypotheses with varying conditions of CGANs and clustering, instead of sampling from a probability distribution. In our approach, the number of conditions is limited and discrete, since the hypotheses space is limited and discrete.

3) We evaluate throughout the experiments that our system is suitable for multiple hypotheses generation for 3D human mesh and pose not only from image datasets of the laboratory and real scenarios, but also from Internet images without networks retraining.

2 Related work

3D mesh reconstruction The reconstruction methods of 3D human body mesh from a single image have made great progress in recent years along with the development of deep learning technique. These methods can be roughly divided into two categories: one is the human body shape reconstruction method based on parametric models [2, 25, 37, 13], the other is the human body shape reconstruction methods based on non-parameters [43, 19, 9, 51, 41]. The former methods try to encode a human mesh into a parametric model. The advantage of this type of methods is that it is easy to reconstruct the complete shape of human body from a single image, even if occlusion happens in the image. The disadvantage is that parametric models suffer from the ability of limited detail representation. A human shape is only represented as a linear combination of 10 shape bases, so details of the human shape may be recovered insufficiently. The model-free methods describe the details better, but require to learn the large amount of variables.

Model-based mesh reconstruction A number of human mesh models have been proposed in the field, such as SCAPE [2], Skinned Multi-Person Linear model (SMPL) [25], ADAM [13] and so on[31, 37, 52]. The most popular SMPL model [25] has 82-dimensional parameters, including 10 shape base parameters and 24×3 joint rotation parameters, and it has been extended into SMPL-H [37], SMPLify [3] and SMPLify-X [31]. SMPLify, a multi-stage optimization method, was proposed in [3], which used DeepCut [36] to estimate 2D pose first, and then optimized reconstruct results. HMR[14] used an end-to-end CNN network to regress SMPL parameters from RGB images with discriminator constraints. TexturePose [32] took texture information to enhance CNN network ability. SPIN [18] combined a learning-based algorithm and optimization algorithm into an iterative framework to obtain SMPL parameters, and achieved state-of-the-art results among the methods of model-based human mesh recovering. SMPLify-X [31] extended the SMPL-X model, which had a human mesh with hands and face, and took 2D human pose to optimize with prior constraints and penalty of mesh collision. MTC [48] optimized ADAM [13] by 3D pose direction vector and 2D pose estimating by CNN-based network.

Model-free mesh reconstruction The model-free shape recovery methods consider a human body consisting of voxel volumes, and learn the body surface directly. BodyNet [43] presented an end-to-end CNN based network to estimate volumetric representation of human with 2D pose, depth and 3D pose. Kolotouros et al. proposed that taking CNN and graph neural network was significantly easy to regress 3D location of human mesh [19]. Gabeur et al. estimated the “visible” and “hidden” depth maps, and combined into a full-body 3D point cloud [9]. Zhu et al. [51] proposed a hierarchical method to capture the detailed human body. Tan et al. presented a self-supervised method to relax the dependance on ground truth data from videos [41].

3D pose estimation While the problem of 2D human pose estimation has been well solved, 3D human pose estimation is still challenging due to 3D reconstruction ambiguity caused by the variation of viewpoint, human body and clothing. Deep learning technique is a popular way to estimate 3D pose from a single image, and these approaches achieve good results. Pavlakos et al. [34] applied the stack hourglass network to estimate every voxel likelihood for each joint. Mehta et al. [28] argued that the algorithm generalizability constrained by available 3D pose datasets, and proposed a benchmark “MPI-INF-3DHP” (MPII-3D for short in this paper) which covered outdoor and indoor scenes. Zhou et al. explored the problem of 3D pose estimation in the wild, and proposed a weakly-supervised transfer learning framework due to the lack of training data [50]. Kacabas et al. proposed EpipolarPose to estimate 3D human poses and camera matrix with the constraints of epipolar geometry [17]. Jahangiri and Yuille addressed the problem to estimate multiple diverse 3D poses in [10], and started initial 3D pose to generate multiple hypotheses according to prior sampling, while Li and Lee [21] generated multiple hypotheses for 3D human pose based on a mixture density network. As transformer technique arises, it is regarded as a powerful backbone in vision field. Zheng et al. [49] and Li et al. [22] proposed to estimate human pose via transformer, and achieve big approvement. However, transformer-based methods always have a big model, which are resource-consuming and have high requirements on GPU, while our approach has less training variables in weakly-supervised way.

3 The proposed approach

In this section, we introduce our proposed weakly-supervised approach in detail. We give the work flow firstly, then explain how to estimate human mesh and pose using CGANs, and how to generate the multiple hypotheses by CGANs by selecting and clustering. Finally we give the implementation details of each parts.

3.1 Overview

We present a cascade framework to recover multiple hypotheses for 3D human shapes that are consistent with the 2D human silhouette and 2D joints heat map extracted from images. As shown in Fig. 2, there are two stages in our pipeline, consisting of three parts $\mathcal{N}_1, \mathcal{N}_2, \mathcal{N}_3$:

$\mathcal{N}_1$: an encoder-decoder structure, takes the original image as input to regress 30 2D joints heat map $\mathcal{F}_J$ and human silhouette S, and learns features $\mathcal{F}$ (Note that 30 joints

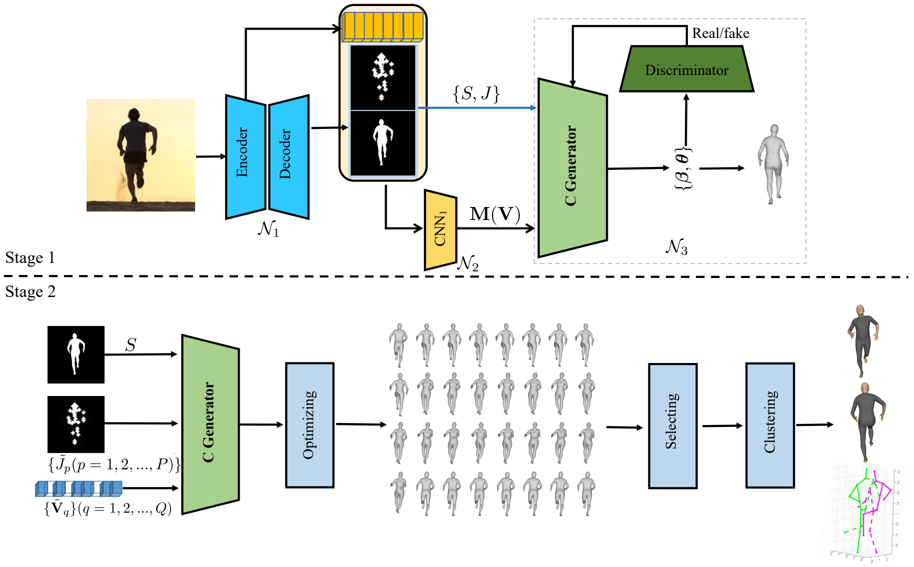

Fig. 2. Schematic diagram of the proposed approach. There are two stages in our pipeline, consisting of three parts: the encoder-decoder $\mathcal{N}_1$ to regress 2D human silhouette and 2D joint heat map, the CNN $\mathcal{N}_2$ to estimate a relative depth vector, the CGANs $\mathcal{N}_3$ to generate human mesh. Firstly 2D human joints, human silhouette learnt by $\mathcal{N}_1$, and the relative depth vector learnt by $\mathcal{N}_2$ are taken as the inputs to train the generator and discriminator of CGANs $\mathcal{N}_3$. In the second stage, multiple hypotheses for human mesh are generated by rationally expanding the relative depth vectors $\{\tilde{\mathbf{V}}_q\}$ and 2D joints sets $\{\tilde{J}_p\}$. Then abnormal 3D human meshes are removed by selecting steps, and the close human meshes are clustered to output the final mesh and pose hypotheses.

including 24 points of SMPL model, 5 points for one nose, two eyes, two ears, and 1 point on head).

$\mathcal{N}_2$: a CNN, takes 2D human silhouette S, joint heat map $\mathcal{F}_J$ and features $\mathcal{F}$ learnt by $\mathcal{N}_1$ as input to estimate a relative depth matrix $\mathbf{M}$. $\mathbf{V}$ is a binary vector form of $\mathbf{M}$, which only has the relative depth between adjacent joints.

$\mathcal{N}_3$: the CGANs, consisting of a conditional generator and a discriminator. The generator takes feature maps of joints, heat map $\mathcal{F}_J$, human silhouette S and the relative depth vector $\mathbf{V}$ to regress the parameters $\boldsymbol{\beta}$ and $\boldsymbol{\theta}$ of SMPL model to generate human mesh. The discriminator distinguishes meshes from $\boldsymbol{\beta}$ and $\boldsymbol{\theta}$ of SMPL model as real or fake.

In the first stage, 2D human joint heat map and human silhouette learnt by $\mathcal{N}_1$ and the relative depth vector learnt by $\mathcal{N}_2$ are taken as inputs to train the generator and discriminator of CGANs $\mathcal{N}_3$. In the second stage, multiple hypotheses for human meshes are generated reasonably by expanding the relative depth vector and the 2D skeleton set. And the optimization is following to make human meshes better. Then abnormal 3D human meshes are rejected by selecting step after discarding human meshes with crossing through parts, and the close human meshes are clustered to output the final mesh hypotheses after detail refining. The whole pipeline is shown in Fig. 2.

Since 2D joints and silhouette for humans are well studied, and a number of 2D skeleton and human segment datasets are published by different institutions, so it is not difficult to learn $\mathcal{N}_1$ and $\mathcal{N}_2$. The implementation details of $\mathcal{N}_1$ and $\mathcal{N}_2$ will be described in Section 3.5. Once $\mathcal{N}_1$ is trained to learn feature map of 2D joints $\mathcal{F}_J$, the silhouette image S, and a 2048D feature vector $\mathcal{F}$, the location set of joints denoted by J is derived by integrating over the corresponding feature map of joints $\mathcal{F}_J$ [39]. $\mathcal{F}_J$, S and $\mathcal{F}$ are taken, as input, to train $\mathcal{N}_2$ to estimate the matrix $\mathbf{M}$ of relative depths. Each element between $[0, 1]$ in $\mathbf{M}_{ij}$, can be thought of as a probability of ith joint farther than jth joint in the camera coordinate [33]. Let $\mathbf{M'}_{ij}$ be a sparse binary matrix, in which $0/1$ represents that the ith joint farther/closer than the jth joint in the camera coordinate. In the following subsections, the key parts are introduced, including human mesh estimation via CGANs, multiple hypotheses generation, and human mesh selecting and clustering.

3.2 Human mesh estimation using CGANs

In our CGANs, human mesh is encoded by SMPL parametric model, and the model parameters – $\boldsymbol{\beta}$ (shape) and $\boldsymbol{\theta}$ (pose) are learnt to represent human mesh from human silhouette S constrained by J and $\mathbf{M}'$. The discriminator distinguishes human meshes from $\boldsymbol{\beta}$ and $\boldsymbol{\theta}$ as real or fake. The modified objective function for CGANs is minimized as follows,

$$\min_{\mathcal{G}} \max_{\mathcal{D}} L(\mathcal{D}, \mathcal{G}) = E[\log \mathcal{D}(\boldsymbol{\beta}, \boldsymbol{\theta})] + E[\log(1 - \mathcal{D}(\mathcal{G}(S|J, \mathbf{M}')))] \tag{1}$$

In the formula, the generator $\mathcal{G}$ is conditioned by 2D joints J and the relative depth $\mathbf{M}'$, while $\mathcal{D}$ is a classifier learnt by taking samples from $(\boldsymbol{\beta}, \boldsymbol{\theta})$ generated by $\mathcal{G}$ as negative samples, and samples from Mosh dataset as positive ones. The reason that we do not take condition constraint on discriminator $\mathcal{D}$ is because we found it is sufficient to take $(\boldsymbol{\beta}, \boldsymbol{\theta})$ for classifying real or fake one. The generator is built on ResNet-18, and the discriminator is just a multi-layer perceptron net. Please refer the details in Section 3.5. Then the generator $\mathcal{G}$ and the discriminator $\mathcal{D}$ are trained under the GANs framework in an alternative manner. As the discriminator becomes stronger, the generator gets stronger. The generator is trained under the control of J and $\mathbf{M}'$ in weak supervision of 2D joints J and the human silhouette S, and the generator enables us to generate multiple hypotheses by varying the conditions.

Further, the loss for the conditional generator is minimized over $\boldsymbol{\beta}$, $\boldsymbol{\theta}$ as follows:

$$L_{\mathcal{G}}(\boldsymbol{\beta}, \boldsymbol{\theta}) = L_{2D} + L_{dep} + L_{reg} \tag{2}$$

The $L_{\mathcal{G}}$ consists of three terms, 2D constraint loss L_{2D}, the relative depth constraint loss L_{dep}, and the regularization term L_{reg}. They have to be balanced in the training.

Each term is defined as follows,

$$
\begin{aligned}
L_{2D} &= ||\mathfrak{J}(\mathcal{M}(\boldsymbol{\beta},\boldsymbol{\theta})) - J_{2D}||_2 + ||\mathfrak{S}(\mathcal{M}(\boldsymbol{\beta},\boldsymbol{\theta})) - S||_2 \\
L_{dep} &= \sum_{l=1}^{|J_{2D}|} \sum_{J_k \in \mathbb{N}(J_l)} max((J_l - J_k)\cdot(2\cdot \mathbf{M'}_{l,k} - 1), 0) \\
L_{reg} &= \gamma_{\boldsymbol{\beta}}||\boldsymbol{\beta}||_2 + \gamma_{\boldsymbol{\theta}}||\boldsymbol{\theta}||_2
\end{aligned}
\qquad (3)
$$

In which, $\mathcal{M}$ is an operation of recovering a mesh from $(\boldsymbol{\beta},\boldsymbol{\theta})$, $\mathfrak{J}$ is an operation to project 3D joints of mesh into 2D, and $\mathfrak{S}$ is an operation to render a silhouette binary image from the generated mesh [16]. $\mathbb{N}$ denotes the neighbourhood set of a joint, since only the adjacent joints are considered in our approach. $\mathbf{M}'_{l,k}$ in L_{dep} estimated in the first stage, which is 1 or 0, would constrain the joint J_l is closer or farther than the joint J_k. And $\gamma_{\boldsymbol{\beta}}$ and $\gamma_{\boldsymbol{\theta}}$ are the balance factors in L_{reg}. Once CGANs are trained in the first stage, the generator is exploited to produce multiple possible human meshes and poses in the second stage.

3.3 Condition expansion for multiple hypotheses

Now we explain how to generate multiple hypotheses for human mesh and pose from a single image. We only care about the relative depth between adjacent joints, and transform matrix $\mathbf{M}'$ into a vector form $\mathbf{V}$. Each element in $\mathbf{V}$ represents the relative depth relationship between to adjacent joints, and a limited space is generated with the relative depth vector $\mathbf{V}_q$ $(q = 1, 2, ..., Q)$ and 2D joints sets J_p $(p = 1, 2, ..., P, P <= 4)$ to help multiple human mesh and pose recovery. Since the relative depth estimation and 2D joints detection are not always correct, so it should be tolerant reasonably. As has been observed, each element in $\mathbf{M}_{ij}$ describes an estimate confidence of ith joint farther than jth joint in the camera coordinates. When $\mathbf{M}_{ij}$ is close to 1, which means the ith joint is more likely to be closer than the jth one, and when $\mathbf{M}_{ij}$ is close to 0, which means the ith joint is more likely to be farther than the jth one. When $\mathbf{M}_{ij}$ is near around 0.5, it is ambiguous to determine which one is closer. Let δ be an ambiguity capacity factor, then we have

$$
\tilde{\mathbf{M}}_{ij} = \begin{cases} 0, & \mathbf{M}_{ij} \in [0, 0.5-\delta), \\ 0 \; or \; 1, & \mathbf{M}_{ij} \in [0.5-\delta, 0.5+\delta], \\ 1, & \mathbf{M}_{ij} \in (0.5+\delta, 1]. \end{cases} \qquad (4)
$$

$\tilde{\mathbf{M}}$ is also transformed into a vector form $\mathbf{V}$. Assume that $\tilde{\mathbf{M}}$ has Q possible values when $\mathbf{M}_{ij}(\in [0.5-\delta, 0.5+\delta])$, then it leads to $\tilde{\mathbf{V}}_q, q = 1, 2, ..., Q$. Note that we only care about the relative depth between the adjacent joints in the vector. In this way, a single $\mathbf{V}$ in the first stage is expanded into a set space $\{\tilde{\mathbf{V}}_1, \tilde{\mathbf{V}}_2, ..., \tilde{\mathbf{V}}_Q\}$, which makes the relative depth be tolerant to a certain estimation error.

Further, the initial estimation of 2D pose J shows much ambiguous as lighting, occlusion or self-occlusion as well. The human mesh may be confused with the body of the right and left parts due to the fact that a human body is a symmetrical object.

There are three extra cases: 1) the right leg is confused with the left leg; 2) the right arm is confused with the left arm; 3) the right body is confused with the left body. So the initial estimation of 2D joints is expanded into a space of 4 cases, that is $\{\tilde{J}_1, ..., \tilde{J}_P\}$ and $P = 4$, 2D joints of right leg is flipped with ones of left leg, the right arm is flipped with joints on the left arm. However, we know if a human to be recovered is facing forward or unknown by the detection from the relative depth between eyes and ears. If the human faces forward confidently, then one of four cases is ignored, that is $\{\tilde{J}_1, ..., \tilde{J}_P\}$ and $P = 3$. The expansions of relative depth and 2D skeleton are taken as the reconstruction condition for the conditional generator to produce multiple hypotheses for human mesh, which enable our approach be tolerant to the estimation error and be resistant to the ambiguity of body structure.

All generated human meshes are optimized, and the objective function is the same as Equ. 2, minimized over $\boldsymbol{\beta}$ and $\boldsymbol{\theta}$ by gradient descend as follows,

$$\min_{\boldsymbol{\beta},\boldsymbol{\theta}} \quad L_{\mathcal{G}} \tag{5}$$

The optimized $\boldsymbol{\beta}$ and $\boldsymbol{\theta}$ are used to generate more accurate human mesh.

3.4 Human mesh selecting and clustering

By varying $\{\tilde{\mathbf{V}}_1, ..., \tilde{\mathbf{V}}_Q\}$ and $\{\tilde{J}_1, ..., \tilde{J}_P\}$, QP 3D human meshes (each one has 6890 $3D$ points) are generated in the previous subsection. Obviously, besides of the positive effect of withstanding generation ambiguity, the expansions may cause some redundancy, and generate abnormal meshes.

The abnormal meshes include crossing through parts, abnormal posture and so on. Crossing through parts would be rejected partially by body collision detection [31]. Also the rest meshes have to be classified by classifier in [4], and abnormal posture would be rejected and removed from the human mesh set. Assume that N human body meshes are passed through the selection as normal, and they are clustered using mean-shift, each category has N_k human meshes, $k = 1, 2, ..., K$. K is the number of classes from mean-shift clustering. Obviously, the method of mean-shift is not necessary to cluster samples into a fixed number of class. Take the one with the least projection error of 2D silhouette and 2D human joints to represent each class, the objective function is as follows:

$$\min_{n_k \in [1,...,N_k]} \quad ||\mathfrak{J}(\mathcal{M}_{n_k}) - J||_2 + \gamma||\mathfrak{S}(\mathcal{M}_{n_k}) - S||_2 \tag{6}$$

where γ is a balance number, and need to be set in advance. Once K human meshes are clustered, they output the 3D human poses by determining 3D joints from human mesh.

3.5 Implementation details

Four different parts $\mathcal{N}_1, \mathcal{N}_2, \mathcal{N}_3$ are trained separately. The $\mathcal{N}_1$ takes the pre-trained ResNet-50 as the encoder, and 9 layers of deconvolution as the decoder to regress the 2D silhouettes/segments S and feature maps of 2D human joints $\mathcal{F}_J$. ResNet-18 is used in the second CNN $\mathcal{N}_2$ to predict the relative depth of adjacent human joints $\mathbf{V}$ and the distance from the camera T. The conditional generator, a modified ResNet, is a stack of

ResNet-18 and four fully connected layers, and end up with a regressor like in [18]. The discriminator, which is a multi-layer perceptron based network with 1024 neurons. The generator and discriminator in $\mathcal{N}_3$ are trained alternatively by minimizing the objective Equ. 2. In order to overcome the challenge of unbalanced laboratory and real scene datasets, the similar number of images are sampled from the laboratory dataset and the real scene dataset in each epoch as [6]. In the optimization step, each optimizing only run ten times. δ is set as 0.2 through all experiments. The learning rate for conditional generator is $1e-4$, and learning rate for discriminator is $2e-4$. All networks are trained with GPU of RTX TITAN.

4 Experimental results

4.1 Datasets

A variety of datasets are assembled to train different deep networks in our experiments. All 2D pose human datasets are all made to be consistent with our 30 points human skeleton, the occlusion ones will be ignored. LSP [11] and LSPET [12] are 2D human pose datasets of 14 joint points, while MPII [1] is a 2D human pose dataset of 16 joint points. The UP [20] and COCO [23] dataset consist of 2D human pose and 2D segmentation. MTC [45] and MPII-3D [28] are laboratory multi-view 3D pose datasets, while the former has a large range of viewing angle, and the latter has more than 1.3M images in total. The Mosh [24] is a synthetic 3D human mesh dataset, while SURREAL [44] renders the original human meshes in Mosh dataset into real scenes. 3DPW [26] is a 3D pose dataset captured from real scenes. 2D pose datasets are utilized to train 2D human skeleton detection. Because our method is a weakly-supervised algorithm, we only use 2D pose and silhouette to weakly supervise the training process of $\mathcal{N}_3$.

LSP, LSPET, MPII, COCO, MPII-3D, MTC, SURREAL, 3DPW datasets are used to train $\mathcal{N}_1$. MPII-3D, MTC, SURREAL, 3DPW datasets are used to train $\mathcal{N}_2$. The datasets including LSP, LSPET, MPII-3D, SURREAL, 3DPW are used for $\mathcal{N}_3$. Although MPII-3D, MTC, SURREAL, 3DPW are both 2D and 3D pose datasets, we do not use 3D pose but 2D pose only through all experiments.

4.2 Qualitative results

We not only reconstruct multiple human meshes from a single image, but also recover multiple 3D human poses. Fig. 3 shows more qualitative results on challenging images from the laboratory dataset (MPII-3D), real scene datasets (LSP, LSPET, MPII), and Fig. 4 shows more results on Internet images. The first column shows the original input images, the second column shows an example of 2D projection of 3D human mesh and detected 2D pose. Four hypotheses of human meshes in two different view generated from our approach are shown from the third to tenth column. The last two columns show all 3D human poses in the same coordinates to compare the difference between poses in two different views. The dash line represents the right part of a human body, the solid line represents the left part of a human body.

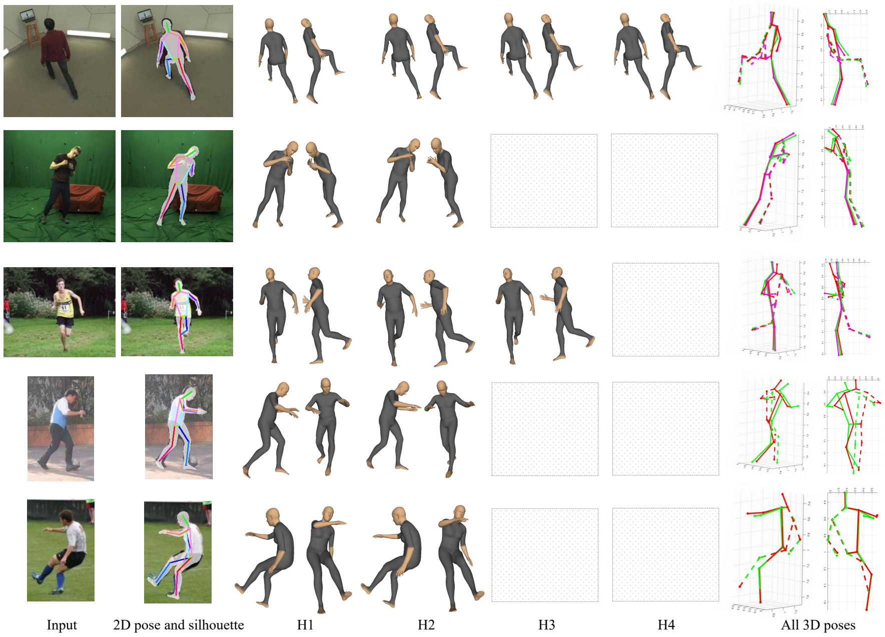

Fig. 3. Results from images of laboratory and real scenes dataset. The first column shows the original input images, the second column shows an example of 2D projection of 3D human mesh and 2D pose. The hypotheses for human mesh and 3D pose generated from our approach are shown from the third to tenth column in two different view. The last two columns show all 3D human poses in the same coordinates to compare the difference between poses in two different views. The first image has four hypotheses generation of human mesh, the third image has three hypotheses generation, while the second, fourth and fifth have two hypotheses generation.

4.3 Statistic analysis of multiple hypotheses

We also do statistic analysis on LSP, MPII-3D and 3DPW datasets, and count the number of generation hypotheses for all images in these datasets. Statistic analysis is reported in Fig. 5. Most of images have 2 human mesh hypotheses generated by our method on all datasets. More than 85% images have $1-3$ hypotheses, and few images (about 2% or less) have more than 5 hypotheses. In order to show the statistic figure better, the number of images in MPII-3D and 3DPW is reduced to one tenth just for better viewing. We investigate that these images with more than 5 hypotheses are often caused by the low-quality generation for human mesh results from the generator, which have extreme large 2D projection error.

4.4 Quantitative comparison

Our approach is evaluated quantitatively with respect to errors of human mesh reprojection and 3D joints location on three datasets, which are popular ways in the field. Since our method has generated multiple human meshes and 3D poses, the best one

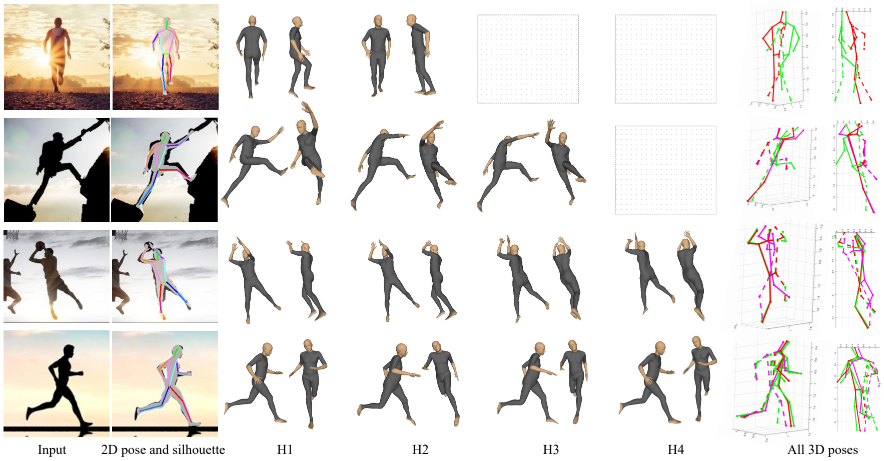

Fig. 4. Results from Internet images. The first column shows the original input images, the second column shows an example of 2D projection of 3D human mesh and 2D pose. The hypotheses for human mesh and 3D pose generated from our approach are shown from the third to tenth column in two different view. The last two columns show all 3D human poses in the same coordinates to compare the difference between poses in two different views. The first image has two hypotheses generation of human mesh, the second image has three hypotheses generation, while the third and fourth images have four hypotheses generation. The dash line represents the right part of a human body, the solid line represents the left part of a human body.

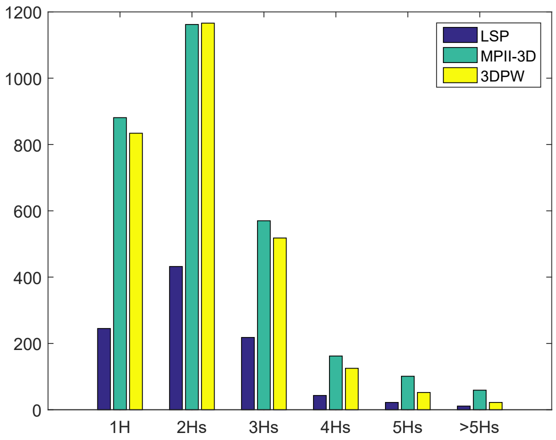

Fig. 5. Statistic analysis of hypotheses generation after selecting and clustering by our approach on three datasets. The number is the number of images with different generation hypotheses in each dataset.

is selected to compare with the ground truth. The test dataset of LSP has 972 images, MPII-3D has 24111 test images, while 3DPW has 26549 test images totally, we use all test images to do a quantitative evaluation.

We compare with SMPLify and its variations [3, 20, 35], HMR [14] and SPIN [18] in terms of the silhouette Intersection over Union (sil-IoU) on LSP test dataset, which measures the matching rate of the projected silhouette of the predicted 3D human mesh and image human segment, and report the results in Table 1. Our method shows best results in both accuracy and F1 measurement on foreground-background and part segmentation on LSP dataset, and increases about 2% compared with SPIN in terms of accuracy. We also compare with PoseNet3D [42], HMR[14], HMR-Video[15], CMR[19], HM-LGD [38], SPIN [18], PARE [5], ROMP [47], ProHMR [30] and [40, 7], with respect to 3D joint error on 3DPW in Table 2. And we report the comparison results on MPII-3D datasets in Table 3 with PoseNet3D [42], HMR[14], SPIN[18], Vnect[8], DenseRaC[46] and [27]. 3D joint error is measured by the least mean per joint position error (MPJPE) between the generated 3D poses and the ground truth before and after rigid alignment. Because our method is a weakly-supervised algorithm, we do not use 3D human pose in training process, and all data are unpaired through our pipeline.

Table 1. Evaluation on foreground-background and six-part segmentation on LSP test set.

Methods	FB Segments		Part Segments	
	Acc	F1	Acc	F1
SMPLify [3]	91.89	0.88	87.71	0.64
SMPLify oracle [20]	92.17	0.88	88.82	0.67
SMPLify+anchor [35]	92.17	0.88	88.24	0.64
HMR [14]	91.67	0.87	87.12	0.60
SPIN [18]	91.83	0.87	89.41	0.68
Our-*Stage1*	92.01	0.87	89.16	0.67
Ours-*es1*	93.83	0.90	91.00	0.72
Ours-*es2*	93.75	0.90	90.78	0.71
Ours	**93.94**	**0.91**	**91.92**	**0.72**

Ablative analysis We examine that if we only have the generator to produce human meshes in the first stage. And the key point of our method is that multiple hypotheses J_p and $\mathbf{V}_q$ are used to guide the generator to create multiple human meshes and poses, which help to improve the human mesh accuracy. Here we try to study how multiple 2D joints J_p and multiple relative depth $\mathbf{V}_q$ are helpful to generate human meshes. 2D human joints, the relative depth and 2D silhouette are all needed as the inputs for the generator, so we fix one when examining the other. They are denoted by two experimental setups as follows,

Experimental setup 1 (es1): When the effect of multiple hypotheses of $\mathbf{V}$ is tested, J is fixed to the initial estimation of J', which is the estimation from the encoder-decoder network $\mathcal{N}_1$.

Table 2. Evaluation on 3DPW dataset. The numbers are MPJPE after rigid alignment.

Supervision	Methods	3D data	PAMPJPE
Strong	HMR [14]	paired, with 3D	81.3
	HM-LGD [38]	paired, with 3D	55.9
	SPIN-static fits [18]	with $\{\boldsymbol{\beta}, \boldsymbol{\theta}\}$	66.3
	SPIN in the loop [18]	with $\{\boldsymbol{\beta}, \boldsymbol{\theta}\}$	59.2
	Doersch et al. [7]	with 3D	74.7
	HMR-Video [15]	with 3D	72.6
	CMR [19]	with 3D	70.2
	Sun et al. [40]	with 3D	66.3
	PARE [5]	with 3D and $\{\boldsymbol{\beta}, \boldsymbol{\theta}\}$	57.1
	ROMP [47]	with 3D and $\{\boldsymbol{\beta}, \boldsymbol{\theta}\}$	56.8
	ProHMR [30]	with 3D and $\{\boldsymbol{\beta}, \boldsymbol{\theta}\}$	59.8
Weak	PoseNet3D [42]	No	63.2
	Our-*stage1*	No	64.35
	Ours-*es1*	No	61.48
	Ours-*es2*	No	64.73
	Ours	No	59.78

Experimental setup 2 (es2): When the effect of multiple hypotheses of J is tested, $\mathbf{V}$ is fixed to $\mathbf{V}'$, which is the original estimation from $\mathcal{N}_2$.

The quantitative results are reported in Table 1, Table 2 and Table 3 on different datasets as well. It can be seen that the space expansion of 2D skeleton and relative depth are helpful for generating more accurate human segments and 3D joints.

4.5 Failure case

Fig. 6 shows two failure examples. Most of failure cases happen for these special images. The human in these two images makes some unusual pose, which is less learnt by CGANs from the dataset, and even it is difficult for 2D human joints detection. Then the generator is not able to generate an reasonable initial human mesh. So training better networks for 2D joint detection and CGANs with more various data is still a fundamental task.

5 Conclusion

We propose a two-stage weakly-supervised pipeline to generate multiple hypotheses via CGANs for 3D human mesh and pose in this paper. The CGANs are trained to generate a single human mesh by taking 2D silhouette as input under the control of 2D joints and relative depth without any 3D supervision. With a reasonable assumption, the conditions of inputs are expanded into a bigger discrete space for generating multiple hypotheses. Then the generated abnormal meshes are rejected by collision detection and classifier, redundant human meshes are clustered. The benefit of our system is that

Table 3. Evaluation on MPII-3D dataset. The numbers are 3D Percentage of Correct Keypoints (PCK) and mean per joint position error (MPJPE) before and after rigid alignment.

Supervision	Methods	3D data	Absolute		Rigid alignment	
			PCK	MPJPE	PCK	PAMPJPE
Strong	HMR[14]	paired,with 3D,$\{\boldsymbol{\beta}, \boldsymbol{\theta}\}$	72.9	124.2	86.3	89.8
	SPIN	paired, with $\{\boldsymbol{\beta}, \boldsymbol{\theta}\}$	76.4	105.2	92.5	67.5
	DenseRaC[46]	with 3D	–	–	89.0	83.5
	Mehta et al. [27]	with 3D	75.7	117.6	–	–
Weak	HMR[14]	unpaired, with 3D,$\{\boldsymbol{\beta}, \boldsymbol{\theta}\}$	59.6	169.5	77.1	113.2
	Vnect [8]	No (but video)	76.6	124.7	83.9	98
	PoseNet3D [42]	No	–	–	81.9	102.4
	Our-*stage1*	No	64.68	148.51	84.45	92.58
	Ours-*es1*	No	68.68	140.96	87.24	87.32
	Ours-*es2*	No	58.84	163.2	79.72	104.14
	Ours	No	70.81	133.91	88.58	84.88

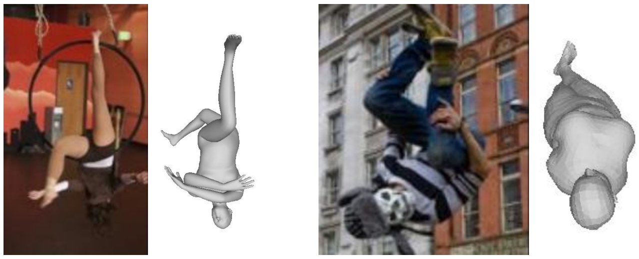

Fig. 6. Failure cases in our experiments.

the background influence is removed, and it is easy to generalize to Internet images. The main limitation is that, it generates multiple hypotheses for human mesh and pose, but the details of meshes are still not sufficient, and some unusual meshes and poses are still hard to reconstruct. This is our future work to discover more elaborate human meshes.

Acknowledgement This work is supported by the National Natural Science Foundation of China (NSFC No. 61971106). Authors would like to thank all reviewers for their valuable and meticulous comments.

References

1. Andriluka, M., Pishchulin, L., Gehler, P., Schiele, B.: 2d human pose estimation: New benchmark and state of the art analysis. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2014)
2. Anguelov, D., Srinivasan, P., Koller, D., Thrun, S., Rodgers, J., Davis, J.: Scape: shape completion and animation of people. In: ACM SIGGRAPH 2005 Papers (2005)

3. Bogo, F., Kanazawa, A., Lassner, C., Gehler, P., Romero, J., Black, M.J.: Keep it smpl: Automatic estimation of 3d human pose and shape from a single image. In: European Conference on Computer Vision (ECCV) (2016)
4. Bouritsas, G., Bokhnyak, S., Ploumpis, S., Bronstein, M., Zafeiriou, S.: Neural 3d morphable models: Spiral convolutional networks for 3d shape representation learning and generation. In: International Conference on Computer Vision (ICCV) (2019)
5. Bouritsas, G., Bokhnyak, S., Ploumpis, S., Bronstein, M., Zafeiriou, S.: Pare: Part attention regressor for 3d human body estimation. In: International Conference on Computer Vision (ICCV) (2021)
6. Cao, Z., Hidalgo, G., Simon, T., Wei, S.E., Sheikh, Y.: OpenPose: realtime multi-person 2D pose estimation using Part Affinity Fields. IEEE Transactions on Pattern Analysis and Machine Intelligence **43**(1), 172–186 (2021)
7. Doersch, C., Zisserman, A.: Sim2real transfer learning for 3d human pose estimation: motion to the rescue. In: Advances in Neural Information Processing Systems (NIPS) (2019)
8. Dushyant, M., Srinath, S., Oleksandr, S., Helge, R., Mohammad, S., Hans-Peter, S., Weipeng, X., Dan, C., , Christian, T.: Vnect: Real-time 3d human pose estimation with a single rgb camera. ACM Transactions on Graphics (TOG) **36**(4), 33–51 (2017)
9. Gabeur, V., Franco, J.S., Martin, X., Schmid, C., Rogez, G.: Moulding humans: Non-parametric 3d human shape estimation from single images. In: IEEE International Conference on Computer Vision (ICCV) (2019)
10. Jahangiri, E., Yuille, A.L.: Generating multiple diverse hypotheses for human 3d pose consistent with 2d joint detections. In: IEEE International Conference on Computer Vision (ICCV) (2017)
11. Johnson, S., Everingham, M.: Clustered pose and nonlinear appearance models for human pose estimation. In: The British Machine Vision Conference (BMVC) (2010)
12. Johnson, S., Everingham, M.: Learning effective human pose estimation from inaccurate annotation. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2011)
13. Joo, H., Simon, T., Sheikh, Y.: Total capture: A 3d deformation model for tracking faces, hands, and bodies. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2018)
14. Kanazawa, A., Black, M.J., Jacobs, D.W., Malik, J.: End-to-end recovery of human shape and pose. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2018)
15. Kanazawa, A., Zhang, J.Y., Felsen, P., Malik, J.: Learning 3d human dynamics from video. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2019)
16. Kato, H., Ushiku, Y., Harada, T.: Neural 3d mesh renderer. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2018)
17. Kocabas, M., Karagoz, S., Akbas, E.: Self-supervised learning of 3d human pose using multi-view geometry. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2019)
18. Kolotouros, N., Pavlakos, G., Black, M.J., Daniilidis, K.: Learning to reconstruct 3d human pose and shape via model-fitting in the loop. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2019)
19. Kolotouros, N., Pavlakos, G., Daniilidis, K.: Convolutional mesh regression for single-image human shape reconstruction. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2019)
20. Lassner, C., Romero, J., Kiefel, M., Bogo, F., Black, M.J., Gehler, P.V.: Unite the people: Closing the loop between 3d and 2d human representations. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2017)
21. Li, C., Lee, G.H.: Generating multiple hypotheses for 3d human pose estimation with mixture density network. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2019)

22. Li, W., Liu, H., Tang, H., Wang, P., Gool, L.V.: Mhformer: Multi-hypothesis transformer for 3d human pose estimation. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2022)
23. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: European Conference on Computer Vision (ECCV) (2014)
24. Loper, M., Mahmood, N., Black, M.J.: Mosh: Motion and shape capture from sparse markers. ACM Transactions on Graphics (TOG) **33**(6), 1–13 (2014)
25. Loper, M., Mahmood, N., Romero, J., Pons-Moll, G., Black, M.J.: Smpl: A skinned multi-person linear model. ACM transactions on graphics (TOG) **34**(6), 1–16 (2015)
26. Marcard, V.T., Henschel, R., Black, M.J., Rosenhahn, B., Pons-Moll, G.: Recovering accurate 3d human pose in the wild using imus and a moving camera. In: European Conference on Computer Vision (ECCV) (2018)
27. Mehta, D., Rhodin, H., Casas, D., Fua, P., Sotnychenko, O., Xu, W., , Theobalt, C.: Monocular 3d human pose estimation in the wild using improved cnn supervision. In: International Conference on 3D vision (3DV) (2017)
28. Mehta, D., Rhodin, H., Casas, D., Fua, P., Sotnychenko, O., Xu, W., Theobalt, C.: Monocular 3d human pose estimation in the wild using improved cnn supervision. In: International Conference on 3D Vision (3DV) (2017)
29. Mirza, M., S., O.: Conditional generative adversarial nets. In: https://arxiv.org/abs/1411.1784 (2014)
30. Nikos Kolotouros, Georgios Pavlakos, D.J., Daniilidis, K.: Probabilistic modeling for human mesh recovery. In: International Conference on Computer Vision (ICCV) (2021)
31. Pavlakos, G., Choutas, V., Ghorbani, N., Bolkart, T., Osman, A.A., Tzionas, D., Black, M.J.: Expressive body capture: 3d hands, face, and body from a single image. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2019)
32. Pavlakos, G., Kolotouros, N., Daniilidis, K.: Texturepose: Supervising human mesh estimation with texture consistency. In: IEEE International Conference on Computer Vision (ICCV) (2019)
33. Pavlakos, G., Zhou, X., Daniilidis, K.: Ordinal depth supervision for 3d human pose estimation. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2018)
34. Pavlakos, G., Zhou, X., Derpanis, K.G., Daniilidis, K.: Coarse-to-fine volumetric prediction for single-image 3d human pose. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2017)
35. Pavlakos, G., Zhu, L., Zhou, X., Daniilidis, K.: Learning to estimate 3d human pose and shape from a single color image. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2018)
36. Pishchulin, L., Insafutdinov, E., Tang, S., Andres, B., Andriluka, M., Gehler, P.V., Schiele, B.: Deepcut: Joint subset partition and labeling for multi person pose estimation. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2016)
37. Romero, J., Tzionas, D., Black, M.J.: Embodied hands: Modeling and capturing hands and bodies together. ACM Transactions on Graphics (ToG) **36**(6), 245 (2017)
38. Song, J., Chen, X., Hilliges, O.: Human body model fitting by learned gradient descent. In: European Conference on Computer Vision (ECCV) (2020)
39. Sun, X., Xiao, B., Wei, F., Liang, S., Wei, Y.: Integral human pose regression. In: European Conference on Computer Vision (ECCV) (2018)
40. Sun, Y., Ye, Y., Liu, W., Gao, W., Fu, Y., Mei, T.: Human mesh recovery from monocular images via a skeleton disentangled representation. In: International Conference on Computer Vision (ICCV) (2019)

41. Tan, F., Zhu, H., Cui, Z., Zhu, S., Pollefeys, M., Tan, P.: Self-supervised human depth estimation from monocular videos. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2020)
42. Tripathi, S., Ranade1, S., Tyagi, A., Agrawal, A.: Posenet3d: Learning temporally consistent 3d human pose via knowledge distillation. In: International Conference on 3D Vision (IC3DV) (2020)
43. Varol, G., Ceylan, D., Russell, B., Yang, J., Yumer, E., Laptev, I., Schmid, C.: Bodynet: Volumetric inference of 3d human body shapes. In: European Conference on Computer Vision (ECCV) (2018)
44. Varol, G., Romero, J., Martin, X., Mahmood, N., Black, M.J., Laptev, I., Schmid, C.: Learning from synthetic humans. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2017)
45. Xiang, D., Joo, H., Sheikh, Y.: Monocular total capture: Posing face, body, and hands in the wild. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2019)
46. Xu, Y., Zhu, S.C., Tung, T.: Denserac: Joint 3d pose and shape estimation by dense render-and-compare. In: IEEE International Conference on Computer Vision (ICCV) (2019)
47. Yu Sun, Qian Bao, W.L.Y.F.M.J.B., Mei, T.: Monocular, one-stage, regression of multiple 3d people. In: International Conference on Computer Vision (ICCV) (2021)
48. Zanfir, A., Marinoiu, E., Sminchisescu, C.: Monocular 3d pose and shape estimation of multiple people in natural scenes-the importance of multiple scene constraints. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2018)
49. Zheng, C., Zhu, S., Mendieta, M., Yang, T., Chen, C., Ding, Z.: 3d human pose estimation with spatial and temporal transformers. In: International Conference on Computer Vision (ICCV) (2021)
50. Zhou, X., Huang, Q., Sun, X., Xue, X., Wei, Y.: Towards 3d human pose estimation in the wild: a weakly-supervised approach. In: IEEE International Conference on Computer Vision (ICCV) (2017)
51. Zhu, H., Zuo, X., Wang, S., Cao, X., Yang, R.: Detailed human shape estimation from a single image by hierarchical mesh deformation. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2019)
52. Zuffi, S., Black, M.J.: The stitched puppet: A graphical model of 3d human shape and pose. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2015)