DaNet: Decompose-and-aggregate Network for 3D Human Shape and Pose Estimation

Hongwen Zhang[1,2] Jie Cao[1,2] Guo Lu[3] Wanli Ouyang[4] Zhenan Sun[1,2*]
[1]CRIPAC & NLPR, Institute of Automation, Chinese Academy of Sciences, Beijing, China
[2]University of Chinese Academy of Sciences, Beijing, China [3]Shanghai Jiao Tong University, Shanghai, China
[4]The University of Sydney, SenseTime Computer Vision Research Group, Sydney, Australia
{hongwen.zhang,jie.cao}@cripac.ia.ac.cn,luguo2014@sjtu.edu.cn,wanli.ouyang@sydney.edu.au,znsun@nlpr.ia.ac.cn

ABSTRACT

Reconstructing 3D human shape and pose from a monocular image is challenging despite the promising results achieved by most recent learning based methods. The commonly occurred misalignment comes from the facts that the mapping from image to model space is highly non-linear and the rotation-based pose representation of the body model is prone to result in drift of joint positions. In this work, we present the Decompose-and-aggregate Network (DaNet) to address these issues. DaNet includes three new designs, namely UVI guided learning, decomposition for fine-grained perception, and aggregation for robust prediction. First, we adopt the UVI maps, which densely build a bridge between 2D pixels and 3D vertexes, as an intermediate representation to facilitate the learning of image-to-model mapping. Second, we decompose the prediction task into one global stream and multiple local streams so that the network not only provides global perception for the camera and shape prediction, but also has detailed perception for part pose prediction. Lastly, we aggregate the message from local streams to enhance the robustness of part pose prediction, where a position-aided rotation feature refinement strategy is proposed to exploit the spatial relationship between body parts. Such a refinement strategy is more efficient since the correlations between position features are stronger than that in the original rotation feature space. The effectiveness of our method is validated on the Human3.6M and UP-3D datasets. Experimental results show that the proposed method significantly improves the reconstruction performance in comparison with previous state-of-the-art methods. Our code is publicly available at https://github.com/HongwenZhang/DaNet-3DHumanReconstrution.

CCS CONCEPTS

• **Computing methodologies** → **Computer vision**; *Shape inference*; *Reconstruction*; • **Human-centered computing**;

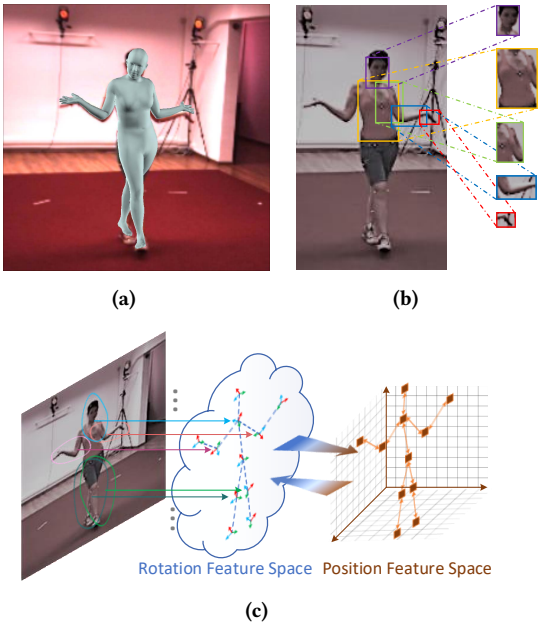

Figure 1: (a) A human image with the reconstructed 3D shape. The rotation-based pose representation of the body model is prone to result in drift of joint positions. (b) Local visual cues are helpful for part pose perception. (c) Our DaNet has multiple local streams for fine-grained perception of the part rotation status and aggregate them into position feature space to exploit the spatial relationship.

KEYWORDS

Decompose-and-aggregate Network; 3D human shape and pose estimation; position-aided rotation feature refinement

ACM Reference Format:
Hongwen Zhang, Jie Cao, Guo Lu, Wanli Ouyang, and Zhenan Sun. 2019. DaNet: Decompose-and-aggregate Network for 3D Human Shape and Pose Estimation. In *Proceedings of the 27th ACM International Conference on Multimedia (MM '19), October 21–25, 2019, Nice, France.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3343031.3351057

1 INTRODUCTION

Reconstructing human shape and pose from a monocular image is an appealing yet challenging task, which typically involves the prediction of the camera and parameters of a statistical body model (e.g. the most commonly used SMPL [27] model). Fig. 1(a) shows an example of the reconstructed result. The challenges of this task

*Corresponding author: Zhenan Sun.

MM '19, October 21–25, 2019, Nice, France

ACM ISBN 978-1-4503-6889-6/19/10…$15.00
https://doi.org/10.1145/3343031.3351057

(a) (b) (c) (d)

Figure 2: Comparison of (a) raw RGB image, (b) silhouette, (c) segmentation, and (d) UVI map.

come from the fundamental depth ambiguity, the complexity and flexibility of human bodies, and variations in clothing and viewpoint, etc. Traditional approaches [5, 23] fit the SMPL model to 2D evidence such as 2D body joints or silhouettes in images, which involve complex non-linear optimization and iterative refinement. Recently, learning based approaches [19, 34, 37, 50] integrate the SMPL model within neural networks and predict model parameters directly in an end-to-end manner.

A main obstacle for this task is that the direct prediction of the body model from the image space is complex and difficult even for deep neural networks. In this work, we propose to adopt UVI maps as an intermediate representation to facilitate the learning of the mapping from image to model. As depicted in Fig. 2, compared with other 2D representations [34, 37, 50], the UVI map could provide more rich information, because it encodes the dense correspondence between foreground pixels on 2D image and vertexes on 3D mesh. Such a densely semantic map not only contains essential information for shape and pose estimation from the RGB images, but also eliminates interference of unrelated factors such as appearance, clothing, and illumination variations.

The representation of 3D body model can be factorized into the shape and pose parameters of SMPL [27], depicting the model at different scales. The shape parameters give an overall description about the model such as the height and weight, while the pose parameters provide the more detailed descriptions about the rotation status of each body joint. Previous learning-based methods [19, 34] typically predict them simultaneously using the global information from the last layer of the neural network. We observe that the detailed pose of body parts should be captured by local visual cues instead of global information. As shown in Fig. 1(b), we can estimate the rotation status of those visible body joints only based on local visual cues, while the information from other body joints and background regions would be irrelevant.

For the rotation-based pose representation, small rotation errors accumulated along the kinematic chain could lead to large drift of position at the leaf joint. Moreover, the rotation estimation is error-prone for those occluded body parts since the perception of local body parts is less reliable under occlusions. Hence, it is crucial to utilize information from visible body parts and the prior about the structure of human bodies. As shown in previous work [7–9], leveraging the structural information at feature level is helpful to obtain more robust and accurate pose estimation results. However, it is non-trivial to apply these feature refinement methods to our case due to the weak correlation between rotation-based part poses. For instance, the shoulder, elbow and wrist are three consecutive body joints, and one can hardly infer the relative rotation of wrist w.r.t. the elbow given the relative rotation of elbow w.r.t. the shoulder. On the other hand, we observe that the 3D locations of body joints have stronger correlations than the rotation of body joints. For instance, the positions of shoulder, elbow and wrist are strongly constrained by the length of the arm.

Based on the observations above, we propose a Decompose-and-Aggregate Network (DaNet) for 3D human shape and pose estimation. The DaNet utilizes UVI maps as the intermediate information for the task. In the DaNet, we decompose the task into one global and multiple local streams in consideration that the prediction of different parameters requires different sizes of the receptive field. In order to robustly predict the rotation of 3D body joints, DaNet aggregates the message from local streams and refines the rotation feature via an auxiliary position feature space to exploit the spatial relationship between body parts, as shown in Fig. 1(c).

The main contributions in this work are summarized as follows.

- We introduce the UVI maps as the intermediate representation for the task of 3D human pose and shape estimation. Such a densely semantic map contains essential information for shape and pose estimation while eliminating interference of other unrelated factors, which greatly facilitates the learning of the mapping from image to body model.
- We decompose the reconstruction task into one global and multiple local streams so that the prediction of different aspects of the task can utilize different information sources. This enables the network to provide global perception for the camera and shape prediction and detailed perception for pose prediction of each body part.
- We propose a position-aided rotation feature refinement strategy to aggregate the message from local streams for robust part pose prediction. The rotation features are gathered and converted into a position feature space where the features of body joints refine each other along the kinematic chain. It is more efficient to exploit the spatial relationship between body parts in the position feature space since the correlations between position feature are stronger than that in the original rotation feature space.

2 RELATED WORK

Intermediate Representation for 3D Pose Recovery: The recovery of 3D human pose from a monocular image is challenging. Common strategies use intermediate estimations as the proxy representation to alleviate the difficulty. For 3D human pose estimation, two-stage methods [6, 24, 30, 31, 33, 33, 39, 48] typically perform 2D keypoint estimations at first and then lift the 2D estimation to 3D pose. These methods can benefit from existing state-of-the-art 2D pose estimation algorithms. One-stage methods in literature adopt volumetric representation [36], joint heat map [47] or 3D orientation fields [29] as intermediate representations to facilitate the learning task. Similarly, for 3D human shape and pose estimation, silhouette [37], joint heatmap [37, 50], segmentation [34] and 3D orientation field [54] have also been exploited in literature as proxy representations for estimating the 3D human shape and pose. Though the aforementioned representations are helpful for the task, detailed information contained within body parts is missing

in these coarse 2D representations, which becomes the bottleneck for the subsequent prediction. Recently, DensePose [2] regresses the UVI maps directly from images, which provides the dense correspondence mapping from the image to the human body model. However, the 3D pose cannot be directly retrieved from such a 2.5D projection. In our work, we propose to adopt such a dense semantic map as the intermediate representation for the task of 3D human shape and pose estimation. To the best of our knowledge, we are the first to investigate learning the human shape and pose from UVI maps via CNN. In concurrent work, [22] obtains UVI predictions using a pretrained network of DensePose [2], while [12] leverages UVI predictions for refinement. Very recently, [58] uses the UV position map as a representation of 3D human body. These efforts are all different from ours.

3D Human Pose and Shape Estimation: Compared to the problem of predicting sparse 3D joint position, the recovery of human pose and shape from a monocular image has received much less attention. Early pioneering works [11, 40] fit the body model SCAPE [4] with the requirement of ground truth silhouettes or manual initialization. Bogo et al. [5] introduce the optimization method SMPLify and make the first attempt to automatically fit the SMPL model to 2D body joints by leveraging multiple priors. Lassner et al. [23] extend this method and improve the reconstruction performance by incorporating the silhouette information in the fitting procedure. These optimization based methods typically rely on accurate 2D observations and the prior terms imposed on the shape and pose parameters, making the procedure time-consuming and sensitive to the initialization. Alternatively, there are several attempts to employ the neural network for predicting the SMPL parameters directly and learn the priors in a data-driven manner. Tan et al. [46] develop an encoder-decoder based framework where the decoder learns the SMPL-to-silhouette mapping from synthetic data and the encoder learns the image-to-SMPL mapping with the fixed decoder. Tung et al. [50] predict SMPL parameters from video frames by integrating several re-projection losses against 2D keypoints, silhouettes and optical flow. Kanazawa et al. [19] present an end-to-end framework to reconstruct the SMPL model directly from images using a single CNN with an iterative regression module. To alleviate the learning of highly non-linear mapping, 2D estimations are exploited as proxy representation during the learning procedure. For instance, Pavlakos et al. [37] propose to predict the shape and pose parameters from the estimated silhouettes and joint heatmaps respectively. Omran et al. [34] propose to use segmentation as proxy representation and show it is more helpful to 3D shape/pose estimation compared with the raw RGB images or silhouettes. In addition to using 2D estimations, 3D volumetric representation is also adopted in [16, 51] to facilitate the reconstruction of human body shape. All aforementioned learning-based methods predicting the pose in a global manner. In contrast, our DaNet predicts part poses from multiple streams, hence the visual cues could be captured in a fine-grained manner. Additionally, existing approaches for jointly estimating 3D pose and shape do not consider feature refinement, while our DaNet uses the feature refinement for better pose estimation under the rotation-based pose representation in the SMPL model. We believe our framework could also be extended to other expressive body models [18, 35].

Structured Feature Learning for Human Pose Estimation: Leveraging the articulated structure information is crucial for accurate human pose estimation. Early work utilized the spatial relationships between body joints through graphical models such as pictorial structure [38] and mixture-of-parts [57]. Recent state-of-the-art methods [7–9, 32, 49, 52, 55] employ convolution networks for better feature extraction and incorporate the structured feature learning in the architecture design. Among them, Chu et al. [8] investigate learning relationship among body parts at the feature level. They further extend their work in [9] to a CRF-CNN framework to model structures in both output and hidden feature layers within CNN. All these methods exploit the relationship between the *position features* of body parts and these feature refinement strategies are only validated on the position-based pose estimation problem. Our approach is complementary to them by investigating the refinement strategy for *rotation features* under the context of rotation-based pose representation. We further show that the spatial relationship between body parts is a good intermediate space for refining the rotation features. Our approach aggregates the rotation features into the position feature space, where the aforementioned structural feature learning approaches could be easily applied.

Pose priors at output level. For 3D human pose, different types of pose prior [1, 43, 56, 59, 60] are also employed as the constraint in the learning procedure for a more geometrically reasonable prediction. For instance, Akhter and Black [1] learn the pose prior in the form of joint angle constraints. Sun et al. [43] design handcrafted constraints such as limb-lengths and their proportions. Similar constraints are exploited in [59] under the weakly-supervised setting. For the rotation-based pose representation in SMPL model, though it inherently satisfies structure constraints such as limb proportions, the pose prior is still essential for better reconstruction performance. SMPLify [5] imposes several penalizing terms on predicted poses to prevent unnatural results. Kanazawa et al. [19] introduce an adversarial prior for guiding the prediction to be realistic. All these methods consider the pose prior at the *output level*. In our work, we will exploit the relationship at the *feature level* for better 3D pose estimation in SMPL model.

3 SMPL MODEL AND UVI MAP

SMPL Model. The Skinned Multi-Person Linear model (SMPL) [27] is one of the widely used statistical human body models, which represents the body mesh with two sets of parameters, i.e. the shape and pose parameters. The shape indicates the model's height, weight and limb proportions while the pose indicates how the model deforms with the rotated skeleton joints. Such decomposition of shape and pose makes it convenient for algorithms to focus on one of these two factors independently. In the SMPL model, the shape parameters $\boldsymbol{\beta} \in \mathbb{R}^{10}$ denotes the coefficients of the PCA basis of body shape. The pose parameters $\boldsymbol{\theta} \in \mathbb{R}^{3K}$ denotes the axis-angle representations of the relative rotation of K skeleton joints with respect to their parents in the kinematic tree, where $K = 23 + 1$ in the SMPL model, including the root joint. Given the pose and shape parameters, the model deforms accordingly and generates a triangulated mesh with $N = 6890$ vertices $M(\boldsymbol{\theta}, \boldsymbol{\beta}) \in \mathbb{R}^{3\times N}$. The deformation process $M(\boldsymbol{\theta}, \boldsymbol{\beta})$ is differentiable with respect to the pose $\boldsymbol{\theta}$ and shape $\boldsymbol{\beta}$, which means that the SMPL model could be

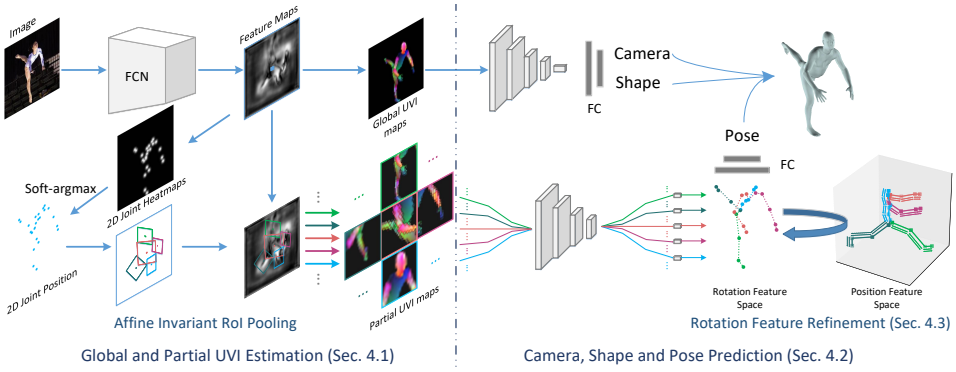

Figure 3: Overview of the proposed Decompose-and-aggregate Network (DaNet).

integrated within a neural network as a typical layer without any learnable weights.

UVI Map. Reconstructing the 3D object model from a monocular image is ambiguous, but there's a determinate correspondence between pixels on 2D image and vertexes on 3D surface. Such correspondence could be represented in the form of UV map, which is an image with each foreground pixel containing the UV coordinate values. In this way, the pixels on the foreground could be projected back to vertexes on the template mesh according to a predefined bijective mapping between the 3D surface space and the 2D UV space. For the human body model, the correspondence could have finer granularity by introducing the index I of the body parts [2, 3], which results in the UVI representation. In each body part, the UV space is independent so that the representation could be more fine-grained. Currently, the only dataset providing UVI annotations is the DensePose-COCO [2] dataset, which is an extended version of 2D human pose dataset.

Preparation of UVI map for 3D human pose dataset. Currently, there is no 3D human pose dataset providing UVI annotations. In this work, for those datasets providing SMPL parameters with human images, we augment their annotations by adding the corresponding ground-truth UVI maps. Specifically, given images and the corresponding camera and SMPL parameters, the ground-truth UVI maps could be obtained by using existing rendering algorithms such as [20, 28]. For each face in the triangulated mesh, the texture value used for rendering is a triplet vector (u, v, i) denoting the corresponding U, V and I values. The UVI mapping to the SMPL model adopts the same protocol provided in DensePose [2].

4 METHODOLOGY

As illustrated in Fig. 3, our DaNet decomposes the prediction task into a global stream for the camera and shape prediction and multiple local streams for part pose prediction. The overall pipeline involves two consecutive stages, where the UVI maps are firstly estimated from the fully convolution network and then taken as inputs for subsequent parameter prediction.

In the first stage, the UVI maps are estimated from global and local perspectives in consideration of the different sizes of the receptive fields required by the prediction of different parameters.

In the second stage, the global and local UVI maps are used for separate tasks. The global UVI maps are used for extracting global features, which are directly used to predict camera and body shape. The partial UVI maps are used for extracting the rotation features, which are further refined and then used to predict part poses.

Overall, our objective function is the combination of three objectives:

$$\mathcal{L} = \mathcal{L}_{inter} + \mathcal{L}_{target} + \mathcal{L}_{refine}, \tag{1}$$

where $\mathcal{L}_{inter}$ is the objective for estimating the intermediate representation (Sec. 4.1), $\mathcal{L}_{target}$ is the objective for predicting the camera and SMPL parameters (Sec. 4.2), $\mathcal{L}_{refine}$ is the objective involving in the feature refinement procedure (Sec. 4.3). In the following subsections, we will present the technical details and rationale of our method.

4.1 Global and Partial UVI Estimation

The first stage in our method aims to estimate corresponding UVI maps from input images for subsequent prediction tasks. Specifically, a fully convolutional network is employed to produce $K + 1$ sets of UVI maps, including one set of global UVI maps and K sets of partial UVI maps for the corresponding K body parts. The global UVI maps are aligned with the original image through up-sampling, while the partial UVI maps center around the body joints. The feature maps outputted from the last layer of the FCN would be shared by the estimation tasks of both global and partial UVI maps. The estimation of the global UVI maps is quite straightforward since they could be obtained by simply feeding these feature maps into a convolutional layer. For the estimation of each set of partial UVI maps, the affine invariant RoI pooling would be first applied on these feature maps to extract an appropriate sub-region, which results in partial feature maps. Then, the K sets of partial UVI maps would be estimated independently from the resulting K sets of partial feature maps. Now, we will give details about the proposed affine invariant RoI pooling.

Affine Invariant RoI Pooling. Spatial alignment or normalization strategies are widely employed to reduce variations for down-stream tasks such as face recognition [45, 53] and human pose estimation [10, 13, 41]. In our approach, a similar mechanism is proposed for better perception of part poses.

For the pose parameters in the SMPL model, they represent the relative rotation of each body joint with respect to its parent in the kinematic tree. Hence, the perception of part poses should also be invariant to the global scale, translation and rotation. Moreover, the ideal scale factor for the perception of part pose should vary from one part to another since the proportions of body parts are different.

To this end, we introduce the affine invariant RoI pooling for partial UVI estimation. Particularly, for each body part, a sub-region of the feature maps are extracted and spatially transformed to a fixed resolution for subsequent partial UVI map estimation and part pose prediction.

The affine transformation parameters, i.e. scale, translation and rotation, are calculated individually for each sub-region (RoI), in order that the partial UVI maps could cover two connected bones, center around corresponding body parts, and be rotated such that the one of the two bones consistently has the same orientation. Such a strategy serves as an *attention* for each body part such that the perception of part pose is adaptable to spatial variations caused by global scales and orientations. In comparison with the Spatial Transformer Networks (STNs) [17], the pooling process in our network is learned under an explicit supervision manner.

As illustrated in Fig. 4, the transformation parameters used for spatial transformation of each RoI are calculated from the 2D joint positions. Specifically, 2D joint heatmaps are estimated along with the global UVI maps in a multi-task learning manner, and the 2D joint positions are retrieved from heatmaps using the soft-argmax [44] operation. Without loss of generality, let $\boldsymbol{j}_k$ denote the position of the k-th body joint, and let functions $p(k)$ and $c(k)$ return the index of the parent and child joint for the k-th body joint respectively. Then, for the k-th set of partial UVI maps, the center $\boldsymbol{c}_k$, scale s_k and rotation angle r_k used for spatial transformation could be calculated from $\boldsymbol{j}_{p(k)}$, $\boldsymbol{j}_k$ and $\boldsymbol{j}_{c(k)}$, i.e. the positions of the k-th body joint itself and its parent and child joints. Specifically, the center $\boldsymbol{c}_k$ is the positions of the target joint. The scale s_k is proportional to the maximum lengths of adjacent bones. The rotation angle r_k is calculated according to the orientation of the bone pointing from $\boldsymbol{j}_{p(k)}$ to $\boldsymbol{j}_k$. These transformation parameters can be formulated as

$$\begin{aligned} \boldsymbol{c}_k &= \boldsymbol{j}_k, \\ s_k &= \alpha_k \max\left(\left\|\boldsymbol{j}_{p(k)} - \boldsymbol{j}_k\right\|_2, \left\|\boldsymbol{j}_k - \boldsymbol{j}_{c(k)}\right\|_2\right) + \beta_k, \\ r_k &= \arccos \frac{\left(\boldsymbol{j}_{p(k)} - \boldsymbol{j}_k\right) \cdot \boldsymbol{e}_\perp}{\left\|\boldsymbol{j}_{p(k)} - \boldsymbol{j}_k\right\|_2}, \end{aligned} \tag{2}$$

where α_k and β_k are two constants, $\boldsymbol{e}_\perp$ denotes the unit vector pointing the vertical downward direction. After obtaining these parameters, the feature maps extracted from the last layer of fully convolutional network are spatially transformed to a fixed resolution and use to estimate the partial UVI maps, where the corresponding ground-truth partial UVI maps are extracted from the ground-truth global UVI maps using the same pooling process. In our experiments, the scale and rotation adjustments are only applied to those visible limb joints in consideration that the localization of torso and hidden joints are relatively unreliable.

Loss Functions. A classification loss and several regression losses are involved in the training of this stage. For both global and partial UVI maps, the loss is calculated in the same manner and denoted as $\mathcal{L}_{uvi}$. Specifically, a classification loss is imposed on the index I channels of UVI maps, where the $K+1$-way cross-entropy loss is employed to classify a pixel belonging to either background or one among the K body parts. For the UV channels of UVI maps, an L_1 based regression loss is adopted, and is only

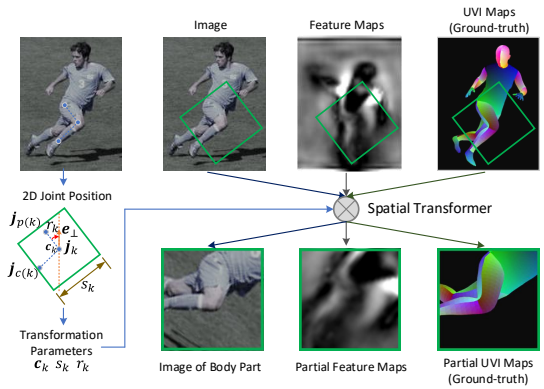

Figure 4: Illustration of the affine invariant RoI pooling.

taken into account for those pixels on the foreground. For the 2D joint heatmaps and 2D joint positions estimated for RoI pooling, an L_1 based regression loss is adopted and denoted as $\mathcal{L}_{roi}$. Overall, the objective in the UVI estimation stage involves two main losses and is denoted as

$$\mathcal{L}_{inter} = \lambda_{uvi}\mathcal{L}_{uvi} + \lambda_{roi}\mathcal{L}_{roi}. \tag{3}$$

4.2 Camera, Shape and Pose Prediction

After obtaining the global and partial UVI maps, the camera and shape parameters would be predicted in the global stream, while pose parameters would be predicted in the local streams.

The global stream consists of a ResNet [14] as the backbone network and a fully connection layer added at the end with 13 outputs, corresponding to the camera scale $s \in \mathbb{R}$, translation $\boldsymbol{t} \in \mathbb{R}^2$ and the shape parameters $\boldsymbol{\beta} \in \mathbb{R}^{10}$. In the local stream, a tailored ResNet acts as the backbone network shared by all body parts and is followed by K residual layers for rotation feature extraction individually. For the k-th body part, the extracted rotation features would be refined (see Sec. 4.3) and then used to predict the rotation matrix $\boldsymbol{R}_k \in \mathbb{R}^{3\times3}$ via a fully connection layer. Here, we follow previous work [34, 37] to predict the rotation matrix representation of the pose parameters $\boldsymbol{\theta}$ rather than the axis-angle representation defined in the SMPL model. An L_1 loss is imposed on the predicted camera, shape and pose parameter, and we denote it as $\mathcal{L}_{smpl}$.

Following previous work [19, 34, 37], we also add additional constraint and regression objective for better performance. For the predicted rotation matrix, it is necessary to make it lie on the manifold of rotation matrices. In our method, we impose an orthogonal constraint loss on the predicted rotation matrix to guarantee its orthogonality. The orthogonal constraint loss for predicted rotation matrices $\{\boldsymbol{R}_k\}_{k=1}^K$ is denoted as $\mathcal{L}_{orth}$ and could be written as

$$\mathcal{L}_{orth} = \sum_{k=1}^{K} \left\|\boldsymbol{R}_k\boldsymbol{R}_k^{\mathsf{T}} - \boldsymbol{I}\right\|_2. \tag{4}$$

Given the predicted SMPL parameters, the performance could be further improved by adding supervision explicitly on the resulting model $M(\boldsymbol{\theta}, \boldsymbol{\beta})$. Specifically, we use three L_1 based loss functions to measure the difference between the ground-truth position and the predicted one, and the corresponding losses are denoted as $\mathcal{L}_{vert}$ for vertexes on 3D mesh, $\mathcal{L}_{3Dkp}$ for sparse 3D human keypoints

and $\mathcal{L}_{reproj}$ for the reprojected 2D human keypoints respectively. For the sparse 3D human keypoints, the predicted position could be obtained by a pre-trained linear regressor to map the mesh vertices to 3D human keypoints defined in human pose datasets. Overall, the objective in this prediction stage involves multiple losses and is denoted as

$$\begin{aligned}\mathcal{L}_{target} &= \lambda_{smpl}\mathcal{L}_{smpl} + \lambda_{orth}\mathcal{L}_{orth} \\ &+ \lambda_{point}\left(\mathcal{L}_{vert} + \mathcal{L}_{3Dkp} + \mathcal{L}_{reproj}\right).\end{aligned} \tag{5}$$

4.3 Rotation Feature Refinement

In our approach, a position-aided rotation feature refinement strategy is proposed to exploit spatial relationships among body parts. As illustrated in Fig. 5, the rotation refinement procedure includes three consecutive steps, namely rotation feature to position feature mapping, position feature refinement, and refined feature aggregation. Specifically, the rotation features are first aggregated and converted to the position feature space where the feature refinement is performed. After that, the rotation feature refinement is accomplished by aggregating the messages from the refined position features.

Step 1: rotation feature to position feature mapping. The rotation features extracted independently from partial UVI maps are viewed as sequential data along the kinematic chain. This is inspired by the fact that the human could act in a recurrent manner according to the kinematic tree. Given the position of a body joint, the position of its child joint can be calculated according to the relative rotation and the bone length. At the feature level, such mapping is learned by the bilinear unit [30]. Formally, let $\{\boldsymbol{x}_k\}_{k=1}^K$ denote the rotation features extracted from K sets of partial UVI maps. After accumulating the information from rotation features according to the kinematic tree, the position features of all joint are generated, which are denoted as $\{\boldsymbol{v}_k\}_{k=1}^K$. For the k-th body joint, a bilinear unit learns the mapping function $f_k(\cdot)$ such that it takes the rotation feature $\boldsymbol{x}_{p(k)}$ and position feature $\boldsymbol{v}_{p(k)}$ as input and output the position feature $\boldsymbol{v}_k$, i.e.

$$\boldsymbol{v}_k = f_k(\boldsymbol{x}_{p(k)}, \boldsymbol{v}_{p(k)}). \tag{6}$$

The position feature of the root body joint is initialized as its rotation feature.

Step 2: position feature refinement. Since there is strong correlation of the spatial relationship among body joints belonging to a kinematic chain, utilizing such rich information could effectively improve features learned at each joint. Towards this goal, an LSTM-based feature refinement scheme is utilized to pass spatial information between joints along the kinematic chain. Specifically, let C_i denote the set containing the indices of the body joints belonging to the i-th chain. The position features $\{\boldsymbol{v}_k\}_{k\in C_i}$ are viewed as sequential data. A bi-directional LSTM takes them as input and then outputs the refined features $\{\hat{\boldsymbol{v}}_k\}_{k\in C_i}$, where $\hat{\boldsymbol{v}}_k$ is the concatenated features for the k-th body joint refined from forward and backward directions. The refined position features $\hat{\boldsymbol{v}}_k$ are then used to predict the corresponding 3D joint position. An L_1 loss is imposed on the predicted 3D joint position, which composes the objective $\mathcal{L}_{refine}$ involved in the refinement procedure.

Step 3: refined feature aggregation. Since the rotation and position of body joints are two mutual representation of 3D human pose, after the refinement of position feature, the rotation feature can be refined accordingly. Specifically, for the k-th body joint, its rotation features can be refined by aggregating messages from the refined position feature of three consecutive body joints, i.e. the joint itself and its parent and child joints. Likewise, the mapping from position features to rotation features is also learned by the bilinear unit. Formally, the mapping function $g_k(\cdot)$ takes the features of three consecutive body joints as input and outputs the features in the rotation feature space. These features are added with original rotation features in a residual manner, resulting in the refined rotation features $\hat{\boldsymbol{x}}_k$ for the final prediction of part pose parameters, i.e.

$$\hat{\boldsymbol{x}}_k = \boldsymbol{x}_k + g_k(\hat{\boldsymbol{v}}_{p(k)}, \hat{\boldsymbol{v}}_k, \hat{\boldsymbol{v}}_{c(k)}). \tag{7}$$

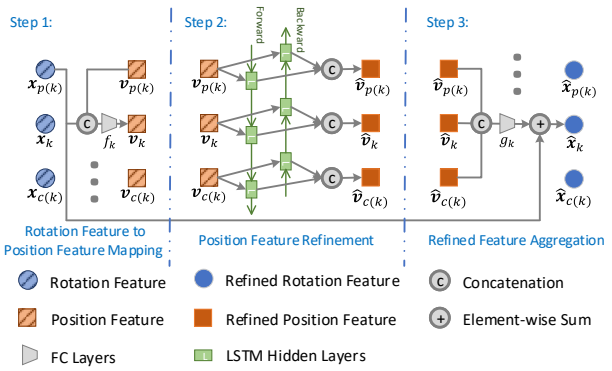

Figure 5: Illustration of the position-aided rotation feature refinement.

5 EXPERIMENTS

5.1 Implementation Details

The FCN for UVI estimation in our framework adopts the architecture of HRNet-W48 [42], one of the most recent state-of-the-art networks for dense prediction tasks. The FCN receives the 224×224 input and produces 56×56 feature maps for estimating the global and local UVI maps, which have the same resolution of 56×56. Two ResNet-18 [14] are employed as the backbone networks for global and rotation feature extraction. The pyper-parameters λs are selected in order to make values of objectives have similar scales. αs and βs in Eq. 2 can be learned using ground-truth UVI maps as inputs. During training, data augmentation techniques, including rotation $\pm30°$, color jittering ($\pm30\%$ channel-wise) and flipping, are applied randomly to input images. The FCN is initialized with the model pre-trained on the COCO keypoint detection dataset [25] for 2D human pose estimation, which is essential for robust 2D joint position localization and partial UVI estimation. The UVI estimation task is first trained for 5k iterations before involving the parameter prediction task. We adopt the ADAM [21] optimizer with an initial learning rate of 1×10^{-4} to train the model, and reduce the learning rate to 1×10^{-5} after 30k iterations. The training process converges after around 60k iterations. During testing, due to the fundamental depth-scale ambiguity, we follow previous work [19, 34] to center the person within the image and perform scaling such that the inputs have the same setting as training. For our DaNet, a single forward to infer the shape and pose from an image takes about

170ms on a single TITAN Xp GPU. More details could be found in the publicly available code.

5.2 Datasets and evaluation metrics

Human3.6M [15] is a large-scale dataset which consists of 3.6 millions of video frames captured in the controlled environment, and currently the most commonly used benchmark dataset for 3D human pose estimation. Kanazawa et al. [19] generated the ground truth SMPL parameters by applying MoSH [26] to the sparse 3D MoCap marker data. Following the common protocols [19, 36, 37], we use five subjects (S1, S5, S6, S7, S8) for training and two subjects (S9, S11) for evaluation. We also down-sample the original videos from 50fps to 10fps to remove redundant frames, resulting in 312,188 frames for training and 26,859 frames for testing. For evaluation, the Mean Per Joint Position Error (MPJPE) and the MPJPE after rigid alignment of the prediction with ground truth using Procrustes Analysis (MPJPE-PA) are used as the evaluation metrics.

UP-3D [23] is a collection dataset of existing 2D human pose datasets, containing 5703 images for training, 1423 images for validation, and 1389 images for testing. The SMPL parameter annotations of these real-world images are augmented in a semi-automatic way by using an extended version of SMPLify [23]. Following previous work [37], we evaluate the reconstruction performance using the mean per-vertex error between the predicted and ground truth body mesh.

5.3 Comparison with state-of-the-art methods

For Human3.6M, we evaluate the 3D human pose estimation performance for quantitative comparison. Table 1 reports the comparison results with previous methods that output more than sparse 3D keypoint position. Among them, HMR [19] adopts a single CNN and an iterative regression module to produce all parameters. Pavlakos et al. [37] decompose the shape and pose prediction tasks, while their pose parameters are predicted from 2D joints positions. NBF [34] adopts segmentation as the intermediate representation and learns all parameters from it. CMR [22] directly regresses 3D shapes with a graph-based convolutional network. All these methods except [12] estimate pose parameters through a single stream and our method outperforms them significantly. Concurrent work [12] predicts pose parameters using a part-based model and has similar results with ours. Example results of the proposed method on Human3.6M are shown in Fig. 6. Benefit from the decomposition design, our DaNet could capture more detailed part poses and produce accurate reconstruction results.

We further evaluate the reconstruction performance of our method on the UP-3D dataset. We report quantitative evaluation on the per-vertex error of the reconstructed mesh of our method in Table 2. In comparison with previous methods, our method outperforms them across all subsets of UP-3D by a large margin. As our closest competitor, BodyNet [51] uses both 2D and 3D estimation as the intermediate representation, which is much more time-consuming than ours. Example results of our method on UP-3D are shown in Fig. 7. It can be seen that our DaNet could produce satisfactory results under challenging scenarios, which could be attributed to the proposed aggregation design for rotation feature refinement.

Table 1: Quantitative comparison on Human3.6M.

Method	MPJPE	MPJPE-PA
Zhou et al. [60]	107.3	-
Tung et al. [50]	-	98.4
SMPLify [5]	-	82.3
SMPLify++ [23]	-	80.7
Pavlakos et al. [37]	-	75.9
HMR [19]	88	56.8
NBF [34]	-	59.9
Xiang et al. [54]	65.6	-
CMR [22]	-	50.1
HoloPose [12]	64.3	50.6
DaNet	**61.5**	**48.6**

Table 2: Quantitative comparison on UP-3D.

Method	LSP	MPII	FashionPose	Full
SMPLify++ [23]	174.4	184.3	108	169.8
Pavlakos et al. [37]	127.8	110.0	106.5	117.7
BodyNet [51]	102.5	-	-	-
DaNet	**90.4**	**83.0**	**61.8**	**83.7**

Table 3: Validation of the UVI intermediate representation.

Method	MPJPE	MPJPE-PA
ConvFeat	80.4	58.9
Segmentation	75.1	57.5
UVI	**73.3**	**56.6**

5.4 Ablation study

To evaluate the efficacy of the key components proposed in our method, we conduct ablation experiments on Human3.6M under various settings.

Intermediate Representation. The UVI map acts as a bridge between pixels on 2D images and vertexes on 3D meshes and facilitates the learning task of the network. To validate its effectiveness, we use alternative representations as input for the subsequent parameter prediction. For experiments in this part, we remove the local stream in our method, and use only the global stream to predict all parameters. In ablation approaches, the UVI maps are replaced by the feature maps outputted form the last layer of the FCN or the part segmentation (Index channels of UVI maps). As observed from Table 3, the approach using the UVI maps outperforms other ablation approaches using feature maps or segmentation as intermediate representations. In our experiment, we found that the approach using feature maps is more prone to overfitting to the training set.

Decomposed Perception. We conduct experiments to validate the effectiveness of the decomposed perception. Performances of the approaches adopting one-stream (Global) and multiple streams (Global+Local) are reported in Table 4. For fair comparison, the one-stream approach adopts ResNet50 [14] for parameter prediction such that their model sizes are comparable. As can be seen, using multiple streams brings a significant improvement over the approach using one stream.

In our affine invariant RoI pooling mechanism, the scale and rotation are adaptable to spatial variations caused by global scales and orientation, which contributes more stable perception of local

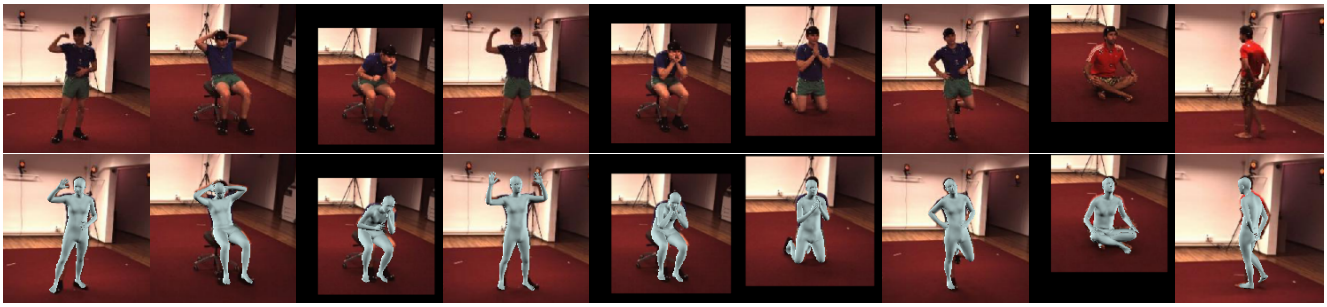

Figure 6: Example results on the Human3.6M dataset.

Figure 7: Example results on the UP-3D dataset.

visual cues for part pose prediction. To validate this claim, we fix either the scale or the rotation in the pooling mechanism. Specifically, for all body parts, the scales $\{s_k\}_{k=1}^K$ are fixed as 0.3 which accounts for around half of the body height, while the rotations $\{r_k\}_{k=1}^K$ are simply fixed as 0. As can be seen from the 3-rd and 4-th row in Table 4, fixing either the scale or the rotation degrades the performance.

Table 4: Comparison of different perception strategies.

Method	MPJPE	MPJPE-PA
Global	73.3	56.6
Global+Local	**65.6**	**52.2**
Global+Local (fixed scale)	66.4	52.7
Global+Local (fixed rotation)	66.2	52.8

Table 5: Comparison of different feature refinement strategies.

Method	MPJPE	MPJPE-PA
Baseline (w/o Refinement)	65.6	52.2
Direct	64.4	50.5
Position-aided	**61.5**	**48.6**

Position-aided Rotation Feature Refinement. The feature refinement is essential for better pose estimation. A straight-forward strategy to refine the feature would be conducting the refinement between the rotation features directly. In this strategy, the rotation features are fed to bi-LSTM for feature refinement and then used to predict the part poses. We report results of the approach using such a strategy in Table 5 and make a comparison to the proposed one. As can be seen, direct refinement of rotation features brings much less improvement. The reason is that the correlation between rotation features is weak, and the message from the adjacent rotation feature is generally irrelevant to refine the current rotation feature. Our aggregation strategy builds an auxiliary position feature space for feature refinement, making it much more efficient than that in the original rotation feature space.

6 CONCLUSION

In this work, we propose a Decompose-and-aggregate Network (DaNet) for 3D human shape and pose estimation. First, the UVI maps are adopted as the intermediate representation to facilitate the learning of image-to-model mapping. The reconstruction task is decomposed into one global and multiple local streams so that the network could provide global perception for the camera and shape prediction and detailed perception for pose prediction of each body part. The affine invariant RoI pooling mechanism is further introduced for a more fine-grained and stable perception of the part poses. Lastly, a position-aided rotation feature refinement strategy is proposed for aggregating messages from body parts to enhance the robustness of pose prediction. It is more efficient to exploit the spatial relationship between body parts in the position feature space since the correlations between position features are stronger than that in the original rotation feature space. The decomposition and aggregation designs contribute to the accurate and robust human shape and pose estimation performance of our method.

ACKNOWLEDGMENTS

This work was supported in part by the National Natural Science Foundation of China under Grant 61427811, Grant U1836217, and Grant 61806197.

REFERENCES

[1] Ijaz Akhter and Michael J Black. 2015. Pose-conditioned joint angle limits for 3D human pose reconstruction. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 1446–1455.

[2] Rıza Alp Güler, Natalia Neverova, and Iasonas Kokkinos. 2018. Densepose: Dense human pose estimation in the wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 7297–7306.

[3] Riza Alp Guler, George Trigeorgis, Epameinondas Antonakos, Patrick Snape, Stefanos Zafeiriou, and Iasonas Kokkinos. 2017. Densereg: Fully convolutional dense shape regression in-the-wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 6799–6808.

[4] Dragomir Anguelov, Praveen Srinivasan, Daphne Koller, Sebastian Thrun, Jim Rodgers, and James Davis. 2005. SCAPE: shape completion and animation of people. In *ACM Transactions on Graphics*, Vol. 24. ACM, 408–416.

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

[6] Ching-Hang Chen and Deva Ramanan. 2017. 3d human pose estimation= 2d pose estimation+ matching. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 7035–7043.

[7] Xianjie Chen and Alan L Yuille. 2014. Articulated pose estimation by a graphical model with image dependent pairwise relations. In *Advances in Neural Information Processing Systems*. 1736–1744.

[8] Xiao Chu, Wanli Ouyang, Hongsheng Li, and Xiaogang Wang. 2016. Structured feature learning for pose estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 4715–4723.

[9] Xiao Chu, Wanli Ouyang, Xiaogang Wang, et al. 2016. Crf-cnn: Modeling structured information in human pose estimation. In *Advances in Neural Information Processing Systems*. 316–324.

[10] Hao-Shu Fang, Shuqin Xie, Yu-Wing Tai, and Cewu Lu. 2017. Rmpe: Regional multi-person pose estimation. In *Proceedings of the IEEE International Conference on Computer Vision*. 2334–2343.

[11] Peng Guan, Alexander Weiss, Alexandru O Balan, and Michael J Black. 2009. Estimating human shape and pose from a single image. In *Proceedings of the IEEE International Conference on Computer Vision*. IEEE, 1381–1388.

[12] Riza Alp Guler and Iasonas Kokkinos. 2019. HoloPose: Holistic 3D Human Reconstruction In-The-Wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 10884–10894.

[13] Albert Haque, Boya Peng, Zelun Luo, Alexandre Alahi, Serena Yeung, and Li Fei-Fei. 2016. Towards viewpoint invariant 3d human pose estimation. In *European Conference on Computer Vision*. Springer, 160–177.

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

[15] Catalin Ionescu, Dragos Papava, Vlad Olaru, and Cristian Sminchisescu. 2014. 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 (2014), 1325–1339.

[16] Aaron S Jackson, Chris Manafas, and Georgios Tzimiropoulos. 2018. 3d human body reconstruction from a single image via volumetric regression. In *Proceedings of the European Conference on Computer Vision*.

[17] Max Jaderberg, Karen Simonyan, Andrew Zisserman, et al. 2015. Spatial transformer networks. In *Advances in Neural Information Processing Systems*. 2017–2025.

[18] Hanbyul Joo, Tomas Simon, and Yaser Sheikh. 2018. Total capture: A 3d deformation model for tracking faces, hands, and bodies. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 8320–8329.

[19] Angjoo Kanazawa, Michael J Black, David W Jacobs, and Jitendra Malik. 2018. End-to-end recovery of human shape and pose. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 7122–7131.

[20] Hiroharu Kato, Yoshitaka Ushiku, and Tatsuya Harada. 2018. Neural 3d mesh renderer. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 3907–3916.

[21] Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *International Conference on Learning Representations* (2014).

[22] Nikos Kolotouros, Georgios Pavlakos, and Kostas Daniilidis. 2019. Convolutional Mesh Regression for Single-Image Human Shape Reconstruction. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 4501–4510.

[23] Christoph Lassner, Javier Romero, Martin Kiefel, Federica Bogo, Michael J Black, and Peter V Gehler. 2017. Unite the people: Closing the loop between 3d and 2d human representations. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 6050–6059.

[24] Kyoungoh Lee, Inwoong Lee, and Sanghoon Lee. 2018. Propagating lstm: 3d pose estimation based on joint interdependency. In *Proceedings of the European Conference on Computer Vision*. 119–135.

[25] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *European Conference on Computer Vision*. Springer, 740–755.

[26] Matthew Loper, Naureen Mahmood, and Michael J Black. 2014. MoSh: Motion and shape capture from sparse markers. *ACM Transactions on Graphics* 33, 6 (2014), 220.

[27] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J Black. 2015. SMPL: A skinned multi-person linear model. *ACM Transactions on Graphics* 34, 6 (2015), 248.

[28] Matthew M Loper and Michael J Black. 2014. OpenDR: An approximate differentiable renderer. In *European Conference on Computer Vision*. Springer, 154–169.

[29] Chenxu Luo, Xiao Chu, and Alan L. Yuille. 2018. OriNet: A Fully Convolutional Network for 3D Human Pose Estimation. In *British Machine Vision Conference 2018*. 92.

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

[31] Francesc Moreno-Noguer. 2017. 3d human pose estimation from a single image via distance matrix regression. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2823–2832.

[32] Alejandro Newell, Kaiyu Yang, and Jia Deng. 2016. Stacked hourglass networks for human pose estimation. In *European Conference on Computer Vision*. Springer, 483–499.

[33] Bruce Xiaohan Nie, Ping Wei, and Song-Chun Zhu. 2017. Monocular 3D human pose estimation by predicting depth on joints. In *Proceedings of the IEEE International Conference on Computer Vision*. IEEE, 3467–3475.

[34] Mohamed Omran, Christoph Lassner, Gerard Pons-Moll, Peter Gehler, and Bernt Schiele. 2018. Neural body fitting: Unifying deep learning and model based human pose and shape estimation. In *International Conference on 3D Vision*. IEEE, 484–494.

[35] Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed AA Osman, Dimitrios Tzionas, and Michael J Black. 2019. Expressive body capture: 3d hands, face, and body from a single image. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 10975–10985.

[36] Georgios Pavlakos, Xiaowei Zhou, Konstantinos G Derpanis, and Kostas Daniilidis. 2017. Coarse-to-fine volumetric prediction for single-image 3D human pose. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 7025–7034.

[37] Georgios Pavlakos, Luyang Zhu, Xiaowei Zhou, and Kostas Daniilidis. 2018. Learning to estimate 3D human pose and shape from a single color image. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 459–468.

[38] Leonid Pishchulin, Mykhaylo Andriluka, Peter Gehler, and Bernt Schiele. 2013. Poselet conditioned pictorial structures. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 588–595.

[39] Varun Ramakrishna, Takeo Kanade, and Yaser Sheikh. 2012. Reconstructing 3d human pose from 2d image landmarks. In *European Conference on Computer Vision*. Springer, 573–586.

[40] Leonid Sigal, Alexandru Balan, and Michael J Black. 2008. Combined discriminative and generative articulated pose and non-rigid shape estimation. In *Advances in Neural Information Processing Systems*. 1337–1344.

[41] Ke Sun, Cuiling Lan, Junliang Xing, Wenjun Zeng, Dong Liu, and Jingdong Wang. 2017. Human pose estimation using global and local normalization. In *Proceedings of the IEEE International Conference on Computer Vision*. 5599–5607.

[42] Ke Sun, Bin Xiao, Dong Liu, and Jingdong Wang. 2019. Deep High-Resolution Representation Learning for Human Pose Estimation. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (2019).

[43] Xiao Sun, Jiaxiang Shang, Shuang Liang, and Yichen Wei. 2017. Compositional human pose regression. In *Proceedings of the IEEE International Conference on Computer Vision*. 2602–2611.

[44] Xiao Sun, Bin Xiao, Fangyin Wei, Shuang Liang, and Yichen Wei. 2018. Integral human pose regression. In *Proceedings of the European Conference on Computer Vision*. 529–545.

[45] Yaniv Taigman, Ming Yang, Marc'Aurelio Ranzato, and Lior Wolf. 2014. Deepface: Closing the gap to human-level performance in face verification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 1701–1708.

[46] Vince Tan, Ignas Budvytis, and Roberto Cipolla. 2017. Indirect deep structured learning for 3D human body shape and pose prediction. In *British Machine Vision Conference*.

[47] Bugra Tekin, Pablo Márquez-Neila, Mathieu Salzmann, and Pascal Fua. 2017. Learning to fuse 2d and 3d image cues for monocular body pose estimation. In *Proceedings of the IEEE International Conference on Computer Vision*. 3941–3950.

[48] Denis Tome, Chris Russell, and Lourdes Agapito. 2017. Lifting from the deep: Convolutional 3d pose estimation from a single image. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2500–2509.

[49] Jonathan J Tompson, Arjun Jain, Yann LeCun, and Christoph Bregler. 2014. Joint training of a convolutional network and a graphical model for human pose estimation. In *Advances in Neural Information Processing Systems*. 1799–1807.

[50] Hsiao-Yu Tung, Hsiao-Wei Tung, Ersin Yumer, and Katerina Fragkiadaki. 2017. Self-supervised learning of motion capture. In *Advances in Neural Information Processing Systems*. 5236–5246.

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

[52] Shih-En Wei, Varun Ramakrishna, Takeo Kanade, and Yaser Sheikh. 2016. Convolutional pose machines. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 4724–4732.

[53] Wanglong Wu, Meina Kan, Xin Liu, Yi Yang, Shiguang Shan, and Xilin Chen. 2017. Recursive spatial transformer (rest) for alignment-free face recognition. In *Proceedings of the IEEE International Conference on Computer Vision*. 3772–3780.

[54] Donglai Xiang, Hanbyul Joo, and Yaser Sheikh. 2019. Monocular total capture: Posing face, body, and hands in the wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 10965–10974.

[55] Wei Yang, Wanli Ouyang, Hongsheng Li, and Xiaogang Wang. 2016. End-to-end learning of deformable mixture of parts and deep convolutional neural networks for human pose estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 3073–3082.

[56] Wei Yang, Wanli Ouyang, Xiaolong Wang, Jimmy Ren, Hongsheng Li, and Xiaogang Wang. 2018. 3d human pose estimation in the wild by adversarial learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 5255–5264.

[57] Yi Yang and Deva Ramanan. 2011. Articulated pose estimation with flexible mixtures-of-parts. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 1385–1392.

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

[59] Xingyi Zhou, Qixing Huang, Xiao Sun, Xiangyang Xue, and Yichen Wei. 2017. Towards 3d human pose estimation in the wild: a weakly-supervised approach. In *Proceedings of the IEEE International Conference on Computer Vision*. 398–407.

[60] Xingyi Zhou, Xiao Sun, Wei Zhang, Shuang Liang, and Yichen Wei. 2016. Deep kinematic pose regression. In *European Conference on Computer Vision*. Springer, 186–201.