A Survey on Deep Learning-Based Approaches to Estimation of 3D Human Pose and Shape from Images for the Smart Environments

Sh. Maleki Arasi[1] and E. Seyedkazemi Ardebili[2](✉)

[1] Department Electrical Engineering, Sahand University of Technology, Tabriz, Iran
shadi.maleki93@gmail.com
[2] Department Computer Engineering, Kocaeli University, 41001 İzmit, Kocaeli, Turkey
seyed.esk@gmail.com

Abstract. Nowadays, on the one hand, due to the increase of equipment and smart environments, increasing attention to intelligence and the need to know more about the pose and shape of humans in these smart environments, and on the other hand the increase of being integrated the virtual world with the real world caused that the proper representation of humans in the virtual world has a great importance. Hence, human analysis of images has become very important. However, this work of human analysis not only goes beyond estimating a two-dimensional pose for one or multiple person, but it also goes beyond estimating a simple three-dimensional skeleton.

The estimation of 3D human pose and shape of images has received special attention due to various applications in the real world.

After studying and reviewing the papers in this field, it can be concluded that the existing approaches to obtain these estimates can be broadly grouped into two main approaches. An optimization-based approach and a deep learning-based approach are presented in which deep learning-based approaches are made in two methods: parametric and non-parametric.

Optimization-based approaches provide the most reliable solution for obtaining these three-dimensional estimates. However, optimization-based approaches are slow to implement, sensitive to appropriate initialization, and often fail due to weak local minimums. So, the focus is on deep learning approaches that regress poses and shapes directly from images. But on the drawbacks, can be said of these deep learning-approaches require a lot of training data and time-consuming; in the other words, their execution time is also high and produce low-resolution 3D predictions. So after reviewing both available approaches, it is found that both methods are challenged to obtain acceptable results. In this paper we have mentioned the optimization-based approaches in general and our main focus is on the deep learning-based approaches.

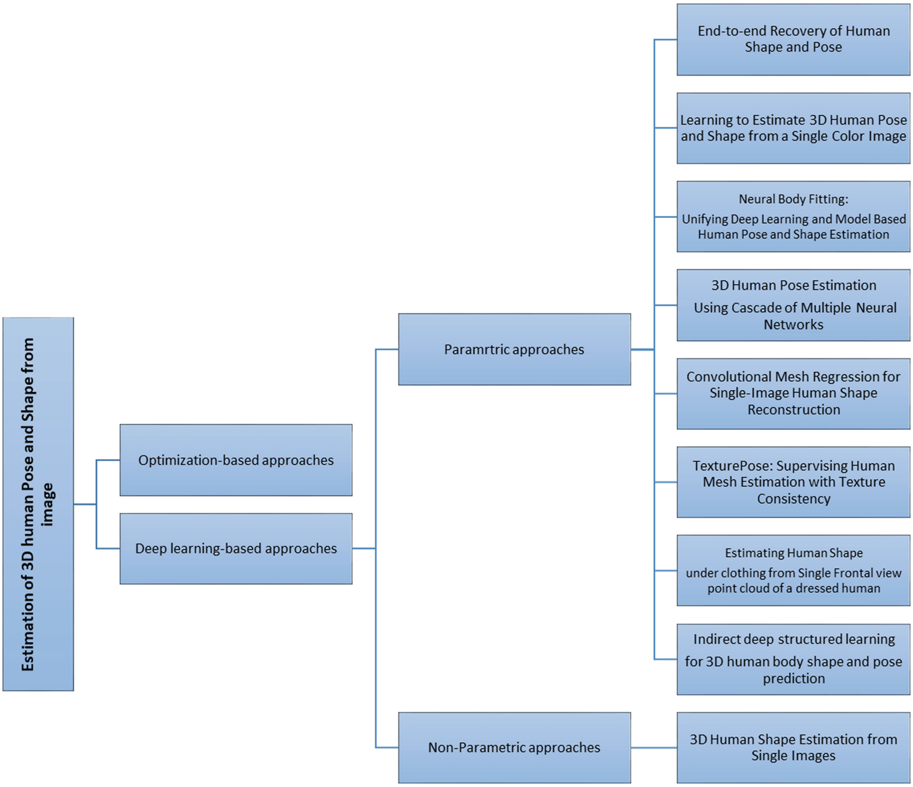
Keywords: Deep learning · Convolutional Neural Network (ConvNet) · Skinning Multi-person Liner (SMPL) · Active Shape Model (ASM) · Neural Body Fitting (NBF)

M. Ben Ahmed et al. (Eds.): SCA 2020, LNNS 183, pp. 891–916, 2021.
https://doi.org/10.1007/978-3-030-66840-2_68

1 Introduction

Smarting is one of the most attractive topics that technology manufacturers and technology customers have shown interest in recent years. A survey on the concept of smarting can be found in a personal vehicle, smart cars, smart home, smart building, and even in a smart city. For more detailed and perhaps more specialized examination of smartening, an environment such as a home, a hospital, or an office with a smart management center (or decision center) is considered. After studies and efforts to increase the smartening of this environment, it is concluded that in these environments, to increase smartening, one of the basic needs is to increase the ability of the management center to communicate effectively with the environment and the elements in those environments.

Table 1. The table above shows the main approaches for estimation 3D human pose and shape


Estimation of 3D human Pose and Shape from image
Optimization-based approaches
Deep learning-based approaches
Paramrtric approaches
Non-Parametric approaches
End-to-end Recovery of Human Shape and Pose
Learning to Estimate 3D Human Pose and Shape from a Single Color Image
Neural Body Fitting: Unifying Deep Learning and Model Based Human Pose and Shape Estimation
3D Human Pose Estimation Using Cascade of Multiple Neural Networks
Convolutional Mesh Regression for Single-Image Human Shape Reconstruction
TexturePose: Supervising Human Mesh Estimation with Texture Consistency
Estimating Human Shape under clothing from Single Frontal view point cloud of a dressed human
Indirect deep structured learning for 3D human body shape and pose prediction
3D Human Shape Estimation from Single Images

Given the most important element of any environment can be considered human, so recognizing more and more human poses and shapes leads to a better understanding of the environment, and as a result in many cases leads to increased smart decision-making power. Therefore, with the increasing integration of virtual environments with humans as the most important element in the scene and to increase the relationship between virtual environments and everyday life, one of the most basic and challenging tasks is human perception of available two-dimensional images and its analysis. Estimation of 2D and 3D human pose and shape (mesh) from images is done for the purpose of understanding and analysis. Due to the comprehensiveness outputs of these estimating, it can be immediately used in animation, correction, measurement, manipulate and reload to be used. However, their most important application is in the smartening of environments and buildings that deal with humans in some way.

In general, the proposed approaches for estimation 3D human pose and shape from images categorized in optimization-based and deep learning-based which is shown in the Table 1. Each of these approaches has advantages and disadvantages, which according to trade-off of speed and accuracy, one of the methods of these two approaches has been used. All our efforts are to introduce the best methods for reconstructing and estimating the human pose and shape from in-the-wild images so that we can use them to have smarter environments. Therefore, in this article, we try to present the approaches and methods for those who want to choose an approach and method according to their datasets and environment by reviewing the approaches and summarizing the available methods. Our main aim is on presenting the deep learning-based approaches, so the optimization-based approaches are mentioned in general.

2 Optimization-Based Approaches

In optimization approaches [12], the best answer (according to a set of criteria) is selected from a set of possible answers for a specific problem. The goal is minimizing or maximizing of a Real Function. In general, the term optimization refers to a process that aims to find the best values of one (or more) functions in a defined domain. That is, the problem is finding the best answer from a set of possible candidate answers. The methods that use an intermittent optimization scheme to update parameters locally are sensitive to initial values.

The work of Zhou et al. In 2015 [12] is a convex relation approach which estimate 3D Shape from 2D Landmarks. This method uses an augmented shape-space model, in which a shape is represented as a linear combination of rotating base shapes. It can show a linear representation of both intrinsic shape deformation and exterior viewpoint changes. Convex relaxation of orthogonality constraint to convert the entire problem into a spectral norm regularized linear inverse problem, which is a convex program. So this convex relaxation provides an efficient algorithm for solving global convex applications.

Another method to this approaches is an Improved Method for 3D Shape Estimation Using Active Shape Model (ASM), [20] which presented by Hoang et al. In 2017 [24]. This work uses the active shape model to estimate 3D poses and shapes as a linear combination of predefined basic shapes and fits with the 2D input landmarks. This model has improved the execution time and output accuracy by categorizing the data into sub-spaces.

3 Deep Learning-Based Approaches

Deep Learning-Based Approaches Description

Over the past decade, Deep Learning and Computer Vision have been among the interesting areas of research in Artificial Intelligence and Machine Learning. Therefore, it is normal for researchers in these two fields to pay more attention to the use of deep learning models in computer vision, and the next logical step is to move forward in the field of computer vision. As a result, the field of computer vision is shifting from statistical methods to deep learning. There are still challenging problems in computer vision. Nevertheless, deep learning methods are achieving state-of-the-art results to solve some specific problems. One of the most important models of deep learning for computer vision and its applications in related fields are convolutional neural networks (CNN) which by applying it, state-of-the-art results have been obtained.

In this section, methods based on deep learning that has been done in both parametric and non-parametric ways and their significant impact on the results are expressed.

3.1 Parametric Approaches

A parametric model records all its information about the existent data in its parameters. That means to predict the amount of future data from the current state of the model, only its parameters are needed. For example, linear regression with one variable has two parameters. If these two parameters are available, a new value can be predicted. On the other hand, a non-parametric model can capture more subtle aspects of data.

Parametric approaches, also considered "traditional", require a number of hypotheses. This approaches includes linear regression, logistic regression, linear differentiation analysis and so on.

3.1.1 End-to-End Recovery of Human Shape and Pose

This is an end-to-end framework for the full 3D Human Mesh Recovery (HMR) of the human body by a single RGB image [1]. This work describes the shape and angles of parameterized 3D joints. The main purpose of which is minimizing the projection function of key-points, (Due to its existing network, this approach makes it possible to train the model with real images that only include 2D Ground truth interpretations. As a result, it eliminates the need for costly 3D ground truth. It has also used an adversarial training to check the reality of production parameters. To evaluate the production meshes in an adversarial network, there is a database of 3D meshes of the human body with various shapes and poses. These meshes do not necessarily need the corresponding images, so this data is expressed as unpaired.

In this network, the Skinned Multi-person Linear (SMPL) model [11] is used, which Parameterizes the body mesh by 3D joint angles (pose) and linear shape space (shape) with low dimension. The full 3D mesh of the human body is reconstructed directly with a forward-looking process of a single RGB image.

In this method Convolutional features of the image are sent to the iterative 3D regression module whose objective is to infer the 3D human body and the camera such that its 3D joints project onto the annotated 2D joints. The inferred parameters are also sent to an adversarial [11] discriminator network whose task is to determine if the 3D parameters are real meshes from the unpaired data [7]. This encourages the network to output 3D human bodies that lie on the manifold of human bodies and acts as a weak-supervision for in-the-wild images without ground truth 3D annotations. Due to the rich representation of the 3D mesh model, this data-driven prior can capture joint angle limits, anthropometric constraints (e.g. height, weight, bone ratios), and subsumes the geometric priors used by models that only predict 3D joint locations.

More concretely, the shape and pose resulting from SMPL [12] decomposition are mirrored, and train a discriminator for shape and pose independently. The pose is based on a kinematic tree, so the pose discriminators are decomposed and train one for each joint rotation. An overview of this framework is shown in Fig. 1. HMR has the ability to train with or without the use of any paired 2D-to-3D supervision. Because of that during the training, all images are with 2D ground truth joint interpretations, and in some cases, 3D interpretations are considered. When Ground truth 3D information is available, it is used as an intermediate loss. The overview of proposed framework:

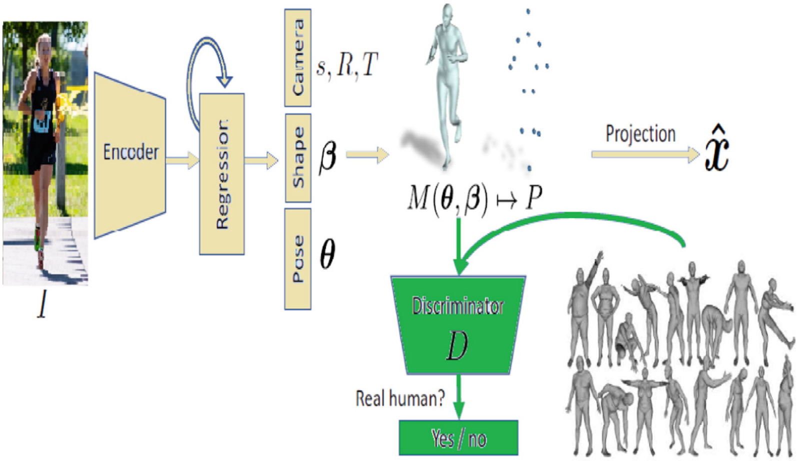


Fig. 1. An image I is passed through a convolutional encoder. Then it sent to an iterative 3D regression module that infers the latent 3D representation of the human that minimizes the joint reprojection error. The 3D parameters also are sent to the discriminator D, whose goal is to tell us if these parameters come from a real human shape and pose or not [1].

In this approach, a model without any 3D monitoring pairs is also trained. All methods approximately rely on direct 3D supervision and cannot train without it. Given these challenging learning setting, the results of this method are very competitive.

3.1.2 Learning to Estimate 3D Human Pose and Shape from a Single Color Image

Conv-Net approaches do not work well for estimation due to the lack of training data and low-resolution 3D predictions. Therefore, iterative optimization approaches prevail in this way, despite the high execution time and their common failures due to local minimums.

The new solution is a direct prediction of the pose and shape of the color image, and aims to bridge this gap and provide an effective Conv-Net-based approach. This method is a two-step approach that is the main part of the SMPL statistical body shape model combination approach in an end-to-end framework. Advantages of this method Fully accurate 3D estimates, requiring a small number of parameters and direct network prediction is possible, accurate and therefore easier using only 2D key points and silhouettes. As a result, the limiting assumption of a lack of natural images with 3D Grand truth for training is weakened. And while parametric model examples are used to teach 2D to 3D inference, the available 2D image interpretations can be used to teach 2D inference. One of the important advantages of using this parametric model is that its structure allows the use of a 3D loss at each vertex of the estimated 3D mesh at the time of training, and optimizes it directly for the surface. This loss correlates better with the 3D head-to-head error typically used for evaluation, and improves training compared to parametric regression.

Finally, a separable rendering is used for the 3D mesh effect generated by the 2D image, which makes it possible to adjust a differentiable renderer to project the generated 3D mesh back to the 2D image. A schematic framework of this method is shown in Fig. 2. Schematic framework of the method:

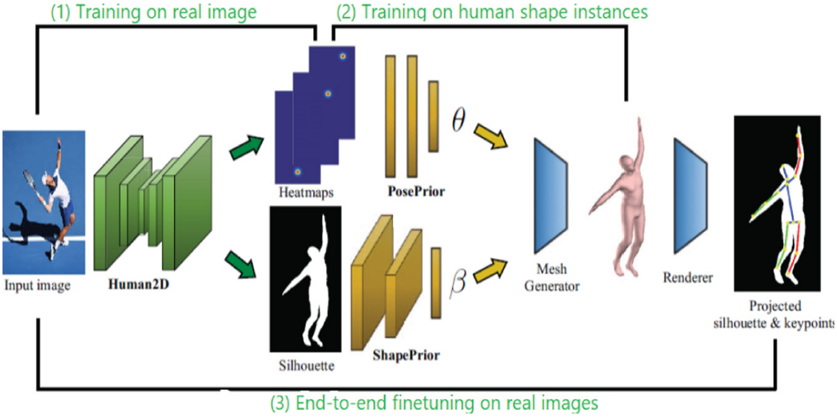

(1) Training on real image
(2) Training on human shape instances
Input image
Human2D
Heatmaps
PosePrior
θ
Silhouette
ShapePrior
β
Mesh Generator
Renderer
Projected silhouette & keypoints
(3) End-to-end finetuning on real images

Fig. 2. (1) An initial Conv-Net predicts heat-maps, and 2D masks using 2D pose data to train. (2) The two networks estimate the parameters of the SMPL statistical model using examples of parametric models for training. (3) The framework can be adjusted end-to-end without the need for 3D grand truth images [11].

Instead of using two Conv-Nets, one Conv-Net is trained as Human2D, which follows a stacked hourglass design [2], (using 2 hourglasses) that deals well between accuracy and execution time. And it has 2 outputs, one for key points and the other for silhouette. The output is a key point in the form of heat maps, and the Silouette outputs has two body and background channels using a pixel-wise binary cross-entropy. The second step of the work requires estimating the pose and 3D shape of the whole body from these key points and 2D silhouettes. This mapping can also be learned from the data for which 2 components of the network are trained:

1) Pose-Prior:

Its inputs are 2D key-point locations with the confidence of the detections (realised by the maximum value of each heat-map) and its outputs are estimates of 72 the pose coefficients θ.

2) Shape-Prior:

Its inputs are the silhouette and its outputs are estimates of 10 the shape coefficients β.

This method creates a modular path (i.e. updating Pose-Prior without retraining the entire network). The type of these inputs and outputs allows a large amount of training data to be generated by producing SMPL model samples with different 3D poses and shapes.

3.1.3 Neural Body Fitting: Unifying Deep Learning and Model Based Human Pose and Shape Estimation

This model-based method estimates the parameters from a single color image by a statistical body model. Traditional model-based approaches often have a goal function that measures the fit between the model and image observations. They do not require 3D training data but must be initialized. While forward-looking models such as CNNs, which directly predict key points, do not require initialization, images with 3D pose interpretations should be available. The CNN architecture provides a link to take advantage of both methods and does not require initialization and large amounts of 3D training data. The Neural Body Fitting (NBF) approach [13] integrates a statistical body model [12] within a CNN, leveraging reliable bottom-up semantic body part segmentation and robust top-down body model constraints.

This work makes several principled steps towards a full integration of parametric 3D human pose models into deep CNN architectures, and use a region-based 2D representation, namely a 12-body-part segmentation, as an intermediate step prior to the mapping to 3D shape and pose. This segmentation provides full spatial coverage of a person as opposed to the commonly used sparse set of key-points, while also retaining enough information about the arrangement of parts to allow for effective lifting to 3D. The NBF method is a linked architecture that integrates a human body model into a deep learning architecture (CNN) and uses body partitioning as an intermediate representation. From a color image or meaningful image segmentation, it directly predicts the model parameters, and these parameters are transferred to the flexible and realistic SMPL body model to produce 3D mesh, and then to evaluate the cost function in 2D space, 3D joints are projected to 2D images. NBF therefore accepts both full 3D supervision (in the model or 3D Euclidean space) and weak 2D supervision (if images with only 2D annotations are available).

The goal is to build a simple processing path with components that can be optimized in isolation to reduce both the number of hyperparameters and interactions and to train the components of the model sequentially.

There are 2 main step in this architecture:

1) Segment the predicted body parts from color images (first receives the input cut (512 × 512) and produces a component segmentation. A RefineNet model (based on ResNet-101) is used. This component partition is in the form of color code (color-coded) and its size has been changed to (224 × 224) and is given to the second step as an RGB image).
2) Use this segmentation to predict the lower dimension parameters of the mesh (this step itself consists of two parts: a regression network (ResNet-50) whose output is 226 to the SMPL parameter (shape and pose) and a set of non-trainable from the layers that implement the SMPL model and an projection image).

NBF predicts the parameters of the body model from a color-coded part segmentation map I using a CNN-based predictor parameterized by weights w. The SMPL model and a simple 2D projection layer are integrated into CNN estimator. Depending on the kind of supervision used to train, output a 3D mesh, 3D skeleton joint locations or 2D joints. This flexible implementation allows us to experiment with the 3D losses only for parts of the data, moving towards a weakly supervised training scenario that avoids expensive 3D labeled data. With 3D information for only 20% of our training data, we could reach similar performance as with full 3D annotations. This encouraging result is an important finding for the design of future datasets and the development of 3D prediction methods that do not require expensive 3D annotations for training.

3.1.4 3D Human Pose Estimation Using Cascade of Multiple Neural Networks

This method proposes a method called cascade of multiple neural networks (CMNN) [24] in following two steps:

1) Create the initial estimated 3D shape using the Zhou et al. [28] method with a small number of basis shapes,
2) Make this initial shape more alike to the original shape by using the CMNN. In comparing to existing works, the proposed method shows a significant outperformance in both accuracy and processing time.

In this method, the problem of 3D-to-2D compatibility has been done according to the ASM method. The way to use the CMNNs to estimate 3D shapes:

First, from the input 2D shape, the initial estimated 3D shape is created by using Zhou et al. method. Then, this shape is adjusted by the CMNN [25] to make it more resemblance to the real 3D shape. This proposed method uses the Zhou et al. method with a small number of predefined basis shapes to make sure it can be used in the real-time application.

Network input: includes the coordinates (x, y) of the 2D input (X), and the z coordinates of $\mathrm{S}^{-\,(t-1)}$. Network output: An update vector to generate the z coordinate of $\mathrm{S}^{-\,(t)}$. The structure if cascade is shown in Fig. 3. The cascade consists of T stages

C = {C1, . . ., CT}. Each stage Ct includes L neural networks. Each neural network predicts an update vector for q ∈ [1, p] joints, and consists of one input layer with 3 × P nodes, two hidden layers each containing 20 nodes, and an output layer with q nodes. The update vector at stage t is the combination of the outputs of all neural networks in this stage.

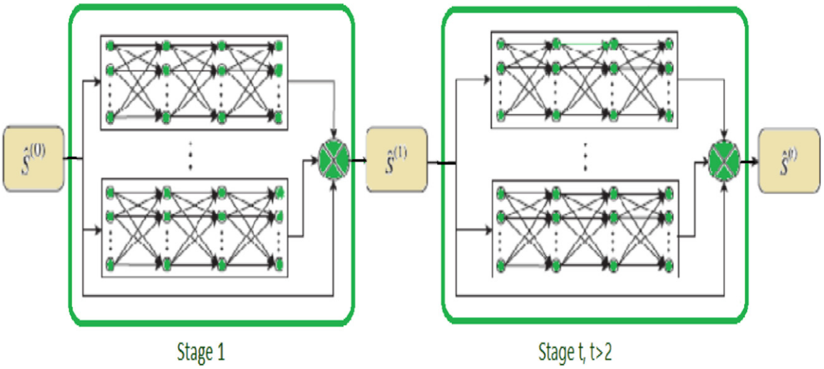


Fig. 3. The cascade consists of T stages. There are L neural networks in each stage. The combination of outputs of L neural networks is the update vector for the current estimated shape ˆ S. Each neural network has four layers: an input layer, two hidden layers, and an output layer [25].

At the training step after learning the dictionary of basis 3D shapes, the 2D shapes of all 3D shapes in the training data are created by projecting the 3D joints to a 2D plane. Then, the initial-estimated shapes 3D shapes ˆ S(0)s of these 2D shapes are initialized by using the method proposed by Zhou et al. The target to train the neural network is the difference between the (z) coordinate of the current estimated shapes ˆ S(t − 1)s and the ground truth shapes Ss of q joints corresponding to this neural network.

The overview of the CMNN:

3.1.5 Convolutional Mesh Regression for Single-Image Human Shape Reconstruction

The purpose of this method is to address the problem of estimating posture and shape by trying to reduce reliance on the parametric model, which is usually SMPL. In this method, the poses and shapes are regressed directly from the images. This approach [16] proposed to take a more hybrid route towards pose and shape regression. While maintaining the SMPL mesh topology, for an input image, instead of directly predicting the model parameters, the positions of the 3D mesh vertices are first estimated. To achieve this, the Graph-CNN architecture is proposed, which explicitly encodes the mesh structure and, while the regression target is for each vertex of its 3D location, processes the image properties attached to its vertices. Each typical CNN is used to extract the features attached to the coordinates of the vertex of the pattern mesh, and the processing on the graph structure defined for Graph-CNN continues, and finally each vertex deforms its 3D position in the mesh to finds targets.

This makes it possible to retrieve the full 3D geometry of the human body without the explicit need for a predefined parametric space, and after estimating the 3D position for each vertex, if the existing prediction is required to match a particular model, the parameters can be regressed it from mesh geometry. The first part of the work is an image-based CNN that extracts the general feature of the input representation and follows the Resnet-50 architecture whose final fully connected layer is ignored and only the 2048D feature vector after the pulling layer. Is kept. To regress the 3D coordinates of the vertices of the mesh from the CNN graph, this method starts from a template human mesh with N vertices as depicted in Fig. 4. The architecture starts from a patterned human mesh with N vertices, and according to the extracted feature vector 2048D, these attributes are attached to the 3D coordinates of each vertex in the pattern mesh.

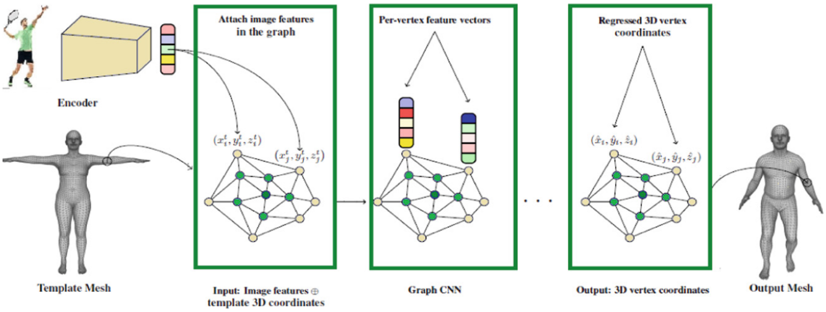

Encoder
Template Mesh
Attach image features in the graph
(x_i^t, y_i^t, z_i^t)
(x_j^t, y_j^t, z_j^t)
Input: Image features ⊕ template 3D coordinates
Per-vertex feature vectors
Graph CNN
. . .
Regressed 3D vertex coordinates
$(\hat{x}_i, \hat{y}_i, \hat{z}_i)$
$(\hat{x}_j, \hat{y}_j, \hat{z}_j)$
Output: 3D vertex coordinates
Output Mesh

Fig. 4. Depending on an input image, CNN will encode the image into a low-dimensional feature vector. This feature vector is embedded by placing it in three-dimensional coordinates of each vertex i in the graph specified by the template human mesh. It then processes through a series of convolutional Layers and processes the coordinates of the three-dimensional vertices of the deformed mesh [16].

The CNN graph uses the 3D coordinates of each vertex along with the input properties as input, and the purpose of estimating the 3D coordinates for each vertex in the output is the deformation mesh. The existing processes are by graph convolution layers, in which the formulas of the approach of Kipf et al. [17] are used. For the graph convolution layers, this work makes use of residual connections as they help in speeding up significantly the training and also lead in higher quality output shapes. Also, Batch Normalization [19] is replaced by Group Normalization [32]. Batch Normalization leads to unstable training and poor test performance, whereas with no normalization the training is very slow and the network can get stuck at local minima and collapse early during training. Besides the 3D coordinates for each vertex, this Graph CNN also regresses the camera parameters for a weak perspective camera model. Following Kanazawa et al. [12], this work predicts a scaling factor s and a 2D translation vector t. Since the prediction of the network is already on the camera frame, so there is no need to regress an additional global camera rotation. The camera parameters are regressed from the graph embedding and not from the image features directly. This way gets a much more reliable estimate that is consistent with the output shape.

In general, this hybrid approach is comparable to model-based approaches and is not largely sensitive to the type of inputs. And it allows us to connect features extracted from RGB pixels, segmentation of meaningful segmentations or even dense correspondence. The overview of proposed Framework:

It should be noted that model-based approaches create precise meshes of naked bodies under human clothing, but are unsuccessful in estimating the details and elements of the model, such as hair or clothing. On the other hand Non-parametric volumetric approaches estimate complete shapes but are limited in resolution and partial estimates.

3.1.6 Texture-Pose: Supervising Human Mesh Estimation with Texture Consistency

As mentioned, due to the lack of natural images with three-dimensional shape grand truth for training, the main challenge is reliable resources. This work [4] has relied on more clues that are present in natural images without the need for additional interpretations or changes in network architecture and are often ignored. Texture-Pose is a neural network training approach for model-based human pose estimation, with direct supervision of natural images. This method uses the conclusion that a person's appearance does not change significantly in a short film or for multi-view images.

This seemingly insignificant and often overlooked cue goes a long way for model-based pose estimation. This parametric model is used to calculate the texture map for each frame, assuming it is fixed. Which makes each point of the texture map have the same value in all frames. Due to the texture transfer in the space of this map, there is no need to calculate the camera movement and assume that the frames are smooth. This general formulation makes the approach flexible and practical, especially in video images and duplicate images. The parametric model used in this work is also SMPL. The joints of the body X are the linear composition of the vertices of the mesh, so using a pre-trained linear regressor W, the mesh can be mapped to the desired joint (X = WM). The overview of this work is shown in Fig. 5.

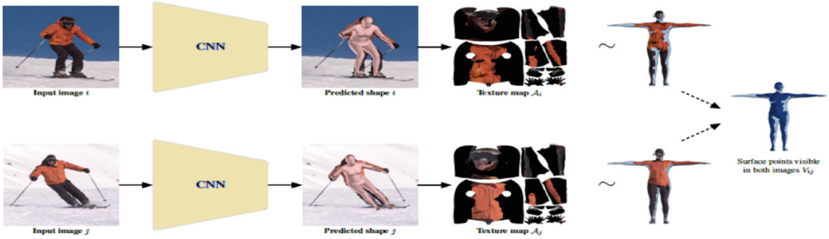


Fig. 5. Here, for simplicity, the input during the training contains two j, i images of the same person. The basic assumption is that a person's appearance does not change dramatically in the input images (i.e., the frames are made from a single film, or from synchronized multimedia cameras). The deep network works on both the image and estimates the shape of the person. After that, the projected shape is scattered on the image and after creating visibility for any point on the surface, texture maps Ai and Aj are created [4].

SMPL production meshes are modifications of the original T pattern. The corresponding UV map guides the pattern surface onto an A image, a texture map of each pixel t called texel. By making the mapping between the textiles and the mesh, the surface coordinates become fixed and independent of the changes in the surface geometry in 3D. The goal is to learn a predictor f that is perceived by a deep network and maps a single I input image to the parameters of a person's pose and shape on the image. The network output specifies the SMPL position and shape parameters. This deep network, with the exception of the output, regresses the 3D rotations with the representation provided by Zhou et al. [31]. The overview of the proposed texture consistency supervision:

The important observation, (that the person's appearance remains constant translates to a texture consistency loss), forces the two texture maps to be equal for all points on the V_{ij} surface that can be seen in both images. This lass acts as a network monitor and complements other weak lasses commonly used in training.

According to the parameters of the pose and the shape of the mesh M and the corresponding 3D joints, X is generated.

The mesh can be projected to the image using the estimated camera parameters. Through efficient computation (MPI-IS. Mesh processing library. https://github.com/MPI-IS/mesh.), this work can infer the visibility for each point on the surface, and as a result, for every texel_t of the texture map A. To guarantee that this method gets a valid 3D shape, this method used the adversarial prior, which factorizes the model parameters into: (i) pose parameters θ, (ii) shape parameters β, and (iii) per-part relative rotations, that is one 3D rotation for each of the 23 joints of SMPL. In the end, it trains a discriminator D_k for each factor of the body model. When there is access to multiple views i and j of a subject at the same time instance, then the main additional constraint it needs to enforce is that the pose of the person is the same across all viewpoints. This could be incorporated, by simply forcing all the pose parameters to have the same value. This generic formulation makes this approach particularly flexible and applicable in monocular video and multi-view images alike.

3.1.7 Estimating Human Shape Under Clothing from Single Frontal View Point Cloud of a Dressed Human

In general, model-based methods are not practical for estimating loose clothing, but non-modeled, free-change methods, because they are not limited to the naked body space, can hold clothing or other surface details but cannot shape Estimate the actual clogged with clothing.

In this approach, the advantages of both methods are combined and a personalized statistical body model is presented that describes the clothes as deviations from the parametric model of naked man. This approach is the first method to accurately estimate the parameters of the naked body shape from the depth or point cloud images of the uniformed front view of a dressed man, which has been proposed to deal with the situation of comfortable clothes, and a new target function has been designed that offers the benefits of model-based shape estimation and free variations for comfortable clothing. Provides free changes to deal with casual wear. The task is to estimate the naked shape parameters of a human wearing casual clothing from single-frame point cloud. Depth images of humans are captured by only one Microsoft Kinect Sensor v2. So the point clouds generated from depth images only contain part of clothed human surface which is visible to the depth camera. The overview of this method is displayed in Fig. 6. The overview of proposed method:

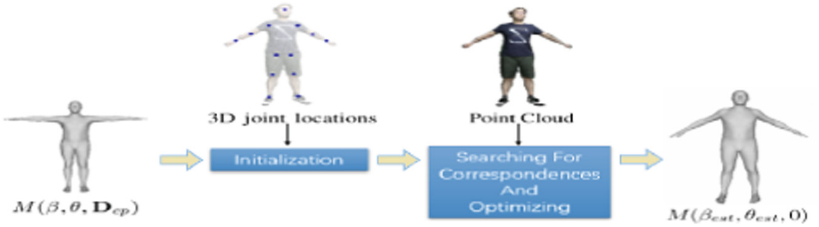

3D joint locations
Point Cloud
Initialization
Searching For Correspondences And Optimizing
$M(\beta, \theta, \mathbf{D}_{cp})$
$M(\beta_{est}, \theta_{est}, 0)$

Fig. 6. At first, according to 3D joint locations which are automatically detected by algorithm [8] integrated in Kinect, the shape and pose parameters of model are initialized. Then, according to input front-view point cloud, multi-step searching for correspondences and optimizing are applied, and finally, the estimated shape parameters and estimated model are obtained [6].

In order to personalize the original SMPL model [11] for dealing with the task of estimating shape under casual clothes, a set of auxiliary variables D_{cp} applied to model template T is used to describe personalized deviation for more accurate shape estimation. Shape and pose parameters are initialized according to 3D articulated locations automatically by the algorithm in the approach of Alldieck [22], Video-based remake on Kinect. If the human height is known, the shape parameters are controlled to make the model height more realistic, and if it is unknown, the human height can be calculated with 3D joint locations. Then, due to the drastic changes in the human condition, the focus is more on the global orientation of the body. Finally, the corresponding pairs of vertices (vi, pi) are found, where v is the vertex of the body model and p belongs to the point cloud. To be specific, they are achieved by calculating the rigid transformation of torso, and 3D rotation of limbs respectively. The important part of this work is its objective function which is minimized in the last step. In this work, the Microsoft Kinect v2 sensor is used for training. And the performance of the work is compared with other methods that have different objective functions and the existing method is selected as the most effective method.

3.1.8 Indirect Deep Structured Learning for 3D Human Body Shape and Pose Prediction

This method [21] is used for indirect training of deep networks, for structural prediction of three-dimensional human shape and pose, and has been proposed due to the need to reduce reliance on expensive three-dimensional Grand truth labels. Unlike most modern approaches, this method of training in real-world images does not require hard-to-obtain 3D human-shape tags, but instead uses the trained decoding power of artificial data. To achieve this goal, an encryption-decryption network (Auto Encoder) is trained using the two-step method described below.

In the first step, an encoder is trained to predict a body silhouette using SMPL parameters (a statistical body shape model) as input. In the second step, the entire network is trained on the actual image and the corresponding silhouette pairs, while the encoder is held constant. As a result, this method allows indirect learning of body shape and posture from real images without the need for Grand truth parametric data. In this work the encoder and decoder split into three units each, serving particular purposes as described in Fig. 7. As a result, this method allows us for an indirect learning of body shape and pose parameters from real images without the need for any ground truth parameter data. Components of proposed encoder-decoder network:

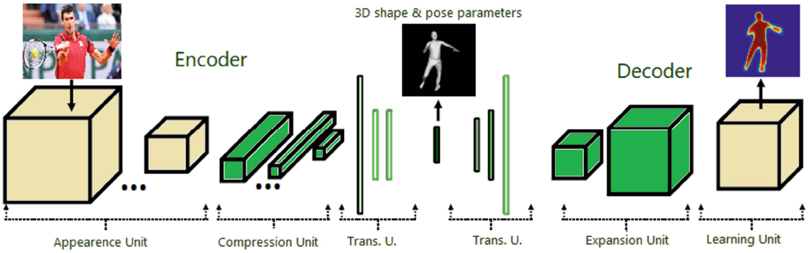


Fig. 7. The figure above shows the main components of the encoder-decoder network of the existing approach [21].

For ease of explanation, the encoder is divided into appearance, compression and transmission units. The appearance unit teaches convolutional filters for human silhouette and background separation. The compression unit further compresses the output of the appearance unit to a vector of dimensions 64 × 1 × 1. The transfer unit then converts this vector into shape and pose parameters using three fully connected layers. Similarly, decoders are divided into units of transmission, expansion, and learning. The transfer unit converts 3D shape and pose parameters into a low-dimensional image (9 × 9), an 8-channel image through three fully connected layers, and a deformation layer. This method has shown high accuracy in artificial images. While the accuracy of the method in real-world images decreases, even when using training methods, by not exposing any real image and a pair of corresponding shape and pose parameters, it can regain a close fit to Grand truth. On the other hand, by implementing more complex architectures in the network of this method and additional higher quality training data, it is possible to enable the proposed method to compete with modern direct learning approaches.

3.2 Non-parametric Approach

The non-parametric model allows more information to be provided from the current data set so that future data can be predicted. Usually these parameters can express the properties of the data much better than the parametric models, have a greater degree of freedom and are more flexible. For example, a Gaussian mixed model has more flexibility. If more data is observed, future data can be predicted even better. So knowing only the parameters is enough for a parametric model to predict new data. In the case of a non-parametric model, the prediction of future data is based not only on the parameters but also on the current state of the observed data.

3.2.1 Moulding Humans: Non-parametric 3D Human Shape Estimation from Single Images

While the recent progress in convolutional neural networks has allowed impressive results for 3D human pose estimation, estimating the full 3D shape of a person is still an open issue. Model-based approaches can output precise meshes of naked under-cloth human bodies but fail to estimate details and un-modelled elements such as hair or clothing. On the other hand, non-parametric volumetric approaches can potentially estimate complete shapes but, in practice, they are limited by the resolution of the output grid and cannot produce detailed estimates.

This method [23] uses a binary depth map representation to show and encode the 3D shape. To reconstruct the full 3D human shape, there are two depth maps, a depth map that records the visible surface elements that are directly visible in the image, and a hidden depth map that records the blocked surface of the estimate.

In general, this method designs an encoder decoder architecture that takes the single image as input and simultaneously creates an estimate for both depth maps. These depth maps are then combined to obtain a full 3D surface point cloud that can be easily reconstructed using Poisson reconstruction. And produce high-resolution outputs with the same amount of image input but much smaller dimensions than vertex-based volume representations (O (N2) compared to O (N3) where N is the size of the box that restricts humans. Frames in the input image). This (depth or deep) depth-based model also provides a competitive separator to improve the accuracy and humanity of 3D output. To reconstruct the full 3D human shape, there are two depth maps, a depth map that records the visible surface elements that are directly visible in the image, and a hidden depth map that records the blocked surface of the estimate, which is shown in Fig. 8. Non-parametric representation for human 3D shape:

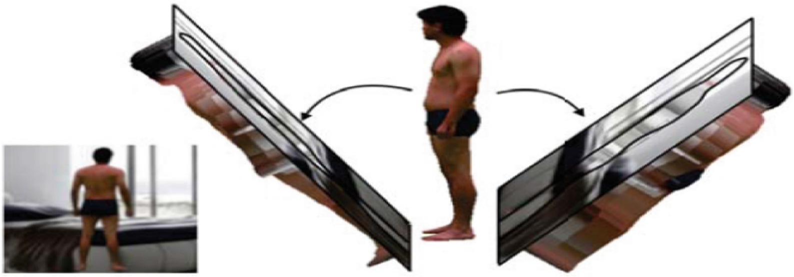


Fig. 8. With a single image, "visible" and "hidden" depth maps are estimated from the camera. Two depth maps can be viewed as two halves of a virtual "pattern" template [23].

Two 2D depth maps z $_{vis}$ and z $_{hid}$ according to a 3D mesh, obtained by animating a 3D human model or by reconstructing a real person from multiple views, and according to a camera hypothesis, i.e. location and parameters, by ray-tracing is introduced. To keep the depth values within a reasonable range and estimate them more accurately, a flat background a distance L behind the subject to define all pixels values in the depths maps in the range [− z $_{orig}$ . . . L]. The method framework is based on the stacked hourglass network proposed by Newell et al. (Alejandro Newell et al. 2016), and designed a 2-stack hourglass architecture that takes as input an RGB image I cropped around the human and outputs the 2 depths maps z_{vis} and z_{hid} aligned with I. See Fig. 9. Each of these modules has a set of convolutional and pooling layers that process features down to a low resolution and then up-sample them until reaching the final output resolution. The error obtained in mould-representating of this method is reduced and converged to a minimum value that corresponds to surface details that cannot be correctly encoded even with high resolution depth maps, i.e. when some rays intersect more than twice with the human surface for particular poses.

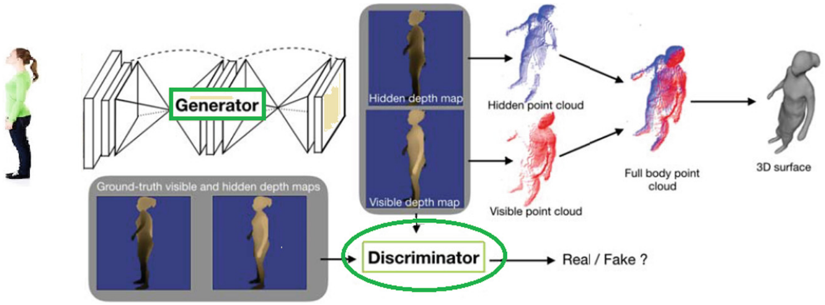


Fig. 9. According to a single image, "visible" and "hidden" depth maps are estimated. 3D dot clouds from these two depth maps are combined to form a 3D dot cloud of the whole body, as if they hold two halves of a pattern [23].

Finally, a competitive training method is followed according to the Generative adversarial network (GAN) [5], which is a framework for estimating productive models through an adversarial process, in which two models simultaneously, a productive G model that distributes the data. Obtains and teaches a discriminant model D that estimates the probability of a sample coming from training data or from generator G. The goal in this section is to accurately distinguish Grand truth depth maps from generated ones. The overview of proposed approach:

The 3D figure is reconstructed using Poisson [14] reconstruction, and to increase the humanity of the estimate, a competitive training with a discriminator has been used. This method can recover detailed surfaces while keeping the output to a reasonable size. This makes the learning stage more efficient. And this architecture can also efficiently incorporate a discriminator in an adversarial fashion to improve the accuracy and "humanness" of the output.

Table 2. The table above shows an overview of the optimization-based and deep learning-based approaches for estimation 3D human pose and shape.

Title	Author date	Techniques	Dataset	Properties
End-to-end recovery of human shape and pose	Kazanawa et al. /2018	GAN	LSP LSP_ extended MPII MS COCO Human 3.6 M MPI_INF_3DHP	1. Infer 3D mesh parameters, directly from image features This avoids the need for two stage training and also avoids throwing away a lot of image information 2. Going beyond skeletons, so output meshes, which are more complex and more appropriate for many applications 3. Its framework is trained in an end-to-end manner 4. It remains open whether increasing the amount of 2D data will significantly increase 3D accuracy

(*continued*)

Table 2. (*continued*)

Title	Author date	Techniques	Dataset	Properties
Learning to estimate 3D human pose and shape from a single color image	Georgios Pavlakos et al. /2018	The conventional ConvNet-based approach	UP-3D SURREAL Human 3.6 M	1. An end-to-end framework 2. Incorporation of a parametric statistical shape model, SMPL, within the end-to-end framework, enabling: – Prediction of the SMPL model parameters from ConvNet-estimated 2D key-points and masks to avoid training on synthetic image examples – Generation of the 3D body mesh at training time and supervision based on the 3D shape consistency – Use of a differentiable renderer for 3D mesh projection and refinement of the network with supervision based on the consistency with 2D annotations 3. Superior performance compared to previous approaches for 3D human pose and shape estimation at significantly faster running time

(*continued*)

Table 2. (*continued*)

Title	Author date	Techniques	Dataset	Properties
Neural body fitting: unifying deep learning and model based human pose and shape estimation	Mohamed Omran et al. /2018	Neural Body Fitting (NBF) a. hybrid architecture	UP-3D Human 3.6 M	1. Directly predicts the parameters of the model 2. Admits both full 3D supervision (in the model or 3D Euclidean space) and weak 2D supervision (if images with only 2D annotations are available) 3. It requires neither initialization nor large amounts of 3D training data 4. Build a simple processing pipeline with parts that can be optimized in isolation and avoiding multiple network heads 5. Analyze: (1) How the 3D model can be integrated into a deep neural network, (2) How loss functions can be combined and, (3) How a training can be set up that works efficiently with scarce 3D data

(*continued*)

Table 2. (*continued*)

Title	Author date	Techniques	Dataset	Properties
3D Human pose estimation using cascade of multiple neural networks	Van-Thanh Hoang et al. /2018	Cascade of multiple neural networks (CMNN)	MoCap Human 3.6 M	1. Create the initial 3D shape by using the method proposed by ASM methods with a small number of predefined basis shapes 2) Make estimated shape more accurate by using the CMNN 3. The proposed method outperforms in both accuracy and processing time 4. Its speed is fast enough to use in the real-time application

(*continued*)

Table 2. (*continued*)

Title	Author date	Techniques	Dataset	Properties
Convolutional mesh regression for single-image human shape reconstruction	Nikos Kolotouros et al. /2019	Graph-CNN architecture	Human 3.6 M UP-3D LSP	1. Reformulate the problem of human pose and shape estimation in the form of regressing the 3D locations of the mesh vertices, to avoid the difficulties of direct model parameter regression 2. Propose a Graph CNN for this task which encodes the mesh structure and enables the convolutional mesh regression of the 3D vertex locations 3. Demonstrate the flexibility of framework by considering different input representations 4. Current limitations (e.g., low resolution of output mesh, missing details in the recovered shape)

(*continued*)

Table 2. (*continued*)

Title	Author date	Techniques	Dataset	Properties
TexturePose: supervising human mesh estimation with texture consistency	Georgios Pavlakos et al. /2019	TexturePose, (an approach to train CNN)	Human 3.6 M MPII 3LSP	1. A novel approach to leverage complementary supervision from natural images through appearance constancy of each human across different frames 2. Demonstrate the effectiveness of texture consistency supervision in cases of monocular video and multi-view capture, consistently outperforming approaches with access to the same or more annotations
Estimating human shape under clothing from single frontal view point cloud of a dressed human	Wang et al. /2019	Using of point cloud	Point Cloud	1. Proposed the first method of estimating 3D naked body shape parameters from a single-frame frontal view point cloud of a dressed human 2. Design a novel objective function that combines the advantages of model-based shape estimation and free deformation method to deal with casual clothes

(*continued*)

Table 2. (*continued*)

Title	Author date	Techniques	Dataset	Properties
Indirect deep structured learning for 3D human body shape and pose prediction	Jun Kai Vince Tan et al. /2017	Autoencoder (encoder-decoder network)	Artificial Images Real Images (the Unite the People)	1. A novel encoder-decoder architecture for 3D body shape and pose prediction, 2. This method does not require hard-to-obtain 3D human shape and pose labels for training on real world images, but instead leverages the power of a decoder trained on artificial data
Moulding Humans: Non-parametric 3D Human Shape Estimation from Single Images	Gabeur et al./2019	Double depth map	3D HUMANS	1. This method can recover detailed surfaces while keeping the output to a reasonable size. This makes the learning stage more efficient 2. This architecture can also efficiently incorporate a discriminator in an adversarial fashion to improve the accuracy of the output 3. This representation allows a higher resolution output, potentially the same as the image input, with a much lower dimension than voxel-based volumetric representaions

4 Discussion and Conclusion

The purpose of this article is to address the existing approaches and methods for the problem of estimation of three-dimensional human pose and shape from images. Existing approaches to this work are categorized into two forms optimization-based and deep learning-based. In this paper, first, methods optimization-based and then in detailed parametric and non-parametric deep learning-based approaches for estimation three-dimensional human meshes were presented, which is shown in Table 2.

For this purpose, optimization-based approaches provide reliable results, but due to the lack of proper initialization and the usual failures due to the weakness of the minimum initialization, run time and slow convergence of the adaptation process, the use of deep learning approaches, which regresses poses and shapes directly from images, is enhanced by their high efficiency and accuracy.

Convolutional networks are not a practical candidate for this problem, due to the need for a lot of training data and low-resolution of 3D predictions. However, by providing a direct and efficient forecasting approach that is better than repetitive optimization methods, it has been shown that convolutional networks can provide an attractive solution to this problem. In general, the goal of all methods is a proper trade-off between the speed and accuracy of the output results.

To achieve the best results in estimating of the human pose and shape in smart environments, the best method must be studied and selected.

In general, it seems that non-parametric methods perform better than parametric methods in these three-dimensional estimates of images, although they also have some drawbacks.

Finally, we would like to thank all those who used to collect these articles, and especially the articles whose photos we have used in this work. We have tried to mention the names of all these people in our references.

References

1. Kanazawa, A., Black, M.J., Jacobs, D.W., Malik, J.: End-to-end recovery of human shape and pose. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT (2018)
2. Newell, A., Yang, K., Deng, J.: Stacked hourglass networks for human pose estimation. In: ECCV/CVPR (2016)
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. Pavlakos, G., Kolotouros, N., Daniilidis, K.: TexturePose: supervising human mesh estimation with texture consistency. In: IEEE/CVF International Conference on Computer Vision (ICCV), Seoul, Korea (South)/ICCV (2019)
5. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial nets. In: NIPS (2014)
6. Wang, J., Lu, Z., Liao, Q.: Estimating human shape under clothing from single frontal view point cloud of a dressed human. In: IEEE International Conference on Image Processing (ICIP), Taipei, Taiwan (2019)

7. Zhu, J.-Y., Park, T., Isola, P. and Efros, A.A.: Unpaired image-to-image translation using cycle-consistent adversarial networks. In: ICCV (2017)
8. Shotton, J., Fitzgibbon, A., Cook, M., Sharp, T., Finocchio, M., Moore, R., Kipman, A., Blake, A.: Real-time human pose recognition in parts from single depth images. Commun. ACM **56**, 116–124 (2013)
9. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: CVPR (2016)
10. Andriluka, M., Pishchulin, L., Gehler, P., Schiele, B.: 2D human pose estimation: new benchmark and state of the art analysis. In: CVPR (2014)
11. Loper, M., Mahmood, N., Tung, H.F., Harley, A.W., Seto, W., Fragkiadaki, K.: Adversarial inverse graphics networks: learning 2D-to-3D lifting and image-to-image translation from unpaired supervision. In: IEEE International Conference on Computer Vision (ICCV), Venice (2017)
12. Loper, M., Mahmood, N., Romero, J., Pons-Moll, G., Black, M.J.: SMPL: a skinned multi-person linear model. ACM Trans. Graph. (TOG) **34**, 1–16 (2015)
13. Omran, M., Lassner, C., Pons-Moll, G., Gehler, P., Schiele, B.: Neural body fitting: unifying deep learning and model based human pose and shape estimation. In: International Conference on 3D Vision (3DV), Verona (2018)
14. Kazhdan, M., Hoppe, H.: Screened poisson surface reconstruction. ACM Trans. Graph. **32**, 1–13 (2013)
15. MPI-IS: Mesh processing library. https://github.com/MPI-IS/mesh
16. Kolotouros, N., Pavlakos, G., Daniilidis, K.: Convolutional mesh regression for single-image human shape reconstruction. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA (2019)
17. Litany, O., Bronstein, A., Bronstein, M., Makadia, A.: Deformable shape completion with graph convolutional autoencoders. In: CVPR (2018)
18. Johnson, S.: Everingham, M.: Clustered pose and nonlinear appearance models for human pose estimation. In: BMVC (2010)
19. Ioffe, S., Szegedy, C.: Batch normalization: accelerating deep network training by reducing internal covariate shift. In: ICML (2015)
20. Cootes, T.F., Taylor, C.J., Cooper, D.H., Graham, J.: Active shape models-their training and application. Comput. Vis. Image Underst. **61**, 38–59 (1995)
21. Tan, J., Budvytis, I., Cipolla, R.: Indirect deep structured learning for 3D human body shape and pose prediction. In: British Machine Vision Conference 2017. BMVC (2017). https://doi.org/10.17863/CAM.21421
22. Alldieck, T., Magnor, M., Xu, W., Theobalt, C., Pons-Moll, G.: Video based reconstruction of 3D people models. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (2018)
23. Gabeur, V., Franco, J., Martin, X., Schmid, C., Rogez, G.: Moulding humans: non-parametric 3D Human shape estimation from single images. In: IEEE/CVF International Conference on Computer Vision (ICCV), Seoul, Korea (South) (2019)
24. Hoang, V., Jo, K.: An improved method for 3D shape estimation using active shape model. In: 10th International Conference on Human System Interactions (HSI), Ulsan, (2017)
25. Hoang, V., Jo, K.: 3-D human pose estimation using cascade of multiple neural networks. In: IEEE Transactions on Industrial Informatics (2019)
26. Ramakrishna, V., Kanade, T., Sheikh, Y.: Reconstructing 3D human pose from 2D image landmarks. In: Computer Vision–ECCV (2012)
27. Sun, X., Shang, J., Liang, S., Wei, Y.: Compositional human pose regression. In: IEEE International Conference on Computer Vision, ICCV (2017)

28. Zhou, X., Zhu, M., Leonardos, S., Daniilidis, K.: Sparse representation for 3D shape estimation: a convex relaxation approach. IEEE Trans. Pattern Anal. Mach. Intell. **39**(8), 1648–1661 (2017)
29. Zhou, X., Huang, Q., Sun, X., Xue, X., Wei, Y.: Weaklysupervised transfer for 3D human pose estimation in the wild. In: IEEE International Conference on Computer Vision, ICCV (2017)
30. Zhou, X., Leonardos, S., Hu, X., Daniilidis, K.: 3D shape estimation from 2D landmarks: a convex relaxation approach. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston (2015)
31. Zhou, Y., Barnes, C., Lu, J., Yang, J., Li, H.: On the continuity of rotation representations in neural networks. In: CVPR (2019)
32. Wu, Y., He, K.: Group normalization. In: ECCV (2018)