DC-GNet: Deep Mesh Relation Capturing Graph Convolution Network for 3D Human Shape Reconstruction

Shihao Zhou[1#], Mengxi Jiang[1#], Shanshan Cai[1], Yunqi Lei[1*]
[1]Department of Computer Science, School of Informatics, Xiamen University, 361005,
Xiamen, Fujian Province, China
{shzhou,jiangmengxi,sscai}@stu.xmu.edu.cn,yqlei@xmu.edu.cn

ABSTRACT

In this paper, we aim to reconstruct a full 3D human shape from a single image. Previous vertex-level and parameter regression approaches reconstruct 3D human shape based on a pre-defined adjacency matrix to encode positive relations between nodes. The deep topological relations for the surface of the 3D human body are not carefully exploited. Moreover, the performance of most existing approaches often suffer from domain gap when handling more occlusion cases in real-world scenes.

In this work, we propose a **D**eep Mesh Relation **C**apturing **G**raph Convolution **Net**work, DC-GNet, with a shape completion task for 3D human shape reconstruction. Firstly, we propose to capture deep relations within mesh vertices, where an adaptive matrix encoding both positive and negative relations is introduced. Secondly, we propose a shape completion task to learn prior about various kinds of occlusion cases. Our approach encodes mesh structure from more subtle relations between nodes in a more distant region. Furthermore, our shape completion module alleviates the performance degradation issue in the outdoor scene. Extensive experiments on several benchmarks show that our approach outperforms the previous 3D human pose and shape estimation approaches.

CCS CONCEPTS

• **Computing methodologies** → **Computer vision**.

KEYWORDS

3D human shape reconstruction, Graph Convolution Network, Deep mesh relation capturing

ACM Reference Format:
Shihao Zhou[1#], Mengxi Jiang[1#], Shanshan Cai[1], Yunqi Lei[1]. 2021. DC-GNet: Deep Mesh Relation Capturing Graph Convolution Network for 3D Human Shape Reconstruction. In *Proceedings of the 29th ACM International Conference on Multimedia (MM '21), October 20–24, 2021, Virtual Event, China.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3474085.3475242

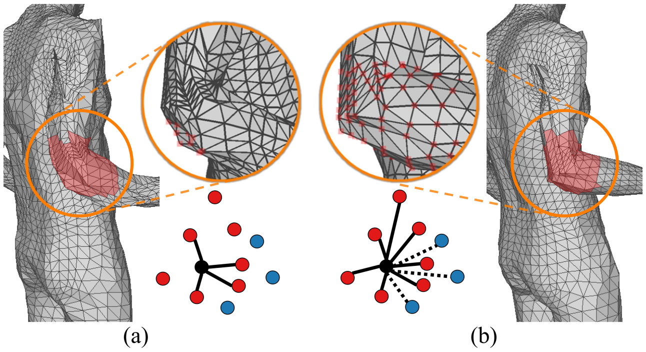


Figure 1: Illustration of different strategies to reason local structure. (a) Previous popular approaches are based on a pre-defined adjacency matrix that encodes only positive relations with physically connected nodes. (b) Our approach learns deep relations (*i.e.*, positive and negative) between nodes in a more distant region. The inference node, positively related node and negatively related node are shown in the black, red and blue circle, respectively. The solid line denotes a positive relationship, while the dashed line denotes a negative relationship.

1 INTRODUCTION

3D human pose and shape estimation is a fundamental yet challenging task in computer vision. There are plenty of approaches proposed to accurately capture 2D pose and even 3D joint locations [9, 26, 43, 46, 57, 58]. Since sparse joints alone cannot provide enough information for analyzing humans [22], incremental recent works interest in recovering the 3D mesh of a human body, where the 3D surface is defined.

To obtain 3D mesh for a human being in an image, optimization-based approaches generate a reliable human body fitting [5, 25]. Unfortunately, their slow inference speed and sensitivity to initialization have shifted the focus to regression-based approaches, which directly regresses mesh coordinates [7, 22, 51] or the parameters [17, 33, 37] of the human body model (*e.g.*, SCAPE [2] and SMPL(-X) [29, 35, 40]). Although regression-based methods achieve clear performance improvement in constrained environments, there are still limitations hinders better performance under the real scenario.

Firstly, most of existing regression-based approaches including vertex-level regression approaches [7, 22] and parameter regression approaches [17, 37] are based on a fixed adjacency matrix to encode the inherent shape nodes relations, which ignores deep relations between shape nodes and focus only on physically connected nodes,

[*] Corresponding Author.
[#] Both authours contributed equally to this research.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.
MM '21, October 20–24, 2021, Virtual Event, China
© 2021 Copyright held by the owner/author(s). Publication rights licensed to ACM.
ACM ISBN 978-1-4503-8651-7/21/10…$15.00
https://doi.org/10.1145/3474085.3475242

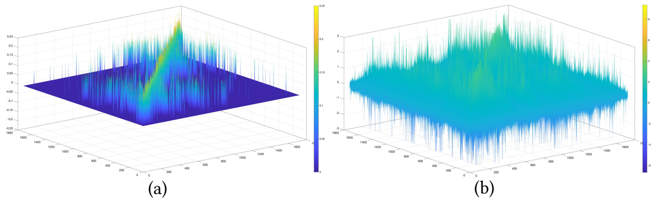


Figure 2: Visualization of the different adjacency matrix. (a) A pre-defined adjacency matrix with encoding only positive relations between physically connected nodes. (b) Our adjacency matrix learns subtle relations (*i.e.*, positive and negative) between nodes in a more distant region.

as shown in Figure 1(a). As a result, models often can not fully explore the spatial relations within the human body, which owns a highly related structure. Previous works [8, 57] attempt to explore a joint-to-joint topological structure for skeleton joints in the 3D pose estimation task. However, compared to the 3D pose estimation task with skeleton representation, a full 3D human shape reconstruction task needs to infer node-to-surface relations rather than a simple joint-to-joint. Thus, these approaches cannot be applied directly to the 3D shape estimation. To the best of our knowledge, exploiting the deep topological relations for the surface of the 3D human body is an unexplored yet important problem to the current 3D human shape reconstruction approaches.

Secondly, the full human body information is usually difficult to obtain in real scenarios due to various occlusions. Existing 3D annotation datasets that are collected in the limited indoor environment not carefully simulate such cases of body partially missing, which forms a gap issue of appearance domain (*i.e.*, the performance degradation in real-world scenes). Therefore, the generalization of existing shape reconstruction approaches trained on indoor 3D data are often poor when handling with more occlusion cases in real-world scenes, resulting in the performance degradation.

In this paper, to alleviate the above issues, we propose a Deep Mesh Relation Capturing Graph Convolution Network, namely DC-GNet, with a shape completion task. Firstly, to capture deep relation among mesh vertexes of body shape, we impose an adaptive adjacency matrix to learn both positive and negative relationships between nodes in a more distant region, as shown in Figure 2(b). Base on this learnable matrix, our network can aggregate subtle information from not only physically connected nodes but also nodes with long-distance, as shown in Figure 1(b). Secondly, we propose a shape completion task to alleviate the gap issue of appearance domain. Specifically, we first fabricate artificial holes on the surface of the training body shape data. Then, we force the network to recover a full body shape to learn prior about various kinds of human body part missing case. The overview of our approach is shown in Figure 3. Concretely, our network is started with the initial estimation, which is the feature extraction stage. In the pretrain process, the image features are inputted into DC-GNet with the proposed shape completion task, where the network learns the prior for occlusion cases. This network is further trained in the main inference process for the 3D human shape estimation.

Extensive evaluated experiments are conducted on several public benchmarks [14, 31], and the results show that our proposed approach outperforms the previous state-of-the-art 3D pose and shape estimation methods. Moreover, qualitative results on in-the-wild datasets [1, 27] show that our approach has learned better representation and generalization ability through making better use of the topology structure of the human body.

We summarize our contributions as follows:

- We propose a Deep Mesh Relation Capturing Graph Convolution Network, DC-GNet, to reconstruct 3D human shape from a single RGB image. It is the first attempt to learn deep relations between nodes among human mesh vertices and consider reasoning from more than partial structure, which boosts the model capturing complex local deformation.
- We first propose a shape completion module as an auxiliary task to alleviate the appearance domain gap issue between indoor and outdoor scenes, and thus our model can benefit from the rich indoor data source to learn the inference paradigm that can be well applied in natural environments.
- Extensive experimental results across several benchmarks demonstrate the effectiveness of exploring deep relations among mesh vertices, through comparing with the previous state-of-the-art 3D human shape reconstruction methods.

2 RELATED WORK

Although numerous approaches have been proposed to boost the topic of 3D pose estimation in the form of a skeleton in the last few years [23, 42, 45, 59, 60, 62], we will focus on closely-related works reconstructing the whole shape and pose in this Section [11, 21, 47].

2.1 3D Human Pose and Shape Estimation

Instead of a skeleton, recovering the shape of human body is a more challenging task. Bogo *et al.* [5] firstly introduced the fully automatic model-based approach, SMPLify, to estimate 3D human shape and pose from 2D pose by fitting a classical human body model SMPL. After that, Lassner *et al.* [25] applied SMPLify for building a dataset with fairly successful 3D fits. Beyond SMPLify, many different model-based approaches were proposed to explore including adversarial prior [17, 20], temporal information [3, 18], or even dealing with multiple humans [15, 54]. More recently, instead of predicting parameters of the model, model-free methods that directly regress each vertex were proposed to avoid representation issues [30, 37]. Venkat *et al.* [50] captured the local deformation by learning the “implicitly structure" of human mesh. Similarly, Kolotouros *et al.* [22] directly regressed the vertices of a template mesh to explore the topological structure explicitly. Moon and Lee [32] proposed an image-to-lixel network to model the prediction uncertainty for each mesh vertex. More interesting, many works [4, 12, 61] tried to obtain clothed mesh.

2.2 Graph Convolution Network for 3D shape Reconstruction

Recently, there has been a surge of approaches considering Graph Convolution Network(GCN) for capturing the graph structure of mesh, due to the graph-like nature of human mesh. Choi *et al.* [7]

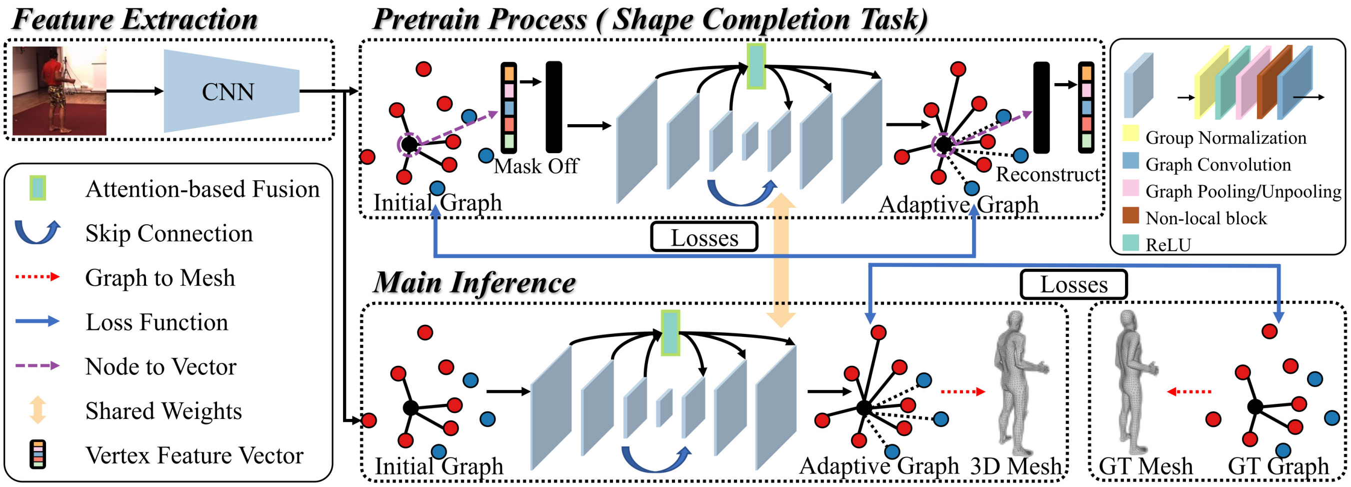


Figure 3: Overview of DC-GNet. The workflow contains three parts, a feature extraction stage to generate the initial graph from a single image, a pretrain process to learn an adaptive graph with a shape completion task and the main inference phase to reconstruct the 3D mesh.

recovered 3D mesh from the 2D input with GCN in a coarse-to-fine fashion. Kolotouros *et al.* [22] explored mesh structure and leveraged spatial locality via applying GCN to directly regress the vertices of the SMPL model. Hugo *et al.* [13] proposed to generate 3D clothed human with graph convolution variational Auto-Encoder (GCVAE). Simultaneously, some works [10, 50] intended to deal with mesh vertices as point clouds for capturing deformation.

Similarly, we also adopt GCN to process the mesh structure. Different from previous approaches, instead of simply applying convolution operation to aggregate information from direct neighbors, we incorporate an adaptive adjacent matrixto obtain local structure from both physically connected nodes and distant ones with deep relations.

2.3 Relations Capture via Learnable Adjacency Matrix

The adjacency matrix is a necessary component in GCN. In the 3D shape Reconstruction task, existing regression-based approaches usually use a pre-defined adjacency matrix, in which only positive relations between physically connected nodes are encoded. Such a pre-defined adjacency matrix is not able to capture complex local surface deformation, as shown in Figure 1(a). Replacing a pre-defined adjacency matrix with learnable ones is a common strategy in the other GCN-based computer vision applications, such as skeleton estimation [8, 57] or action recognition task [44, 53]. These works use different updating strategies to learn adjacency matrix for different purposes. Zhao *et al.* [57] learns adjacent matrix for describing subtle semantic relations within human skeleton joints. Doosti *et al.* [8] keeps the connectivity of the graph structure in joint hand and object pose estimation task. Against the action recognition task, Yan *et al.* [53] captures the body skeletons dynamics information to meet the specific demands in skeleton modeling. Shi *et al.* [44] learns the topology structure of the graph and skeleton samples for the flexibility of model.

These strategies only focus on joint-to-joint relations, which can not capture the node-to-surface relations existing in a full 3D human shape reconstruction task. In this paper, against the 3D shape reconstruction, we first propose a novel updating strategy for the learnable adjacency matrix to explore node-to-surface relations.

3 PROBLEM SETUP

Our input is a cropped image, which is centered around a person. For each input image, an image-based convolutional network is applied as a feature extractor and outputs a 2048-D feature vector for every single vertex in the graph. Our network is started with the initial estimation, which is the feature extraction stage shown in the Figure 3.

Previous approaches had already adopted GCN to process graph-like human mesh. The network is composed of basic graph convolution operations [19], which is defined as:

$$X_{out} = \sigma(\boldsymbol{A}\boldsymbol{X}_{in}\boldsymbol{W}), \qquad (1)$$

where $\boldsymbol{A} \in \mathbb{R}^{N\times N}$ is a pre-defined adjacency matrix of the graph, $\boldsymbol{X}_{in} = \{x_i\}_{i=1}^{N} \in \mathbb{R}^{N\times k}$ is the input feature matrix, $\boldsymbol{W} \in \mathbb{R}^{k\times h}$ is the trainable weight matrix, $\boldsymbol{X}_{out} \in \mathbb{R}^{N\times h}$ is the output feature matrix, and σ is the activation function. Specifically, N is nodes of the input graph, k and h are the input features and output features for each node, separately.

The graph convolution described in Eq. (1) is calculated based on a pre-defined adjacency matrix, which only encodes positive relations between physically connected nodes. As a result, the complex local structure can not be carefully captured.

4 PROPOSED APPROACH

In this Section, we present our approach. First, in Subsection 4.1, we describe the proposed network for obtaining effectively local structure information. Next, in Subsection 4.2, we describe our proposed shape completion task that alleviates the gap issue of appearance

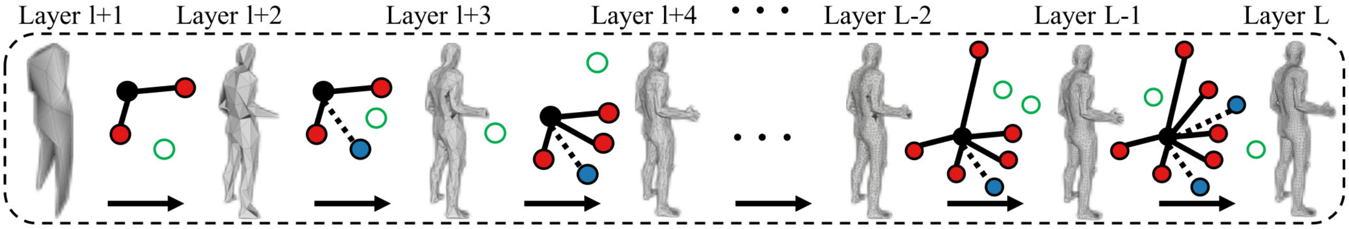


Figure 4: Visualization of the mesh from different layers in the decoder part of the U-Net. The refining process generates the final prediction from a coarse graph by adding nodes (the green hollow circle).

domain. Finally, Subsection 4.3 present the loss functions that we used for training.

4.1 Relations Capture for the 3D Shape Reconstruction

Instead of using a pre-defined adjacency matrix, we propose to use an adaptive adjacency matrix to learn subtle relationships between nonadjacent nodes. To achieve this, we rewrite Eq. (1) as:

$$X_{out} = \sigma(\hat{A} X_{in} W). \tag{2}$$

where $\hat{A} = A + I$ is a learnable adjacency matrix. I is the identity matrix. Based on Eq. (2), our network is able to infer local structure from nodes with subtle relations in a more distant region. Specifically, not only relations between nodes belong to the same semantic part (*e.g.*, nodes on arm and elbow that belong to the same limb can be leveraged for inference), but nodes with far distance (*e.g.*, one node on the left elbow can be related to the node on the right elbow) are encoded into the network. Moreover, following [6], we introduce a non-local block [52] to facilitate a holistic processing of the full body.

By training the network with Eq. (2), we can obtain a learnable adjacency matrix. However, such stacking of the convolution operation with adaptive adjacency matrix requires the expensive training computational cost. Therefore, we design a classical hierarchical U-shaped network architecture including encoder and decoder parts, to simplify the calculations and achieve our shape reconstruction pipeline.

In the encoder part, we introduce sampling operations [39] to simplify the calculations. The process of the encoder part can be formulated as:

$$Y_l = f(Y_{l-1}), \tag{3}$$

where $Y_l \in \mathbb{R}^{\tilde{N} \times h}$ is the processed feature matrix with $\tilde{N}$ nodes, and f demonstrates a fully-connected layer. $l \in \{1, 2, ...l-1, l, l+1, ..., L-1, L\}$ is the l-th layer of the network. Indeed, by downsampling the mesh data with a pre-defined factor, the high redundancy in the original scale and memory requirements are both dramatically reduced.

In the decoder part, we combine the features to boost the understanding of human body in a coherent way. Similarly, we can denote the decoder part as:

$$Y_{l+1} = f([f(Y_l); f(m(Y_1, ..., Y_l)); Y_{L-l}]), \tag{4}$$

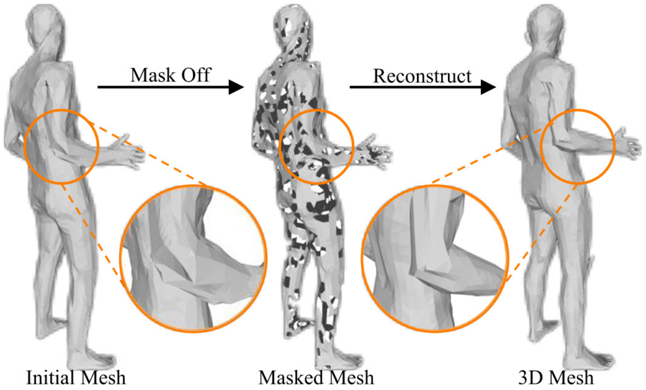


Figure 5: Illustration of the proposed shape completion task. With an initial mesh as input, we fabricate artificial holes on the surface of mesh. In order to recover the missing information, the network is forced to reason from the neighborhood in a more distant region. Moreover, we highlight the same part during different phases to show the effectiveness of this module.

where Y_{l+1} is a linear combination of above three parts, $m(Y_1, ..., Y_l) = \bigcup_{l=1}^{l} \sigma(E_l Y_l W_l)$ denotes the feature obtained from our feature fusion module, and Y_{L-l} is the previous feature in each level of the encoder part in symmetrical module of the decoder part. More specifically, $\bigcup$ represents concatenation connection, $E_l \in \mathbb{R}^{\tilde{N} \times \tilde{N}}$ is the calculated attention coefficients matrix and $W_l \in \mathbb{R}^{h \times p}$ is a shared linear transformation towards p features for each node, which are detailed described in [49].

Essentially, the modeling of the decoder part formulated by Eq. (4) explicitly fuse multi-level topology information, which alleviates the semantic gap and different spatial resolution [38]. In the process of the decoder part, the body shape is gradually refined when more features are fused, as visualized in Figure 4.

4.2 Shape Completion Task

Due to various complex occlusions (*e.g.*,self-occlusion or be sheltered) in an in-the-wild scenario, the human body part information is often missing. Since the training data are collected from simple human actions in a clear indoor environment, the information missing issue is rare. In order to enable the network to learn a generic adjacency matrix for various occlusion cases, we propose a shape completion task in which a mask off and a reconstruction part are included, as shown in Figure 5.

In the mask off part, we simulate the occlusion cases by fabricating artificial holes on the surface of the initial human mesh. Specifically, we randomly mask the partial mesh information of a given full human mesh, which can be formulated as

$$\hat{X}_{in} = X_{in} \cdot M, \tag{5}$$

where $\hat{X}_{in}$ denotes a masked human mesh. $M \in \mathbb{R}^{N \times k}$ is a matrix of ones except c row set zeros, and randomly shuffled before dot product.

Then, to force the network to recover the missing information for the masked human mesh,

we replace the X_{in} in Eq. (2) as a marked human mesh $\hat{X}_{in}$,

$$X_{in} = \sigma(\hat{A}\hat{X}_{in}W). \tag{6}$$

Note that in Eq. (6), the output of the network is settled as the initial full human mesh X_{in}.

4.3 Loss Functions

We use three loss functions to train our network. We first calculate per-vertex L_1 loss between the estimated and ground truth shape, which is denoted as $\mathcal{L}_{vertex}$. Additionally, we include joint-wise loss for further aligning mesh with keypoints. Specifically, we implement L_1 losses between the projected coordinates and the ground truth keypoints in 2D and 3D space (J_{2D} and J_{3D}). Finally, the complete training objective is:

$$\mathcal{L} = \mathcal{L}_{vertex} + \mathcal{L}_{3d} + \mathcal{L}_{2d} \tag{7}$$

We provide a more detailed description of the loss function in the supplementary material.

5 EXPERIMENT

In this Section, we concern with the experimental analysis of the proposed approach. First, we present the datasets that we use for evaluation (Section 5.1) and the implementation details of the proposed pipeline (Section 5.2). Then, we discuss the comparison approaches (Section 5.3) and ablation studies (Section 5.4). Finally, comparison with the-state-of-the-art approaches (Section 5.5) and qualitative analysis (Section 5.6) are provided.

5.1 Datasets and Evaluation Metrics

Datasets. In this paper, we present extensive experiments of our approach on several standard benchmarks including Human3.6M [14], UP-3D [25], MPI-INF-3DHP [31], COCO [27] and LSP [16]. For training, we apply benchmarks with 3D annotations, including Human3.6M, UP-3D and MPI-INF-3DHP. Additionally, similar to [21], we incorporate other 2D datasets, *i.e.*, COCO and LSP. For evaluation, we use MPI-INF-3DHP and Human3.6M. For the evaluation on Human3.6M, two popular evaluation protocols can be found. The first one, denoted as P1, includes the subjects S1, S5, S6, S7 and S8 for training, and the subjects S9 and S11 for testing. The second protocol, denoted as P2, tests only on the frontal camera with the same train/test sets. A more detailed description of the datasets can be found in the supplementary material.

Evaluation Metrics. For the MPI-INF-3DHP and Human3.6M datasets, following the evaluation in the most approaches [7, 17, 55], we report Mean Per Joint Position Error (MPJPE) and mean reconstruct error. MPJPE is defined as 3D joint errors, which are the projected coordinates from mesh data. While mean reconstruct error is the same calculation with MPJPE but with a rigid alignment. For MPI-INF-3DHP, in addition to MPJPE and mean reconstruct error, we further report Area Under the Curve (AUC) over a range of Percentage of Correct Keypoints (PCK) thresholds [31], which are also used in many approaches[17, 21, 48].

Table 1: Comparison on Human3.6M (Protocol 1 and 2) of our proposed approach with different components (*i.e.*, adaptive adjacency matrix and U-shaped Net, denoted as A and U separately). The numbers are MPJEP and mean reconstruct errors in mm. We conduct experiments with several models using CMR [22] and HMR [17] as the pretrained feature extractors. Best results are in bold.

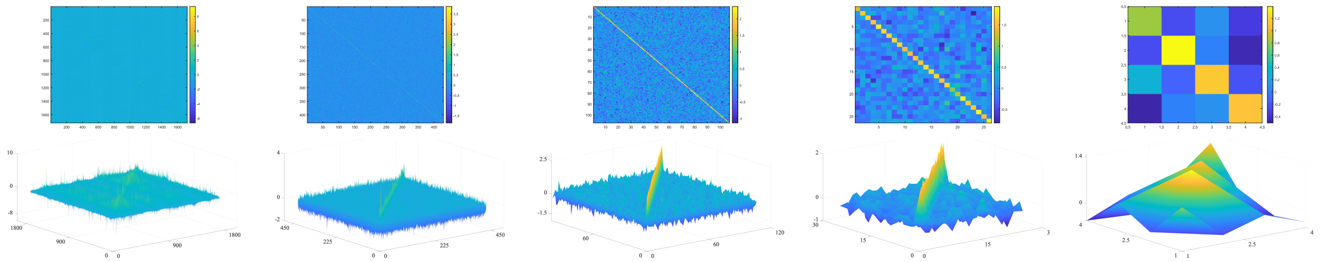
Method	MPJPE		Rec.Error	
	P1	P2	P1	P2
CMR [22]	77.3	73.5	51.2	49.6
Ours (U)	74.7	71.0	49.0	46.5
Ours (A)	73.4	69.8	49.1	45.5
Ours (A+U)	**72.3**	**68.8**	**48.5**	**45.3**
HMR [17]	91.2	89.1	61.8	59.5
Ours (U)	89.2	87.2	57.9	55.5
Ours (A)	88.2	85.3	56.3	54.1
Ours (A+U)	**87.8**	**85.1**	**55.6**	**53.9**

Table 2: Comparison on MPI-INF-3DHP of our proposed Shape Completion task with different configurations. The numbers are PCK, AUC, and MPJEP in mm. We conduct experiments with CMR [22] as pretrained feature extractor. We report results with the different number of nodes that are masked off. Best results are in bold.

Method	Absolute		
	PCK ↑	AUC ↑	MPJPE ↓
w/o Shape Completion	62.2	25.0	136.2
Mask off - 50	63.3	25.6	134.6
Mask off - 100	64.0	27.9	131.3
Mask off - 200	**66.3**	**30.4**	**128.5**
Mask off - 400	64.6	28.7	129.7

5.2 Implementation Details

Our model is implemented with PyTorch [34]. As shown in Figure 3, we first train our network with a shape completion task, and then in the second step we train the model in an end-to-end fashion. Noted the network trained with shape completion task shares the parameters with the model in the main inference process. In each stage, following [22] we subsample original mesh by a factor of 4 and upsample it back at the end of the network with [39]. For the training process, we utilize Adam optimizer with a mini-batch size of 16, where the learning set is set to 3e-4. In the pre-trained process, only Human3.6M dataset is used, while in the main inference process, we first train our model from Human3.6M and UP-3D for



A_0 A_1 A_2 A_3 A_4

Figure 6: Examples of the learned adjacency matrix. For each column, the images from top-to-bottom correspond to the visualization of the learned matrix, the surface plot of the matrix.

Table 3: Comparison with state-of-the-art models on MPI-INF-3DHP and Human3.6M datasets (P2). The numbers are MPJEP and mean reconstruct errors in mm, and AUC. DC-GNet achieves a comparable result on Human3.6M dataset and beyond all state-of-the-art approaches on more challenging in-the-wild MPI-INF-3DHP dataset. "-" means the corresponding results are not available. † indicates that extra temporal infromation is leveraged. Best results are in bold.

Method	MPI-INF-3DHP			Human3.6M	
	AUC ↑	MPJPE ↓	Reconst.Error ↓	MPJPE ↓	Reconst.Error ↓
HMR [17] (CVPR'18)	36.5	124.2	89.8	-	56.8
†HMMR [18] (CVPR'19)	-	-	-	-	56.9
†Arnab *et al.* [3] (CVPR'19)	-	-	-	77.8	54.3
CMR [22] (CVPR'19)	24.3	152.0	83.8	71.9	50.1
†TexturePose [36] (ICCV'19)	-	-	-	-	49.7
SPIN [21] (ICCV'19)	37.1	105.2	67.5	-	41.1
DaNet [56] (ACM MM'19)	-	-	-	61.5	48.6
Jiang *et al.* [15] (CVPR'20)	-	-	-	-	52.7
Kundu *et al.* [24] (ECCV'20)	-	-	-	-	48.1
Pose2Mesh [7] (ECCV'20)	-	-	-	64.9	47.0
†VIBE [20] (CVPR'20)	-	97.7	63.4	65.9	41.5
DecoMR [55] (CVPR'20)	-	102.0	65.9	**60.6**	**39.3**
DC-GNet	**40.7**	**97.2**	**62.5**	63.9	42.4

30 epochs and then impose more data (*i.e.*, COCO, MPI-INF-3DHP, *etc.*) for greater image diversity. We use a single NVIDIA RTX 2080 Ti GPU for training and our model inference for a single image takes 55ms (including time (33ms) for feature extractor), which is nearly real-time.

5.3 Comparison Approaches

Following the common-used comparison setting in the literature [7, 20, 55], we first compare with two recent baselines for regression-based approaches (the vertex-level regression and parameter regression approaches, *i.e.* HMR [17] and CMR [22]). As mentioned earlier, both the above two approaches are based on a pre-defined adjacency matrix. Moreover, several of recent state-of-the-art methods [3, 15, 18, 21, 22, 24, 36, 55, 56], are considered in the comparison, including vertex-level regression approaches [22, 55] and parameter regression ones [17, 21].

5.4 Ablation Studies

Firstly, to put our approach into perspective, we conduct ablation studies on our approach. Following the literature [20], we also use two pre-trained feature extractors CMR [22] and HMR [17], respectively. We construct three different settings: (a) U-shaped network without adjacency matrix. (b) Only adaptive adjacency matrix (without sampling and features fusion operations). (c) U-shaped network with an adaptive adjacency matrix. These three settings are denoted as "U", "A", and "U+A" in the comparison. The ablation experiments results are reported in Table 1, in which the results are organized into two groups in terms of two different feature extractors.

Since the adaptive adjacency matrix captures the topological structure of human body. It significantly improves the reconstruction performance, as we observed in the Table 1. The proposed U-shaped network also boosts the performance due to the better understanding of human body in a coherent way. The best result is always achieved by "U+A" setting whichever feature extractor is used. Figure 6 visualizes the learned adjacency matrix. It shows that deep relations between nodes (*i.e.*, positive and negative) are encoded.

Moreover, we also study the effectiveness of the proposed shape completion task. In the experiment, we train different networks

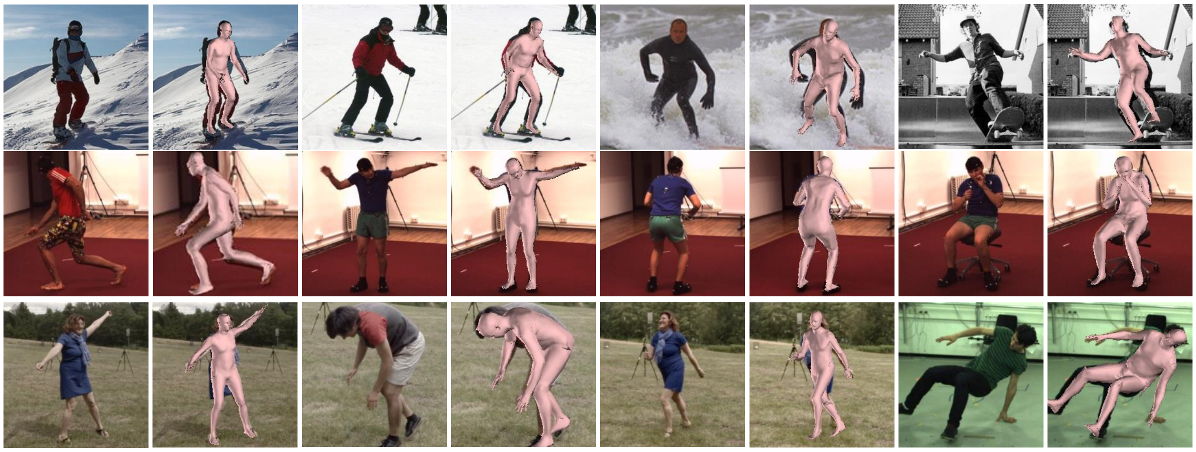


Figure 7: Examples of successful reconstructions. COCO (row 1), H36M (row 2) and MPI-INF-3DHP (row 3).

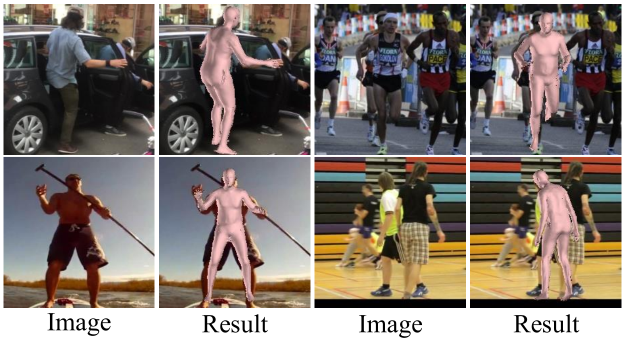


Figure 8: Examples of erroneous reconstructions. Typical failure cases may be caused by severe occlusions, rare viewpoint, or interactions among multiple people.

with different number of nodes that are masked off on Human3.6M and UP-3D datasets, and evaluate on MPI-INF-3DHP. The results are obtained by using CMR [22] as a feature extractor. Since the different number of masked nodes leads to different performance, we set the number of masked nodes as 0, 50, 100, 200, 400. As we clearly observed in Table 2, with the number of masked nodes increases from 0 to 200, the performance also begins to increase, which demonstrates the effectiveness of the shape completion task. We observe degradation occurs when the number of masked nodes is settled at 400, which may exceed the ability to recover mesh.

5.5 Comparison to State-of-the-Art Results

We report MPJPE and mean reconstruct error of DC-GNet on Human3.6M, and additionally AUC over a range of 3D-PCK threshholds (150mm) on MPI-INF-3DHP. For the fair comparison, following the same setting in the literature [20], we use HMR as a feature extractor pretrained by [21].

We first conduct a comparable result on the challenging in-the-wild MPI-INF-3DHP dataset. As shown in Table 3, comparing to the baseline methods, DC-GNet achieves more than 21% (HMR) and 36% (CMR) improvement on average MPJPE, respectively. Similarly, more than 30% (HMR) and 25% (CMR) improvements are achieved by ours on average reconstruct error. Compared to other considered approaches, we still achieve the best performance in all metrics. It seems that our model only achieves sightly performance improvement than [20] under MPJPE, note that it exploits video temporal information while we use only a single image.

In the Human3.6M dataset, DC-GNet still shows its superiority. As shown in Table 3, DC-GNet outperforms the baseline approaches with a wide margin (more than 25% (HMR) and 15% (CMR) improvement on reconstruction error metric). It seems that our approach is sightly inferior to approaches[20, 21, 55]. Note that Human3.6M is an indoor dataset with pre-defined action categories in both train and test sets. As analyzed by [7], the performance drop (*i.e.*, performs well on Human3.6M while meets degradation on the in-the-wild dataset) may be attributed to an overfitting issue.

Note that different methods are trained with different training data. Detailed training data comparisons can be found in the supplementary material. To show the superiority of our approach, we compare their best results reported in the original literature.

5.6 Qualitative Evaluation

This section presents qualitative evaluations. In the qualitative experiments, following the same strategy in [17, 21, 22], we leverage Mosh [28] and SPIN [21] to generate pseudo-groundtruth on Human3.6M and in-the-wild datasets, respectively.

Firstly, we conduct the qualitative results of our approach from different datasets. The success and failure cases are also reported, as shown in Figure 7 and Figure 8, respectively. Typical failure cases may be caused by severe occlusions, rare view-point, or interactions among multiple people.

Moreover, we further provide a qualitative comparison with the recent competitive vertex-level regression approaches (*i.e.*, CMR [22] and DecoMR [55]) and parameter regression approaches [17, 21], as shown in Figure 9 and Figure 10, respectively.

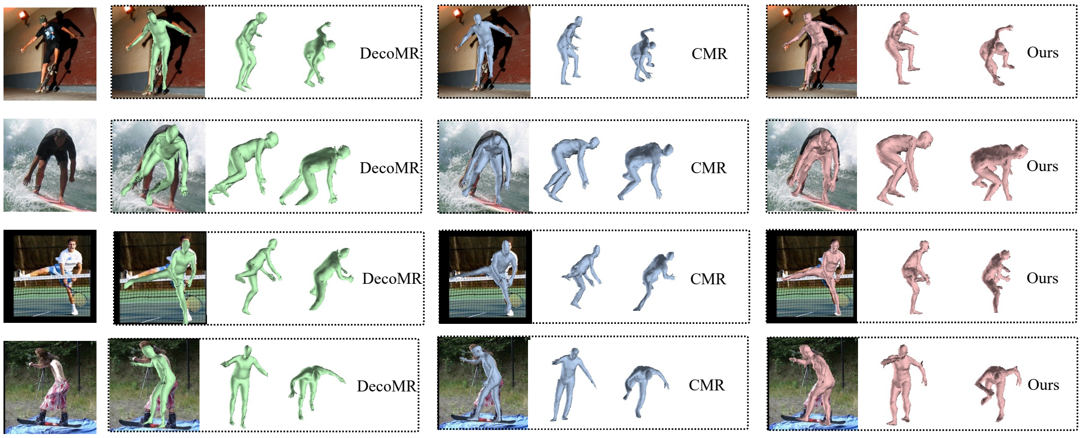


Figure 9: Comparison between our approach and other vertex-level regression methods.

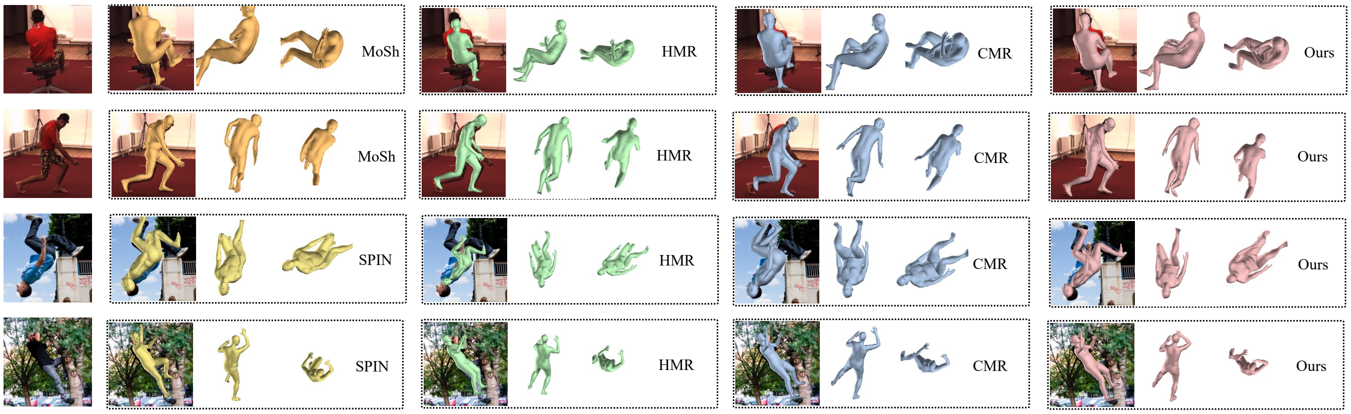


Figure 10: Comparison between our approach and other parameter regression methods.

As shown in the Figure 9, compared to vertex-level regression approaches, DC-GNet generates much more pleasant mesh results that reconstruct details and retain the whole topological structure.

We further report qualitative comparisons with the parameter representation,as shown in Figure 10. We also achieve more reasonable reconstruction results. Note that CMR and our approach are vertex-level regression approaches while they can also be implemented as a parameter regression by using Multi-Layer Perceptron (MLP) [41].

More qualitative results can be found in the supplementary material.

6 CONCLUSION

The aim of this paper is to explore deep relation among mesh vertices for 3D human pose and shape reconstruction, by encoding both positive and negative relations. We incorporate these relations into a graph convolution network with a shape completion module for complex topological structure learning cross domain. Moreover, the comparison with a series of state-of-the-art approaches shows the superiority of our approach. More specifically, the extensive experiments conducted on wild datasets demonstrate that the proposed strategies are crucial to make our approach of practical use for in-the-wild scene. Future work may explore using denser cues (*e.g.*, video input or optical flow) and consider extending our approach for multiple people.

ACKNOWLEDGMENTS

This research was supported by the National Natural Science Foundation of China (Grant no 61671397).

REFERENCES

[1] Mykhaylo Andriluka, Leonid Pishchulin, Peter Gehler, and Bernt Schiele. 2014. Human Pose Estimation: New Benchmark and State of the Art Analysis. In *Computer Vision and Pattern Recognition (CVPR)*. 3686–3693.

[2] Dragomir Anguelov, Praveen Srinivasan, Daphne Koller, Sebastian Thrun, Jim Rodgers, and James Davis. 2005. SCAPE: Shape Completion and Animation of People. *ACM Transactions on Graphics* 24 (2005), 408–416.

[3] Anurag* Arnab, Carl* Doersch, and Andrew Zisserman. 2019. Exploiting temporal context for 3D human pose estimation in the wild. In *Computer Vision and Pattern Recognition (CVPR)*. 3395–3404.

[4] Bharat Lal Bhatnagar, Cristian Sminchisescu, Christian Theobalt, and Gerard Pons-Moll. 2020. Combining Implicit Function Learning and Parametric Models for 3D Human Reconstruction. In *European Conference on Computer Vision (ECCV)*.

[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 (ECCV)*. 561–578.

[6] Y. Cai, L. Ge, J. Liu, J. Cai, T. Cham, J. Yuan, and N. M. Thalmann. 2019. Exploiting Spatial-Temporal Relationships for 3D Pose Estimation via Graph Convolutional Networks. In *International Conference on Computer Vision (ICCV)*. 2272–2281.

[7] Hongsuk Choi, Gyeongsik Moon, and Kyoung Mu Lee. 2020. Pose2Mesh: Graph Convolutional Network for 3D Human Pose and Mesh Recovery from a 2D Human Pose. In *European Conference on Computer Vision (ECCV)*.

[8] Bardia Doosti, Shujon Naha, Majid Mirbagheri, and David J. Crandall. 2020. HOPE-Net: A Graph-Based Model for Hand-Object Pose Estimation. In *Computer Vision and Pattern Recognition (CVPR)*. 6608–6617.

[9] Matteo Fabbri, Fabio Lanzi, Simone Calderara, Stefano Alletto, and Rita Cucchiara. 2020. Compressed Volumetric Heatmaps for Multi-Person 3D Pose Estimation. In *Computer Vision and Pattern Recognition (CVPR)*. 7204–7213.

[10] Valentin Gabeur, Jean-Sébastien Franco Franco, Xavier Martin, Cordelia Schmid, and Gregory Rogez. 2019. Moulding Humans: Non-parametric 3D Human Shape Estimation from Single Images. In *International Conference on Computer Vision (ICCV)*. 2232–2241.

[11] Riza Alp Guler and Iasonas Kokkinos. 2019. HoloPose: Holistic 3D Human Reconstruction In-The-Wild. In *Computer Vision and Pattern Recognition (CVPR)*. 10884–10894.

[12] Marc Habermann, Weipeng Xu, Michael Zollhofer, Gerard Pons-Moll, and Christian Theobalt. 2020. DeepCap: Monocular Human Performance Capture Using Weak Supervision. In *Computer Vision and Pattern Recognition (CVPR)*. 5052–5063.

[13] Bertiche Hugo, Madadi Meysam, and Escalera Sergio. 2020. CLOTH3D: Clothed 3D Humans. In *European Conference on Computer Vision (ECCV)*.

[14] C Ionescu, D Papava, V Olaru, and C Sminchisescu. 2014. Human3.6M: Large Scale Datasets and Predictive Methods for 3D Human Sensing in Natural Environments. *Transactions on Pattern Analysis and Machine Intelligence* 36, 7 (2014), 1325–1339.

[15] Wen Jiang, Nikos Kolotouros, Georgios Pavlakos, Xiaowei Zhou, and Kostas Daniilidis. 2020. Coherent Reconstruction of Multiple Humans From a Single Image. In *Computer Vision and Pattern Recognition (CVPR)*. 5579–5588.

[16] Sam Johnson and Mark Everingham. 2010. Clustered Pose and Nonlinear Appearance Models for Human Pose Estimation. In *British Machine Vision Conference (BMVC)*. 12.1–12.11.

[17] Angjoo Kanazawa, Michael J. Black, David W. Jacobs, and Jitendra Malik. 2018. End-to-end Recovery of Human Shape and Pose. In *Computer Vision and Pattern Recognition (CVPR)*. 7122–7131.

[18] A. Kanazawa, J. Y. Zhang, P. Felsen, and J. Malik. 2019. Learning 3D Human Dynamics From Video. In *Computer Vision and Pattern Recognition (CVPR)*. 5607–5616.

[19] Thomas N. Kipf and Max Welling. 2017. Semi-Supervised Classification with Graph Convolutional Networks. In *International Conference on Learning Representations (ICLR)*.

[20] Muhammed Kocabas, Nikos Athanasiou, and Michael J. Black. 2020. VIBE: Video Inference for Human Body Pose and Shape Estimation. In *Computer Vision and Pattern Recognition (CVPR)*. 5252–5262.

[21] Nikos Kolotouros, Georgios Pavlakos, Michael J. Black, and Kostas Daniilidis. 2019. Learning to Reconstruct 3D Human Pose and Shape via Model-fitting in the Loop. In *International Conference on Computer Vision (ICCV)*. 2252–2261.

[22] Nikos Kolotouros, Georgios Pavlakos, and Kostas Daniilidis. 2019. Convolutional Mesh Regression for Single-Image Human Shape Reconstruction. In *Computer Vision and Pattern Recognition (CVPR)*. 4501–4510.

[23] Ilya Kostrikov and Juergen Gall. 2014. Depth Sweep Regression Forests for Estimating 3D Human Pose from Images. In *British Machine Vision Conference (BMVC)*. 1–13.

[24] Jogendra Nath Kundu, Mugalodi Rakesh, Varun Jampani, Rahul M Venkatesh, and R. Venkatesh Babu. 2020. Appearance Consensus Driven Self-Supervised Human Mesh Recovery. In *European Conference on Computer Vision (ECCV)*.

[25] 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 *Computer Vision and Pattern Recognition (CVPR)*. 4704–4713.

[26] Chen Li and Gim Hee Lee. 2019. Generating Multiple Hypotheses for 3D Human Pose Estimation With Mixture Density Network. In *Computer Vision and Pattern Recognition (CVPR)*. 9887–9895.

[27] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollar, and Larry Zitnick. 2014. Microsoft COCO: Common Objects in Context. In *European Conference on Computer Vision (ECCV)*. 740–755.

[28] Matthew Loper, Naureen Mahmood, and Michael J. Black. 2014. MoSh: Motion and Shape Capture from Sparse Markers. *ACM Transactions on Graphics* 33, 6, Article 220 (Nov. 2014), 13 pages.

[29] 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 (Oct. 2015), 248:1–248:16.

[30] J. Martinez, R. Hossain, J. Romero, and J. J. Little. 2017. A Simple Yet Effective Baseline for 3d Human Pose Estimation. In *International Conference on Computer Vision (ICCV)*. 2659–2668.

[31] Dushyant Mehta, Helge Rhodin, Dan Casas, Pascal Fua, Oleksandr Sotnychenko, Weipeng Xu, and Christian Theobalt. 2017. Monocular 3D Human Pose Estimation In The Wild Using Improved CNN Supervision. In *International Conference on 3D Vision (3DV)*. 506–516.

[32] Gyeongsik Moon and Kyoung Mu Lee. 2020. I2L-MeshNet: Image-to-Lixel Prediction Network for Accurate 3D Human Pose and Mesh Estimation from a Single RGB Image. In *European Conference on Computer Vision (ECCV)*.

[33] Mohamed Omran, Christoph Lassner, Gerard Pons-Moll, Peter V. 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 (3DV)*. 484–494.

[34] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. 2017. Automatic differentiation in PyTorch. In *Neural Information Processing Systems Workshop (NIPSW)*.

[35] G. Pavlakos, V. Choutas, N. Ghorbani, T. Bolkart, A. A. Osman, D. Tzionas, and M. J. Black. 2019. Expressive Body Capture: 3D Hands, Face, and Body From a Single Image. In *Computer Vision and Pattern Recognition (CVPR)*. 10967–10977.

[36] Georgios Pavlakos, Nikos Kolotouros, and Kostas Daniilidis. 2019. TexturePose: Supervising Human Mesh Estimation with Texture Consistency. In *International Conference on Computer Vision (ICCV)*. 803–812.

[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 *Computer Vision and Pattern Recognition (CVPR)*. 459–468.

[38] Xuebin Qin, Zichen Zhang, Chenyang Huang, Masood Dehghan, Osmar Zaiane, and Martin Jagersand. 2020. U2-Net: Going Deeper with Nested U-Structure for Salient Object Detection. *Pattern Recognition* 106, 107404.

[39] Anurag Ranjan, Timo Bolkart, Soubhik Sanyal, and Michael J. Black. 2018. Generating 3D Faces Using Convolutional Mesh Autoencoders. In *European Conference on Computer Vision (ECCV)*. 725–741.

[40] Javier Romero, Dimitrios Tzionas, and Michael J. Black. 2017. Embodied Hands: Modeling and Capturing Hands and Bodies Together. *ACM Transactions on Graphics* 36, 6 (Nov. 2017), 245:1–245:17. http://doi.acm.org/10.1145/3130800.3130883

[41] D. E. Rumelhart and J. L. McClelland. 1987. *Learning Internal Representations by Error Propagation.* 318–362.

[42] Nikolaos Sarafianos, Bogdan Boteanu, Bogdan Ionescu, and Ioannis A. Kakadiaris. 2016. 3D Human pose estimation: A review of the literature and analysis of covariates. *Computer Vision and Image Understanding* 152 (2016), 1 – 20.

[43] Saurabh Sharma, Pavan Teja Varigonda, Prashast Bindal, Abhishek Sharma, and Arjun Jain. 2019. Monocular 3D Human Pose Estimation by Generation and Ordinal Ranking. In *International Conference on Computer Vision (ICCV)*. 2325–2334.

[44] L. Shi, Y. Zhang, J. Cheng, and H. Lu. 2019. Two-Stream Adaptive Graph Convolutional Networks for Skeleton-Based Action Recognition. In *Computer Vision and Pattern Recognition (CVPR)*. 12018–12027.

[45] Suman, Sedai, Mohammed, Bennamoun, Du, Q, and Huynh. 2013. A Gaussian process guided particle filter for tracking 3D human pose in video. *Transactions on Image Processing* 22, 11 (2013), 4286–4300.

[46] Ke Sun, Bin Xiao, Dong Liu, and Jingdong Wang. 2019. Deep High-Resolution Representation Learning for Human Pose Estimation. In *Computer Vision and Pattern Recognition (CVPR)*. 5693–5703.

[47] Yu Sun, Yun Ye, Wu Liu, Wenpeng Gao, Yili Fu, and Tao Mei. 2019. Human Mesh Recovery From Monocular Images via a Skeleton-Disentangled Representation. In *International Conference on Computer Vision (ICCV)*. 5349–5358.

[48] Wei Tang, Pei Yu, and Ying Wu. 2018. Deeply Learned Compositional Models for Human Pose Estimation. In *European Conference on Computer Vision (ECCV)*. 197–214.

[49] Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Liò, and Yoshua Bengio. 2018. Graph Attention Networks. In *International Conference on Learning Representations (ICLR)*.

[50] Abbhinav Venkat, Chaitanya Patel, Yudhik Agrawal, and Avinash Sharma. 2019. HumanMeshNet: Polygonal Mesh Recovery of Humans. In *International Conference on Computer Vision Workshop (ICCVW)*. 2178–2187.

[51] K. Wang, J. Xie, G. Zhang, L. Liu, and J. Yang. 2020. Sequential 3D Human Pose and Shape Estimation From Point Clouds. In *Computer Vision and Pattern Recognition (CVPR)*. 7273–7282.

[52] Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He. 2018. Non-Local Neural Networks. In *Computer Vision and Pattern Recognition (CVPR)*. 7794–7803.

[53] Sijie Yan, Yuanjun Xiong, and Dahua Lin. 2018. Spatial Temporal Graph Convolutional Networks for Skeleton-Based Action Recognition. In *AAAI Conference on Artificial Intelligence(AAAI)*. 7444–7452.

[54] Andrei Zanfir, Elisabeta Marinoiu, and Cristian Sminchisescu. 2018. Monocular 3D Pose and Shape Estimation of Multiple People in Natural Scenes - The Importance of Multiple Scene Constraints. In *Computer Vision and Pattern Recognition (CVPR)*. 2148–2157.

[55] W. Zeng, W. Ouyang, P. Luo, W. Liu, and X. Wang. 2020. 3D Human Mesh Regression With Dense Correspondence. In *Computer Vision and Pattern Recognition (CVPR)*. 7052–7061.

[56] 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 *ACM International Conference on Multimedia(ACM MM)*. 935–944.

[57] Long Zhao, Xi Peng, Yu Tian, Mubbasir Kapadia, and Dimitris N. Metaxas. 2019. Semantic Graph Convolutional Networks for 3D Human Pose Regression. In *Computer Vision and Pattern Recognition (CVPR)*. 3425–3435.

[58] Kun Zhou, Xiaoguang Han, Nianjuan Jiang, Kui Jia, and Jiangbo Lu. 2019. HEMlets Pose: Learning Part-Centric Heatmap Triplets for Accurate 3D Human Pose Estimation. In *International Conference on Computer Vision (ICCV)*. 2344–2353.

[59] Shihao Zhou, Mengxi Jiang, Qicong Wang, and Yunqi Lei. 2021. Towards Locality Similarity Preserving to 3D Human Pose Estimation. In *Computer Vision – ACCV 2020 Workshops*. 136–153.

[60] Xiaowei Zhou, Menglong Zhu, Georgios Pavlakos, Spyridon Leonardos, Konstantinos G Derpanis, and Kostas Daniilidis. 2019. MonoCap: Monocular Human Motion Capture using a CNN Coupled with a Geometric Prior. *Transactions on Pattern Analysis and Machine Intelligence* 41, 4 (2019), 901–914.

[61] Hao Zhu, Xinxin Zuo, Sen Wang, Xun Cao, and Ruigang Yang. 2019. Detailed Human Shape Estimation From a Single Image by Hierarchical Mesh Deformation. In *Computer Vision and Pattern Recognition (CVPR)*. 4491–4500.

[62] S. Zuffi and M. J. Black. 2015. The stitched puppet: A graphical model of 3D human shape and pose. In *Computer Vision and Pattern Recognition (CVPR)*. 3537–3546.