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

This supplementary material presents details that are not included in the main manuscript due to space constraints. Firstly, in Section 1, we provide details of our proposed network structure. Next, in Section 2, we present the training data used by different approaches that mentioned in the paper. Then, in Section 3, we give a detailed description of the loss functions that we use. Finally, in Section 4, we show additional experimental results for comparison.

## 1 NETWORK ARCHITECTURE

Figure 1 shows the detailed architectures of DC-GNet. Our network takes the initial human mesh with the size of $B \times M_0 \times 3$ as input and the output is the processed human mesh with the same size. Here, 'GN', 'FC' and 'GAT' are short for Group Normalization [27], Fully Connected layer and Graph Attention [26], respectively. $B$ is the batch size and $M_i$ is the node number of the $i$-th level for graph resolution, with $M_0$ = 1723, $M_1$ = 430, $M_2$ = 107, $M_3$ = 26, $M_4$ = 4, $M_5$ =1. Moreover, 'GCN unit' is composed of a Group Normalization operation with ReLU activation and a Graph Convolution layer [15].

## 2 TRAINING DATA

As we mentioned in the Section 5.5, different training data are used by various methods. In this section, we give a detailed description of the training data, and comparison in Table 1. We first describe the datasets that we use.

**Human3.6M.** Human3.6M [10] is an indoor dataset with 3D annotations. It consists of several subjects performing actions like "Direction", "Sitting" and "Waiting". Adhering to the typical setting [13], we use the subjects S1, S5, S6, S7 and S8 for training, and we report the results on subjects S9 and S11. Noted that, the dataset is collected with a motion capture system, and there is no 3D mesh ground-truth. As a result, we use pseudo-groundtruth from MoSh [21], following previous works [13, 17, 18]. We present results with two popular protocols (P1 and P2, as defined in [13]) and two evaluation metrics (MPJPE and Reconstruction error, as defined in [30]). **COCO.** COCO [20] is one of the most widely used 2D datasets. It contains considerable poses with various scales in natural environments. We use this dataset for training, which is processed by Kolotouros *et al.* [17].

**UP-3D.** UP-3D [19] is a dataset created by employing SMPLify [5] on well-designed 2D benchmarks, like FashionPose [8] and MPII [3]. It comprises images with high-quality 3D shape fits that are selected by human annotators. Following [18], we use this dataset for training.

**MPI-INF-3DHP.** MPI-INF-3DHP [23] is a dataset that contains 3D pose ground-truth captured with multi-view cameras under both indoor and outdoor environments. The ground-truth 3D pose seems to be less accurate due to no markers are used. Following [17], we use this dataset for training and we report the results on it. Noted that, in addition to MPJPE, we further report Area Under the Curve (AUC) over a range of Percentage of Correct Keypoints (PCK) thresholds [23].

**LSP.** LSP [11] is a standard 2D dataset. We use this dataset that processed by Kolotouros *et al.* [17] for training.

**MPII.** MPII [3] is a 2D in-the-wild dataset. We use this dataset that processed by Kolotouros *et al.* [17] for training.

Then we will introduce the related datasets that are used by other methods.

**LSP-extended.** LSP-extended [12] is a 2D dataset that contains 10,000 images

**CMU.** CMU [9] motion capture dataset consists of 2605 motions of about 140 people performing various actions.

**PosePrior.** PosePrior [1] is a dataset used for learning pose-dependent joint angle limits which formulate a prior for the human pose.

**PennAction.** PennAction [29] dataset contains 2326 video sequences and covers 15 different actions. Each video involves an action class label and human joint annotations.

**InstaVariety.** InstaVariety [14] is an in-the-wild dataset consist of 28272 videos downloaded from Instagram.

**PoseTrack.** PoseTrack [2] used for single-frame pose estimation, multi-person pose estimation, and multi-person pose tracking is extended from the MPII dataset. It provides more than 500 video sequences with over 22000 labeled frames and over 150000 annotated poses.

**AMASS.** AMASS [22] is a large 3D human motion capture dataset that contains 40 hours of motion data,344 topics, and more than 11,000 actions.

**Kinetics-400.** Kinetics-400 [6] is a large-scale and high-quality dataset that contains 400 human action classes, each action class covering 600 video clips.

## 3 LOSS FUNCTIONS

We use three loss functions to train our network.

**Vertex Loss.** Since we aim to regress the mesh of human body, a vertex-wise $L_1$ loss is applied between the estimated and ground truth shape:

$$\mathcal{L}_{vertex} = \sum_{i=1}^{N} \parallel y_i - \hat{y}_i \parallel_1, \tag{1}$$

where $\hat{y}_i \in \mathbb{R}^3$ is the ground truth vertex and $y_i \in \mathbb{R}^3$ is the predicted vertex.

**Joint 3D Loss.** Additionally, we include joint-wise loss for further aligning mesh with keypoints. We regress the 3D keypoints from the estimated mesh. Then the loss can be formulated as:

$$\mathcal{L}_{3d} = \sum_{i=1}^{D} \parallel J_{3D_i} - \hat{J}_{3D_i} \parallel_1, \tag{2}$$

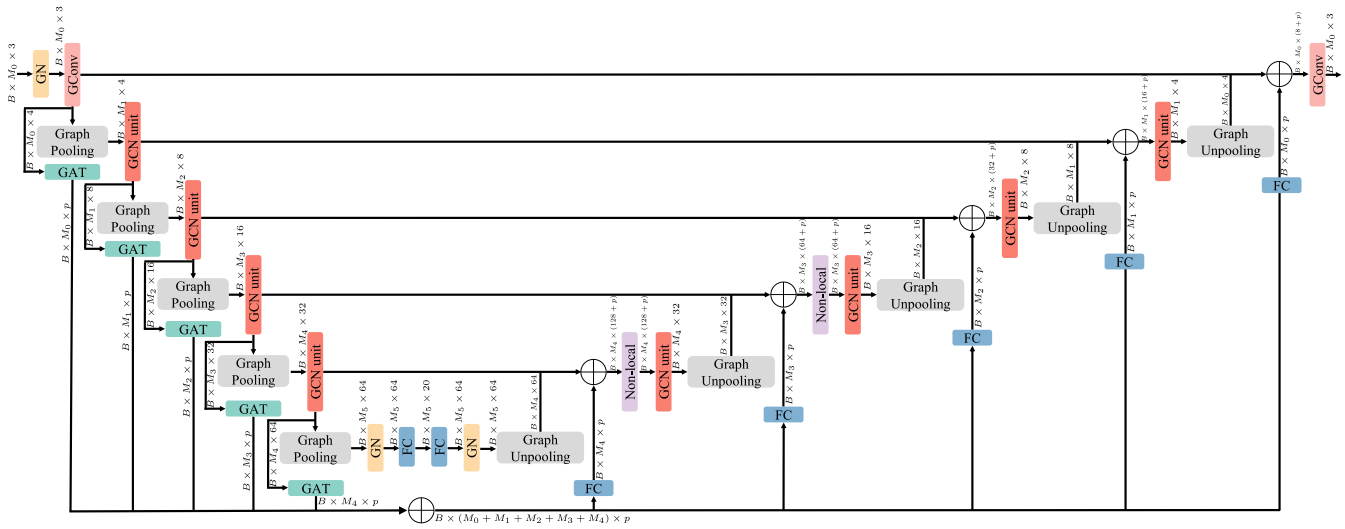


**Figure 1: Detailed architecture of DC-GNet. Here, 'GN', 'FC' and 'GAT' are short for Group Normalization [27], Fully Connected layer and Graph Attention [26], respectively. $B$ is the batch size and $M_i$ is the node number of the $i$-th level for graph resolution, with $M_0 = 1723$, $M_1 = 430$, $M_2 = 107$, $M_3 = 26$, $M_4 = 4$, $M_5 = 1$. Moreover, 'GCN unit' is composed of a Group Normalization operation with ReLU activation and a Graph Convolution layer [15].**

**Table 1: The datasets used by different approaches for training when evaluated on MPI-INF-3DHP.**

| Datasets | HMR [13] | CMR [18] | SPIN [17] | VIBE [16] | DecoMR [28] | Ours |
|---|---|---|---|---|---|---|
| Human3.6M [10] | √ | √ | √ | √ | √ | √ |
| UP-3D [19] | | √ | | | √ | √ |
| MPI-INF-3DHP [23] | √ | | √ | √ | | √ |
| COCO [20] | √ | | √ | | √ | √ |
| LSP [11] | √ | | √ | | √ | √ |
| LSP-extended [12] | √ | | √ | | | |
| MPII [3] | √ | | √ | | √ | √ |
| CMU [9] | √ | | | | | |
| PosePrior [1] | √ | | | | | |
| PennAction [29] | | | | √ | | |
| InstaVariety [14] | | | | √ | | |
| PoseTrack [2] | | | | √ | | |
| AMASS [22] | | | | √ | | |
| Kinetics-400 [6] | | | | √ | | |

**Table 2: Comparison with state-of-the-art methods on MPI-INF-3DHP dataset. The numbers are PCK, AUC and MPJPE in mm. Best results are in bold. ⋆ indicates methods that output only 3D joints.**

| Method | Absolute | | | Rigid Alignment | | |
|---|---|---|---|---|---|---|
| | PCK↑ | AUC↑ | MPJPE↓ | PCK↑ | AUC↑ | MPJPE↓ |
| ⋆Mehta *et al.* [23] (3DV'17) | 75.7 | 39.3 | 117.6 | - | - | - |
| ⋆VNect [24] (TOG'17) | 76.6 | 40.4 | 124.7 | 83.9 | 47.3 | 98.0 |
| HMR [13] (CVPR'18) | 72.9 | 36.5 | 124.2 | 86.3 | 47.8 | 89.8 |
| CMR [18] (CVPR'19) | 56.7 | 24.3 | 155.2 | 85.9 | 47.7 | 85.5 |
| SPIN [17] (CVPR'19) | 76.4 | 37.1 | 105.2 | 92.5 | 55.6 | 67.5 |
| DecoMR [28] (CVPR'20) | - | - | 102.0 | - | - | 65.9 |
| DC-GNets | **80.4** | **40.7** | **97.2** | **93.8** | **58.8** | **62.5** |

**Table 3: Comparison with state-of-the-art methods on 3DPW dataset. The number is mean reconstruct error in mm. Best results are in bold.**

| Method | Reconst.Error |
|---|---|
| HMR [13] (CVPR'18)* | 81.3 |
| CMR [18] (CVPR'19)* | 70.2 |
| Arnab *et al.* [4] (CVPR'19) | 72.2 |
| SPIN [17] (CVPR'19) | 59.2 |
| Sun *et al.* [25] (ICCV'19) | 69.5 |
| DecoMR [28] (CVPR'20 | 61.7 |
| DC-GNet | **59.1** |

where $J_{3D} \in \mathbb{R}^{D\times3}$ is the regressed 3D keypoints, $\hat{J}_{3D} \in \mathbb{R}^{D\times3}$ is the ground truth 3D keypoints, and $D$ is the predefined number of keypoints in the skeleton.

**Joint 2D Loss.** Similarly, we implement joint-wise loss in the 2D space. By simply projecting 3D joints on the image plane, we formulate this loss as:

$$\mathcal{L}_{2d} = \sum_{i=1}^{D} \parallel J_{2D_i} - \hat{J}_{2D_i} \parallel_1, \tag{3}$$

where $J_{2D} \in \mathbb{R}^{D\times2}$ is the projected 2D keypoints, and $\hat{J}_{2D} \in \mathbb{R}^{D\times3}$ is the ground truth 2D keypoints.

Finally, the complete training objective is:

$$\mathcal{L} = \mathcal{L}_{vertex} + \mathcal{L}_{3d} + \mathcal{L}_{2d} \tag{4}$$

# 4 ADDITIONAL RESULTS

## 4.1 Quantitative Analysis

First, we provide a detailed comparison on MPI-INF-3DHP in various metrics in Table 2. Our DC-GNet outperforms previous state-of-the-art methods in all metrics, which shows the priority of our proposed approach.

Then, in Table 3 we present a comparison with other approaches on 3DPW dataset, which is an in-the-wild dataset with 3D pose and mesh ground-truth. It leverages IMU sensors to obtain accurate pose and shape annotations under complex scenes. Following the previous works [7, 17], we only perform an evaluation with the test set. The mean reconstruction error is reported for comparison. Note that, we outperform all compared approaches and do not include the training data of LSP-extended as SPIN. Moreover, considering to combine our approach with the in-the-loop optimization in SPIN may further boost the performance [28].

## 4.2 Qualitative Analysis

In this section, we provide more qualitative results in Figure 2. Several datasets are leveraged for evaluation, including LSP, MPII, COCO, Human3.6M, 3DPW and MPI-INF-3DHP.

# REFERENCES

[1] I. Akhter and M. J. Black. 2015. Pose-conditioned joint angle limits for 3D human pose reconstruction. In *Computer Vision and Pattern Recognition (CVPR)*. 1446–1455.
[2] M. Andriluka, U. Iqbal, E. Insafutdinov, L. Pishchulin, A. Milan, J. Gall, and B. Schiele. 2018. PoseTrack: A Benchmark for Human Pose Estimation and Tracking. In *Computer Vision and Pattern Recognition (CVPR)*. 5167–5176.
[3] 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.
[4] 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.
[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] J. Carreira and A. Zisserman. 2017. Quo Vadis, Action Recognition? A New Model and the Kinetics Dataset. In *Computer Vision and Pattern Recognition (CVPR)*. 4724–4733.
[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] M. Dantone, J. Gall, C. Leistner, and L. Van Gool. 2014. Body Parts Dependent Joint Regressors for Human Pose Estimation in Still Images. *Transactions on Pattern Analysis and Machine Intelligence* 36, 11 (2014), 2131–2143.
[9] The Carnegie Mellon University (CMU) Graphics Laboratory Motion Capture Database. http://mocap.cs.cmu.edu. 2000.
[10] 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.
[11] 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.
[12] S. Johnson and M. Everingham. 2011. Learning effective human pose estimation from inaccurate annotation. In *Computer Vision and Pattern Recognition (CVPR)*. 1465–1472.
[13] 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.
[14] 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.
[15] Thomas N. Kipf and Max Welling. 2017. Semi-Supervised Classification with Graph Convolutional Networks. In *International Conference on Learning Representations (ICLR)*.
[16] 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.
[17] 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.
[18] 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.
[19] 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.
[20] 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.
[21] 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.
[22] Naureen Mahmood, Nima Ghorbani, Nikolaus F. Troje, Gerard Pons-Moll, and Michael J. Black. 2019. AMASS: Archive of Motion Capture As Surface Shapes. In *International Conference on Computer Vision (ICCV)*. 5442–5451.
[23] 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.
[24] Dushyant Mehta, Srinath Sridhar, Oleksandr Sotnychenko, Helge Rhodin, Mohammad Shafiei, Hans-Peter Seidel, Weipeng Xu, Dan Casas, and Christian Theobalt. 2017. VNect: Real-Time 3D Human Pose Estimation with a Single RGB Camera. *ACM Transactions on Graphics* 36, 4, Article 44 (July 2017), 14 pages.
[25] 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.
[26] Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Liò, and Yoshua Bengio. 2018. Graph Attention Networks. In *International Conference on Learning Representations (ICLR)*.
[27] Yuxin Wu and Kaiming He. 2018. Group Normalization. In *European Conference on Computer Vision (ECCV)*. 3–19.

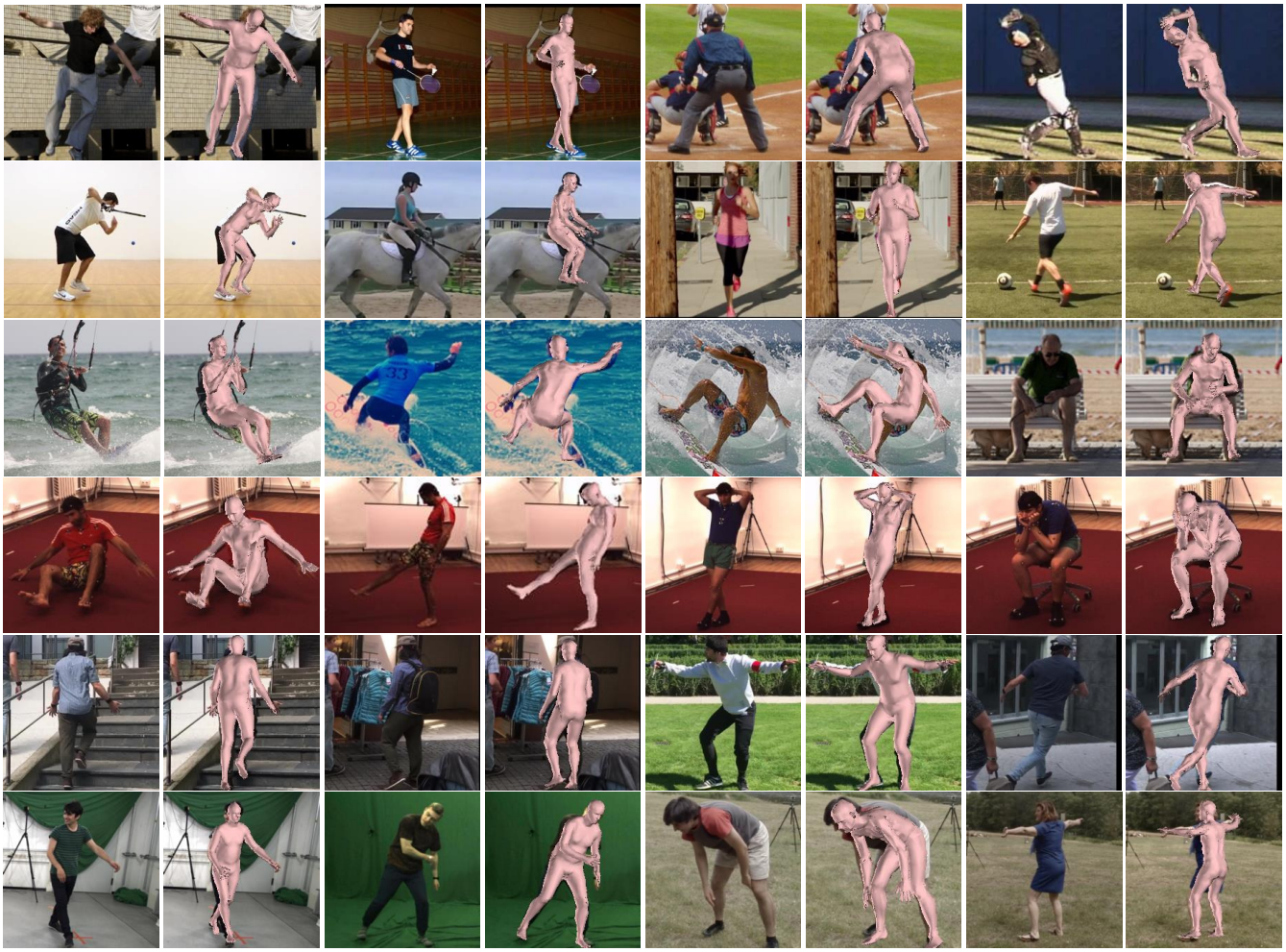

**Figure 2: Additional qualitative results. LSP (row 1), MPII (row 2), COCO (row 3), H36M (row 4), 3DPW (row 5), MPI-INF-3DHP (row 6).**

[28] 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.

[29] W. Zhang, M. Zhu, and K. G. Derpanis. 2013. From Actemes to Action: A Strongly-Supervised Representation for Detailed Action Understanding. In *Computer Vision and Pattern Recognition (CVPR).* 2248–2255.

[30] 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.