A Sequential Learning-based Approach for Monocular Human Performance Capture Supplementary Material

1. Appendix A: SMPL Body Shape and Pose Estimation

In this section, we describe our detailed approach to estimate SMPL body shape and pose in Sec. 3. Given a video with length L, the deep neural networks [8,9,18] provides an initial estimate of body pose ($\bar{\boldsymbol{\beta}}$) and shape $\{\boldsymbol{\theta}\}_1^L$. Then, we refine the predicted SMPL parameters jointly with global translations $\{\mathbf{t}\}_1^L$ and camera intrinsic parameters $\mathbf{K}$ (i.e. focal length $\mathbf{f_x}, \mathbf{f_y}$ in perspective camera) by a reconstruction loss. The camera intrinsics and extrinsics (i.e. global translation) project the 3D body mesh to the image plane. By minimizing the following cost function [15], we fit the projected body mesh into the input image to recover an accurate and temporal coherent SMPL body mesh.

$$\min_{\mathbf{K},\boldsymbol{\beta},\{\boldsymbol{\theta}\}_1^L,\{\boldsymbol{t}\}_i^L} E^b = E^b_{2d} + E^b_{dp} + E^b_{sil} + E^b_{pof} + E^b_{reg} \quad (1)$$

E^b_{2d} [2] minimizes the L2 distance between Openpose [3] detected 2D keypoints and projected SMPL joints. E^b_{dp} [6] minimizes the L2 distance between the location of 2D pixels inside the body and projection of corresponding vertices of SMPL body predicted by DensePose [7]. E^b_{sil} [15] maximizes the Intersection-over-Union between differentiably rendered silhouette of SMPL body and 0-1 mask obtained from 2D human parser [5]. E^b_{pof} minimizes the difference between the SMPL joint orientation and predicted Part Orientation Field [14]. E^b_{reg} regularizes $\{\theta\}_1^t$ with the Gaussian prior, β with the L2 loss and the temporal consistency of the SMPL vertices over time. Concretely,

$$E^b_{2d} = \sum_{joint\ i} \omega_i \rho(\Pi_{\mathbf{K}}(\boldsymbol{R_\theta}(J(\boldsymbol{\beta})_i)) - \mathbf{J}_{est,i}) \quad (2)$$

$$E^b_{dp} = \sum_{pixel\ i} ||x_i - \Pi_{\mathbf{K}}(\boldsymbol{M_\phi}(u(x_i)))||_2 \quad (3)$$

$$E^b_{sil} = \sum_{pixel\ i} (1 - \mathbf{IOU}(\mathbf{S_i}, \mathbf{S}_{est,i})) \quad (4)$$

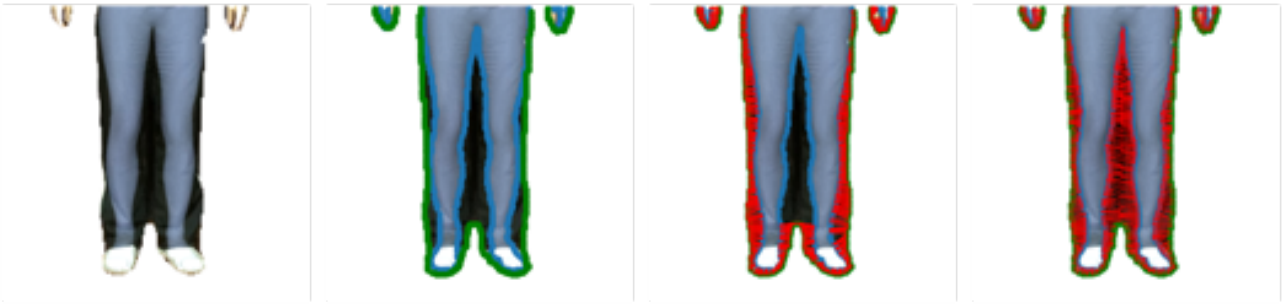
$$E^b_{pof} = \sum_{part\ i} (1 - P_\theta(J(\boldsymbol{\beta}))_i \cdot \mathbf{P}_{est,i}) \quad (5)$$

$$E^b_{reg} = -\log \sum_j (g_j \mathcal{N}(\boldsymbol{\theta}; \mu_{\boldsymbol{\theta},j}, \Sigma_{\boldsymbol{\theta},j})) + \boldsymbol{\beta}^T \Sigma_{\boldsymbol{\beta}}^{-1} \boldsymbol{\beta} + ||\boldsymbol{\theta}^{(n)} - \boldsymbol{\theta}^{(n-1)}||_2 \quad (6)$$

where ω_i is the keypoint confidence value predicted by Openpose [3]; $\mathbf{R_\theta}$ is the articulated transformation given pose $\boldsymbol{\theta}$; ρ is the Geman-McClure robust loss function [4]; $\mathbf{M_\theta}$ is the UV-mapping from UV coordinate $u(x_i)$ to 3D coordinate on predicted SMPL body mesh; $\boldsymbol{S_i}$ is the rendered silhouette map from predicted SMPL body mesh; P_θ is a unit vector that indicates the orientation of skeleton part defined by joint locations; $g_j, \mu_{(\boldsymbol{\theta},j)}, \Sigma_{\boldsymbol{\theta},j}$ are Gaussian Mixture Model parameters for pose $\boldsymbol{\theta}$ and $\Sigma_{\boldsymbol{\beta}}$ is PCA parameters for body shape $\boldsymbol{\beta}$ learned from SMPL training set.

2. Appendix B: Searching Correspondences of Boundary Points in 2D Silhouette Maps

In this section, we introduce a correspondence searching algorithm to align our rendered silhouette map to target silhouette map from segmentation [5], which is used in Sec. 5.1 and 5.2. Instead of using Closest Point to align boundaries of two silhouette maps [17], our main motivation is to cope with the corner cases where the boundary of silhouette degenerates due to self-occlusion. Fig. 1b demonstrates a scenario where the green boundary of target silhouette map degenerates. Compared with ICP methods that registers two sets silhouette boundary, we register the boundary of our rendered silhouette map to cover the disparity regions between two silhouette maps. Specifically, we detect the edge of our rendered silhouette by a laplacian filter [13] as source point set $\mathcal{S}$ and record the 2D coordinate of all misaligned pixels as target point set $\mathcal{T}$. For $i \in \mathcal{S}$, Alg. 1 searches the corresponding point $j \in \mathcal{T}$ over all misaligned region. Fig. 1c and 1d shows the advantage of our algorithm in finding correct 2D correspondence for boundary points of rendered silhouette map. Since the human parsing approach [5] also provides semantic part labels, we apply our correspondence search algorithm independently for upper and lower cloth, which better generates the boundary between two separate 3D clothes.



(a) Input Image and Rendered Mesh (b) Boundaries of Silhouette Maps (c) ICP Correspondence (d) Correspondence from Alg. 1

Algorithm 1 2D Correspondence Searching Algorithm for Silhouette Map Alignment.

Input Source point set $\mathcal{S}$, Target point set $\mathcal{T}$
Output Correspondence Assignment Matrix $\mathcal{M}$

$\mathcal{D} = \infty, \mathcal{M} = 0$
for $t_j \in \mathcal{T}$ **do**
  $s_k = \texttt{ClosestPoint}(\mathcal{S}, t_j)$
  $\mathcal{D}_{k,j} = ||s_k - t_j||^2$
end for
for $s_i \in \mathcal{S}$ **do**
  $\mathcal{M}_{i,\texttt{argmax}(\mathcal{D}_i)} = 1$
end for
return $\mathcal{M}$

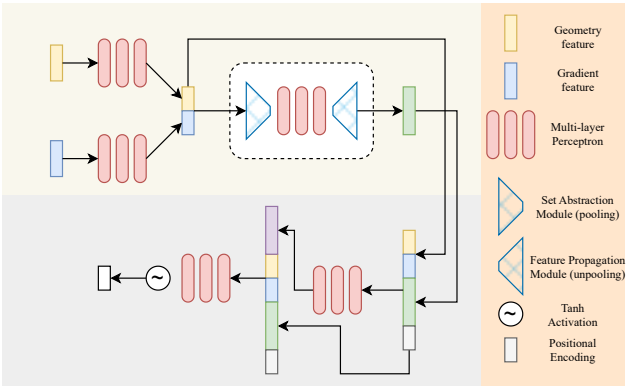


Figure 2. General architecture of our proposed *gradient rectification network* $\mathcal{F}$.

3. Appendix C: Network Architecture of *gradient rectification network*

This section introduces the network architecture we used in Sec. 4,5. Overall, we exploit an encoder-decoder structure for GRN. The encoder consists of two separate branches for input geometry features $\mathbf{X}, n_{\mathbf{X}}$ and gradient features $\frac{\partial E^c}{\partial \mathbf{X}}$. We leverage a PointNet++ [12] structure to encode the input gradient and a coordinate-based MLPs to decode the multi-scale feature to output gradient. Given $\mathbf{X}$, $\frac{\partial E^c}{\partial \mathbf{X}}$, the encoder is consist of two separate MLPs for geometry feature and gradient feature. Geometry feature simply contains 3D coordinate of $\mathbf{X}$ and normal $n_{\mathbf{X}}$. For Sec. 5.1 and Sec. 5.2, the gradient feature are typically sparse, since 2D energy term E_c is from silhouette map. Therefore, we also consider the symmetric assumption of garments motivated by [1]. To achieve this, each vertex further takes the "input gradient" of its x-symmetric and z-symmetric vertex as auxiliary feature. We set two levels of downsampling for PointNet++. Since the input point clouds are from canonical T-posed SMPL+D body with fixed topology, the clusters for downsampling in *Set Abstraction Module* and the interpolation weights for upsampling in *Feature Propagation Module* in PointNet++ are pre-defined. Particularly, the vertices are devided into 104 patches according to UV-map and 24 components according to SMPL-joint location for two levels of downsampling. The multiscale feature learned from PointNet++ is concatenated with the positional-encoding [10] of 3D vertex coordinates to predict the output gradient. Following the fashion from [10, 11], the 8-layer MLP decoder has a skip connection. In order to constraint the output range in each step, we add a *tanh* activation function in the last layer. The architecture of our network is visualized in Fig. 2.

4. Appendix D: Limitations

The major limitations of our method are: (1) Due to the underlying SMPL+D model, our method can not reconstruct certain types of clothing such as dresses or multilayered outfits. To extend to these cases, we need a more expressive body model. (2) As it is a deformation-based approach, in some challenging cases, we can not perfectly align the silhouette as in the case of PIFu methods do (e.g., trousers leg in the last row of Fig.6). As a trade-off, our method is more robust in generating clean and compact surfaces without outliers.

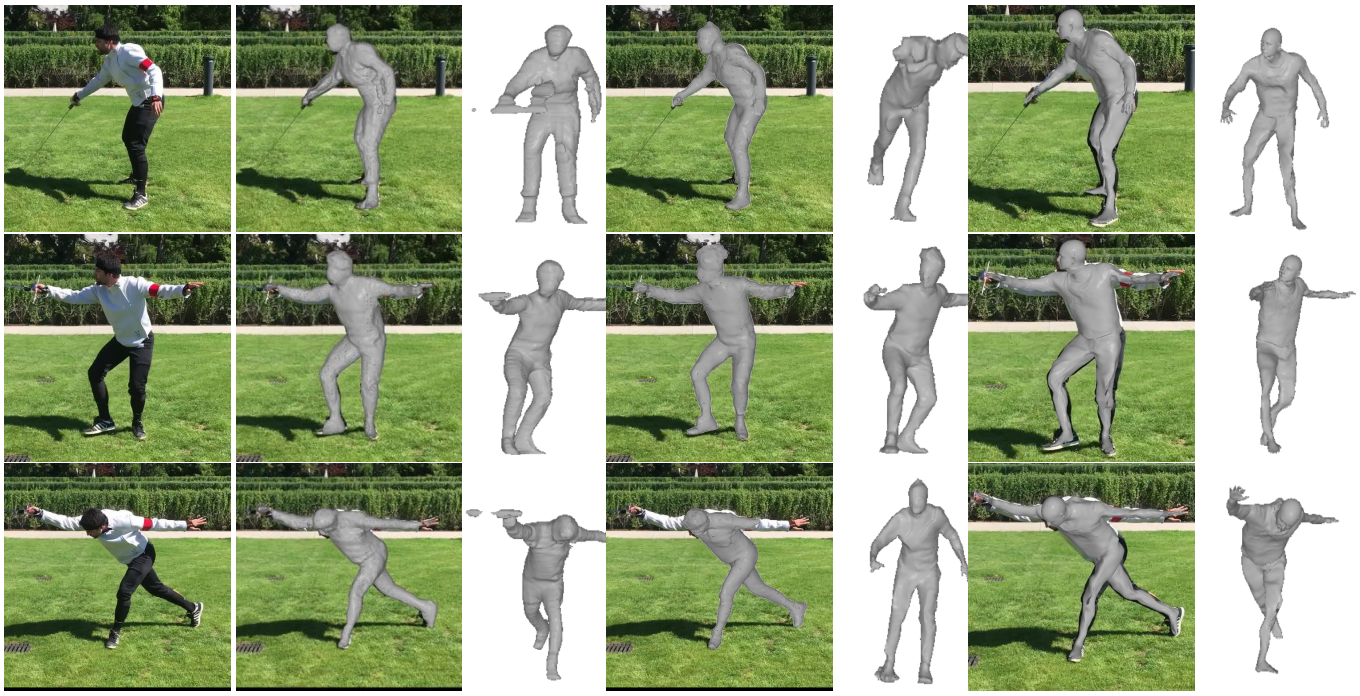


Figure 3. Qualitative results on human performance capture from 3DP-W dataset.

References

[1] Thiemo Alldieck, Marcus Magnor, Weipeng Xu, Christian Theobalt, and Gerard Pons-Moll. Video based reconstruction of 3d people models. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8387–8397, 2018. 2

[2] Federica Bogo, Angjoo Kanazawa, Christoph Lassner, Peter Gehler, Javier Romero, and Michael J. Black. Keep it SMPL: Automatic estimation of 3D human pose and shape from a single image. In *Computer Vision – ECCV 2016*, Lecture Notes in Computer Science. Springer International Publishing, Oct. 2016. 1

[3] Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Realtime multi-person 2d pose estimation using part affinity fields. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7291–7299, 2017. 1

[4] Stuart Geman. Statistical methods for tomographic image reconstruction. *Bull. Int. Stat. Inst*, 4:5–21, 1987. 1

[5] Ke Gong, Yiming Gao, Xiaodan Liang, Xiaohui Shen, Meng Wang, and Liang Lin. Graphonomy: Universal human parsing via graph transfer learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7450–7459, 2019. 1

[6] Riza Alp Guler and Iasonas Kokkinos. Holopose: Holistic 3d human reconstruction in-the-wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10884–10894, 2019. 1

[7] Rıza Alp Güler, Natalia Neverova, and Iasonas Kokkinos. Densepose: Dense human pose estimation in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7297–7306, 2018. 1

[8] Muhammed Kocabas, Nikos Athanasiou, and Michael J. Black. Vibe: Video inference for human body pose and shape estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020. 1

[9] Muhammed Kocabas, Chun-Hao P. Huang, Otmar Hilliges, and Michael J. Black. PARE: Part attention regressor for 3D human body estimation. In *Proc. International Conference on Computer Vision (ICCV)*, pages 11127–11137, Oct. 2021. 1

[10] Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. In *European conference on computer vision*, pages 405–421. Springer, 2020. 2

[11] Jeong Joon Park, Peter Florence, Julian Straub, Richard Newcombe, and Steven Lovegrove. Deepsdf: Learning continuous signed distance functions for shape representation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 165–174, 2019. 2

[12] Charles Ruizhongtai Qi, Li Yi, Hao Su, and Leonidas J Guibas. Pointnet++: Deep hierarchical feature learning on point sets in a metric space. *Advances in neural information processing systems*, 30, 2017. 2

[13] Mohsen Sharifi, Mahmood Fathy, and Maryam Tayefeh Mahmoudi. A classified and comparative study of edge detection algorithms. In *Proceedings. International conference*

on information technology: Coding and computing, pages 117–120. IEEE, 2002. 1

[14] Donglai Xiang, Hanbyul Joo, and Yaser Sheikh. Monocular total capture: Posing face, body, and hands in the wild. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10965–10974, 2019. 1

[15] Donglai Xiang, Fabian Prada, Chenglei Wu, and Jessica Hodgins. Monoclothcap: Towards temporally coherent clothing capture from monocular rgb video. In *2020 International Conference on 3D Vision (3DV)*, pages 322–332. IEEE, 2020. 1

[16] Yuliang Xiu, Jinlong Yang, Dimitrios Tzionas, and Michael J Black. Icon: implicit clothed humans obtained from normals. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13286–13296. IEEE, 2022. 3

[17] Weipeng Xu, Avishek Chatterjee, Michael Zollhöfer, Helge Rhodin, Dushyant Mehta, Hans-Peter Seidel, and Christian Theobalt. Monoperfcap: Human performance capture from monocular video. *ACM Transactions on Graphics (ToG)*, 37(2):1–15, 2018. 1

[18] Hongwen Zhang, Yating Tian, Xinchi Zhou, Wanli Ouyang, Yebin Liu, Limin Wang, and Zhenan Sun. Pymaf: 3d human pose and shape regression with pyramidal mesh alignment feedback loop. In *Proceedings of the IEEE International Conference on Computer Vision*, 2021. 1

[19] Zerong Zheng, Tao Yu, Yebin Liu, and Qionghai Dai. Pamir: Parametric model-conditioned implicit representation for image-based human reconstruction. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021. 3