# BodyMap: Learning Full-Body Dense Correspondence Map Supplementary Material

Anastasia Ianina[1*], Nikolaos Sarafianos[3], Yuanlu Xu[3], Ignacio Rocco[2], Tony Tung[3]
[1]Moscow Institute of Physics and Technology, [2]Meta AI, [3]Meta Reality Labs Research, Sausalito
[1]yanina@phystech.edu, [2,3]{nsarafianos, yuanluxu, irocco, tonytung}@fb.com

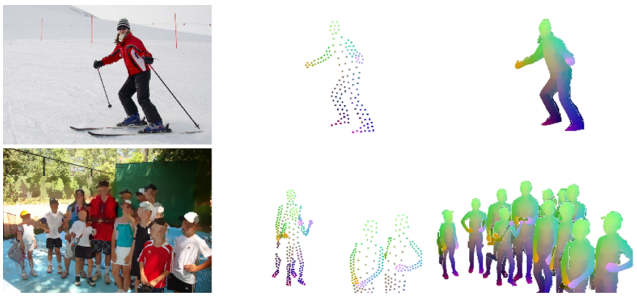

Figure 1. **Sparse and dense ground truth for DensePose-COCO dataset**. Sparse ground truth: annotation available in the dataset. Dense ground truth: pseudo ground-truth estimates are generated on the fly by extrapolating both the available ground-truth annotations but also the CSE initialization such that they cover the whole estimated silhouette of the human

| Method \ Window (px) | 5 | 10 | 20 |
|---|---|---|---|
| DensePose [13] | 49.23 | 55.75 | 59.71 |
| CSE [14] | 58.10 | 60.34 | 64.14 |
| BodyMap: no ft | 40.51 | 52.18 | 55.22 |
| BodyMap: ft 1 | 56.12 | 60.02 | 63.15 |
| BodyMap: ft 2 | **65.34** | **68.22** | **73.88** |

Table 1. **Ablation study on the effectiveness of finetuning: metrics in 2D image space.** We illustrate the accuracy in 2D space on DensePose-COCO (the percentage of pixels correctly matched within the established error window) before fine-tuning (**no ft**), and after fine-tuning with: (i) sparse annotations with 100 labeled points per person (**ft 1**), and (ii) dense annotations, obtained heuristically by interpolating annotations within silhouettes **ft 2**). We include the performance of DensePose and CSE for comparison.

In this supplementary material we provide information regarding the broader impact of our method (Sec. A) additional details regarding fine-tuning our proposed BodyMap on real data (Sec. B) and an in-depth discussion on applications to neural re-rendering and cloth swapping with several qualitative examples (Sec. C). Additional qualitative evaluations and results are shown in the supplemental video.

## A. Broader Impact

The positive impact of our technology is further discussed in the applications section, including novel view synthesis and appearance swapping. While every human-related technology may raise concerns of fraudulent activities, we should note that our approach does not reconstruct facial or any other features used for personality identification. Thus, it is not possible to identify a person using our method, which makes our technology safe.

## B. Fine-tuning BodyMap on real data

While BodyMap is trained with mostly synthetic data, we further fine-tune it with DensePose-COCO dataset. Available DensePose-COCO annotations are extremely sparse (around 100 annotated pixels per person, also annotations are not present for small people in the background). Thus, we leverage a heuristically made dense pseudo ground truth generation scheme. Given an image from the dataset, we generate pseudo ground-truth estimates on the fly by extrapolating both the available ground-truth annotations but also the CSE initialization such that they cover the whole estimated silhouette of the human (Figure 1). In that way we can fine-tune our model on real-data with denser supervision and utilize losses in both 2D and 3D spaces.

We experimented with two ways of fine-tuning on real data: (1) using only available sparse annotations (sparse fine-tuning); (2) using the generated dense pseudo ground-truth estimates (dense fine-tuning). The results of the both schemes are presented in Tables 1 and 2, showing that dense fine-tuning allows to get the best results beating competitive methods such as HumanGPS, DensePose and CSE.

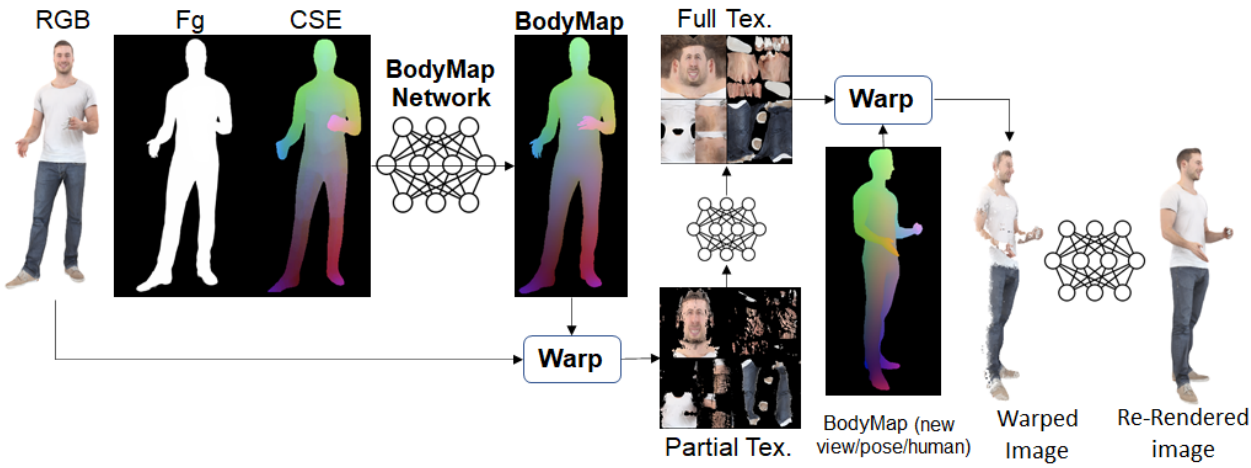


Figure 2. **Novel view rendering with BodyMap**. Given an RGB image we first obtain its CSE [14] estimates and feed both to BodyMap network to get refined per-pixel predictions covering the whole silhouette. With such correspondences we can obtain partial texture maps that are very well aligned with the complete texture map which are then utilized to train a texture completion network to fill in the missing information. At last, we use the BodyMap as a mapping function back to the image space where we train a neural re-rendering framework that generates the photorealistic renders.

| Method | $AP$ | $AP_{50}$ | $AP_{75}$ | $AP_M$ | $AP_L$ | $AR$ | $AR_{50}$ | $AR_{75}$ | $AR_M$ | $AR_L$ |
|---|---|---|---|---|---|---|---|---|---|---|
| DP [2] | 55.3 | 85.6 | 60.1 | 48.3 | 58.2 | 66.8 | 90.1 | 68.2 | 50.1 | 66.1 |
| CSE [14] | 72.8 | 95.7 | 84.2 | 65.7 | 73.1 | 78.2 | 97.3 | 87.5 | 67.2 | 78.0 |
| BodyMap: no ft | 50.1 | 82.3 | 58.7 | 45.7 | 57.5 | 65.1 | 88.1 | 65.4 | 49.2 | 64.5 |
| BodyMap: ft 1 | 70.3 | 94.2 | 83.1 | 63.2 | 72.1 | 77.8 | 96.5 | 85.1 | 65.9 | 76.1 |
| BodyMap: ft 2 | **79.5** | **97.8** | **90.5** | **72.3** | **79.4** | **85.3** | **98.1** | **92.5** | **73.4** | **84.5** |

Table 2. **Ablative studies on the effectiveness of finetuning: metrics in 3D space.** We illustrate Average Precision (AP) and Average Recall (AR) over GPS scores on DensePose-COCO dataset before fine-tuning (**no ft**), and after fine-tuning with: (i) sparse annotations with 100 labeled points per person (**ft 1**), and (ii) dense annotations, received heuristically by interpolating annotations within silhouettes **ft 2**). We also include the performance of DensePose and CSE for comparison.

# C. Applications

## C.1. Rendering people in arbitrary poses/views

One of the straightforward applications of our proposed BodyMap framework is re-rendering people in novel poses/views. Previous attempts to do so include Texformer [21] — Transformer-based framework for 3D human texture estimation from a single image. They utilize pre-computed color encoding of the UV space obtained by mapping the 3D coordinates of a standard human body mesh to the UV space as a Query, feeding it to the Transformer. We also rely on specifically designed Transformer-based architecture to retrieve dense correspondences that can lately be used as a mapping function between 2D image space and 3D space. While we can do novel view rendering without any extra efforts simply warping the texture using BodyMap estimates as a mapping function, using additional neural renderer significantly increases the quality of the final render. Next we describe in the detail the proposed pipeline for generating novel views consisting of two main steps: texture completion network and neural rendering. The pipeline is also depicted in Figure 2.

### C.1.1 Texture Completion Network

Given the BodyMap estimates and the foreground RGB image we define a warping function that maps each foreground pixel of the image space to the UV-map space using BodyMap estimates as the mapping function. Note that in the UV map space, every point on the mesh surface of a human body template is represented by its coordinates on this UV map. We then train a texture completion neural network that takes as input partial textures at $1024 \times 1024$ resolution and completes the missing information by producing the full texture. The fact that partial texture is so well aligned on the UV map with full texture enables us to utilize a U-Net-like architecture since the skip connections can transfer the aligned input from the encoder to the decoder layers without adding an additional overheard to the decoder. A challenge that arises when dealing with high-resolution inputs is that only a few samples can fit into the GPU memory during training. Instance normalization blocks have widely been used in such cases to avoid collecting batch statistics that can be inaccurate due to the small sample size [1, 9], but our experimental investigation indicated that instance normalization produces completed textures with distorted colors in non-visible regions. To overcome this challenge we propose to utilize synchronized batch normalization which differs from previous methods in the way the statistics are computed over all training samples distributed on multiple

devices. This enables us to learn more accurate batch statistics that can then be used at test-time with traditional batch normalization blocks. Thus, given pairs of partial and full textures ($\{P_T, F_T\} \sim T$) coming from the data distribution we train the texture completion network with the following losses:

- Hinge version [10, 23] of the **adversarial loss**, along with a multi-scale discriminator as used in Pix2PixHD [4]:

$$L_G = -\mathbb{E}_{P_T \sim p_{P_T}, F_T \sim T} D(G(P_T), F_T) \tag{1}$$

$$\begin{aligned} L_D = &-\mathbb{E}_{\{P_T, F_T\} \sim T}[0, -1 + D(P_T, F_T)] \\ &- \mathbb{E}_{P_T \sim p_{P_T}, F_T \sim T}[0, -1 - D(G(P_T), F_T)] \end{aligned} \tag{2}$$

- **Perceptual loss.** We utilize a pre-trained VGG [20] network and compute the $L_1$ loss between the completed texture estimate and the ground-truth texture map at the activations of five different layers of the network. Perceptual similarity losses help the network to generate fine-level details which are common in the textures of clothed humans.

- **Total variation loss.** We add a total variation (TV) loss [12] with a small weight in order to encourage spatial smoothness in the generated textures and remove some artifacts that are quite common in image-to-image translation networks. The TV loss is formulated as follows:

$$\begin{aligned} L_{TV} = \sum_u \sum_v &|F_T(u, v+1) - F_T(u, v)| \\ &+ |F_T(u+1, v) - F_T(u, v)| \end{aligned} \tag{3}$$

### C.1.2 Neural Re-rendering

We present an additional step to obtain photo-realistic human renders after texture completion. While we can assume access to 3D geometry for our synthetic data and perform rendering with tools such as Blender Cycles [3] or PyTorch3D [15, 17], this is not the case for real-world examples. One approach to tackle this problem and obtain a 3D geometry would be to estimate the 3D human body pose [18] and shape from a single image by using any of the recent state-of-the-art methods [5–8, 16, 22, 24, 25]. However, all these methods estimate the body under the clothing which is usually relatively slimmer and with pose inaccuracies (e.g., the body bends forward) due to the depth ambiguity which makes it unsuitable for rendering the texture of a clothed human on top. Thus, we propose a model for neural re-rendering which aims at learning a function $\widetilde{X} = R(BodyMap, F_T)$ that given the complete texture map $F_T$ and the estimated BodyMap generates a photorealistic render $\widetilde{X}$ in the image space. The advantage of our approach compared to prior work [19] is that BodyMap not only provides a proper silhouette in the image space that needs to be rerendered but also serves as a mapping function between the UV and the image space. This enables us to warp the texture map back to the image space using the BodyMap to obtain an initial estimate which is then fed to an encoder-decoder network that generates the final output render. The fact that BodyMap provides accurate per-pixel foreground estimates makes the warped image well aligned with a target output which simplifies the learning process of the neural renderer. Finally, since during the warping process some texture information can be lost due to warping inaccuracies [11], we pass $F_T$ through an encoder network to generate a lower dimensional tensor representation which is then fed via the bottleneck to the decoder of the neural render. We train this network with the same losses described that we used for texture completion network and in addition, we employ the following two losses:

- **Feature Matching Loss.** We use a feature matching loss in the discriminator layers [4] to obtain high-frequency details such as wrinkles and cloth patterns which is defined as: $L_{FM} = \sum_{l=1}^{3} ||D_l(x) - D_l(G(x))||_1$.

- **Reconstruction Loss in the face region.** Using the segmentation mask of the face region which are then used to employ additional reconstruction losses in that area in order to force the network to estimate more photo-realistic faces and fix artifacts around the eyes and the mouth.

### C.1.3 Results

Expanding Figure 4 from the main paper, we present several more results of re-rendered people in 3. It shows that even before re-rendering (column 6) novel views demonstrate a decent level of details including textile patterns, hairstyles and fingers. Neural re-rendering helps to get rid of occasional artifacts, smooth out the final result and indicate even more fine-grained details.

## C.2. Appearance swapping

Additionally to generating people in novel views and/or poses, our approach allows to redress people providing renders of them in different clothes. Using BodyMap estimates as a mapping function, we show several examples of appearance swapping in Figure 4.


Figure 3. Novel view synthesis results before and after neural re-rendering (NR).


Figure 4. **Applications on virtual dressing and motion imitation:** given a single RGB image and a variety of target poses our method can dress the target subjects with the input hoodie at different poses with the hands being potentially occluded while preserving fine-level details. Our approach also generates photorealistic renders of the input human given target images of other people with different poses and viewpoints. Our method has not been trained specifically for any of these tasks but it can still generalize and produce crisp results.

## References


[1] Thiemo Alldieck, Gerard Pons-Moll, Christian Theobalt, and Marcus Magnor. Tex2shape: Detailed full human body geometry from a single image. In *ICCV*, 2019. 2

[2] Rıza Alp Güler, Natalia Neverova, and Iasonas Kokkinos. Densepose: Dense human pose estimation in the wild. In *CVPR*, 2018. 2

[3] Blender Online Community. *Blender - A 3D modelling and rendering package*. 3

[4] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *CVPR*, 2017. 3

[5] Angjoo Kanazawa, Michael J Black, David W Jacobs, and Jitendra Malik. End-to-end recovery of human shape and pose. In *CVPR*, 2018. 3

[6] Muhammed Kocabas, Nikos Athanasiou, and Michael J


Figure 5. Warping results with BodyMap estimates as a mapping function vs. competitors.

Black. Vibe: Video inference for human body pose and shape estimation. In *CVPR*, 2020. 3

[7] Nikos Kolotouros, Georgios Pavlakos, Michael J Black, and Kostas Daniilidis. Learning to reconstruct 3D human pose and shape via model-fitting in the loop. In *ICCV*, 2019. 3

[8] Nikos Kolotouros, Georgios Pavlakos, and Kostas Daniilidis. Convolutional mesh regression for single-image human shape reconstruction. In *CVPR*, 2019. 3

[9] Verica Lazova, Eldar Insafutdinov, and Gerard Pons-Moll. 360-degree textures of people in clothing from a single image. In *3DV*, 2019. 2

[10] Jae Hyun Lim and Jong Chul Ye. Geometric gan. *arXiv preprint arXiv:1705.02894*, 2017. 3

[11] Wen Liu, Zhixin Piao, Jie Min, Wenhan Luo, Lin Ma, and Shenghua Gao. Liquid warping GAN: A unified framework for human motion imitation, appearance transfer and novel view synthesis. In *ICCV*, 2019. 3

[12] Aravindh Mahendran and Andrea Vedaldi. Understanding deep image representations by inverting them. In *CVPR*, 2015. 3

[13] Natalia Neverova, Riza Alp Guler, and Iasonas Kokkinos. Dense pose transfer. In *ECCV*, 2018. 1

[14] Natalia Neverova, David Novotny, Vasil Khalidov, Marc Szafraniec, Patrick Labatut, and Andrea Vedaldi. Continuous surface embeddings. *arXiv preprint arXiv:2011.12438*, 2020. 1, 2

[15] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *NeurIPS*, 32, 2019. 3

[16] Georgios Pavlakos, Luyang Zhu, Xiaowei Zhou, and Kostas Daniilidis. Learning to estimate 3D human pose and shape from a single color image. In *CVPR*, 2018. 3

[17] Nikhila Ravi, Jeremy Reizenstein, David Novotny, Taylor Gordon, Wan-Yen Lo, Justin Johnson, and Georgia Gkioxari. Accelerating 3d deep learning with pytorch3d. *arXiv:2007.08501*, 2020. 3

[18] Nikolaos Sarafianos, Bogdan Boteanu, Bogdan Ionescu, and Ioannis A Kakadiaris. 3D human pose estimation: A review of the literature and analysis of covariates. *Computer Vision and Image Understanding*, 2016. 3

[19] Kripasindhu Sarkar, Dushyant Mehta, Weipeng Xu, Vladislav Golyanik, and Christian Theobalt. Neural re-rendering of humans from a single image. In *European Conference on Computer Vision*, pages 596–613. Springer, 2020. 3

[20] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In *ICLR*, 2015. 3

[21] Xiangyu Xu and Chen Change Loy. 3d human texture estimation from a single image with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13849–13858, 2021. 2

[22] Yuanlu Xu, Song-Chun Zhu, and Tony Tung. Denserac: Joint 3D pose and shape estimation by dense render-and-compare. In *ICCV*, 2019. 3

[23] Han Zhang, Ian Goodfellow, Dimitris Metaxas, and Augustus Odena. Self-attention generative adversarial networks. In *ICML*, 2019. 3

[24] Fang Zhao, Shengcai Liao, Kaihao Zhang, and Ling Shao. Human parsing based texture transfer from single image to 3D human via cross-view consistency. *NeurIPS*, 2020. 3

[25] Zerong Zheng, Tao Yu, Yixuan Wei, Qionghai Dai, and Yebin Liu. Deephuman: 3D human reconstruction from a single image. In *ICCV*, 2019. 3