BodyMap: Learning Full-Body Dense Correspondence Map

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

Abstract

Dense correspondence between humans carries powerful semantic information that can be utilized to solve fundamental problems for full-body understanding such as in-the-wild surface matching, tracking and reconstruction. In this paper we present BodyMap, a new framework for obtaining high-definition full-body and continuous dense correspondence between in-the-wild images of clothed humans and the surface of a 3D template model. The correspondences cover fine details such as hands and hair, while capturing regions far from the body surface, such as loose clothing. Prior methods for estimating such dense surface correspondence i) cut a 3D body into parts which are unwrapped to a 2D UV space, producing discontinuities along part seams, or ii) use a single surface for representing the whole body, but none handled body details. Here, we introduce a novel network architecture with Vision Transformers that learn fine-level features on a continuous body surface. BodyMap outperforms prior work on various metrics and datasets, including DensePose-COCO by a large margin. Furthermore, we show various applications ranging from multi-layer dense cloth correspondence, neural rendering with novel-view synthesis and appearance swapping.

1. Introduction

Several fundamental problems related to human understanding in images can be addressed by labeling every pixel covering the human body with semantic information. This enables numerous applications including video analysis, image editing, texture generation and style transfer. From a single RGB image of a human, the literature has proposed methods to extract sparse information such as 2D body keypoints (e.g., face, hands, body joints), or 2D segmentation masks (e.g., for full body, clothes, hair or skin), and also 3D body pose and shape parameters defined by a template body model [6, 8, 13, 33, 35, 44], while work on dense surface correspondence has further enabled pixel-level understanding by establishing unique correspondences between 2D pixels covering the visible regions of the human body and 3D points on the surface of a body template.

Figure 1. We introduce **BodyMap** — a method that establishes accurate dense correspondences between a 2D image and the surface of a 3D clothed human with high precision. Our approach handles loose clothes, different hairstyles and various accessories, like hats and bags, providing crisp silhouettes, and works well in multi-person cases with occlusions.

In the seminal work DensePose [26], correspondences are estimated between image pixels belonging to the human body and points in disjoint parts of a human body template located using UV coordinates, similar to local texture mapping. The method is trained on the large in-the-wild dataset DensePose-COCO and is robust to human pose variability, image resolution, diversity in clothing, and occlusions. However, it has some inherent limitations that impact methods that rely on it (e.g., for clothed-human applications) [1, 21, 46]. First, the discretization generated by

*This work was conducted during an internship at Meta RL Research.

dividing the body into disjoint parts produces clearly visible seams and discontinuities between them that are not optimal for learning models. Second, the DensePose estimates suffer from inaccuracy as reported in prior work [26, 28, 39], mainly due to the difficulty in acquiring ground-truth annotations for the task [2, 23]. Follow-up methods have tackled some of its shortcomings and a few recent works addressed the discontinuity of UV maps [4, 27, 47]. HumanGPS [41] proposes to predict per-pixel embeddings using geodesic distances between corresponding points on the surface of a 3D human scan and does not produce an explicit mapping. None of the proposed approaches has established high-definition correspondences for areas with finer details such as hair and hands (with fingers), with generalization to clothed humans, especially with loose clothing.

In this work we introduce a novel technique to establish high-definition full-body and continuous dense correspondence between images of *clothed* humans and the human body surface. Our method, which we term as BodyMap, takes as input an RGB image of a human and outputs accurate per-pixel continuous correspondence estimates for each foreground pixel (*i.e*. including the full body, with clothes and hair). We designed a transformer-based architecture that learns appearance-based and Continuous-Surface-Embeddings-based representations to infer accurate dense surface correspondence for the depicted human. Our variant of Vision Transformer [11] as a computational block of the encoder brings its advantageous properties for dense prediction tasks. The vector dimension is kept constant throughout all processing stages as well as global receptive field for every stage. With these properties, our network is well designed for dense correspondence prediction.

Furthermore, we capitalize on the power of synthetic data. Since no real-world dataset provides ground-truth annotation at the quality we aim for (fingers, clothes, hair), we created a synthetic dataset of animated 3D clothed human scans. In that way, we obtained ground-truth dense correspondence for a large variety of humans with diverse clothing, in different poses and from different viewpoints. A differentiating factor of our framework is that it is not tied to a human body with topology constraints, and can handle layered representations such as humans with separate cloth geometries. To summarize, our key contributions are:

- BodyMap is the first method to establish dense continuous correspondence for every foreground pixel of clothed humans, whether that is fingers, hair, or clothes that are displaced from the human body with high-precision — something that all prior works fail to achieve.
- A novel transformer-based architecture designed specifically for this task that when trained in a multi-task learning manner with per-pixel classification losses for each channel significantly outperforms prior works across several datasets and tasks.
- We achieve state-of-the-art results on DensePose COCO by a large margin. We show our approach can be applied to real-world applications such as novel view synthesis. Our method can be extended to learn layered representations with clothed humans and predict per-geometry surface correspondences.

2. Related Work

Dense Surface Correspondences. One of the most widely used approaches in this topic is DensePose [26], where classification and regression branches were trained to obtain per-pixel body part and UV estimates. The body parts constitute the I channel which takes one of 25 values (including the background) and the UV estimates which are continuous numbers mapped to $[0, 255]$. However, its output is discretized resulting in seams between body parts. This problem is alleviated in Continuous Surface Embeddings (CSE) [27], which for each pixel learns a positional embedding of the corresponding vertex in the object mesh. In CSE correspondences are learned without being constrained on specific geometry types (*e.g*., humans), and show the effectiveness of their approach on other deformable object categories, like animal classes which was later extended by discovering correspondences between different object classes automatically [29]. HumanGPS [41] maps each pixel to a feature space, where the feature distances reflect geodesic distances among vertices of a 3D body model corresponding to every pixel. Similarly to CSE, for every image pixel they produce an embedding capable of differentiating visually similar parts and aligning different subjects into an unified feature space. Zeng *et al*. [47] introduced, a model-free 3D human mesh estimation framework, which explicitly establishes the dense correspondences between the mesh and the local image features in the UV space. They solve human body estimation problem relying on dense local features transferred to the UV space. Getting enough labeled data (especially non-synthetic) to learn dense correspondences is a challenging task. SimPose [49] proposed to alleviate the problem by using simulated multi-person datasets and a specific training strategy with multi-task objectives to learn dense UV coordinates. They obtain favourable results using only simulated human UV labels. The intricacy of getting dense and accurately annotated correspondences is further explored in UltraPose [45]. They provide a dense synthetic benchmark focusing on faces, containing around 1.3 billion corresponding points as well as data generation system based on novel decoupling 3D model.

Architecture Designs for Dense Correspondences. There have been a couple of approaches in terms of network archi-

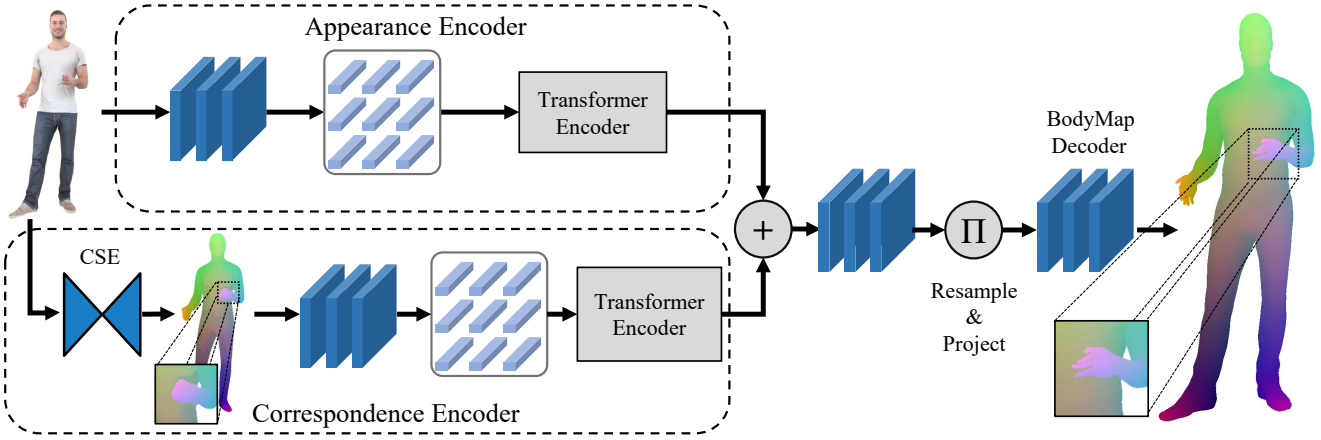

Figure 2. **BodyMap architecture**. Given an RGB image we first obtain its CSE [27] estimates and feed both to their corresponding encoders. We utilize vision transformers specifically designed for this task to learn to extract accurate high-dimensional representations that are then fed to the BodyMap decoder that predicts per-pixel dense correspondences.

tectures to extract dense human correspondences. In DensePose [26] a Mask-RCNN [15] with Feature Pyramid Features [22] is utilized to obtain accurate image features. SimPose [49] opted for a ResNet-101 backbone trained with losses adjusted to each of their tasks (*e.g.*, human pose, segmentations, normals, UVs). Another simple yet effective choice employed by HumanGPS [41] is an Encoder-Decoder architecture such as U-Net [36]. Our investigation indicated that while one can achieve satisfactory results with the aforementioned approaches they are all unable to capture finer-level details in the depicted human as usually the extracted features are too coarse. To alleviate this we turned into transformer architectures due to their ability to learn these discriminative features necessary for either downstream computer vision tasks or reconstruction applications. Originating from Natural Language Processing, a Transformer architecture [43] has shown its effectiveness within a wide range of Computer Vision tasks: image recognition and classification [11], image retrieval [12], image generation [31] and image captioning [16]. We capitalize upon prior work on vision transformers for dense prediction tasks [34] (*e.g.*, depth estimation) and introduce a new architecture explicitly design for predicting dense surface correspondences for humans.

3. The Proposed Method: BodyMap

The main goal of the proposed approach is to establish dense surface correspondence between a single RGB image and 3D body model. Our method takes as input a single RGB image, foreground mask and coarse correspondences retrieved using Continuous Surface Correspondences (CSE) [27]. CSE serves as a sufficient initialization which our method refines by providing more accurate estimates for the areas covering loose clothes, hair, fingers, etc. Thus, BodyMap provides per-pixel estimates for the foreground image resulting in much more accurate representations and crisp silhouettes. The necessity of foreground mask stems not only from the foreground silhouette that we aim to complete with our estimates but also the image-level features that we prove to be essential in Section 4.

3.1. Continuous Correspondences

Continuous correspondences have significant advantages over their discretized counterparts. First, a continuous representation provides no seams between body parts. Second, it is conceptually simpler as there is no need to explicitly encode and later predict the body part. The benefits of utilizing a continuous representation for surface correspondences have already been discussed in a few prior works [27, 41, 47]. We follow a similar direction with [27] and design a continuous UV map that is then warped to body models in different poses, providing ground truth correspondences for our approach. The color scheme for correspondences used in the paper is unique color-wise: we chose different colors for every vertex of the parametric body model. Given the colored 3D body model we transform its surface into a 4K UV map, which is then utilized during rendering over a body model in a determined pose and from a desired view point. In that way, we obtain ground truth for the synthetic data used for training.

3.2. Surface Embedding Transformers

A classic architectural for a network predicting Dense Correspondences from an RGB image is an Encoder-Decoder (*e.g.*, U-net). While simple convolutional backbones in the encoder can usually provide sufficient results, we observed that the right choice of the encoder architecture may significantly boost the whole pipeline performance. Compared to convolutions, transformer-based architectures do not suffer from limited receptive fields, resulting in more expressivity. Moreover, transformers avoid explicit downsampling of the input image embedding leading to more ac-

curate and refined final representations.

As illustrated in Fig. 2, we build upon the work of Ranftl *et al*. [34] for monocular depth estimation and introduce a simple yet novel transformer-based architecture designed explicitly for the task of predicting dense surface correspondences of humans. We transform the RGB image and its CSE estimate into tokens by extracting non-overlapping patches and then linearly projecting resulting flattened representations. Similarly to text-transformers, we add a specific token to the set, that aggregates the global knowledge about an image. The image and CSE embeddings are supplemented with positional embeddings and fed to separate vision transformer backbones with separate weights to retrieve dense features for each input. Later we refer to these blocks as appearance and correspondence transformers (Fig. 2). Positional encoding in Visual Transformers is essential to capture sequence ordering of input tokens instead of transforming the image into "bag-of-patches" omitting its relative order and global spatial consistency.

The transformer outputs are fused forming an intermediate representation which is first resampled and then projected via residual convolutional units. It is then fed into the convolutional decoder where the representation is upsampled to generate a fine-grained correspondence prediction. Finally, the network outputs per-pixel RGB values that encode correspondences according to our coloring scheme discussed in the previous section.

3.3. Supervision in the Image Space

For each pixel p in the foreground image, we predict 3-channeled (RGB) color $p' \in \mathbb{Z}^3$ which represents the correspondence (the colors in such a representation are unique which makes subsequent warping easy). Thus, we treat the whole problem as a multi-task classification problem where each task (predictions for the R, G and B channels) is trained with the same set of losses:

Per-pixel classification loss L_{cls}. For every color channel, we predict the per-pixel classification label $l \in [0, 255]$.

BodyMap provides raw, unnormalized per-pixel scores for each of the classes in each of the three color channels and L_{cls} measures the cross-entropy between the prediction and the ground truth. Since we noticed that it is quite challenging to predict correspondences of realistic gestures, we further define a loss weight for each pixel based on the body part segmentation. We set a higher weight for *hands* and *head* while a lower weight for the rest of the body to encourage fine-grained correspondence estimation.

Silhouette loss L_{sil}. We penalize the model for non-accurate silhouette predictions by calculating the IoU between the predicted and ground truth foreground masks.

3.4. Supervision in the 3D Geometry

Geodesic loss L_{geo}. While per-pixel cross-entropy classification losses supervise our predictions in 2D image space, we expand our supervision scheme to 3D by utilising geodesic distances on the surface of the body model. Geodesic losses have been instrumental in the literature for enforcing supervisions in the 3D space. We design a loss that pushes features between non-matching pixels apart, depending on the geodesic distance. We calculate geodesic distances between vertices predicted with correspondences for every foreground pixel and their ground truth counterparts. Theoretically, such a supervision eliminates imperfection of the proposed coloring scheme for correspondences: distant vertices may have resembling colors (green head and shoulders, blue arm and right thigh). Thus, the geodesic loss provides extra knowledge about the 3D geometry comparing distances between predicted vertices vs. ground truth ones.

$$L_{geo}(I_{pred}, I_{gt}) = \sum_{x} \mathcal{D}_g(V^{(I_{pred}(x))}, V^{(I_{gt}(x))}), \quad (1)$$

where $V^{(I(x))}$ denotes the vertex corresponding to pixel location x in the image I and $\mathcal{D}_g(\cdot,\cdot)$ denotes the geodesic distance between two 3D points on the body surface.

3.5. Regularization and Final Loss

Consistency loss L_{con}. We further add a regularization term to enforce the smoothness of the predictions in the neighboring regions. Specifically, we constrain the predictions from neighboring pixels to be geodesically close to each other, *i.e*.,

$$L_{con}(I_{pred}) = \sum_{p \in I} \log\left(1 + \exp\left(\frac{\mathcal{D}_g(p_r, p)}{\sigma_{geo}} - \frac{\|p_r - p\|_1}{\sigma_{col}}\right)\right), \quad (2)$$

where p_r is a randomly chosen pixel within the foreground silhouette, $\mathcal{D}_g(p_1, p_2)$ the geodesic distance between vertices corresponding to pixels p_1 and p_2, σ_{geo} is the normalizing constant for geodesic distances (maximum possible distance between points in the body model), σ_{col} the normalizing constant for RGB colors, respectively. On each iteration we calculate this loss 100 times for different randomly chosen pixels later averaging the resulting values.

Final loss. The final loss is a weighted sum of all the terms:

$$L_{train} = \lambda_{cls}L_{cls} + \lambda_{sil}L_{sil} + \lambda_{geo}L_{geo} + \lambda_{con}L_{con}, \quad (3)$$

where a loss weight λ corresponds to each loss term in order to balance them.

3.6. Training Details

The BodyMap network is first trained on synthetic data to learn surface correspondences for every foreground pixel. Given an RGB image, we obtain foreground mask and CSE estimates which serve as an initialization for the correspondences. However, if we were to test this model di-

rectly on DensePose-COCO that comprises multiple people, heavy occlusions and low-resolution images, then the results would be unsatisfactory. The annotations provided in this dataset are sparse and noisy with ~100 pixel-SMPL vertex correspondences for each person in the image. To bridge this domain gap, we fine-tune our model on the training set of DensePose-COCO but with a key change that ended up having a significant impact. Given an image from this 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. In that way we can fine-tune our model on real-data with denser supervision and utilize losses in both 2D and 3D spaces.

To further boost the generalization capabilities of BodyMap, we introduce several augmentations. First, we do specific crops in order to get upper-body samples. Second, we generate frames with multiple synthetic people in it to simulate crowds and diminish the gap between synthetic and real data. Third, we do a standard set of augmentations, like rotations, slight hue and saturation changes.

4. Experiments

Datasets. Our proposed approach is trained mainly on synthetic data with the exception of the experiments reported on DensePose-COCO where we utilize the provided training set. We opted for the RenderPeople dataset [10] which has been used extensively in the literature [1, 5, 7, 17, 18, 21, 30, 32, 37, 41, 50] for various human reconstruction and generation tasks. We used 1000 scans which are watertight meshes wearing a variety of garments and in some cases holding objects such as mugs or bags. Since the scans are static we wanted to introduce additional pose variations and as a result we performed non-rigid registration, rigged them for animation and used a motion collection that provides 3D human animations from which we collect a set of $2,446$ 3D animation sequences covering wide action categories of daily activities and sports. With a large set of scans and motions we randomly sample scan-motion pairs and render them with Blender Cycles from different views with uniform lighting to obtain the RGB sequences as well as the corresponding UV ground-truth. We perform a 90/10 train/test split based on identities. This large-scale dataset represents an effort to cover a wide range of motions, poses, and body shapes, captured from multiple views with people that can move towards the camera our even outside the frame and enables us to train our BodyMap network without making any explicit assumptions.

At test-time BodyMap is evaluated quantitatively and qualitatively on both synthetic as well as real data ranging from COCO, fashion images (DeepFashion [24]) as well as a few 3dMD scans of real people captured with a full-body scanner. We used this solely for testing since we wanted to evaluate to what extent our approach can handle the domain gap between synthetic and real data. These real scans do not include any objects but are noisier with complex facial expressions and enable us to stress-test whether our approach can handle such complex inputs.

Method	AP	AP_{50}	AP_{75}	AR	AR_{50}	AR_{75}
AMA-net [14]	64.1	91.4	72.9	71.6	94.7	79.8
DensePose [2]	66.4	92.9	77.9	71.9	95.5	82.6
DensePose-DeepLab [2]	51.8	83.7	56.3	61.1	88.9	66.4
SimPose-Rendppl. [49]	57.3	88.4	67.3	66.4	95.1	77.8
SimPose-SMPL [49]	56.2	87.9	65.3	65.2	95.1	75.2
CSE [27]	67.0	93.8	78.6	72.8	96.4	83.7
CSE-DeepLab [27]	68.0	94.1	80.0	74.3	97.1	85.5
BodyMap RGB-only	71.0	94.3	83.3	75.2	94.3	86.1
BodyMap	**75.2**	**95.8**	**89.7**	**79.8**	**97.3**	**89.7**

Table 1. **Average Precision (AP) and Recall (AR) on DensePose-COCO**. AP and AR are calculated at a number of GPS thresholds ranging from 0.5 to 0.95. Our methods surpasses the state-of-the art methods DensePose [26] and CSE [27]

Baselines and Metrics. We consider two different ways of measuring the quality of correspondences evaluating both in 2D image space by comparing RGB values of the corresponding pixels and in 3D space by measuring geodesic distances between predicted and ground truth vertices.

First, we calculate the accuracy of predictions in the 2D image space by calculating the percentage of pixels colored correctly within a specified threshold. Second, following the evaluation scheme of DensePose that is used widely in the literature [26, 27, 49] we measure average precision and recall over GPS scores. Geodesic point similarity (GPS) score is a correspondence matching score:

$$GPS_j = \frac{1}{|P_j|} \sum\nolimits_{p \in P_j} \exp \frac{-g(i_p, \hat{i_p})^2}{2\kappa^2}, \tag{4}$$

where P_j is the set of points annotated on person instance j, i_p is the vertex estimated by a model at point p, $\hat{i_p}$ is the ground-truth vertex p, and κ is a normalizing parameter. We calculate Average Precision (AP) and Average Recall (AR) metrics considering a vertex prediction as correct if the GPS score is higher than a threshold. Following the evaluation scheme established by prior work [26, 27], GPS thresholds are ranging from 0.5 to 0.95.

Additionally to metrics in 2D and 3D spaces, we evaluate the consistency over time of our predictions in order to estimate quantitatively the amount of flickering. We calculate percentage of positive correspondence matches between frames of the same video for visible vertices.

Using the aforementioned metrics we compare BodyMap quantitatively to the previous works: DensePose [26], CSE [27], SimPose [49] as well as several other baselines. However, calculating AP and AR metrics for HumanGPS is not possible due to the fact, that HumanGPS predicts only embeddings for every foreground pixel, that provide no information on UV coordinates or correspond-

	Synthetic Dataset			DensePose-COCO		
Error Window (px)	5	10	20	5	10	20
DensePose [2]	25.93	46.10	69.91	49.23	55.75	59.71
CSE [27]	44.52	67.51	75.13	58.10	60.34	64.14
BodyMap RGB-only	66.15	73.81	79.80	61.18	65.32	68.52
BodyMap	**71.12**	**79.73**	**96.92**	**65.34**	**68.22**	**73.88**

Table 2. **Accuracy in 2D space**. We show the percentage of pixels correctly matched within the established error window on synthetic dataset and DensePose-COCO. Our methods surpasses the state-of-the art methods DensePose and CSE.

Method	AP	AP_{50}	AP_{75}	AP_M	AP_L	AR	AR_{50}	AR_{75}	AR_M	AR_L
DP-DL [2]	55.3	85.6	60.1	48.3	58.2	66.8	90.1	68.2	50.1	66.1
CSE-DL [27]	72.8	95.7	84.2	65.7	73.1	78.2	97.3	87.5	67.2	78.0
BodyMap RGB-only	75.3	96.1	89.2	69.3	75.2	81.2	97.4	89.2	70.3	80.2
BodyMap	**79.5**	**97.8**	**90.5**	**72.3**	**79.4**	**85.3**	**98.1**	**92.5**	**73.4**	**84.5**

Table 3. **Average Precision (AP) and Recall (AR) over GPS scores in 3D space**. We calculate AP and AR at GPS thresholds ranging from 0.5 to 0.95 on our synthetic dataset. Our method clearly outperforms DensePose-DeepLab and CSE-DeepLab.

ing to pixels SMPL vertices. In their approach warping and appearance swapping is done by nearest neighbors search over embeddings without going to 3D body model space. Thus, we compare with HumanGPS only qualitatively and using temporal consistency metrics.

4.1. Quantitative results

In Tables 2 and 3 we provide a quantitative comparison between BodyMap, DensePose, CSE and HumanGPS on the test set of the aforementioned synthetic dataset. In Tables 1 and 2 we do the same on DensePose-COCO dataset but also provide additional comparisons with prior work. Opposite to the synthetic dataset, for which we have ground truth correspondences for every foreground pixel, for DensePose-COCO we rely only on the available annotated points to calculate the metrics. BodyMap shows a substantial improvement over prior work across all the metrics for both our synthetic and DensePose-COCO datasets. The reasons behind this improvement stem from: i) specifically designed architecture that separates and takes the best out of RGB and CSE inputs; ii) training on well-designed and rendered synthetic data and later fine-tuning on specifically adapted DensePose-COCO with the additional tricks discussed in Sec. 3.6, which helps to bridge the synth2real domain gap; iii) the proposed training scheme that includes supervision both in the image space with per-pixel classification losses as well as the 3D space with geodesic losses.

Temporal Consistency. In Table 4 we test how temporally consistent the dense correspondences of different methods are. We were motivated to run this experiment by observing how jittery DensePose predictions can be on videos. In terms of metrics, we estimate the percentage of positive correspondence matches between the current frame and a frame in the future with an interval in $1, 12, 120$ on a synthetic sequence consisting of $18,000$ frames. BodyMap outperforms prior work by a large margin and establishes accurate correspondences even if the time interval between the 2 frames is substantial. In supplementary we provide demo video showing consistency over time of our results.

Frame Interval	1	12	120
DensePose [26]	77.79	40.86	16.32
CSE [27]	85.55	55.85	18.93
HumanGPS [41]	86.42	65.19	36.17
BodyMap	**88.70**	**74.01**	**46.11**

Table 4. **Temporal consistency**. We estimate the percentage of positive correspondence matches between frames with a different interval on a synthetic sequence of $18,000$ frames.

		Synthetic Dataset			DensePose-COCO		
Error Window (px)		5	10	20	5	10	20
BodyMap (ours)	ResNet	45.12	60.82	79.12	30.41	55.67	61.15
	EffNet	51.22	65.77	82.19	40.25	61.17	70.22
	U-Net	68.42	75.13	94.19	60.82	65.74	70.12
	ViT	**71.12**	**79.73**	**96.92**	**65.34**	**68.22**	**73.88**

Table 5. **Different network backbones**: Ablation study

4.2. Ablation studies

Different Architectures. We experiment with different backbones starting from a simple UNet with skip-connections and then progressing to more complex transformer-based solutions. In Table 5 we provide a comparison in terms of accuracy in the 2D space across all the architectures. An interesting finding is that a simple UNet architecture can get satisfactory results when trained with all the proposed supervisions described in Sec. 3.3 and 3.4. However, our proposed Vision Transformer (ViT) is capable of learning more accurate correspondences in challenging areas like neck, armpits, fingers and hair, making the predicted silhouette clear-cut and crisp. These differences are mostly visible in hard DensePose-COCO examples (with multiple people and occlusions), while on simple synthetic data cases UNet is performing nearly as good as ViT.

We further experiment with the network design, feeding only RGB inputs to the net and omitting the Correspondence Transformer. While RGB-only method performs comparatively worse, it still outperforms existing approaches, e.g.DensePose, CSE or HumanGPS (Tables 1, 2).

Different Losses. We also investigate the impact of the proposed losses in Table 6. While the best score is achieved with the whole set of proposed losses, per-pixel cross-entropy classification losses for color channels contribute the most. Silhouette loss makes the edges of final prediction more accurate and extra supervision in hands and head regions improves correspondences in these ar-

Losses \ Error Window	Synthetic Dataset 5	10	20	DensePose-COCO 5	10	20
L_{cls}	65.16	71.52	85.12	49.37	55.81	59.14
$L_{cls} + L_{sil}$	69.18	75.32	92.31	54.12	60.22	62.17
$L_{cls} + L_{sil} + L_{geo}$	70.23	78.71	95.80	61.83	64.32	68.17
$L_{cls} + L_{sil} + L_{geo} + L_{con}$	**71.12**	**79.73**	**96.92**	**65.34**	**68.22**	**73.88**

Table 6. **Ablation study on the impact of different losses** in the accuracy in 2D space (the percentage of pixels colored correctly within the established error window)

eas. Geodesic losses give tangible improvement only on DensePose-COCO, indicating that simple synthetic one-person-per-frame cases can be handled sufficiently with only image-space supervision. Thus, the model can learn fine-grained body model details even with the first two losses (both supervising in the 2D image space). However, more complicated cases including several people in one frame and significant occlusions require extra supervision in 3D space to obtain satisfactory results.

Different Fine-tuning Schemes: 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 described in Sec. 3.6 (dense fine-tuning). We observed that densifying ground-truth on the fly results in superior performance compared to either no fine-tuning or relying solely on sparse annotations. More results are shown in the supplementary.

Model Complexity: Inference of our model takes $\sim$0.1 seconds on a single Tesla V100-SXM2 for a 1024×1024 image. The model has $\sim 600M$ trainable parameters.

4.3. Qualitative Results

In Figures 1, 3 we show correspondences for a few images from DeepFashion [24] which has lower quality inputs, RenderPeople, DensePose-COCO and finally images from real-people scans captured with a 3dMD system. The silhouettes of the inputs are well covered with our estimates, the hands and fingers are accurately captured and the face is well aligned. Loose clothes, even complicated cases like a long robe in the are well-handled.

In Figure 3 we show qualitative comparisons between BodyMap and competitors: HumanGPS, DensePose and CSE on several examples from DensePose-COCO, RenderPeople and our synthetic dataset. While DensePose and CSE predictions are smooth and consistent, they do not cover the whole silhouette, totally omitting hair and loose clothes. HumanGPS handles silhouettes better, but still struggles with accurate correspondences in challenging scenarios with occlusions or produces blurry patches for back views (Line 4 in Figure 3). We also show in the supplementary, that HumanGPS predictions are not always temporally consistent, jumbling correspondences for right and left arms and legs while the person is rotating.

Figure 3. Qualitative comparison with competitors on DensePose-COCO, our synthetic dataset and RenderPeople.

4.4. Applications & Discussions

Neural Re-rendering. One possible application is re-rendering people from the source frame from another view point and/or in another pose. We introduce a model for neural re-rendering which aims at learning a function that given the complete texture map and the estimated BodyMap correspondences generates a photorealistic render in the image space. Before neural re-rendering it is needed to obtain a complete texture map, which we do in the following way. Given a source and a target view of a person we utilize the predicted BodyMap estimates and defined a warping function W that outputs high-quality neural re-renders at the target viewpoint. We represent W with a neural network that i) warps the input source RGB image to the UV space to obtain a partial texture, ii) learns to complete it to obtain a full texture estimate, and iii) warps it back to the image space using BodyMap and then uses a neural renderer that generates the final output render. Given a source and a target image our neural renderer generates overall higher-fidelity details than prior work as seen in Fig. 4(left), also in the face and hand regions, and does not suffer from color bleeding.

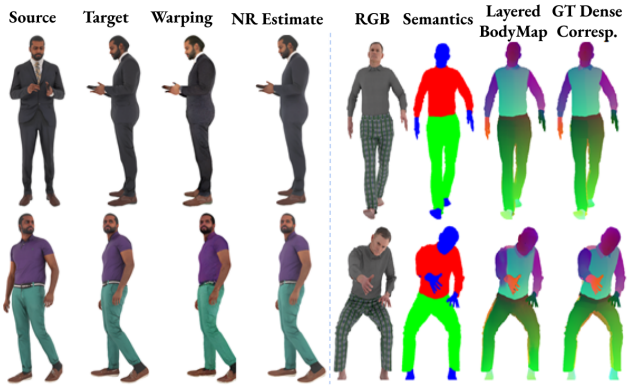

Figure 4. **Applications**. Neural re-rendering (left) and predicting layered correspondences for clothed humans (right).

In the supplementary material we describe in detail this application along with an architecture figure and also present an application to cloth swapping and motion retargeting.

Layered Dense Correspondences. In all prior work dense human correspondences are estimated only for the body surface. That is because a body template (*e.g.*, SMPL [25]) with UV information is available and sparse annotations for COCO exist to accomplish this task. However, when dealing with clothed humans (and especially in loose garments) estimating body correspondences in a single-layer as DensePose or our proposed BodyMap does, can be a challenging task. However, fine-grained clothes details like wrinkles and textile folds can be represented better with decoupling body and clothes correspondences to separate representations. In a first attempt to do so we present an application with a slight BodyMap variation predicting three separate representations for the unclothed body, upper clothes and lower clothes. We named this variation *Layered-BodyMap*. The architecture remains the same besides the three output heads instead of one. To generate ground truth data for such a task, we run cloth simulation for the two garments given various walking and hand-movement motions resulting in 12 sequences of people wearing 3D clothes from our collection. Opposite to BodyMap, where we use RGB together with CSE initialization as input, here we do not have any initialization for the clothes correspondences, and as a result we feed this network with RGB-only inputs, but condition the estimates on semantic segmentation masks. The predicted *layered* correspondences are accurate and cover the whole silhouette (Fig. 4 (right)) which is a promising result that we believe future work will improve upon as more 3D garment libraries become available [3, 38, 42, 48].

Limitations. Our approach relies on foreground human segmentation which makes it susceptible to the performance of that step. We tested different segmentation and matting approaches, [9, 19, 20, 40], and opted for MMSegmentation due to its ability to preserve fine details like fingers and hairstyles. BodyMap was trained on high-resolution mostly full-body images which makes it susceptible to low-res inputs or when only the bottom of the body is visible. This is partly solved by imposing heavy augmentations but occlusions from objects remain a challenge. Moreover, due to the nature of the task most of the training data is synthetic which makes inference on real data challenging. We address that with the fine-tuning scheme described in Sec. 4.2. We show some failure cases in Figure 5, which mostly happen due to bad lightning or severe occlusions and provide additional examples in the supplementary.

Figure 5. **Failure cases**. Most failure cases happen for low resolution images with occlusions and/or bad lightning.

5. Conclusion

We present a novel framework for establishing accurate dense correspondences between an image and the surface of a 3D clothed human. Our key contribution, BodyMap, is a transformer-based architecture that when trained with 2D and 3D supervisions significantly outperforms prior work. BodyMap addresses key limitations of current approaches, such as inability to handle loose clothes, body and garments being represented as a single surface, non-continuity of the correspondences for different body parts. We outperformed prior work by a large margin on synthetic as well as DensePose-COCO datasets and investigated the impact of each of our design selections. Finally, we provided examples of applications such as re-rendering in different poses and extend BodyMap to clothed humans with multiple layers of geometry with promising results.

Acknowledgments. We thank Tuur Stuyck for his help to run cloth simulation and Vasil Khalidov for his help with running the code of the Continuous Surface Embeddings paper.

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. 1, 5

[2] Rıza Alp Güler, Natalia Neverova, and Iasonas Kokkinos. Densepose: Dense human pose estimation in the wild. In *CVPR*, 2018. 2, 5, 6

[3] Hugo Bertiche, Meysam Madadi, and Sergio Escalera. Cloth3d: Clothed 3d humans. In *ECCV*. Springer, 2020. 8

[4] Bharat Lal Bhatnagar, Cristian Sminchisescu, Christian Theobalt, and Gerard Pons-Moll. Loopreg: Self-supervised learning of implicit surface correspondences, pose and shape for 3D human mesh registration. In *NeurIPS*, 2020. 2

[5] Bharat Lal Bhatnagar, Garvita Tiwari, Christian Theobalt, and Gerard Pons-Moll. Multi-garment net: Learning to dress 3D people from images. In *ICCV*, 2019. 5

[6] 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 *ECCV*, 2016. 1

[7] Bindita Chaudhuri, Nikolaos Sarafianos, Linda Shapiro, and Tony Tung. Semi-supervised synthesis of high-resolution editable textures for 3d humans. In *CVPR*, 2021. 5

[8] Vasileios Choutas, Georgios Pavlakos, Timo Bolkart, Dimitrios Tzionas, and Michael J. Black. Monocular expressive body regression through body-driven attention. In *ECCV*, 2020. 1

[9] MMSegmentation Contributors. MMSegmentation: Openmmlab semantic segmentation toolbox and benchmark. 8

[10] RenderPeople Dataset. *http://renderpeople.com/*. 5

[11] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 2, 3

[12] Alaaeldin El-Nouby, Natalia Neverova, Ivan Laptev, and Hervé Jégou. Training vision transformers for image retrieval. *arXiv preprint arXiv:2102.05644*, 2021. 3

[13] Yao Feng, Vasileios Choutas, Timo Bolkart, Dimitrios Tzionas, and Michael J Black. Collaborative regression of expressive bodies using moderation. In *3DV*, 2021. 1

[14] Yuyu Guo, Lianli Gao, Jingkuan Song, Peng Wang, Wuyuan Xie, and Heng Tao Shen. Adaptive multi-path aggregation for human densepose estimation in the wild. In *Proceedings of the 27th ACM International Conference on Multimedia*, pages 356–364, 2019. 5

[15] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, 2017. 3

[16] Sen He, Wentong Liao, Hamed R Tavakoli, Michael Yang, Bodo Rosenhahn, and Nicolas Pugeault. Image captioning through image transformer. In *Proceedings of the Asian Conference on Computer Vision*, 2020. 3

[17] Tong He, Yuanlu Xu, Shunsuke Saito, Stefano Soatto, and Tony Tung. ARCH++: Animation-ready clothed human reconstruction revisited. In *ICCV*, 2021. 5

[18] Zeng Huang, Yuanlu Xu, Christoph Lassner, Hao Li, and Tony Tung. ARCH: Animatable reconstruction of clothed humans. In *CVPR*, 2020. 5

[19] Zhanghan Ke, Kaican Li, Yurou Zhou, Qiuhua Wu, Xiangyu Mao, Qiong Yan, and Rynson W.H. Lau. Is a green screen really necessary for real-time portrait matting? *ArXiv*, abs/2011.11961, 2020. 8

[20] Alexander Kirillov, Yuxin Wu, Kaiming He, and Ross Girshick. Pointrend: Image segmentation as rendering. In *CVPR*, 2020. 8

[21] Verica Lazova, Eldar Insafutdinov, and Gerard Pons-Moll. 360-degree textures of people in clothing from a single image. In *3DV*, 2019. 1, 5

[22] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2117–2125, 2017. 3

[23] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *ECCV*, 2014. 2

[24] Ziwei Liu, Ping Luo, Shi Qiu, Xiaogang Wang, and Xiaoou Tang. Deepfashion: Powering robust clothes recognition and retrieval with rich annotations. In *CVPR*, 2016. 5, 7

[25] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J Black. SMPL: A skinned multi-person linear model. *TOG*, 2015. 8

[26] Natalia Neverova, Riza Alp Guler, and Iasonas Kokkinos. Dense pose transfer. In *ECCV*, 2018. 1, 2, 3, 5, 6

[27] Natalia Neverova, David Novotny, Vasil Khalidov, Marc Szafraniec, Patrick Labatut, and Andrea Vedaldi. Continuous surface embeddings. *arXiv preprint arXiv:2011.12438*, 2020. 2, 3, 5, 6

[28] Natalia Neverova, David Novotny, and Andrea Vedaldi. Correlated uncertainty for learning dense correspondences from noisy labels. In *NeurIPS*, 2019. 2

[29] Natalia Neverova, Artsiom Sanakoyeu, Patrick Labatut, David Novotny, and Andrea Vedaldi. Discovering relationships between object categories via universal canonical maps. In *CVPR*, 2021. 2

[30] Pablo Palafox, Nikolaos Sarafianos, Tony Tung, and Angela Dai. SPAMs: Structured implicit parametric models. In *CVPR*, 2022. 5

[31] Niki Parmar, Ashish Vaswani, Jakob Uszkoreit, Lukasz Kaiser, Noam Shazeer, Alexander Ku, and Dustin Tran. Image transformer. In *International Conference on Machine Learning*, pages 4055–4064. PMLR, 2018. 3

[32] Priyanka Patel, Chun-Hao P. Huang, Joachim Tesch, David T. Hoffmann, Shashank Tripathi, and Michael J. Black. AGORA: Avatars in geography optimized for regression analysis. In *CVPR*, 2021. 5

[33] Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed AA Osman, Dimitrios Tzionas, and Michael J Black. Expressive body capture: 3D hands, face, and body from a single image. In *CVPR*, 2019. 1

[34] René Ranftl, Alexey Bochkovskiy, and Vladlen Koltun. Vision transformers for dense prediction. In *ICCV*, 2021. 3, 4

[35] Yu Rong, Takaaki Shiratori, and Hanbyul Joo. Frankmocap: Fast monocular 3D hand and body motion capture by regression and integration. *arXiv preprint arXiv:2008.08324*, 2020. 1

[36] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *MICCAI*, 2015. 3

[37] Shunsuke Saito, Tomas Simon, Jason Saragih, and Hanbyul Joo. PIFuHD: Multi-level pixel-aligned implicit function for high-resolution 3D human digitization. In *CVPR*, 2020. 5

[38] Igor Santesteban, Nils Thuerey, Miguel A Otaduy, and Dan Casas. Self-supervised collision handling via generative 3d garment models for virtual try-on. In *CVPR*, 2021. 8

[39] Kripasindhu Sarkar, Dushyant Mehta, Weipeng Xu, Vladislav Golyanik, and Christian Theobalt. Neural re-rendering of humans from a single image. In *ECCV*, 2020. 2

[40] Soumyadip Sengupta, Vivek Jayaram, Brian Curless, Steve Seitz, and Ira Kemelmacher-Shlizerman. Background matting: The world is your green screen. In *Computer Vision and Pattern Regognition (CVPR)*, 2020. 8

[41] Feitong Tan, Danhang Tang, Mingsong Dou, Kaiwen Guo, Rohit Pandey, Cem Keskin, Ruofei Du, Deqing Sun, Sofien Bouaziz, Sean Fanello, et al. HumanGPS: Geodesic preserving feature for dense human correspondences. In *CVPR*, 2021. 2, 3, 5, 6

[42] Garvita Tiwari, Bharat Lal Bhatnagar, Tony Tung, and Gerard Pons-Moll. Sizer: A dataset and model for parsing 3d clothing and learning size sensitive 3d clothing. In *ECCV*, 2020. 8

[43] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008, 2017. 3

[44] Yuanlu Xu, Song-Chun Zhu, and Tony Tung. Denserac: Joint 3D pose and shape estimation by dense render-and-compare. In *ICCV*, 2019. 1

[45] Haonan Yan, Jiaqi Chen, Xujie Zhang, Shengkai Zhang, Nianhong Jiao, Xiaodan Liang, and Tianxiang Zheng. Ultrapose: Synthesizing dense pose with 1 billion points by human-body decoupling 3d model. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10891–10900, 2021. 2

[46] Jae Shin Yoon, Kihwan Kim, Jan Kautz, and Hyun Soo Park. Neural 3d clothes retargeting from a single image. *arXiv preprint arXiv:2102.00062*, 2021. 1

[47] Wang Zeng, Wanli Ouyang, Ping Luo, Wentao Liu, and Xiaogang Wang. 3D human mesh regression with dense correspondence. In *CVPR*, 2020. 2, 3

[48] Heming Zhu, Yu Cao, Hang Jin, Weikai Chen, Dong Du, Zhangye Wang, Shuguang Cui, and Xiaoguang Han. Deep fashion3d: A dataset and benchmark for 3d garment reconstruction from single images. In *ECCV*, 2020. 8

[49] Tyler Zhu, Per Karlsson, and Christoph Bregler. Simpose: Effectively learning densepose and surface normals of people from simulated data. In *ECCV*, 2020. 2, 3, 5

[50] Pierre Zins, Yuanlu Xu, Edmond Boyer, Stefanie Wuhrer, and Tony Tung. Data-driven 3D reconstruction of dressed humans from sparse views. In *3DV*, 2021. 5