ConVol-E: Continuous Volumetric Embeddings for Human-Centric Dense Correspondence Estimation

Amogh Tiwari Pranav Manu Nakul Rathore Astitva Srivastava Avinash Sharma

Center for Visual Information Technology (CVIT), IIIT Hyderabad

{amogh.tiwari, pranav.m, nakul.rathore, astitva.srivastava}@research.iiit.ac.in, asharma@iiit.ac.in

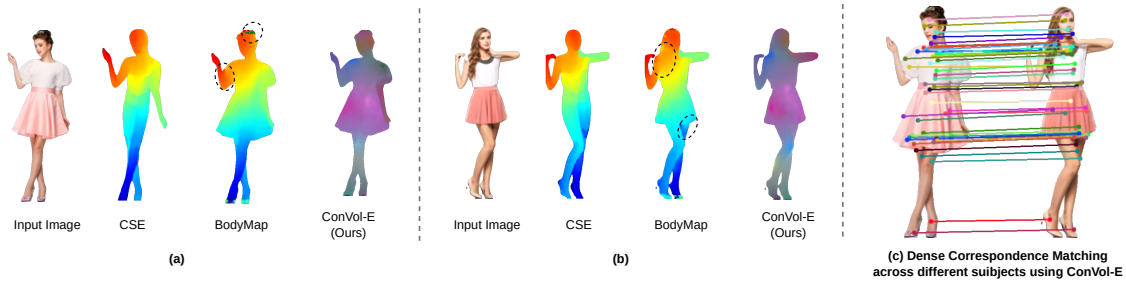

Figure 1. Proposed ConVol-E representation can handle loose clothing details unlike CSE [14] and outperforms BodyMap [8] by avoiding repetitions in embedding space, hence can be used to establish dense correspondence across different subjects.

Abstract

We present Continuous Volumetric Embeddings ***(ConVol-E)****, a novel robust representation for dense correspondence-matching across RGB images of different human subjects in arbitrary poses and appearances under non-rigid deformation scenarios. Unlike existing representations [8, 14], ConVol-E captures the deviation from the underlying parametric body model by choosing suitable anchor/key points on the underlying parametric body surface and then representing any point in the volume based on its euclidean relationship with the anchor points. It allows us to represent any arbitrary point around the parametric body (clothing details, hair, etc.) by an embedding vector. Subsequently, given a monocular RGB image of a person, we learn to predict per-pixel ConVol-E embedding, which carries a similar meaning across different subjects and is invariant to pose and appearance, thereby acting as a descriptor to establish robust dense correspondences across different images of humans. We empirically evaluate our proposed embedding using a novel metric and show superior performance compared to the state-of-the-art for the task of in-the-wild dense correspondence matching across different subjects, camera views, and appearance.*

1. Introduction

Dense pixel-level understanding and labelling of humans in images is a well-attempted, yet challenging research problem in computer vision. Traditionally, it helps estimate body pose & shape, part semantics, dense correspondence/flow with key applications, including instance level segmentation, human tracking, gait analysis, 3D/4D human body reconstruction, virtual try-on, etc. In particular, the dense correspondence estimation can immensely benefit by associating each pixel with body pose/shape/appearance agnostic characterization. The key idea for establishing dense correspondences is to identify a per-pixel feature-based representation that can explain the relationship across different images of humans. The representation should be agnostic to appearance, i.e., it should carry the same meaning across images of different individuals. The formulation of such a representation is non-trivial owing to challenges such as large space of complex pose articulations, significant variations in body shape & size and large camera viewpoint variations. Moreover, the arbitrary and non-rigid nature of the garments causes deformations in the topology, which are extremely hard to model just from an image. The underlying representation should also understand the relationship between the garments and the body, especially under the

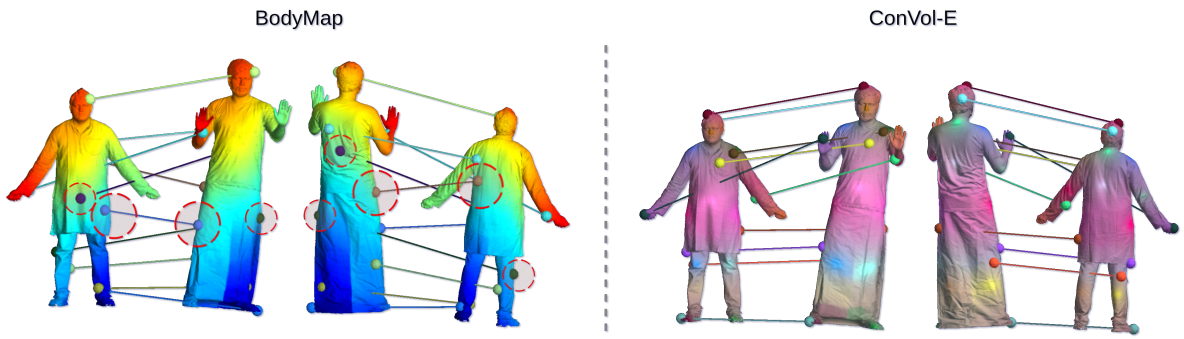

Figure 2. Comparing correspondences on 3D meshes when encoded with BodyMap [8] (left) and ConVol-E (right). Multiple false matching can be seen in the representation of BodyMap whereas, ConVol-E provides robust matching even in presence of loose clothing scenario.

loose clothing setup, which is a highly ill-posed problem.

Continuous Surface Embeddings (CSE) [14] is one such pixel-level representation that leverages the parametric human body model by estimating common embedding space between vertices of the SMPL [12] mesh and the pixels occupied by the humans in RGB image. However, SMPL doesn't capture high-frequency details such as clothing, hair, etc (as show in autoreffig:teaser(a) & (b)). Recently, BodyMap [8] proposed to extend this representation to include these high-frequency details by assigning a three-dimensional embedding to the vertices of a human scan by extrapolating the CSE embedding, represented as simple RGB values, in the UV space based on the geodesic distance. This allows them to establish a relationship between different human scans by estimating similar extrapolated pixel-level embeddings. Subsequently, a network is trained to predict these dense three-dimensional embeddings in the form of color-coded RGB maps from the rendered images of the ground truth scans. However, such extrapolation of the RGB colors in UV space can not prevent distant vertices from having similar colors (as stated by the authors in their original paper [8]), thereby resulting in false matching across different regions of the human body, as shown in Figure 2. Additionally, it doesn't guarantee to produce consistent pixel-wise embedding for loose clothing scenarios as the effect of the geodesic distance will diminish in the far-apart regions of the UV space.

In this paper, we propose *Continuous Volumetric Embeddings* (ConVol-E), a novel representation for establishing dense correspondence across humans in arbitrary poses and appearances. Our representation can handle any arbitrary point in the volume occupied by the human subject, i.e., each point in the 3D space is associated with a continuous value representing its relationship with the underlying parametric body model (SMPL to be specific). To ensure uniqueness and avoiding repetitions of the embeddings, we carefully designate *anchor nodes* on the SMPL surface. The embedding values for the *anchor nodes* are assigned such that the extrapolated embeddings vary significantly across different regions of the body and the volume, thereby ensuring a minimal chance of repeated values for far-away points in different directions (refer to supplementary for a detailed analysis). The anchor nodes are designated for a gender-neutral SMPL model, and their embedding values are extrapolated to all the vertices of a 3D human scan by registering the shape and pose parameters of SMPL with the scan. It is important to note that unlike BodyMap [8], our approach of volumetric extrapolation inherently addresses the challenge of far-away surface deformations (typically caused by loose clothing). Subsequently, we propose to learn dense pixel-wise ConVol-E values using a U-Net [16] encoder-decoder network given an input image of a human with a corresponding ground truth scan, which can later be inferred on in-the-wild images with high accuracy. We perform a thorough evaluation of ConVol-E and compare with existing state-of-the-art methods. Finally, the predicted embeddings are used for dense correspondence matching of ConVol-E across different viewpoints and subjects (as shown in Figure 1(c)) and demonstrate applications like segmentation label transfer and appearance transfer.

2. Related Works

Estimating dense correspondence embeddings across different images of humans is an active area of research, with tons of potentially useful human-centric applications. The problem is well-attempted for general objects, and many solutions exist [4, 15, 21]. However, the complexity and difficulty increase drastically for humans, due to articulation, non-rigid deformation and clothing. Initial attempts were made to first solve the problem in a sparse way using pose estimation [2, 3, 9, 19], which mostly involves fitting either a human joint-skeleton or a parametric model such as SMPL [12]. While such solutions are widely used in the case of 3D human body reconstruction [11, 20, 23] and garment reconstruction from images [1, 17, 24], they can not

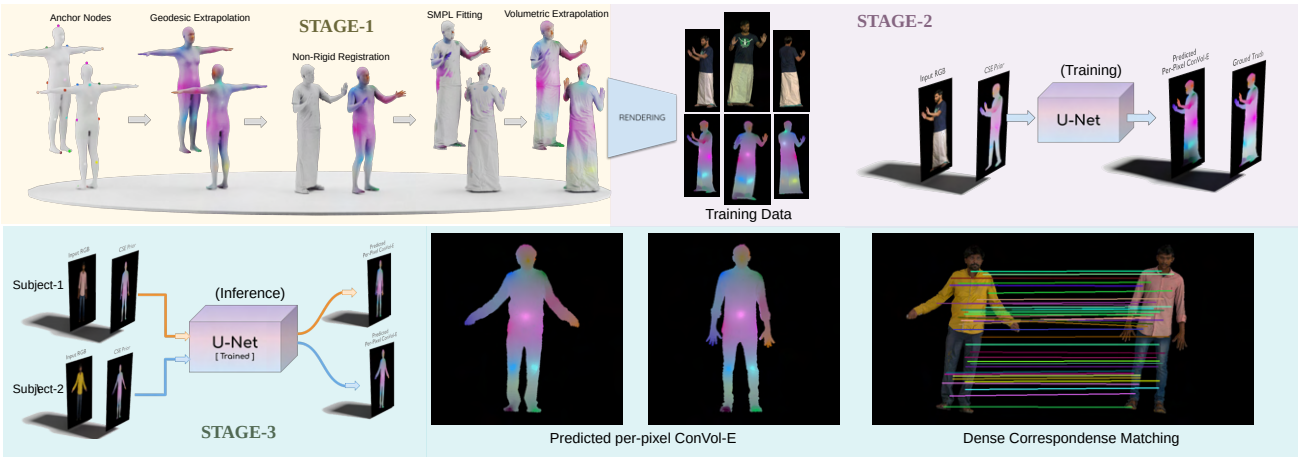

Figure 3. Overview of the three-stage method-pipeline for learning ConVol-E representation on the human images.

provide dense universal correspondences across humans. DensePose [7] aims at establishing dense correspondence from 2D images to a surface-based representation of human body (SMPL), but requires a lot of hand-annotated data. Moreover, it either provides sparse image-to-surface correspondence landmarks (DensePose-COCO) or part-specific UV coordinates on top of the input image (DensePose-RCNN). It doesn't provide pixel-wise unique embedding, which is essential for dense correspondence matching.

Continuous Surface Embeddings (CSE) [14] propose a drop-in replacement of DensePose by introducing a better and more flexible representation of correspondences using learnable positional embeddings. Given a canonical surface model of humans (SMPL), the idea is to estimate the deformation variant identity of any point on the canonical surface, and additionally, train a neural network to predict a per-pixel color-coded embedding corresponding to one such surface points, visible in the given image. Since these embeddings vary smoothly over a 3D manifold, they are continuous in nature. CSE provides a reliable way to match pixel-wise color-coded embeddings across different images of humans, however, it is not guaranteed that every pixel belonging to the human in the image is assigned some unique embedding. However, many pixels are left out, as SMPL does not cover all the intricate details, e.g. hair, clothing, skin deformations, etc. Nevertheless, it provides a universal intrinsic representation applicable to any human body, agnostic to appearance, body shape & pose. HumanGPS [18] tries to circumvent the issues in CSE prediction by proposing to use geodesic distances between corresponding points on the surface of a human scan, but it does not produce an explicit per-pixel mapping from image to scan, and additionally does not generalize well to loose clothing as reported in [8].

Recently, BodyMap [8] proposes to build on top of CSE to include the aforementioned intricate details. The authors propose to extrapolate CSE representation to human scans, by first registering a canonical SMPL mesh to the scan, and then extrapolating the embeddings from SMPL vertices to Scan vertices in the UV space based on the geodesic distance between them. This approach is reasonable as it becomes easier to render images for both the RGB and dense per-pixel correspondences to generate the training data. However, it does not handle loose clothing deformations very well. Modelling extreme deformations that lie far apart from the underlying body can not be achieved in UV space while preserving the uniqueness of the embedding values. Far apart values can have repeated values as they are only influenced by geodesically closer vertices. Although, the authors said that this can be mitigated partially by putting additional constraints on the learning side, however, this is still an inherent flaw in the representation that needs to be addressed.

A very recent work, Virtual Correspondence [13], aims at establishing correspondences across different views of a human subject in a fixed pose, by fitting a common SMPL model to multi-view images of the subject. However, the method does not establish correspondences across different subjects or even the same subject in a different pose. Hence, we propose our appearance agnostic representation that can be used to establish dense correspondences across the different subjects, viewpoints, and mainly to handle loose clothing deformations during the matching.

3. Our Method

We aim to find a novel pixel-wise unique characterization of in-the-wild clothed humans with the goal of establishing appearance-agnostic dense correspondences across multiple

images. Our method consists of three key stages as shown in Figure 3. In Stage-1, we prepare the training data using high-quality human scans registered with corresponding SMPL meshes, which are rendered to generate RGB images and ConVol-E encoded maps. As part of Stage-2, we train a U-Net based encoder-decoder to predict the ConVol-E maps given RGB images as input. Finally, Stage-3 use the trained U-Net to predict ConVol-E maps for unseen images and then perform dense correspondence matching based on predicted embeddings.

3.1. Continuous Volumetric Embeddings

The proposed ConVol-E is the representation of an arbitrary 3D point embedded in the volume of an underlying parametric model. Unlike BodyMap [8], which defines these embedding in 2D UV space, ConVol-E encodes the volume around a parametric model in 3D Euclidean space. **ConVol-E** can be formally defined as a mapping $\mathcal{F} : \mathbb{R}^3 \rightarrow \mathbb{R}^k$ which takes a 3D point $x \in \mathbb{R}^3$ and assigns it an embedding vector $e \in \mathbb{R}^k$ (we choose k=3 in our experiments). The embedding vector precisely captures the information about where a given point x lies in the vicinity of a given parametric human model. More specifically, let $\mathcal{M} = \{\mathcal{V}, \mathcal{F}\}$ be a 3D human mesh scan (obtained either from an off-the-shelf 3D reconstruction solution or using a 3D scanning methods) and $\mathcal{M}^c = \{\mathcal{V}^c, \mathcal{F}^c\}$ be the parametric human mesh (SMPL [12] in our case) in neutral pose and shape. Here, $\mathcal{V}^c$ and $\mathcal{F}^c$ are the fixed number of vertices and faces of the canonical mesh in canonical pose and shape.

First, we select a set of anchor vertices $\mathcal{V}^c_{anchor} \subset \mathcal{V}^c$. These anchor vertices are distributed across the SMPL mesh at 19 key locations like pelvis, shoulder, feet etc. (see supplementary for visualization), and each anchor vertex is assigned a unique value denoted by $\hat{e}_{v^c} \in \mathbb{R}^k$. The embeddings for the remaining vertices $\mathcal{V}^c_{non-anchor} \in \mathcal{V}^c$ (such that $\mathcal{V}^c_{non-anchor} \cap \mathcal{V}^c_{anchor} = \phi$) are computed using the weighted geodesic distance from all the anchor vertices over the canonical mesh. Thus, we can compute embedding for every canonical mesh vertex $v^c_j \in \mathcal{V}_{non-anchor}$ as:

$$e_{v^c_j} = \frac{\sum_{i=1}^{|\mathcal{V}_{anchor}|} w_i * \hat{e}_{v^c_i}}{\sum_{i=1}^{|\mathcal{V}_{anchor}|} w_i} \tag{1}$$

$$w_i = \frac{1}{g(v^c_j, v_i)} \tag{2}$$

where, $g(v^c_j, v^c_i)$ is the geodesic distance between $v^c_j \in \mathcal{V}^c_{non-anchor}$ and $v^c_i \in \mathcal{V}^c_{anchor}$.

It is important to note that, such embedding is only defined for vertices of canonical SMPL mesh $\mathcal{M}^c$. To obtain ConVol-E ebemdding for every vertex of a 3D human mesh $\mathcal{M}$ (in arbitrary pose and shape), we first perform a non-rigid registration with canonical SMPL mesh. This yields aligned SMPL mesh surface close to input 3D human mesh scan. Subsequently, for each vertex $v_j \in \mathcal{V}$ of $\mathcal{M}$, we compute the nearest neighbor set $\mathcal{N}_j \in \mathcal{V}^c$ consisting of $p = 32$ closest vertices of the registered SMPL mesh, and assign the vertex an embedding value using the following equation:

$$e_{v_j} = \frac{\sum_{i=1}^{|\mathcal{N}_j|} w_i * e_{v^c_i}}{\sum_{i=1}^{|\mathcal{N}_k|} w_i} \tag{3}$$

$$w_i = \frac{1}{d(v_j, v^c_i)}, \forall v^c_i \in \mathcal{N}_j \tag{4}$$

where, $d(\cdot, \cdot)$ is the Euclidean distance. The choice of anchor vertices and the embedding values assigned to them is important and empirically chosen to allow highly diverse values during the extrapolation, so that each vertex is assigned a sufficiently unique embedding value. The scripts to generate color-coded embedding for each anchor point and for extrapolating anchor point values to ground-truth scans for data generation will be provided post-acceptance.

Neighborhood Consistency Score : The underlying representation for dense correspondence estimation should be rich and varied enough to avoid repetitions in the feature space when extrapolated, otherwise different body parts would map nearby in the embedding space. More specifically, geodesically far-apart vertices should map far apart in the embedding space and vice-versa. Keeping this idea in mind, in order to quantify the efficacy of the proposed ConVol-E embeddings with other representations, we design a novel metric named Neighborhood Consistency Score(NCS), for each vertex v_i of the scan mesh, and is calculated as follows:

$$NCS_i = (NCS_{near_i} + NCS_{far_i})/2 \tag{5}$$

$$NCS_{near_i} = \frac{1}{q^2} \sum_{i=1}^{q} min(|\mathcal{N}^{rank}_{geo} - \mathcal{N}^{rank}_{emb}|, q) \tag{6}$$

$$NCS_{far_i} = \frac{1}{q^2} \sum_{i=1}^{q} min(|\mathcal{F}^{rank}_{geo} - \mathcal{F}^{rank}_{emb}|, q) \tag{7}$$

where, $\mathcal{N}^{rank}_{geo}$ & $\mathcal{N}^{rank}_{emb}$ denotes the ranks (relative orders) of q-nearest neighbors of v_i in both geodesic and embedding space, and similarly, $\mathcal{F}^{rank}_{geo}$ & $\mathcal{F}^{rank}_{emb}$ denotes the ranks of q-farthest neighbors of v_i in both geodesic and embedding space (q is emprically set as 32). Thus, NCS penalizes the representation if the rank of these nearest/farthest neighbours in geodesic and embedding space

doesn't match, i.e. j-th nearest-neighbor in geodesic space should be j-th nearest-neighbor in embedding space as well. Any neighbor among the q-neighbors of v_i can take maximum rank as q, so we divide by q^2 for normalization. Hence, NCS takes values between 0 and 1 where lower values are preferred. We compare the efficacy of ConVol-E with BodyMap [8] in subsection 4.4.

3.2. Learning Embeddings in Image Space

Given an input image $\mathcal{I}_{rgb}$ and the prior $\mathcal{I}_{cse}$, of size $\mathcal{W}$x$\mathcal{H}$ we train a U-Net [16] style encoder-decoder network to predict the per-pixel embeddings, represented as a three-channel feature map $\mathcal{I}_{\mathcal{E}}$. We train the U-Net by minimizing the $L1$ loss between the foreground pixels of predicted feature map $\mathcal{I}_{\mathcal{E}}$ and the corresponding ground-truth $\hat{\mathcal{I}}_{\mathcal{E}}$ generated in the previous stage.

It should be noted that we estimate the prior $\mathcal{I}_{cse}$ using a pre-trained Densepose-CSE [14] network, however with a key difference that instead of their default per-vertex embedding, we replace it with our proposed ConVol-E embedding. Additionally, BodyMap uses a Vision Transformer (ViT) [5] architecture instead of U-Net for predicting dense pixel-wise embeddings. However, the authors treat U-Net as a baseline and show that the performance of U-Net is on par with the ViT and convergence of ViT is slow and challenging in general. Therefore to avoid overkill, we decide to go with U-Net style encoder-decoder network.

3.3. Dense Correspondence Matching

Let $\mathcal{I}_{rgb_1}$ and $\mathcal{I}_{rgb_2}$ be two input RGB images with the known foreground, where we aim to establish dense correspondences between them. These images can have the same or different human subject, viewpoint, pose, clothing, etc. We predict the respective per-pixel ConVol-E embedding $I_{\mathcal{E}_1}$ and $I_{\mathcal{E}_2}$ using the U-Net trained in the previous stage. Following this, we can establish correspondences by finding, for each pixel $p_1 \in \mathcal{I}_{rgb_1}$, the closest matching pixel $p_2 \in \mathcal{I}_{rgb_2}$, where the matching is established if the absolute difference in embedding values of pixels p_1 & p_2 is below a threshold, i.e. $|I_{\mathcal{E}_1}(p_1) - I_{\mathcal{E}_2}(p_2)| \leq t_{match}$. Further, to ensure more robustness in matching, we provide an additional bi-directional constraint that the matching pixels should mutually be the best matches of each other. More specifically, we consider a correspondence match between pixels p_1 and p_2 to be a valid correspondence if and only if p_1 is the best match of p_2 and p_2 is the best match of p_1. Figure 3 shows the obtained dense correspondences across different subjects and different viewpoints.

4. Experiments and Results

4.1. Dataset Details

We perform quantitative and qualitative evaluation of our method on two publicly available datasets - **3DHumans** [11] and **THUman2.0** [22]. 3DHumans, contains around 180 meshes of people in diverse body shapes in various garments styles and sizes, including a wide variety of clothing styles ranging from loose robed clothing to relatively tight fit clothing, like shirts and trousers. THUman2.0 contains 500 high-quality scans of multiple human subjects in arbitrary clothing and poses. We perform a random 80:20 split for training and testing for both datasets. We render RGB images and corresponding embeddings for each textured scan from 70 viewpoints using a Pre-computed Radiance Transfer (PRT)-based renderer.

4.2. Implementation Details

We adopt Pix2Pix [10] architecture to build our U-Net encoder-decoder network. Since, the final task involves regression and not synthesis, we do not require adversarial training and hence we remove the discriminator from the original Pix2Pix [10] architecture, retaining only the generator network. The generator is a U-Net style encoder-decoder, with 5 convolution and 5 transposed-convolution layers. We train the network to minimize L1 loss, with an initial learning rate of 0.0002 and the standard LR-decay. An input images resized to 512 × 512 resolution before passing through the encoder. For all the experiments, the network is trained with a batch size of 4 for 200 epochs.

4.3. Quantitative Evaluation Metric

Efficacy of ConVol-E: In terms of quantitative evaluation, we first intend to compare the efficacy of ConVol-E representation (i.e., ability to preserve the geodesic neighborhood in the embedding space) in comparison with BodyMap [8] representation. This would indicate the robustness of the underlying representations for the task of correspondence matching in 3D space itself (i.e., on the mesh surface). To this end, we compute Neighborhood Consistency Score (NCS) using Equation 5 for both ConVol-E and BodyMap.

Representation	NCS (3DHumans [11]) ↓	NCS (THUman2.0 [22]) ↓
BodyMap [8]	0.955	0.957
ConVol-E (Ours)	**0.838**	**0.835**

Table 1. Comparison between BodyMap [8] and ConVol-E using the proposed Neighborhood Consistency Score.

Evaluation of Predicted 2D Embedding Maps: Inspired from [18], we develop another metric ***Geodesic Distance Error (GDE)*** to quantitatively evaluate the predicted pixel-level embedding against the ground-truth. Specifically,

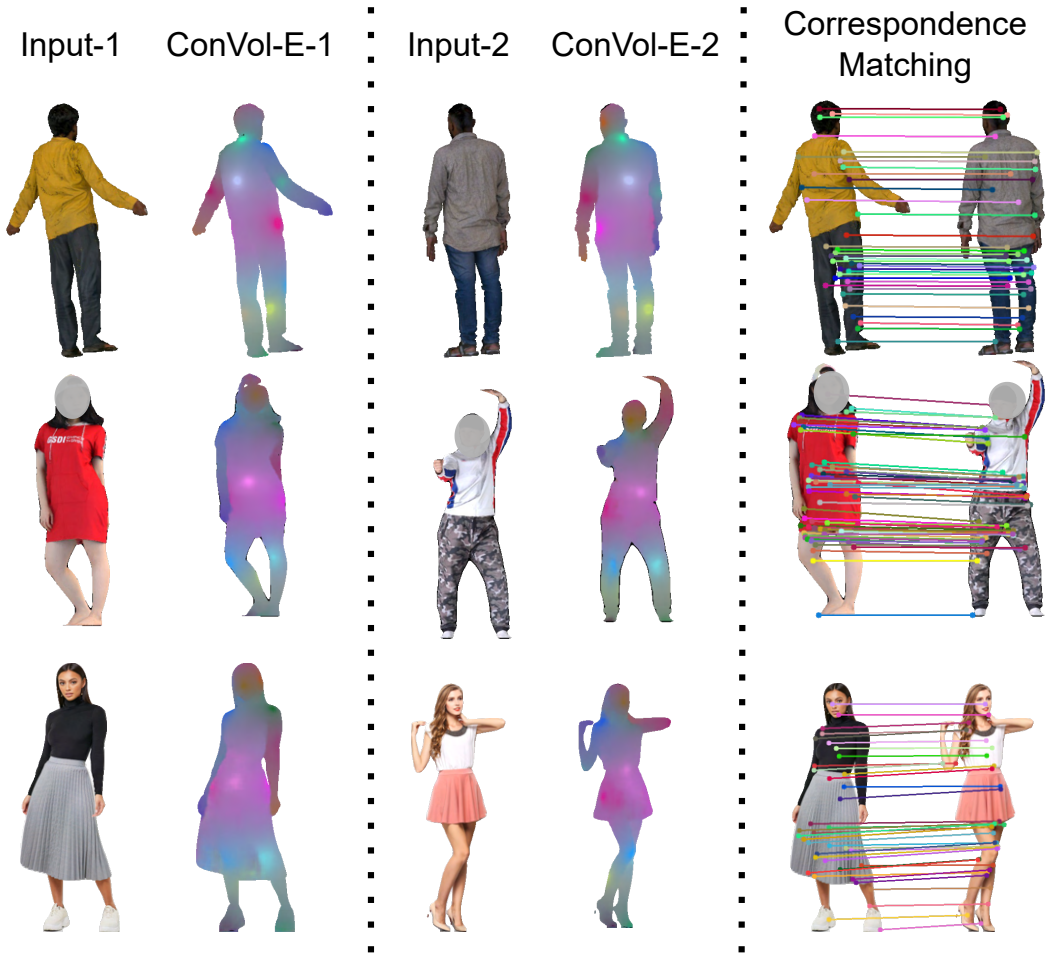

Figure 4. Predicted ConVol-E maps and dense correspondence matching on samples from 3DHumans [11] (first row), THUman2.0 [22] (second row) & internet images (third row) [Some faces have been blurredaccording to the dataset T&C].

we find the geodesic distance between the corresponding ground truth and predicted vertices for each pixel, followed by computing the percentage of pixels having geodesic error less than a particular distance threshold. GDE is computed as:

$$GDE^t = \frac{1}{N}\sum_{i=1}^{N} g(v_i, v_i') < t \tag{8}$$

where t represents the distance threshold, $i \in \{1...N\}$ represents the indices of all foreground pixels and v_i, v_i' represent the corresponding ground-truth and predicted vertices. We compute threshold-specific numbers as that gives us additional information about performance of a method for different thresholds.

4.4. Quantitative Results

We compare our method with the current state of the art method BodyMap [8]. Firstly, we report *NCS* in Table 1 where our ConVol-E representation outperform BodyMap [8] by attaining lower *NCS* score on two datasets.

Table 2 report GDE values where our method significantly outperforms BodyMap across thresholds. The observed improvement in performance is even higher for smaller distance thresholds, indicating that our method is better than BodyMap [8] for correspondence estimation, both overall, and specially, for fine-grain correspondence estimation.

Further, in Table 3 we also report the $L1$ and $L2$ loss between the ground-truth per-pixel embeddings and the predicted per-pixel embeddings for both ConVol-E and BodyMap representations. It can be seen that the $L1$ and $L2$ loss values for ConVol-E are lower than the values for BodyMap for both the cases - RGB only and RGB with CSE prior, respectively. This shows that along with being better than BodyMap in terms of the richness of the representation, our ConVol-E representation is also more easily learnable by a U-Net style encoder-decoder network.

(a) RGB (b) CSE (c) BodyMap (d) Ours

Figure 5. Qualitative comparison between CSE [14], BodyMap [8] and the proposed ConVol-E representation on internet images.

Method	GDE (3DHumans [11])			GDE (THUman2.0 [22])		
	5cm ↑	10cm ↑	15cm ↑	5cm ↑	10cm ↑	15cm ↑
BodyMap [8]: RGB-only	24.85	44.62	58.33	16.37	33.60	48.47
BodyMap [8]: RGB+CSE	25.45	45.02	58.65	21.77	40.76	55.44
ConVol-E (Ours): RGB-only	58.89	68.41	73.30	41.60	53.85	61.72
ConVol-E (Ours): RGB+CSE	63.98	72.40	76.17	51.94	62.35	68.72

Table 2. Comparison between BodyMap [8] and ConVol-E using **GDE** (eq. 8) for varying values of threshold t={5cm,10cm,15cm}.

Method	3DHumans [11]		THUman2.0 [22]	
	L1 ↓	L2 ↓	L1 ↓	L2 ↓
BodyMap [8]: RGB-only	0.064663	0.000713	0.178216	0.001234
BodyMap [8]	0.060392	0.000699	0.088864	0.000847
ConVol-E (Ours): RGB-only	0.046234	0.000212	0.090537	0.000412
ConVol-E (Ours)	0.038513	0.000191	0.061203	0.000280

Table 3. Comparison of L1 and L2 loss between predictions and ground truth across datasets for BodyMap [8] and ConVol-E.

4.5. Qualitative Results

Qualitative results on the test samples from 3DHumans and THUman2.0 datasets, and internet images are shown in Figure 4 and Figure 5. The results demonstrate ConVol-E's ability to generalize on challenging scenarios involving loose clothing deformations, where CSE and BodyMap fail drastically. Please refer to supplementary material for more qualitative results.

4.6. Ablation Study

We perform an ablative study by providing different inputs to our network and compare its performance. Specifically, we use following input setup : (1) Using only RGB images as the input, (2) Using output of Part Grouping Network (PGN) [6], which provides a semantic prior for different human body parts to the network and (3) Using *ConVol-E encoded CSE prior*. The results are reported in Table 4 where we can conclude that RGB + CSE prior outperforms any other setting, and yields lower $L1$ and $L2$ errors. This observation is supported by GDE values reported in Table 5 where RGB + CSE prior input setup outperform other two setups. This is due to the fact that CSE prior (encoded with proposed ConVol-E embedding) provides a good initialization for the network to further refine the intricate details covering hair, clothing, etc. Further, it can also be observed that providing part-segmentation prior from PGN leads to some improvement over the RGB-only case. This is because, part segmentation labels provide the network with information about which pixel corresponds to which body part, and acts as a coarse initialization. However, results obtained with PGN prior are still not as good as those obtained with CSE prior, as CSE provides a more meaningful initialization.

Method	THUman2.0 [22]		3DHumans [11]	
	L1	L2	L1	L2
Ours: RGB-only	0.090537	0.000412	0.046234	0.000212
Ours: RGB + PGN	0.066555	0.000301	0.043039	0.000192
Ours: RGB + CSE	0.061203	0.000280	0.038513	0.000191

Table 4. Effect of different input priors on $L1$ and $L2$ errors between predictions and ground truth of our method.

Method	GDE (3DHumans [11])			GDE (THUman2.0 [22])		
	5cm ↑	10cm ↑	15cm ↑	5cm ↑	10cm ↑	15cm ↑
Ours: RGB-only	58.89	68.41	73.30	41.60	53.85	61.72
Ours: RGB + PGN	61.87	71.06	75.81	49.99	60.83	67.56
Ours: RGB + CSE	63.98	72.4	76.17	51.94	62.35	68.71

Table 5. Effect of different input priors for our method shown using GDE for varying values of threshold t={5,10,15}.

5. Applications

5.1. Segmentation Label Transfer

A potential application of our robust, dense representation is transferring image-based segmentation information across different subjects (given that the style of garments is similar). Given an unlabelled and labelled image of two human subjects I_u and I_l respectively, we can use our method of dense correspondence matching to add labels to the unlabelled image I_u. To do this, we iterate over the pixels of I_u and for each unlabelled pixel $p_u \in I_u$, we identify the *"matching"* pixel $p_l \in I_l$ and label p_u with the same label as p_l. The pixels are *"matched"* using the output of our dense correspondence matching network. Figure 6 shows the result of dense pixel-wise semantic label transfer where the two images have different subjects with significantly different body poses and appearances.

Figure 6. Segmentation label transfer performed with dense correspondences obtained with our method on 3DHumans test samples.

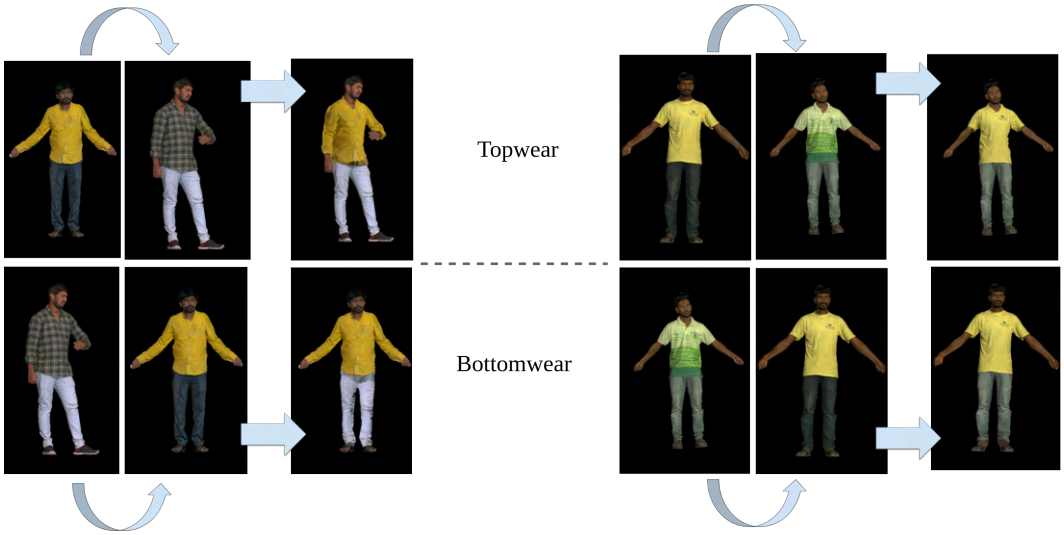

Figure 7. Garment appearance transfer performed with dense correspondences obtained from our method on 3D Humans test samples.

5.2. Garment Appearance Transfer

Another interesting application of the dense correspondence matching includes *garment appearance transfer* i.e., transferring the appearance of a garment worn by one person to another person. Consider an image I_1 with a human H_1 wearing a garment G_1 and another image I_2 with a human H_2 wearing a garment G_2. We want to transfer the appearance of the garment G_1 worn by human H_1 onto human H_2. The task is closely related to the application of virtual try-on, where we would like to see how a garment draped on a mannequin or worn by any other human would look on us.

This appearance transfer is achieved in the same way as semantic label transfer is performed. We first identify the pixels which belong to the garment in images I_1 and I_2. This segmentation can be obtained either by using a PGN [6]-like method on the input images or alternatively, we can also use the method described above to transfer the segmentation labels from the labelled image, if any. Once we have identified the pixels which belong to the required garment(s) in both images, then, for each pixel of interest $p_2 \in I_2$, we find the corresponding pixel in $p_1 \in I_1$ and transfer the RGB value of p_1 to p_2, as shown in Figure 7.

6. Conclusion

We present **ConVol-E**, a robust representation for dense correspondence matching across RGB images of different human subjects in different poses/shapes/appearances. Existing methods fail to capture correspondences for points which do not lie in the vicinity of the body model. Our proposed volumetric representation can model arbitrary deviation from the underlying body model by making use of use of carefully chosen anchor nodes and volumetric extrapolation around the parametric body model. The proposed representation is easily learned with a simple U-Net-based architecture demonstrating superior qualitative and quantitative results. Further, we also show qualitative results on internet images, including, loose clothing scenarios. Finally, we discuss two potential applications of this work. Though the proposed embedding s inherently view-invariant, in future, we would like to model the learning process in a way to provide explicit constraint over multi-view consistency. In future, we can also explore explicit solutions for enforcing the embedding to be temporally consistent.

References

[1] Bharat Lal Bhatnagar, Garvita Tiwari, Christian Theobalt, and Gerard Pons-Moll. Multi-garment net: Learning to dress 3d people from images. In *IEEE International Conference on Computer Vision (ICCV)*. IEEE, oct 2019. 2

[2] Adrian Bulat and Georgios Tzimiropoulos. Human pose estimation via convolutional part heatmap regression. In *European Conference on Computer Vision*, pages 717–732. Springer, 2016. 2

[3] Joao Carreira, Pulkit Agrawal, Katerina Fragkiadaki, and Jitendra Malik. Human pose estimation with iterative error feedback. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4733–4742, 2016. 2

[4] Christopher B Choy, JunYoung Gwak, Silvio Savarese, and Manmohan Chandraker. Universal correspondence network. *Advances in neural information processing systems*, 29, 2016. 2

[5] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. *ArXiv*, abs/2010.11929, 2020. 5

[6] Ke Gong, Xiaodan Liang, Yicheng Li, Yimin Chen, Ming Yang, and Liang Lin. Instance-level human parsing via part grouping network. In *ECCV*, 07 2018. 7, 8

[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. 3

[8] Anastasia Ianina, Nikolaos Sarafianos, Yuanlu Xu, Ignacio Rocco, and Tony Tung. Bodymap: Learning full-body dense correspondence map. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13276–13285, 2022. 1, 2, 3, 4, 5, 6, 7

[9] Eldar Insafutdinov, Leonid Pishchulin, Bjoern Andres, Mykhaylo Andriluka, and Bernt Schiele. Deepercut: A deeper, stronger, and faster multi-person pose estimation model. In *European conference on computer vision*, pages 34–50. Springer, 2016. 2

[10] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *Computer Vision and Pattern Recognition (CVPR), 2017 IEEE Conference on*, 2017. 5

[11] Sai Sagar Jinka, Astitva Srivastava, Chandradeep Pokhariya, Avinash Sharma, and P. J. Narayanan. Sharp: Shape-aware reconstruction of people in loose clothing. *International Journal of Computer Vision*, Dec. 2022. 2, 5, 6, 7

[12] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J. Black. Smpl: a skinned multi-person linear model. *ACM Trans. Graph.*, 34:248:1–248:16, 2015. 2, 4

[13] Wei-Chiu Ma, Anqi Joyce Yang, Shenlong Wang, Raquel Urtasun, and Antonio Torralba. Virtual correspondence: Humans as a cue for extreme-view geometry. 2022. 3

[14] Natalia Neverova, David Novotny, Vasil Khalidov, Marc Szafraniec, Patrick Labatut, and Andrea Vedaldi. Continuous surface embeddings. 2020. 1, 2, 3, 5, 7

[15] Ignacio Rocco, Relja Arandjelovic, and Josef Sivic. Convolutional neural network architecture for geometric matching. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6148–6157, 2017. 2

[16] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. *ArXiv*, abs/1505.04597, 2015. 2, 5

[17] Astitva Srivastava, Chandradeep Pokhariya, Sai Sagar Jinka, and Avinash Sharma. xcloth: Extracting template-free textured 3d clothes from a monocular image. *arXiv preprint arXiv:2208.12934*, 2022. 2

[18] Feitong Tan, Danhang Tang, Mingsong Dou, Kaiwen Guo, Rohit Pandey, Cem Keskin, Ruofei Du, Deqing Sun, Sofien Bouaziz, Sean Fanello, Ping Tan, and Yinda Zhang. HumanGPS: Geodesic PreServing Feature for Dense Human Correspondence. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2021. 3, 5

[19] Shih-En Wei, Varun Ramakrishna, Takeo Kanade, and Yaser Sheikh. Convolutional pose machines. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016. 2

[20] Yuliang Xiu, Jinlong Yang, Dimitrios Tzionas, and Michael J. Black. ICON: Implicit Clothed humans Obtained from Normals. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13296–13306, June 2022. 2

[21] Fan Yang, Xin Li, Hong Cheng, Jianping Li, and Leiting Chen. Object-aware dense semantic correspondence. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2777–2785, 2017. 2

[22] Tao Yu, Zerong Zheng, Kaiwen Guo, Pengpeng Liu, Qionghai Dai, and Yebin Liu. Function4d: Real-time human volumetric capture from very sparse consumer rgbd sensors. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR2021)*, June 2021. 5, 6, 7

[23] Yebin Liu Qionghai Dai Zerong Zheng, Tao Yu. Pamir: Parametric model-conditioned implicit representation for image-based human reconstruction, 2021. 2

[24] Heming Zhu, Lingteng Qiu, Yuda Qiu, and Xiaoguang Han. Registering explicit to implicit: Towards high-fidelity garment mesh reconstruction from single images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3845–3854, June 2022. 2