HumanGPS: Geodesic PreServing Feature for Dense Human Correspondences

Feitong Tan[1,2,*] Danhang Tang[1] Mingsong Dou[1] Kaiwen Guo[1] Rohit Pandey[1] Cem Keskin[1]
Ruofei Du[1] Deqing Sun[1] Sofien Bouaziz[1] Sean Fanello[1] Ping Tan[2] Yinda Zhang[1]
[1] Google [2] Simon Fraser University

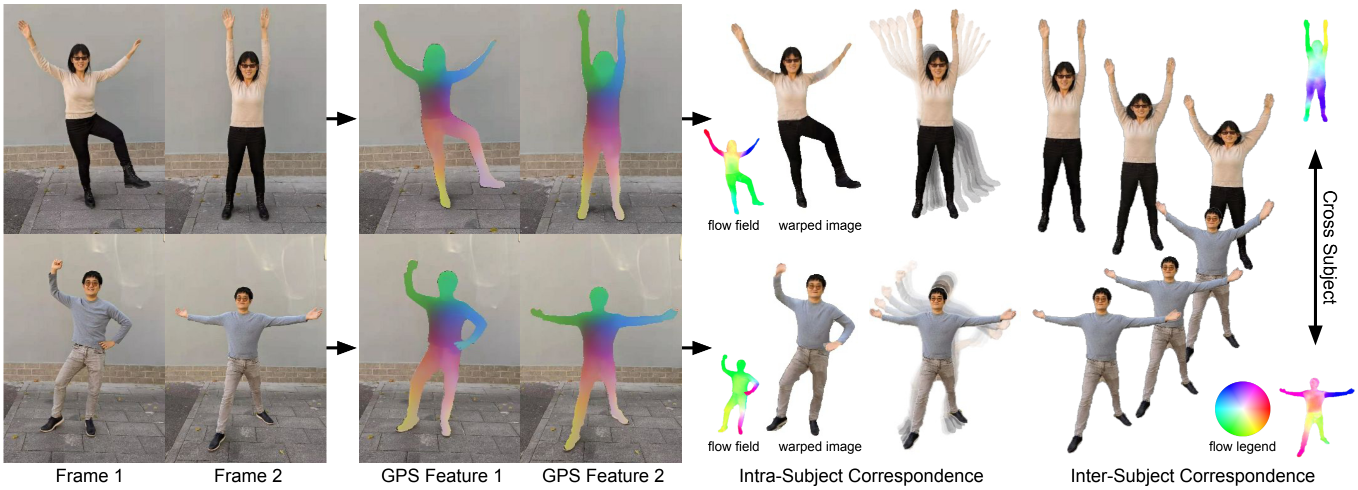

Figure 1. We propose a deep learning framework to learn a Geodesic PreServing (GPS) feature from RGB images that can produce accurate dense human correspondences. From left to right, we show the input images, extracted features visualized in color, the intra-subject correspondences, and the reconstruction of frame 1 using frame 2, the morphing [24, 25] between frames and across subjects, and the inter-subject correspondences. Please refer to the color legend in the bottom right for the flow directions and magnitudes.

Abstract

In this paper, we address the problem of building dense correspondences between human images under arbitrary camera viewpoints and body poses. Prior art either assumes small motion between frames or relies on local descriptors, which cannot handle large motion or visually ambiguous body parts, e.g., left vs. right hand. In contrast, we propose a deep learning framework that maps each pixel to a feature space, where the feature distances reflect the geodesic distances among pixels as if they were projected onto the surface of a 3D human scan. To this end, we introduce novel loss functions to push features apart according to their geodesic distances on the surface. Without any semantic annotation, the proposed embeddings automatically learn to differentiate visually similar parts and align different subjects into an unified feature space. Extensive experiments show that the learned embeddings can produce accurate correspondences between images with remarkable generalization capabilities on both intra and inter subjects. [1]

1. Introduction

Finding correspondences across images is one of the fundamental problems in computer vision and it has been studied for decades. With the rapid development of digital human technology, building dense correspondences between human images has been found to be particularly useful for many applications, such as non-rigid tracking and reconstruction [12, 11, 36, 13], neural rendering [48], and appearance transfer [62, 57]. Traditional approaches in computer vision extract image features on local keypoints and generate correspondences between points with similar descriptors after performing a nearest neighbor search, *e.g.*, SIFT [29]. More recently, deep learning methods [26, 59, 42, 14], replaced hand-crafted components with full end-to-end pipelines. Despite their effectiveness on many tasks, these methods often deliver sub-optimal results when performing dense correspondences search on humans, due to the high variation in human poses and camera viewpoints and visual similarity between body parts. As a result, the existing methods either produce sparse matches, *e.g.*, skeleton joints [7], or dense but imprecise correspondences [15].

[*] Work done while the author was an intern at Google.
[1] Project webpage: https://feitongt.github.io/HumanGPS/

In this paper, we propose a deep learning method to learn a **Geodesic PreServing (GPS)** feature taking RGB images as input, which can lead to accurate dense correspondences between human images through nearest neighbor search (see Figure 1). Differently from previous methods using triplet loss [42, 18], *i.e.* hard binary decisions, we advocate that the feature distance between pixels should be inversely correlated to their likelihood of being correspondences, which can be intuitively measured by the geodesic distance on the 3D surface of the human scan (Figure 2). For example, two pixels having zero geodesic distance means they project to the same point on the 3D surface and thus a match, and the probability of being correspondences becomes lower when they are apart from each other, leading to a larger geodesic distance. While the geodesic preserving property has been studied in 3D shape analysis [43, 23, 33], e.g., shape matching, we are the first to extend it for dense matching in image space, which encourages the feature space to be strongly correlated with an underlying 3D human model, and empirically leads to accurate, smooth, and robust results.

To generate supervised geodesic distances on the 3D surface, we leverage 3D assets such as RenderPeople [1] and the data acquired with The Relightables [16]. These high quality 3D models can be rigged and allow us to generate pairs of rendered images from the same subject under different camera viewpoints and body poses, together with geodesic distances between any locations on the surface. In order to enforce soft, efficient, and differentiable constraints, we propose novel single-view and cross-view dense geodesic losses, where features are pushed apart from each other with a weight proportional to their geodesic distance.

We observe that the GPS features not only encode local image content, but they also have a strong semantic meaning. Indeed, even without any explicit semantic annotation or supervision, we find that our features automatically differentiate semantically different locations on the human surface and it is robust even in ambiguous regions of the human body (*e.g.*, left hand vs. right hand, torso vs. back). Moreover, we show that the learned features are consistent across different subjects, *i.e.*, the same semantic points from other persons still map to a similar feature, without any inter-subject correspondence data provided during the training.

In summary, we propose to learn an embedding that significantly improves the quality of the dense correspondences between human images. The core idea is to use the geodesic distance on the 3D surface as an effective supervision and combine it with novel loss functions to learn a discriminative feature. The learned embeddings are effective for dense correspondence search, and they show remarkable intra- and inter-subjects robustness without the need of any cross-subject annotation or supervision. We show that our approach achieves state-of-the-art performance on both intra- and inter-subject correspondences and that the proposed framework can be used to boost many crucial computer vision tasks that rely on robust and accurate dense correspondences, such as optical flow, human dense pose regression [15], dynamic fusion [36] and image-based morphing [24, 25].

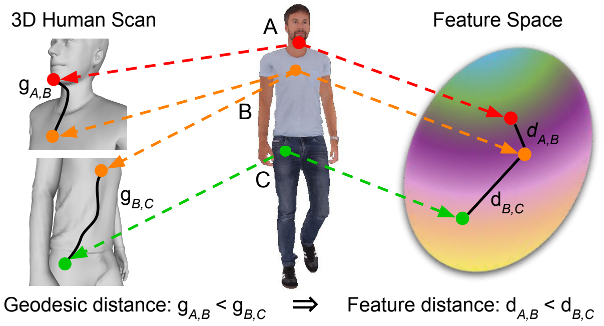
3D Human Scan
A
Feature Space
$g_{A,B}$
B
$d_{A,B}$
C
$g_{B,C}$
$d_{B,C}$
Geodesic distance: $g_{A,B} < g_{B,C}$ ⇒ Feature distance: $d_{A,B} < d_{B,C}$

Figure 2. Core idea: we learn a mapping from RGB pixels to a feature space that preserves geodesic properties of the underlying 3D surface. The 3D geometry is only used in the training phase.

2. Related Work

In this section, we discuss current approaches in the literature for correspondence search tasks.

Hand-Crafted and Learned Descriptors. Traditional approaches that tackle the matching problem between two or more images typically rely on feature descriptors [30, 5, 52, 40] extracted on sparse keypoints, which nowadays are still popular for Structure-From-Motion or SLAM systems when computational budget is limited. More recently, machine learning based feature extractors are proposed for image patches by pre-training on classification tasks [26], making binary decision on pairs [17, 61, 54] or via a triplet loss [59, 4, 32, 51, 31, 18]. Recently, Schuster *et al.* [42] proposed an architecture with stacked dilated convolutions to increase the receptive field. These methods are designed for generic domain and do not incorporate domain specific knowledge, *e.g.*, human body in our case. When the domain is given, a unified embedding can be learned to align objects within a category to enforce certain semantic properties [14, 50, 49, 53]. The resulting intra-domain correspondences are arguably better than previous approaches. While most of methods are built purely on RGB images, 3D data are used to either as the input [56], provide cycle consistency [65], or create normalized label space [53]. In contrast, our method is designed specifically for human correspondences, takes only color images as input, and enforces informative constraints from the 3D geometry according to the geodesic distance on the human surface.

Direct Regression of Correspondences. Orthogonal approaches to correspondence search aim at regressing directly the matches between images. Examples of this trend are optical flow methods that can estimate dense correspondences in image pairs. Early optical flow methods are often built with hand-crafted features and formulated as energy minimization problems based on photometric consistency and spatial smoothness [20, 6, 55].

More recently, deep learning methods have become popular in optical flow [10, 21, 22, 47] and stereo matching [60, 8, 63], where they aim at learning end-to-end correspondences directly from the data. PWC-Net [45] and LiteFlowNet [21] incorporate ideas from traditional methods and present a popular design using a feature pyramid, warping and a cost volume. IRR-PWC [22] and RAFT [47] present iterative residual refinements, which lead to state-of-the-art performance. These methods aim at solving the generic correspondence search problem and are not designed specifically for human motion, which could be large and typically non-rigid.

Human Correspondences. There are plenty of works studying sparse human correspondences by predicting body keypoints describing the human pose in images [38, 37, 7]. For dense correspondences, many works rely on an underlying parametric model of a human, such as SMPL [27], and regress direct correspondences that lie on the 3D model. This 3D model shares the same topology across different people, hence allowing correspondences to be established [58, 15, 66]. DensePose [15] is the first method showing that such correspondences are learnable given a sufficiently large training set, although it requires heavy labor to guarantee the labeling quantity and quality. Follow up work reduces the annotation workload using equivariance [35] or simulated data [66], and extend DensePose to enable pose transfer across subjects [34].

Differently from previous approaches, we show how to learn human specific features directly from the data, without the need of explicit annotations. Our approach learns an embedding from RGB images that follows the geodesic properties of an underlying 3D surface. Thanks to this, the proposed method can be applied to full body images, performs robustly to viewpoint and pose changes, and surprisingly generalizes well across different people without using any inter-subject correspondences during the training.

3. Method

In this section, we introduce our deep learning method for dense human correspondences from RGB images (Figure 3). The key component of our method is a feature extractor, which is trained to produce a Geodesic PreServing (GPS) feature for each pixel, where the distance between descriptors reflects the geodesic distance on surfaces of the human scans. We first explain the GPS feature in detail and then introduce our novel loss functions to exploit the geodesic signal as supervision for effective training. This enhances the discriminative power of the feature descriptors and reduces ambiguity for regions with similar textures.

3.1. Geodesic PreServing Feature

Our algorithm starts with an image $\mathbf{I} \in \mathbb{R}^{H \times W \times 3}$ of height H and width W, where we first run an off-the-shelf segmentation algorithm [9] to detect the person. Then, our feature extractor takes as input this image and maps it into a high-dimensional feature map of the same spatial resolution $\mathbf{F} \in \mathbb{R}^{H \times W \times C}$, where $C = 16$ in our experiments. The dense correspondences between two images $\mathbf{I}_1, \mathbf{I}_2$ can be built by searching for the nearest neighbor in the feature space, *i.e.*, $\text{corr}(\mathbf{p}) = \arg\min_{\mathbf{q} \in \mathbf{I}_2} d(\mathbf{p}, \mathbf{q}), \forall \mathbf{p} \in \mathbf{I}_1$, where d is a distance function defined in the feature space, and $\text{corr}(\mathbf{p})$ is the correspondence for the pixel $\mathbf{p}$ from $\mathbf{I}_1$ to $\mathbf{I}_2$. In our approach, we constrain the feature for each pixel to be a unit vector $\|\mathbf{F}_{\mathbf{I}}(\mathbf{p})\|_2 = 1, \forall \mathbf{p} \in \mathbf{I}$ and use the cosine distance $d(\mathbf{p}, \mathbf{q}) = 1 - \mathbf{F}_{\mathbf{I}_1}(\mathbf{p}) \cdot \mathbf{F}_{\mathbf{I}_2}(\mathbf{q})$.

Since images are 2D projections of the 3D world, ideally $\mathbf{F}$ should be aware of the underlying 3D geometry of the human surface and be able to measure the likelihood of two pixels being a correspondence. We find that the geodesic distance on a 3D surface is a good signal of supervision and thus should be preserved in the feature space, *i.e.*, $d(\mathbf{p}, \mathbf{q}) \propto g(\mathbf{p}, \mathbf{q}), \forall \mathbf{p}, \mathbf{q} \in (\mathbf{I}_1, \mathbf{I}_2)$, where $g(\mathbf{p}, \mathbf{q})$ is the geodesic distance between the projection of two pixels $\mathbf{p}, \mathbf{q}$ to 3D locations on the human surface (Figure 2).

Network Architecture. Theoretically, any network architecture producing features in the same spatial resolution of the input image could be used as backbone for our feature extractor. For the sake of simplicity, we utilize a typical 7-level U-Net [39] with skip connections. To improve the capacity without significantly increasing the model size, we add residual blocks inspired by [64]. More details can be found in supplementary materials.

3.2. Loss Functions

Our model uses a pair of images as a single training example. These images capture the same subjects under different camera viewpoints and body poses. Both images are fed into the network and converted into feature maps. Multiple loss terms are then combined to compute the final loss function that is minimized during the training phase.

Consistency Loss. The first loss term minimizes the feature distance between ground truth corresponding pixels: $L_c(\mathbf{p}) = \sum d(\mathbf{p}, \text{corr}(\mathbf{p}))$. Note however, that training with only L_c will lead to degenerative case where all the pixels are mapped to the same feature.

Sparse Ordinal Geodesic Loss. To prevent this degenerative case, previous methods use triplet loss [42, 19] to increase the distance between non-matching pixels, *e.g.*

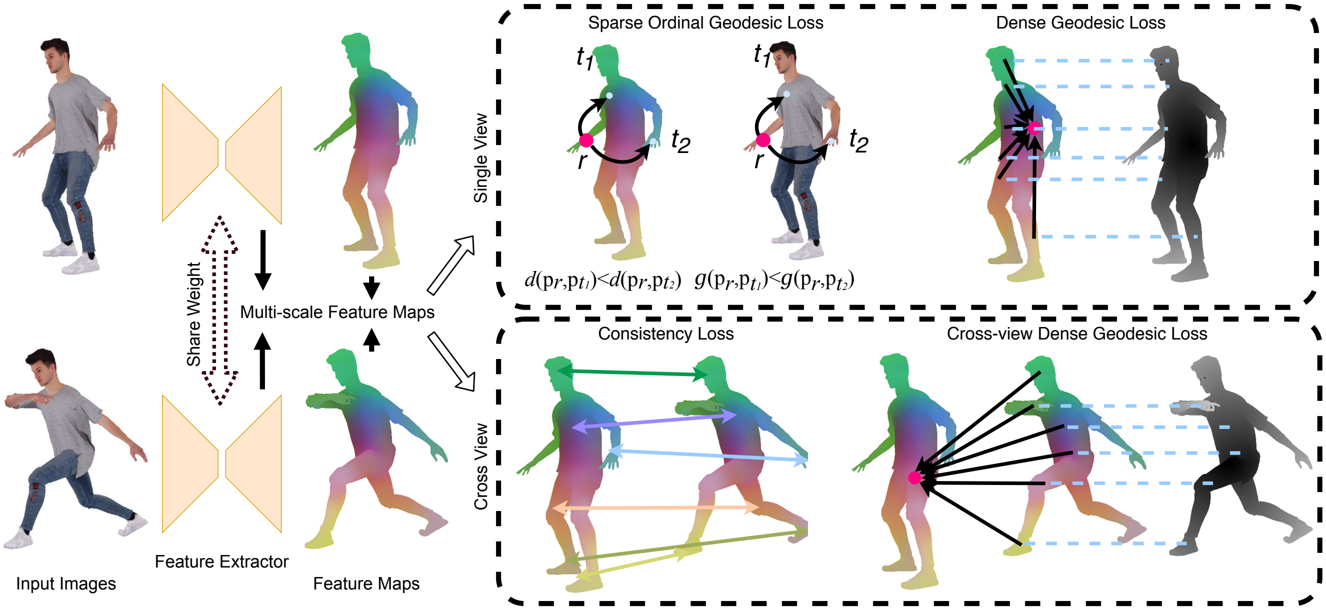

Figure 3. Learning Human Geodesic PreServing Features. We train a neural network to extract features from RGB images. The learned embedding reflects the geodesic distance among pixels projected on the 3D surface of the human and can be used to build accurate dense correspondences. We train our feature extractor with a combination of consistency loss, sparse ordinal geodesic loss and dense intra/cross view geodesic loss: see text for details.

$d(\mathbf{p}, \text{corr}(\mathbf{p})) \ll d(\mathbf{p}, \mathbf{q}), \forall \mathbf{q} \neq \text{corr}(\mathbf{p})$. Whereas the general idea makes sense and works decently in practice, this loss function penalizes all the non-matching pixels equally without capturing their relative affinity, which leads to non-smooth and imprecise correspondences.

An effective measurement capturing the desired behavior is the geodesic distance on the 3D surface. This distance should be 0 for corresponding pixels and gradually increase when two pixels are further apart. To enforce a similar behavior in the feature space, we extend the triplet loss by randomly sampling a reference point $\mathbf{p}_r$ and two target points $\mathbf{p}_{t_1}, \mathbf{p}_{t_2}$, and defining a sparse ordinal geodesic loss:

$$L_s = \log(1 + \exp(s \cdot (d(\mathbf{p}_r, \mathbf{p}_{t_1}) - d(\mathbf{p}_r, \mathbf{p}_{t_2})))), \quad (1)$$

where $s = \text{sgn}(g(\mathbf{p}_r, \mathbf{p}_{t_2}) - g(\mathbf{p}_r, \mathbf{p}_{t_1}))$. This term encourages the order between two pixels with respect to a reference pixel in feature space to be same as measured by the geodesic distance on the surface, and as a result, a pair of points physically apart on the surface tends to have larger distance in feature space.

Dense Geodesic Loss. L_s penalizes the order between a randomly selected pixels pair, which, however, does not produces an optimal GPS feature. One possible reason is due to the complexity of trying to order all the pixels, which is a harder task compared to the original binary classification method proposed for the triplet loss. In theory, we could extend L_s to order all the pixels in the image, which unfortunately is non-trivial to run efficiently during the training.

Instead, we relax the ordinal loss and define a dense version of the geodesic loss between one randomly picked pixel $\mathbf{p}_r$ to all the pixels $\mathbf{p}_t$ in the image:

$$L_d = \sum_{\mathbf{p}_t \in \mathbf{I}} \log\left(1 + \exp(g(\mathbf{p}_r, \mathbf{p}_t) - d(\mathbf{p}_r, \mathbf{p}_t)\right). \quad (2)$$

This loss, again, pushes features between non-matching pixels apart, depending on the geodesic distance. It does not explicitly penalize the wrong order, but it is effective for training since all the pixels are involved in the loss function and contribute to the back-propagation.

Cross-view Dense Geodesic Loss. The features learned with the aforementioned loss terms produces overall accurate correspondence, but susceptible to visually similar body parts. For example in Figure 6 (top row), the feature always matches the wrong hand, since it does not capture correctly the semantic part due to the presence of large motion. To mitigate this issue, we extend L_d and define it between pairs of images such that: given a pixel $\mathbf{p}_1$ on $\mathbf{I}_1$ and $\mathbf{p}_2$ on $\mathbf{I}_2$:

$$L_{cd} = \sum_{\mathbf{p}_2 \in \mathbf{I}_2} \log\left(1 + \exp(g(\text{corr}(\mathbf{p}_1), \mathbf{p}_2) - d(\mathbf{p}_1, \mathbf{p}_2)\right). \quad (3)$$

At its core, the intuition behind this loss term is very similar L_d, except that it is cross-image and provides the network with training data with high variability due to viewpoint and pose changes. We also tried to add a cross-view sparse ordinal geodesic loss but found it not improving.

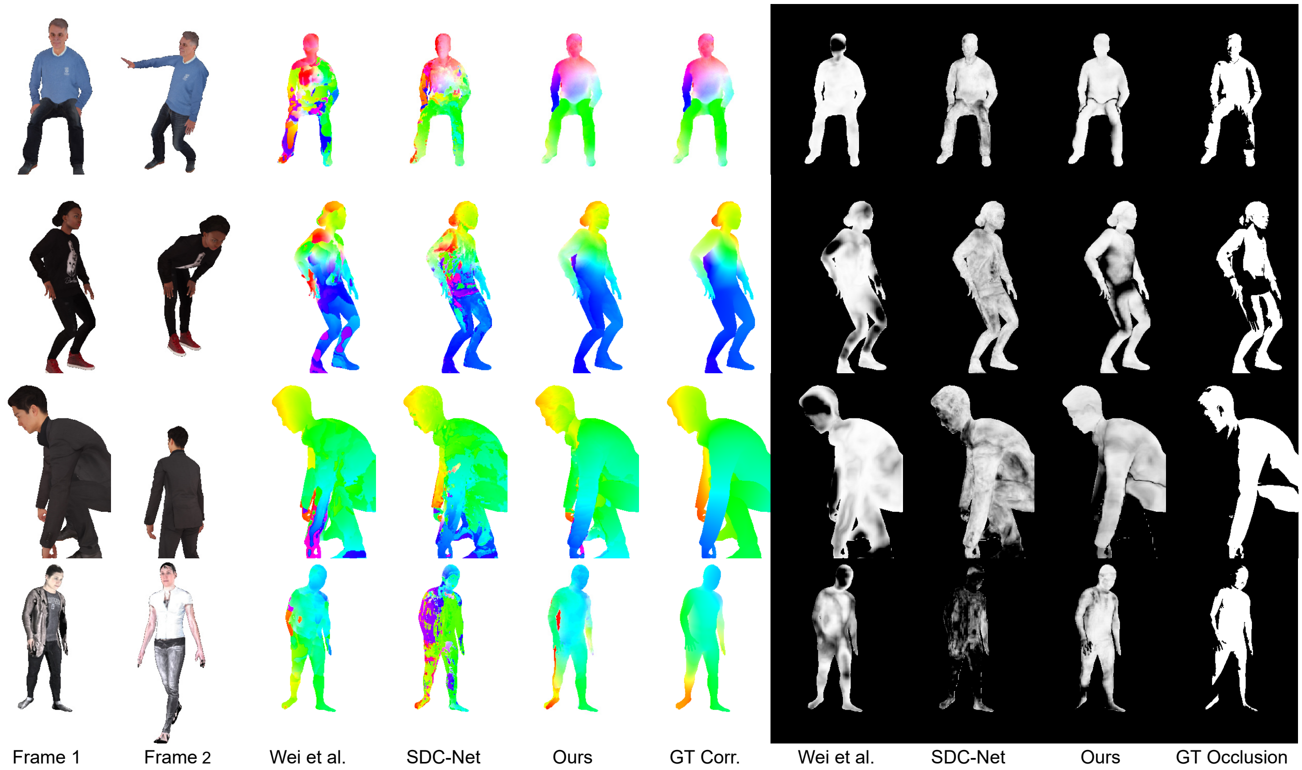

Figure 4. Dense correspondences (visualized as optical flow) built via nearest neighbor search and the predicted visibility masks. Our results are more accurate, smooth, and free from obvious mistakes when compared to other methods. On the right, we show the visibility probability map obtained via the distance to the nearest neighbor. Note that our feature successfully captures occluded pixels (*i.e.*, dark pixels) in many challenging cases. The method is effective for both intra-subjects (rows 1-3) and inter-subjects (row 4).

Total Loss. The total loss function is a weighted sum of the terms detailed above $L_t = w_c L_c + w_s L_s + w_d L_d + w_{cd} L_{cd}$. The weights are set to 1.0, 3.0, 5.0, 3.0 for w_c, w_s, w_d, w_{cd} respectively. The weight for each term is chosen empirically such that the magnitude of gradients from each loss is roughly comparable. To encourage robustness across different scales, we compute this loss at each intermediate level of the decoder, and down-weight the loss to $\frac{1}{8}$. As demonstrated in our ablation studies, we found this increases the overall accuracy of the correspondences.

Our whole system is implemented in TensorFlow 2.0 [2]. The model is trained with batch size of 4 using ADAM optimizer. The learning rate is initialized at 1×10^{-4} and reduces to 70% for every 200K iterations. The whole training takes 1.6 millions iterations to converge.

4. Experiments

In this section, we evaluate the GPS features using multiple datasets and settings. In particular, we compare our approach to other state-of-the-art methods for correspondence search and show its effectiveness for both intra- and inter-subjects problems. Additional evaluations and applications, such as dynamic fusion [36] and image-based morphing [24, 25], can be found in the supplementary materials.

4.1. Data Generation

Since it would be very challenging to fit 3D geometry on real images, we resort to semi-synthetic data, where photogrammetry or volumetric capture systems are employed to capture subjects under multiple viewpoints and illumination conditions. In particular, we generate synthetic renderings using SMPL [27], RenderePeople [1], and The Relightables [16]. These 3D assets are then used to obtain correspondences across views and geodesic distances on the surface. As demonstrated by the previous work [41, 66], training on captured models generalizes well on real images in the wild. A similar idea has been used to create a fully synthetic dataset for optical flow [10], but in this work we focus on human images with larger camera viewpoints and body pose variations. All the details regarding the data generation can be found in the supplementary material.

4.2. Dense Matching with Nearest Neighbor Search

We first evaluate the capability of the GPS features in building dense pixel-wise matches via nearest neighbor search (Section 3.1).

Baseline Methods. We compare to two state-of-the-art descriptor learning methods that use different supervision for

Methods	Intra-Subject						Inter-Subject	
	SMPL [28]		Relightables [16]		RenderPeople [1]		SMPL [28]	
	non	all	non	all	non	all	non	all
SDC-Net [42]	16.96	33.17	17.79	29.14	20.07	39.95	81.48	96.60
Wei *et al.* [56]	18.08	30.59	29.54	43.64	34.42	46.23	18.03	31.55
Ours + Full + Multi-scale	**7.12**	**17.51**	**11.24**	**18.95**	**11.91**	**22.12**	**8.49**	**17.99**
Ours + triplet	9.14	24.34	13.18	25.59	16.84	29.80	21.08	30.75
Ours + classify	9.73	25.80	15.97	33.03	18.33	34.03	11.21	25.72
Ours + L_c + L_s	8.17	19.31	14.61	21.45	14.51	24.21	12.02	24.51
Ours + L_c + L_s + L_d	7.50	18.00	12.24	19.30	12.41	22.73	9.19	18.61
Ours + Full	7.32	17.57	11.50	19.12	12.29	22.48	8.57	**17.87**
Ours + Full + Multi-scale	**7.12**	**17.51**	**11.24**	**18.95**	**11.91**	**22.12**	**8.49**	17.99

Table 1. Quantitative evaluation for correspondences search. We report the average end-point-error (EPE) of non-occluded (marked as non) and all pixels (marked as all) on four test sets created from different sources of 3D assets. Our model significantly outperforms previous methods for descriptor learning on all the datasets. The model trained with the proposed loss is better than all the ablation models trained with other alternatives. We report the results for both intra and inter-subjects.

the learning. 1) SDC-Net [42]: The SDC-Net extracts dense feature descriptors with stacked dilated convolutions, and is trained to distinguish correspondences and non-matching pixels by using threshold hinge triplet loss [3]. 2) Wei *et al.* [56]: This method learns to extract dense features from a depth image via classification tasks on over-segmentations. For fair comparison, we over segment our human models and rendered the segmentation label to generated the training data, i.e. over-segmentation label map. The network is trained with the same classification losses but takes color images as the input.

Data and Evaluation Metrics. For each source of human scans introduced in Section 4.1, we divide the 3D models into train and test splits, and render image pairs for training and testing respectively. Additionally, we also generate a testing set using SMPL to quantitatively evaluate inter subjects performances since the SMPL model provides cross-subject alignment, which can be used to extract inter-subject correspondence ground truth.

As for evaluation metrics, we use the standard average end-point-error (AEPE) between image pairs, computed as the ℓ_2 distance between the predicted and ground-truth correspondence pixels on the image.

Performance on Intra-Subject Data. We first evaluate our method for intra-subject correspondences. All the methods are re-trained on three training sets and tested on each test split respectively. The dense correspondence accuracy of our approach and two state-of-art methods on the each test sets are shown in Table 1 (Intra-Subject). Our method consistently achieves significantly lower error on both non-occluding and all pixels in all three datasets.

Figure 4 (row 1-3) shows the qualitative results of the correspondences built between two images, visualized as flow where hue and saturation indicates the direction and magnitude (See Figure 1 for color legend). Our method generates much smooth and accurate correspondences compared to other methods and makes less mistakes for ambiguous body parts, *e.g.*, hands.

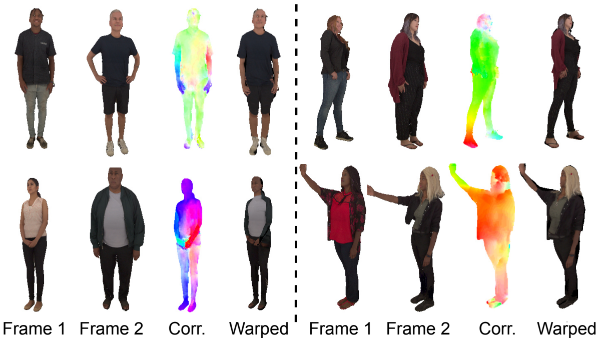

Figure 5. Inter-subject correspondences and warp fields. Note how our approach correctly preserves the shape of the reference Frame 1 while plausibly warping the texture from Frame 2.

Performance on Inter-Subject Data. We also compare our approach on inter-subject data although no explicit cross-subject correspondences are provided during the training. The results are shown in Table 1 (Inter-Subject). SDC-Net [42] does not generalize to inter-subject data since the triplet loss only differentiates positive and negative pairs and does not learn any human prior. Wei *et al.* [56] shows some generalization thanks to the dense classification loss but the average end-point-error is much higher compared to our approach. Comparatively, our method generalizes well to the inter-subject with error slightly higher, but roughly comparable to the intra-subject SMPL test set. Figure 4 (row 4) shows qualitative results of inter-subject correspondences. Again our method significantly outperforms other methods. In Figure 5 we show additional examples where we also build an image warp using the correspondences to map one person to another: notice how we preserve the shape of the body while producing a plausible texture. More results are presented in supplementary materials.

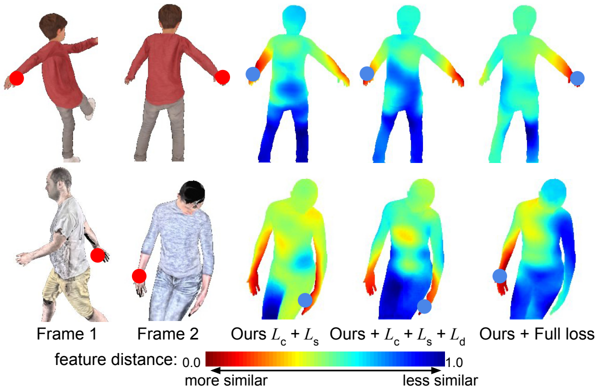

Figure 6. Distance maps on intra- and inter- subjects, using models trained with different loss functions. We visualize the feature distance between the pixel in the left image (marked with a red dot) to all the pixels in the image on the right. The closest correspondence is marked with a blue dot. Our full loss provides the best feature space with less ambiguity, *e.g.*, between left and right hand.

Occluded Pixels As mentioned in Section 3, our feature space learns to preserve the surface geodesic distance, which is a measurement of the likelihood of the correspondence. If a pixel cannot find a matching that is close enough in the feature space, it is likely that the pixel is not visible (*i.e.*, occluded) in the other view. Inspired by this, we retrieve a visibility mask via the distance of each pixel to the nearest neighbor in the other image, *i.e.*, $d_{nn} = \min_{\mathbf{q}\in\mathbf{I}_2} d(\mathbf{p},\mathbf{q}), \forall \mathbf{p} \in \mathbf{I}_1$. Specifically, the visibility is defined as $1 - d_{nn}$ for SDC-Net and our method using cosine distance, and $\frac{1-d_{nn}}{2000}$ for Wei *et al.* using ℓ_2 distance. Figure 4 (Right) visualizes the visibility map, *i.e.*, dark pixels are occluded. Our method effectively detects occluded pixel more accurately than the other methods. More details can be found in supplementary material.

4.3. Ablation Study

In this section, we study the effect of each loss term for learning the GPS features. We add L_s, L_d, L_{cd} gradually into the training and show the performance of correspondences through nearest neighbor search in Table 1 (bottom half). For reference, we also train our feature extractor with losses from SDC-Net [42] and Wei *et al.* [56] for a strict comparison on the loss only (see "Ours+triplet" and "Ours+classify"). Training with our loss is more effective than the baseline methods, and the error on all the test sets, both intra- and inter-subject, keeps reducing with new loss terms added in. This indicates that all the loss terms contributes the learning of GPS feature, which is consistent with our analysis (Section 3.2) that the loss terms improve the embeddings from different aspects.

To further analyse the effect of each loss term, we visualize the feature distance from one pixel in frame 1 (marked by a red dot) to all the pixels in frame 2 (red dot is the ground-truth correspondence) in Figure 6. The blue dots show the pixel with lowest feature distance, *i.e.*, the predicted matching. Training with L_s does not produce clean and indicative distance maps. Adding L_d improves the performance, but the feature is still ambiguous between visually similar parts, *e.g.*, left and right hand. In contrast, the feature trained with all the losses produces the best distance map, and multi-scale supervision further improves the correspondence accuracy as shown in Table 1 (last row).

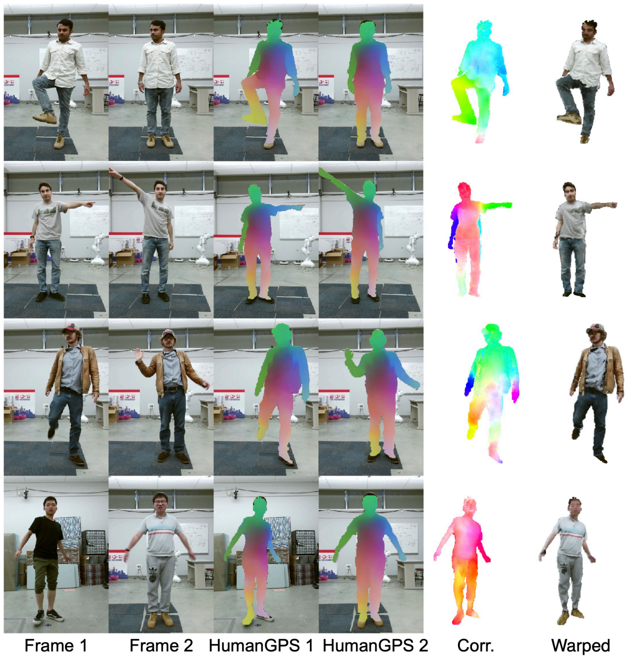

Figure 7. Results on real images. HumanGPS generalizes in the wild and provides accurate correspondences across subjects. The correspondences (column 5) successfully warp frame 2 to frame 1 for both intra- and inter-subject cases (column 6).

4.4. Generalization to Real Images

Our model is fully trained on semi-synthetic data acquired with high-end volumetric capture systems (*e.g.*, RenderPeople [1] and The Relightables [16]), which help to minimize the domain gap. In this section, we assess how our method performs on real data. To our knowledge, there is no real dataset with ground-truth dense correspondences for evaluation. Thus, we compute the cycle consistency across frames on videos in the wild. Specifically, given three frames from a video, we calculate correspondences among them, cascade matching $\mathbf{I}_1 \rightarrow \mathbf{I}_2$ and $\mathbf{I}_2 \rightarrow \mathbf{I}_3$ and measure the average end-point-error (AEPE) to the ones calculated directly with $\mathbf{I}_1 \rightarrow \mathbf{I}_3$. We collected 10 videos of moving performers. The avg. cycle consistency errors of SDC-Net, Wei *et al.* and ours are 17.53, 22.19, and 12.21 respectively. Also, we show the qualitative results on real images from [46]. As shown in Figure 7, our method generalizes reasonably to the real data, producing an accurate feature space, correspondences, as well as warped images

Methods	Intra-Subject						Inter-Subject	
	SMPL [28]		Relightables [16]		RenderPeople [1]		SMPL [28]	
	non	all	non	all	non	all	non	all
PWC-Net [45]	4.51	13.27	3.57	10.01	9.57	18.74	20.06	28.14
PWC-Net*	3.89	12.82	3.42	9.39	7.91	16.99	19.21	23.77
PWC-Net + GPS	**2.73**	**10.86**	**2.99**	**9.01**	**6.89**	**14.72**	**12.08**	**17.92**
RAFT [44]	3.62	12.30	3.27	11.65	6.74	15.90	45.47	53.09
RAFT*	3.24	12.82	2.79	11.39	5.62	14.79	57.82	66.04
RAFT + GPS	**2.13**	**10.12**	**2.27**	**10.52**	**3.95**	**12.68**	**10.76**	**17.61**

Table 2. Quantitative evaluation on dense human correspondences via SoTA optical flow networks - PWC-Net [45] and RAFT [47]. On both architecture, integrating with our GPS feature achieve the best performance. See text for the * models.

using the correspondences. For additional results and evaluation on annotated sparse keypoints, please see Figure 1 and the supplementary material.

4.5. HumanGPS with End-to-end Networks

In this section, we show that our HumanGPS features can be integrated with state-of-art end-to-end network architectures to improve various tasks.

Integration with Optical Flow. We integrate our features to the state-of-the-art optical flow methods PWC-Net [45] and RAFT [47]. Both methods consist of a siamese feature extractor and a follow-up network to regress the flow. We attach an additional feature extractor and pre-train it with our losses to learn our GPS features. The GPS features are then combined with the output of the original feature extractor by element-wise average and then fed into the remaining layers of the network to directly regress a 2D flow vector pointing to the matching pixel in the other image.

The quantitative evaluation is presented in Table 2. All the methods are trained on our training sets. To compare under the same model capacity, we also train both methods with the additional feature extractor but without our loss functions (marked with *). Our GPS features benefits the correspondence learning on both SOTA architectures consistently over all test sets. We also train PWC-Net to output an additional layer to predict the occlusion mask, and again the model trained with GPS features consistently outperforms all the baselines (see supplementary materials).

Integration with DensePose. Our GPS features automatically maps features extracted from different subjects to the same embedding, and thus can benefit cross-subject tasks like dense human pose regression [15]. To show this, we pre-train a GPS feature extractor on our datasets, attach two more MLP layers, and fine-tune on DensePose-COCO dataset [15] to regress the UV coordinates. To make sure the model capacity is comparable with other methods, we use the backbone of previous methods for feature extraction, *i.e.*, a ResNet-101 FCN in DensePose [15], and Hourglass in Slim DensePose [35]. To focus on the matching accuracy, we adopt their same evaluation setup, where ground truth bounding box is given; percentages of pixels with geodesic error less than certain thresholds are taken as the metric; and evaluate on DensePose MSCOCO benchmark [15].

Methods	Accuracy		
	5 cm	10 cm	20 cm
DP ResNet-101 FCN [15]	43.05	65.23	74.17
DP ResNet-101 FCN* [15]	51.32	75.50	85.76
SlimDP HG - 1 stack [35]	49.89	74.04	82.98
Our ResNet-101 FCN	49.09	73.12	84.51
Our ResNet-101 FCN*	53.01	76.77	87.70
Our HG - 1 stack	50.50	75.57	87.18

Table 3. Quantitative evaluation for dense human pose regression on DensePose COCO dataset [15]. Following [15], we assume ground-truth bounding box is given and calculate percentage of pixels with error smaller than thresholds. We also compare models trained on full image and only foreground (marked by *).

Table 3 shows the comparison with previous work. Following DensePose [15], we also train our model with their network backbone on full image and only foreground (marked as *). Our method consistently achieves better performance than previous methods on various setting with different network backbones. Note that the UV coordinates are estimated via only two layers of MLP, which indicates that the GPS features are already effective at mapping different subjects to the same embedding. Please see supplementary material for qualitative results.

5. Conclusion

We presented a deep learning approach to build HumanGPS, a robust feature extractor for finding human correspondences between images. The learned embedding is enforced to follow the geodesic distances on an underlying 3D surface representing the human shape. By proposing novel loss terms, we show that the feature space is able to map human body parts to features that preserve their semantic. Hence, the method can be applied to both intra- and inter-subject correspondences. We demonstrate the effectiveness of HumanGPS via comprehensive experimental results including comparison with the SOTA methods, ablation studies, and generalization studies on real images. In future work, we will extend HumanGPS to work with more object categories and remove the dependency of a foreground segmentation step.

References

[1] Renderpeople. https://renderpeople.com/. 2, 5, 6, 7, 8

[2] Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, et al. Tensorflow: Large-scale machine learning on heterogeneous distributed systems. *arXiv preprint arXiv:1603.04467*, 2016. 5

[3] Christian Bailer, Kiran Varanasi, and Didier Stricker. Cnn-based patch matching for optical flow with thresholded hinge embedding loss. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3250–3259, 2017. 6

[4] Vassileios Balntas, Edward Johns, Lilian Tang, and Krystian Mikolajczyk. Pn-net: Conjoined triple deep network for learning local image descriptors. *arXiv preprint arXiv:1601.05030*, 2016. 2

[5] Herbert Bay, Andreas Ess, Tinne Tuytelaars, and Luc Van Gool. Speeded-up robust features (surf). In *CVIU*, 2008. 2

[6] Thomas Brox, Andrés Bruhn, Nils Papenberg, and Joachim Weickert. High accuracy optical flow estimation based on a theory for warping. In *European conference on computer vision*, pages 25–36. Springer, 2004. 3

[7] Zhe Cao, Gines Hidalgo, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Openpose: realtime multi-person 2d pose estimation using part affinity fields. *IEEE transactions on pattern analysis and machine intelligence*, 43(1):172–186, 2019. 1, 3

[8] Jia-Ren Chang and Yong-Sheng Chen. Pyramid stereo matching network. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 3

[9] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. *CoRR*, abs1802.02611, 2018. 3

[10] Alexey Dosovitskiy, Philipp Fischer, Eddy Ilg, Philip Hausser, Caner Hazirbas, Vladimir Golkov, Patrick Van Der Smagt, Daniel Cremers, and Thomas Brox. Flownet: Learning optical flow with convolutional networks. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2758–2766, 2015. 3, 5

[11] Mingsong Dou, Philip Davidson, Sean Ryan Fanello, Sameh Khamis, Adarsh Kowdle, Christoph Rhemann, Vladimir Tankovich, and Shahram Izadi. Motion2fusion: Real-time volumetric performance capture. *ACM Transactions on Graphics (TOG)*, 36(6):1–16, 2017. 1

[12] Mingsong Dou, Sameh Khamis, Yury Degtyarev, Philip Davidson, Sean Ryan Fanello, Adarsh Kowdle, Sergio Orts Escolano, Christoph Rhemann, David Kim, Jonathan Taylor, et al. Fusion4d: Real-time performance capture of challenging scenes. *ACM Transactions on Graphics (TOG)*, 35(4):1–13, 2016. 1

[13] Ruofei Du, Ming Chuang, Wayne Chang, Hugues Hoppe, and Amitabh Varshney. Montage4D: Real-Time Seamless Fusion and Stylization of Multiview Video Textures. *Journal of Computer Graphics Techniques*, 8(1):1–34, Jan. 2019. 1

[14] Utkarsh Gaur and BS Manjunath. Weakly supervised manifold learning for dense semantic object correspondence. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1735–1743, 2017. 1, 2

[15] Rıza Alp Güler, Natalia Neverova, and Iasonas Kokkinos. Densepose: Dense human pose estimation in the wild. In *CVPR*, 2018. 2, 3, 8

[16] Kaiwen Guo, Peter Lincoln, Philip Davidson, Jay Busch, Xueming Yu, Matt Whalen, Geoff Harvey, Sergio Orts-Escolano, Rohit Pandey, Jason Dourgarian, et al. The relightables: Volumetric performance capture of humans with realistic relighting. *ACM Transactions on Graphics (TOG)*, 38(6):1–19, 2019. 2, 5, 6, 7, 8

[17] Xufeng Han, Thomas Leung, Yangqing Jia, Rahul Sukthankar, and Alexander C Berg. Matchnet: Unifying feature and metric learning for patch-based matching. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3279–3286, 2015. 2

[18] Elad Hoffer and Nir Ailon. Deep metric learning using triplet network. In *International Workshop on Similarity-Based Pattern Recognition*, pages 84–92. Springer, 2015. 2

[19] Elad Hoffer and Nir Ailon. Deep metric learning using triplet network. In *ICLR (Workshop)*, 2015. 3

[20] Berthold KP Horn and Brian G Schunck. Determining optical flow. In *Techniques and Applications of Image Understanding*, volume 281, pages 319–331. International Society for Optics and Photonics, 1981. 3

[21] Tak-Wai Hui, Xiaoou Tang, and Chen Change Loy. Lite-flownet: A lightweight convolutional neural network for optical flow estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8981–8989, 2018. 3

[22] Junhwa Hur and Stefan Roth. Iterative residual refinement for joint optical flow and occlusion estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5754–5763, 2019. 3

[23] Iasonas Kokkinos, Michael M Bronstein, Roee Litman, and Alex M Bronstein. Intrinsic shape context descriptors for deformable shapes. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pages 159–166. IEEE, 2012. 2

[24] Jing Liao, Rodolfo S Lima, Diego Nehab, Hugues Hoppe, and Pedro V Sander. Semi-automated video morphing. In *Computer Graphics Forum*, volume 33, pages 51–60. Wiley Online Library, 2014. 1, 2, 5

[25] Jing Liao, Rodolfo S Lima, Diego Nehab, Hugues Hoppe, Pedro V Sander, and Jinhui Yu. Automating image morphing using structural similarity on a halfway domain. *ACM Transactions on Graphics (TOG)*, 33(5):1–12, 2014. 1, 2, 5

[26] Jonathan L Long, Ning Zhang, and Trevor Darrell. Do convnets learn correspondence? In *Advances in neural information processing systems*, pages 1601–1609, 2014. 1, 2

[27] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J. Black. SMPL: A skinned multi-person linear model. *SIGGRAPH Asia*, 2015. 3, 5

[28] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J Black. Smpl: A skinned multi-

person linear model. *ACM Transactions on Graphics (TOG)*, 34(6):1–16, 2015. 6, 8

[29] David G Lowe. Object recognition from local scale-invariant features. In *Proceedings of the Seventh IEEE International Conference on Computer Vision*, volume 2, pages 1150–1157. Ieee, 1999. 1

[30] David G Lowe. Distinctive image features from scale-invariant keypoints. *International journal of computer vision*, 60(2):91–110, 2004. 2

[31] Zixin Luo, Tianwei Shen, Lei Zhou, Siyu Zhu, Runze Zhang, Yao Yao, Tian Fang, and Long Quan. Geodesc: Learning local descriptors by integrating geometry constraints. In *Proceedings of the European conference on computer vision (ECCV)*, pages 168–183, 2018. 2

[32] Anastasiia Mishchuk, Dmytro Mishkin, Filip Radenovic, and Jiri Matas. Working hard to know your neighbor's margins: Local descriptor learning loss. In *Advances in Neural Information Processing Systems*, pages 4826–4837, 2017. 2

[33] Francesc Moreno-Noguer. Deformation and illumination invariant feature point descriptor. In *CVPR 2011*, pages 1593–1600. IEEE, 2011. 2

[34] Natalia Neverova, Riza Alp Güler, and Iasonas Kokkinos. Dense pose transfer. *ECCV*, 2018. 3

[35] Natalia Neverova, James Thewlis, Rıza Alp Güler, Iasonas Kokkinos, and Andrea Vedaldi. Slim densepose: Thrifty learning from sparse annotations and motion cues. In *CVPR*, 2019. 3, 8

[36] Richard A Newcombe, Dieter Fox, and Steven M Seitz. Dynamicfusion: Reconstruction and tracking of non-rigid scenes in real-time. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 343–352, 2015. 1, 2, 5

[37] George Papandreou, Tyler Zhu, Liang-Chieh Chen, Spyros Gidaris, Jonathan Tompson, and Kevin Murphy. Personlab: Person pose estimation and instance segmentation with a bottom-up, part-based, geometric embedding model. In *ECCV*, 2018. 3

[38] George Papandreou, Tyler Zhu, Nori Kanazawa, Alexander Toshev, Jonathan Tompson, Chris Bregler, and Kevin Murphy. Towards accurate multi-person pose estimation in the wild. In *CVPR*, 2017. 3

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

[40] Ethan Rublee, Vincent Rabaud, Kurt Konolige, and Gary Bradski. Orb: An efficient alternative to sift or surf. In *2011 International Conference on Computer Vision*, pages 2564–2571. Ieee, 2011. 2

[41] Shunsuke Saito, Tomas Simon, Jason Saragih, and Hanbyul Joo. Pifuhd: Multi-level pixel-aligned implicit function for high-resolution 3d human digitization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, June 2020. 5

[42] René Schuster, Oliver Wasenmuller, Christian Unger, and Didier Stricker. Sdc-stacked dilated convolution: A unified descriptor network for dense matching tasks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2556–2565, 2019. 1, 2, 3, 6, 7

[43] Gil Shamai and Ron Kimmel. Geodesic distance descriptors. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6410–6418, 2017. 2

[44] Deqing Sun, Charles Herrmann, Varun Jampani, Michael Krainin, Forrester Cole, Austin Stone, Rico Jonschkowski, Ramin Zabih, William T. Freeman, and Ce Liu. TF-RAFT: A tensorflow implementation of raft. In *ECCV Robust Vision Challenge Workshop*, 2020. 8

[45] Deqing Sun, Xiaodong Yang, Ming-Yu Liu, and Jan Kautz. Pwc-net: Cnns for optical flow using pyramid, warping, and cost volume. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8934–8943, 2018. 3, 8

[46] Sicong Tang, Feitong Tan, Kelvin Cheng, Zhaoyang Li, Siyu Zhu, and Ping Tan. A neural network for detailed human depth estimation from a single image. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 7750–7759, 2019. 7

[47] Zachary Teed and Jia Deng. Raft: Recurrent all-pairs field transforms for optical flow. In *European Conference on Computer Vision*, pages 402–419. Springer, 2020. 3, 8

[48] Ayush Tewari, Ohad Fried, Justus Thies, Vincent Sitzmann, Stephen Lombardi, Kalyan Sunkavalli, Ricardo Martin-Brualla, Tomas Simon, Jason Saragih, Matthias Nießner, Rohit Pandey, Sean Fanello, Gordon Wetzstein, Jun-Yan Zhu, Christian Theobalt, Maneesh Agrawala, Eli Shechtman, Dan B Goldman, and Michael Zollhoefer. State of the art on neural rendering. In *Eurographics*, 2020. 1

[49] James Thewlis, Samuel Albanie, Hakan Bilen, and Andrea Vedaldi. Unsupervised learning of landmarks by descriptor vector exchange. In *International Conference on Computer Vision*. 2

[50] James Thewlis, Hakan Bilen, and Andrea Vedaldi. Unsupervised learning of object frames by dense equivariant image labelling. In *NIPS*. 2017. 2

[51] Yurun Tian, Bin Fan, and Fuchao Wu. L2-net: Deep learning of discriminative patch descriptor in euclidean space. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 661–669, 2017. 2

[52] Engin Tola, Vincent Lepetit, and Pascal Fua. Daisy: An efficient dense descriptor applied to wide-baseline stereo. In *PAMI*, 2010. 2

[53] He Wang, Srinath Sridhar, Jingwei Huang, Julien Valentin, Shuran Song, and Leonidas J. Guibas. Normalized object coordinate space for category-level 6d object pose and size estimation. In *CVPR*, 2019. 2

[54] Shenlong Wang, Sean Ryan Fanello, Christoph Rhemann, Shahram Izadi, and Pushmeet Kohli. The global patch collider. *CVPR*, 2016. 2

[55] Andreas Wedel, Daniel Cremers, Thomas Pock, and Horst Bischof. Structure-and motion-adaptive regularization for high accuracy optic flow. In *2009 IEEE 12th International Conference on Computer Vision*, pages 1663–1668. IEEE, 2009. 3

[56] Lingyu Wei, Qixing Huang, Duygu Ceylan, Etienne Vouga, and Hao Li. Dense human body correspondences using convolutional networks. In *Proceedings of the IEEE Conference*

on Computer Vision and Pattern Recognition, pages 1544–1553, 2016. 2, 6, 7

[57] Zhonghua Wu, Guosheng Lin, Qingyi Tao, and Jianfei Cai. M2e-try on net: Fashion from model to everyone. In *Proceedings of the 27th ACM International Conference on Multimedia*, pages 293–301, 2019. 1

[58] Donglai Xiang, Hanbyul Joo, and Yaser Sheikh. Monocular total capture: Posing face, body, and hands in the wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019. 3

[59] Kwang Moo Yi, Eduard Trulls, Vincent Lepetit, and Pascal Fua. Lift: Learned invariant feature transform. In *ECCV*, 2016. 1, 2

[60] Kwang Moo Yi*, Eduard Trulls*, Yuki Ono, Vincent Lepetit, Mathieu Salzmann, and Pascal Fua. Learning to find good correspondences. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018. 3

[61] Sergey Zagoruyko and Nikos Komodakis. Learning to compare image patches via convolutional neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4353–4361, 2015. 2

[62] Mihai Zanfir, Alin-Ionut Popa, Andrei Zanfir, and Cristian Sminchisescu. Human appearance transfer. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5391–5399, 2018. 1

[63] Jure Zbontar and Yann LeCun. Stereo matching by training a convolutional neural network to compare image patches. *Journal of Machine Learning Research (JMLR)*, 2016. 3

[64] Zhengxin Zhang, Qingjie Liu, and Yunhong Wang. Road extraction by deep residual u-net. *IEEE Geoscience and Remote Sensing Letters*, 15(5):749–753, 2018. 3

[65] Tinghui Zhou, Philipp Krahenbuhl, Mathieu Aubry, Qixing Huang, and Alexei A Efros. Learning dense correspondence via 3d-guided cycle consistency. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 117–126, 2016. 2

[66] Tyler Zhu, Per Karlsson, and Christoph Bregler. Simpose: Effectively learning densepose and surface normals of people from simulated data. In *European Conference on Computer Vision*, pages 225–242. Springer, 2020. 3, 5