Supplementary Material for
“HumanGPS: Geodesic PreServing Feature for Dense Human Correspondences”

Feitong Tan\textsuperscript{1,*} Danny Tang\textsuperscript{2} Mingsong Dou\textsuperscript{2} Kaiwen Guo\textsuperscript{2} Rohit Pandey\textsuperscript{2} Cem Keskin\textsuperscript{2} Ruofei Du\textsuperscript{2} Deqing Sun\textsuperscript{2} Sofien Bouaziz\textsuperscript{2} Sean Fanello\textsuperscript{2} Ping Tan\textsuperscript{1} Yinda Zhang\textsuperscript{2}
\textsuperscript{1} Simon Fraser University \textsuperscript{2} Google Research

In this supplementary material, we provide details about our semi-synthetic dataset, our network architecture to learn the human geodesic preserving feature space, and additional experimental results. Please see the additional webpage for video demos.

A. Semi-synthetic Data

As described in the main paper Section 4.1, our method relies on high quality 3D assets for training. To ensure high diversity and variation in our training set, we created synthetic datasets merging multiple state-of-art acquisition systems. Start from a 3D human scan, we render a pair of images from two different camera viewpoint and ground truths measurements, including 2D correspondences, geodesic distance between pixels and visibility masks. Examples of generated data are shown in Figure A. In each example, we show the input pair of images, the ground truth correspondences with visibility mask, and the geodesic distance map w.r.t. one pixel (marked in red). Row 1-3 shows intra-subject data, and Row 4 shows inter subject data.

A.1. 3D Assets and Candidate Pose

We collected 3D assets and candidate body poses from various sources to ensure good diversity and high realism.

SMPL. We use the SMPL body model \cite{2} and 900 aligned texture maps from SURREAL \cite{27}. These models are less realistic compared to other sources but provide good diversity. Following SURREAL \cite{27}, we randomly sample shape parameters from the distribution of CAESAR subjects \cite{23}, and collect pose parameters by fitting SMPL model using MoSh \cite{15} to the motion capture data from CMU MoCap database \cite{1} which contains 2.6K sequences from 23 high-level action categories.

RenderPeople. Additionally, we acquired 25 rigged 3D human scans from RenderPeople \cite{3}, whose models contains different clothing and hair styles, and the texture map

\textsuperscript{*}Work done while the author was an intern at Google.

are much more detailed than SURREAL. For each human scans, we animate them using pose sequence collected from Mixamo \cite{2}, which includes 27K different poses.

The Relightables \cite{10}. We also use 60 high-fidelity posed human models captured using The Relightables \cite{10} system, since their renderings have higher photorealism compared to the other sources. To minimize rendering artifacts, we do not animate these models and keep them in the original configuration.

A.2. Camera Setup and Pose Selection

We use the pinhole camera model with a resolution of $256 \times 384$ and a focal length of 500 pixels. We randomly sample a pair of camera centers in range [1.5, 3.6] meters away from the person, and control the angle between the camera facing directions no large than 60 degree to produce reasonable overlaps between views. Note this is already much larger camera variations compared to typical optical flow datasets \cite{22, 21}. For the pose, we randomly sample from the pose pool for SMPL and RenderPeople, respectively and, as mentioned above, we fixed the poses for The Relightables scans.

A.3. Ground truth

We generate three kinds of ground truths measurements: 1) 2D correspondences, 2) visibility mask, 3) geodesic distance between pixels. We first render warping fields from one camera to the UV space and from UV space to the other camera. The correspondences and visibility mask between two rendered image can be obtained by cascading two warping operations.

For geodesic distance on 3D surfaces, we adopt the exact method proposed by Mitchell \textit{et al.} \cite{18}. In order to represent surface in a 2D image, we render a triangle index image and a barycentric coordinates image so that each pixel corresponds to a point in the piece-wise linear surface. Given a pixel as the source, we compute geodesic distances to all of the rest pixels in parallel and store them in a distance image.
to support our novel dense geodesic loss.

In total, we generate $280K/2.7K$, $795K/2.9K$, $42K/1.8K$ for training/testing splits from SMPL [17], RenderePeople [3], and The Relightables [10] respectively. Note that inter-subject ground truth correspondences are not available using 3D assets from either RenderePeople [3] or The Relightables [10], since it is non-trivial to align high-fidelity 3D scans from different subjects. Therefore, we generate $2.2K$ cross-subject images from SMPL [17] for inter-subject evaluation purposes only (i.e. no training). Indeed, we tried to add some inter-subject data from SMPL into the training stage, but we found that it did not significantly improve the performances on test cases.

B. Network Architecture

In this section, we introduce our detailed network architecture, including the feature extractor mentioned in main paper Section 3.1 and used in Section 4.2. We also describe how to integrate it with end-to-end architectures for optical flow and DensePose [9].

B.1. Feature Extractor

The architecture of our feature extractor is shown in Figure B. It is 7-level residual U-Net with skip connections. We use residual block to extract feature, and the feature channels of each level are set as 16, 32, 64, 96, 128, 128, 196. In the decoder, bilinear sampling is applied to increase the spatial resolution, and we add a $1 \times 1$ convolution layer followed by a normalization layer after each residual block to produce HumanGPS feature for each level.

B.2. PWC-Net + GPS

As shown in Figure C, we attach our HumanGPS feature extractor along with the original feature extractor of PWC-Net. For each level we fuse the features from both feature extractors to obtain the input to the cost volume module. When fusing the feature, HumanGPS feature is passed to two $1 \times 1$ convolution layers with ReLU as activation function, then the original flow feature and HumanGPS feature are fused by element-wise mean.
Figure B. Proposed architecture. Our method relies on a U-Net, with multiple ResNet blocks and skip connections. C and S are the channel number and stride for the convolution layers.
Figure C. Proposed end-to-end architecture for optical flow. We fuse our GPS feature with the original feature extractor from PWC-Net [25]. As showed in the main paper, this substantially improves the performance even when compared to a PWC-Net with a larger capacity.

B.3. RAFT + GPS

Figure D shows the architecture of RAFT+GPS. Similar to PWC-Net+GPS, a HumanGPS feature extractor is added to the original RAFT framework. The feature from the original feature extractor and our HumanGPS feature extractor are fused before constructing 4D cost volume. Unlike PWC-Net which computes the cost volume in a pyramid, RAFT constructs 4D cost volume at 1/8 resolution, thus we resize the HumanGPS feature map to 1/8 resolution via a stride convolution, then pass it through two $1 \times 1$ convolution layers. Then two feature maps are fused by element-wise mean.

B.4. DensePose + GPS

We use DensePose [9, 19] backbones to extract our GPS feature. To extend our method for UV coordinates regression, we first take the feature from the second last convolution layer; feed it into a normalization layer, and train the whole network with our loss. We then feed the feature before the normalization layer into two fully connected layers with 128 channels, and a regressor to predict the part probability and UV coordinates in each of the 24 parts respectively.

C. Additional Evaluations

In this section, we show more experimental results to evaluate our Human GPS feature for dense human correspondence.

C.1. Correspondence and Visibility Map

Figure E, Figure F show more examples of our predicted correspondences and visibility maps for intra and inter-subject cases respectively. In each figure, we sort all the test cases according to the error metric of our method and randomly pick four from the top and bottom 20% respectively. This gives a demonstration of the full spectrum of the quality on the test set. Our method works consistently well on both easy (the top four) and hard (the bottom four) cases, and outperforms other methods [28, 24] on the visual quality of both the predicted correspondences and visibility maps. There are some regions where all the methods perform equally poorly, however these are mostly occluded regions as shown in the visibility map. Depending on the application, it might be more preferred to mark these pixels as no available matching rather than hallucinate continuous and implausible correspondences.

C.2. Warping via Correspondence

To qualitatively evaluate the correspondences, we show in Figure G the warping results of frame 1 using the texture of the frame 2, leveraging the predicted correspondence field. Our method produces more visually appealing and semantically correct warping results compared to DensePose [9], where we used the predicted UV coordinates to establish correspondences.

C.3. Real Image Results

Figure H shows more results of our method on real images. Although we use semi-synthetic data for training,
our method generalizes well onto various real images in the wild. It is worth mentioning that, often, the foreground computation using off-the-shelf segmentation algorithms [5] may not be accurate, nevertheless, our method is robust against minor errors in practice.

C.4. Occlusion Detection

In the main paper Section 4.2, we show the qualitative results of our occlusion detection. Here we quantitatively evaluate the occlusion detection, following standard evaluation protocol adopted by object detection [8]. We detect occluded pixel as the set of pixels with the visibility score under a threshold. By varying a threshold on the distances, we calculate precision (i.e. percentage of predicted pixels that are truly occluded) and recall (i.e. percentage of occluded pixels that are detected). Finally we report the average precision as the area under the precision-recall curve.

Table A (Top) shows the comparison to other feature descriptor methods [28, 24]. SDC-Net [24] shows better occlusion detection performance, while Wei et al. [28] generalize better to inter-subject data. Overall, our method performs the best over all intra and inter-subject test sets.

In Table A (Bottom), we also show the performance of occlusion detection from neural network architecture designed for optical flow. Taking PWC-Net [25] as example, integrating our HumanGPS feature achieves the best average precision compared to the original PWC-Net even with the augmented encoder. Please check main paper Section 4.5 for explanation of the * version.

C.5. Evaluation on Human Optical Flow Dataset

In the main paper Section 4.5, we showed that our method can improve the human correspondences on our test sets when combined with optical flow network. Here we further evaluate on public human optical flow dataset proposed by Ranjan et al. [22, 21]. Compared with our dataset, they only use SMPL models for data generation, and their camera and pose variations between each pair of images are much smaller than the ones we generated. Note that optical datasets usually contains only small motion and consider both foreground and background, which are not the focus and strength of our approach.

Similar as the experiment setup of the main paper, we augment PWC-Net [25] with an augmented feature extractor and apply loss function to supply HumanGPS feature. The average end-point error on Single-Human Optical Flow dataset (SHOF) [21] is shown in Table C. The PWC-Net integrated with HumanGPS achieves the best performance compared to original PWC-Net with and without augmented feature extractor. This indicates that our method not only provide correspondences for large motion, but it is also effective when the small motion assumption holds. This evaluation is done on both foreground and background, which shows it is straightforward to extend our method on full images without the dependency on segmentation methods.
Figure E. Comparison on intra-subject data. We compare to SDC-Net [24] and Wei et al. [28]. Our method shows consistently better performance on both correspondence (left) and occlusion detection (right). The top and bottom are sampled from the 20% of the test cases with the smallest and largest error respectively.
Figure F. Comparison on inter-subject data. We compare to SDC-Net [24] and Wei et al. [28]. Our method shows consistently better performance on both correspondence (left) and occlusion detection (right). The top and bottom are sampled from the 20% of the test cases with the smallest and largest error respectively.
Figure G. Cross-subject warping results. The left section shows two reference frames. The middle section shows our GPS feature, correspondences, and the warped result of frame 1 using the texture of frame 2. The right section shows the results of DensePose (DP) [9].

Figure H. In-the-wild results. For each pair of images, we show the GPS feature, the established correspondences, and the warped result.
### Methods

#### Intra-Subject

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Wei et al. [28]</td>
<td>32.20</td>
<td>25.78</td>
<td>34.13</td>
<td>32.25</td>
</tr>
<tr>
<td>Ours</td>
<td><strong>71.20</strong></td>
<td><strong>56.08</strong></td>
<td><strong>67.65</strong></td>
<td><strong>69.33</strong></td>
</tr>
<tr>
<td>PWC-Net [25]</td>
<td>90.25</td>
<td>85.16</td>
<td>87.06</td>
<td>72.18</td>
</tr>
<tr>
<td>PWC-Net*</td>
<td>92.20</td>
<td>85.33</td>
<td>89.32</td>
<td>63.06</td>
</tr>
<tr>
<td>PWC-Net + GPS</td>
<td><strong>94.93</strong></td>
<td><strong>87.67</strong></td>
<td><strong>91.38</strong></td>
<td><strong>80.91</strong></td>
</tr>
</tbody>
</table>

#### Inter-Subject

Table A. Quantitative evaluation of occlusion detection. We show the average precision for the occlusion detection on three intra-subject test sets and one inter-subject test set. Methods on the top half directly use feature distance for occlusion detection (see the main paper Section 4.2 for details), and methods in the bottom half use optical flow architecture to regress the occlusion mask. Our feature shows good skill in occlusion detection directly via feature distance, and further improves PWC-Net on this task. Please see the main paper for explanation of the model with *.

#### Architectures

<table>
<thead>
<tr>
<th>Architectures</th>
<th>DensePose [9, 19]</th>
<th>HumanGPS</th>
</tr>
</thead>
<tbody>
<tr>
<td>ResNet-101 FCN [9]</td>
<td>[5cm, 10cm, 20cm]</td>
<td>[5cm, 10cm, 20cm]</td>
</tr>
<tr>
<td>ResNet-101 FCN* [9]</td>
<td></td>
<td></td>
</tr>
<tr>
<td>HG Stack-1 [19]</td>
<td>[5cm, 10cm, 20cm]</td>
<td></td>
</tr>
<tr>
<td>HG Stack-2 [19]</td>
<td>[5cm, 10cm, 20cm]</td>
<td></td>
</tr>
<tr>
<td>HG Stack-8 [19]</td>
<td>[5cm, 10cm, 20cm]</td>
<td></td>
</tr>
</tbody>
</table>

Table B. Quantitative evaluation for dense human pose regression on DensePose COCO dataset [9]. Following previous work [9], we assume ground truth bounding box is given and calculate percentage of pixels with error smaller than thresholds. All the models are trained on images with background, except the one marked with *, which is trained on image with white background following DensePose [9] for comparison.

### C.6. Additional Comparisons with DensePose

In main paper Section 4.5, we showed that using GPS feature can achieve competitive dense human pose regression performance. Here we show comparisons using additional network backbones in Table B. Same as the setup in the main paper Section 4.5, we adopt the same evaluation setup as DensePose [9], 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 [9]. Directly regressing UV from our GPS feature using only 2 layers of MLP consistently achieves competitive performance compared to previous work using similar backbone [9, 19], which indicates the effectiveness of our feature in telling cross-subject correspondences. We also evaluate parametric model fitting based methods [4]. Their errors are 20.73, 40.05, 54.23 for 5cm, 10cm, and 20cm respectively, which is much worse than our method.

### D. Applications

In this section, we show how our dense human correspondence benefits various applications.

#### D.1. Nonrigid Tracking and Fusion

Existing nonrigid tracking/fusion systems [20, 7, 6] have challenges when tracking fast motions. Such a system typically employs the ICP alike method, and it requires a good initialization on the non-rigid deformation parameters to ex-
Figure J. Nonrigid Fusion Comparison. We improve the non-rigid tracking using correspondence extracted via our HumanGPS feature from the color image (top and middle). In the bottom, we show the fusion results without and with our correspondence. The standard dynamic fusion system fails quickly under fast motion, whereas successfully tracks the deformation with our correspondences.

In addition, non-rigid fusion algorithms such as Motion2Fusion [6] improve reconstruction accuracy with learned correspondences. This algorithm takes 6 depth images from different viewpoint and runs non-rigid warping between canonical and live frames. The warping function is solved with additional constraints of learned correspondences from each viewpoint and non-rigid motion is estimated with more accuracy. As shown in Figure J, the standard dynamic fusion system fails quickly under fast motion, whereas successfully tracks the deformation with our correspondences.

Please see the supplementary webpage for video demo.

D.2. Morphing

![Morphing results based on HumanGPS.](image)

Figure K. Morphing results based on HumanGPS. The leftmost and rightmost columns show two input images of intra- or intersubjects. We compute the dense correspondence maps and generate the morphed frames in-between. See the supplementary webpage for more results.
Morphing is a powerful technique to create smooth animation between images. A crucial component to successful image morphing is to create a map that aligns corresponding image elements. [14, 13]. With GPS feature, one can directly establish pixels correspondences to create a smoother video transition between intra- and inter-subjects. Please refer to Fig. K for example morphed results and the supplementary webpage for more morphing videos.

References
