Supplementary Material for
"HumanGPS: Geodesic PreServing Feature for Dense Human Correspondences"

Feitong Tan[1,*] Danny Tang[2] Mingsong Dou[2] Kaiwen Guo[2]
Rohit Pandey[2] Cem Keskin[2] Ruofei Du[2] Deqing Sun[2]
Sofien Bouaziz[2] Sean Fanello[2] Ping Tan[1] Yinda Zhang[2]
[1] Simon Fraser University   [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 [16] and 900 aligned texture maps from SURREAL [27]. These models are less realistic compared to other sources but provide good diversity. Following SURREAL [27], we randomly sample shape parameters from the distribution of CAESAR subjects [23], and collect pose parameters by fitting SMPL model using MoSh [15] to the motion capture data from CMU MoCap database [1] which contains 2.6K sequences from 23 high-level action categories.

RenderPeople. Additionally, we acquired 25 rigged 3D human scans from RenderPeople [3], whose models contains different clothing and hair styles, and the texture map are much more detailed than SURREAL. For each human scans, we animate them using pose sequence collected from Mixamo [2], which includes 27K different poses.

The Relightables [10]. We also use 60 high-fidelity posed human models captured using The Relightables [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 256×384 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 [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 *et al.* [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
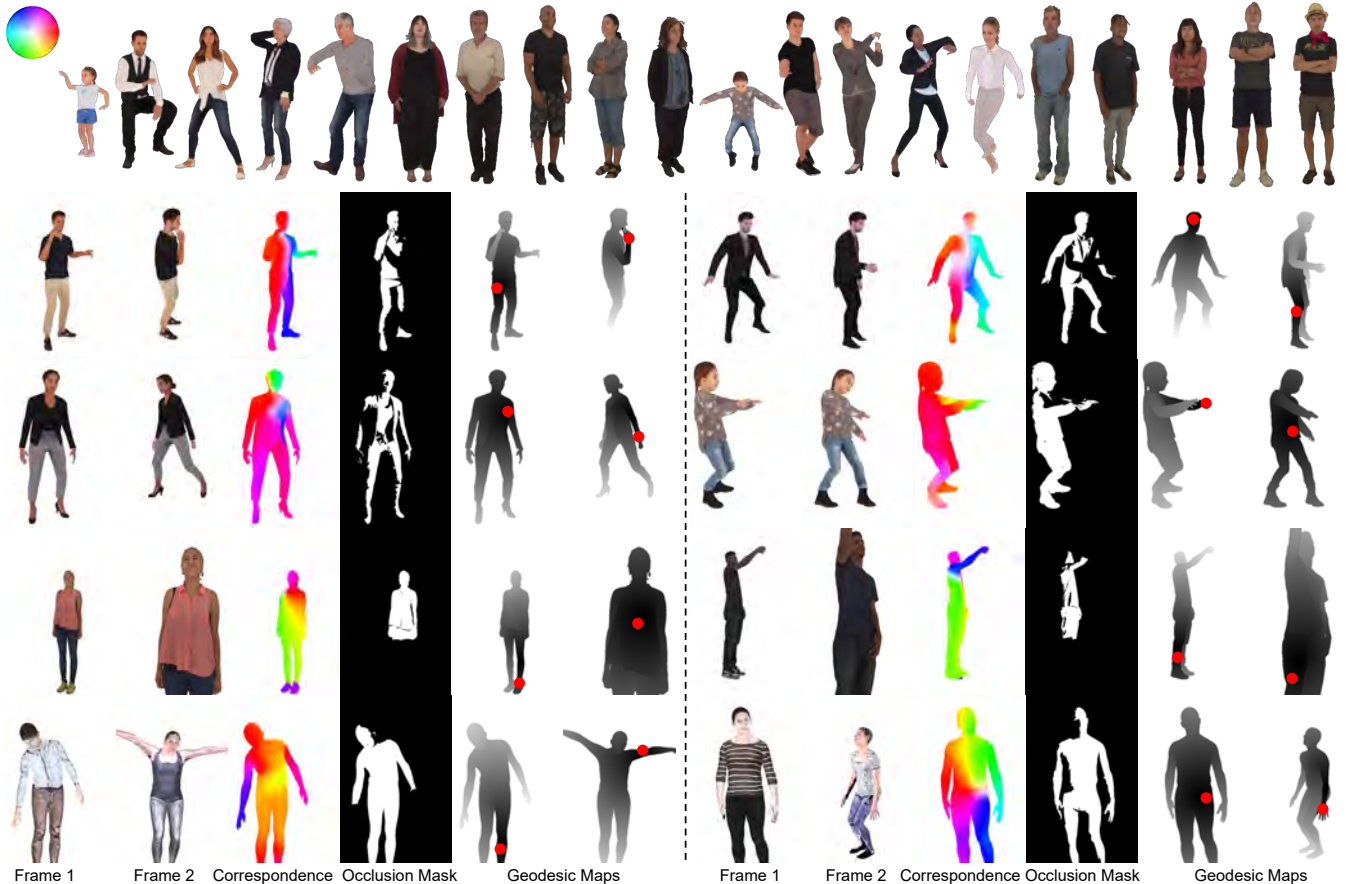

Figure A. Examples of our semi-synthetic data. From left to right, we show: a pair of the images, the ground truth correspondences, visibility mask, and the geodesic distance map w.r.t. one pixel (marked in red). Rows 1-3 show intra-subject data, and Row 4 shows inter subject data. Please refer to the color legend on the top left for the correspondence direction and magnitude.

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 1×1 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 1×1 convolution layers with ReLU as activation function, then the original flow feature and HumanGPS feature are fused by element-wise mean.
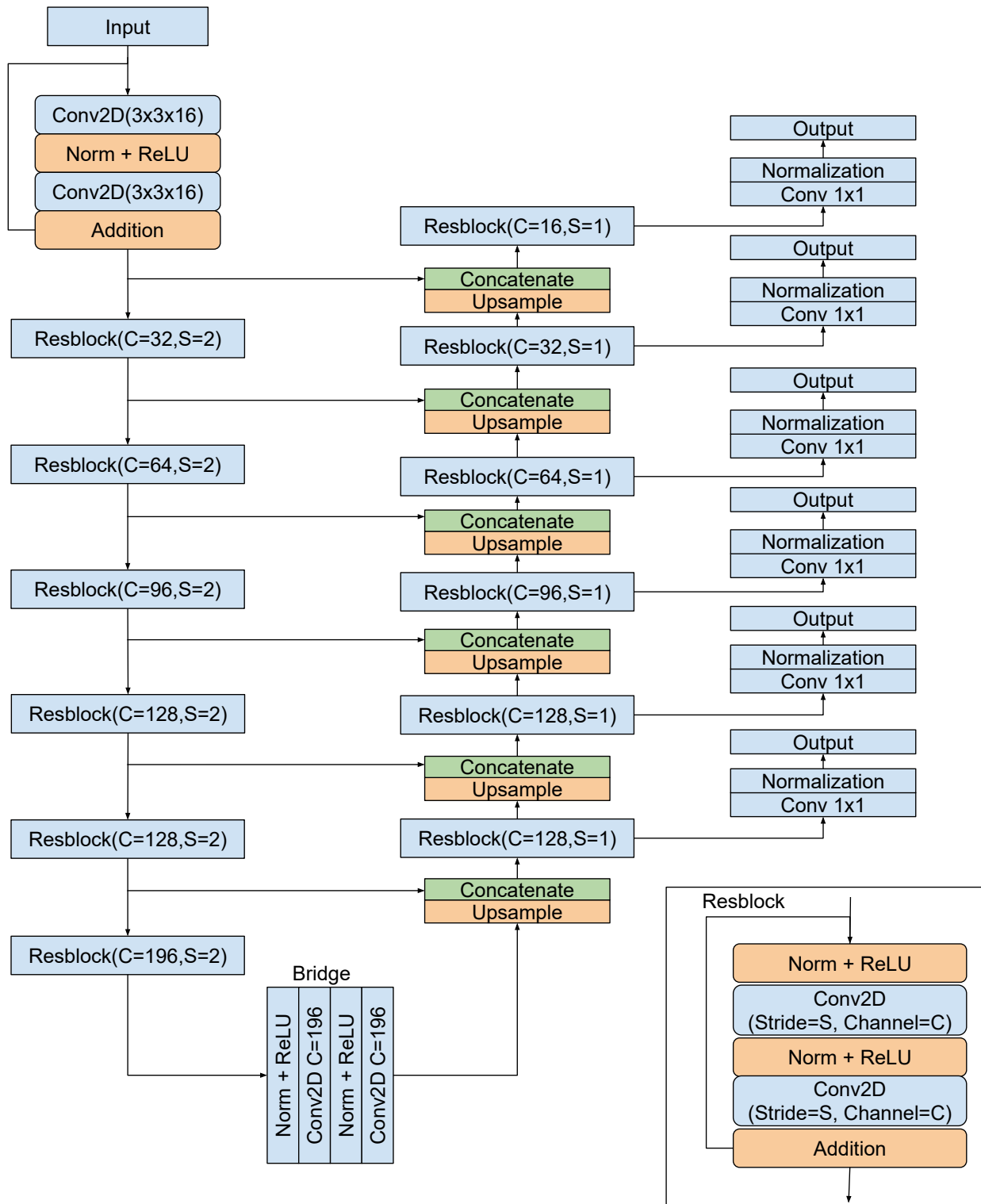
Figure B. Proposed architecture. Our method relies on a U-Net, with multiple ResNet blocks and skip connections. C and S are the channel number and stride for the convolution layers.
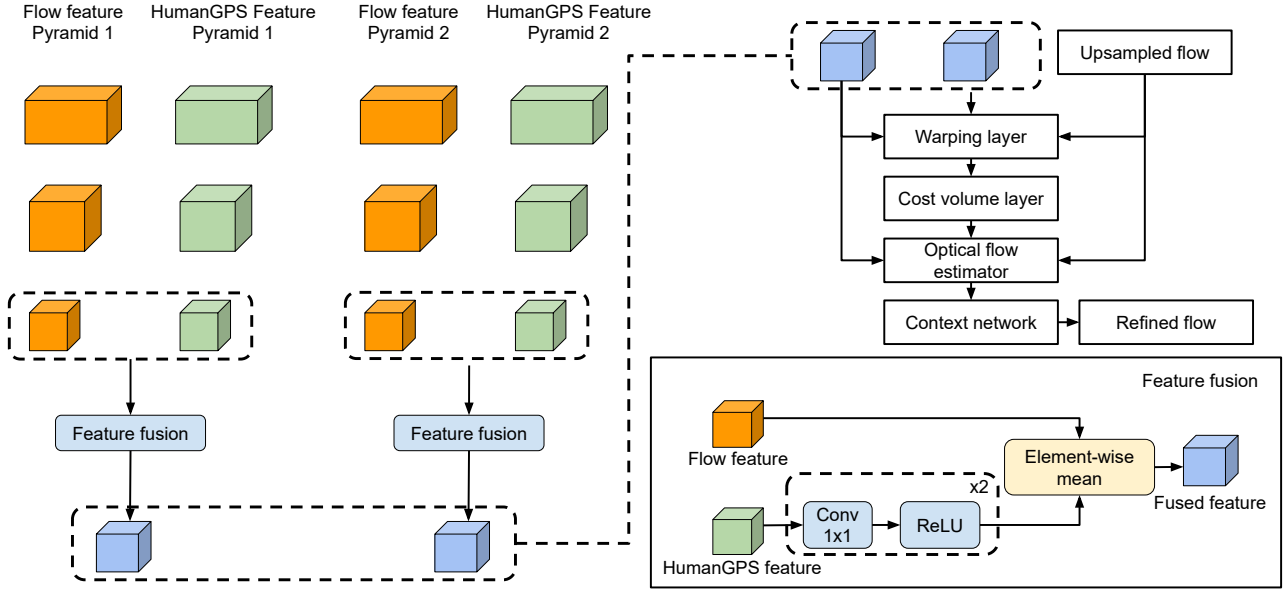
Figure C. Proposed end-to-end architecture for optical flow. We fuse our GPS feature with the original feature extractor from PWC-Net [25]. As showed in the main paper, this substantially improves the performance even when compared to a PWC-Net with a larger capacity.

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 1×1 convoluton 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 performs 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,
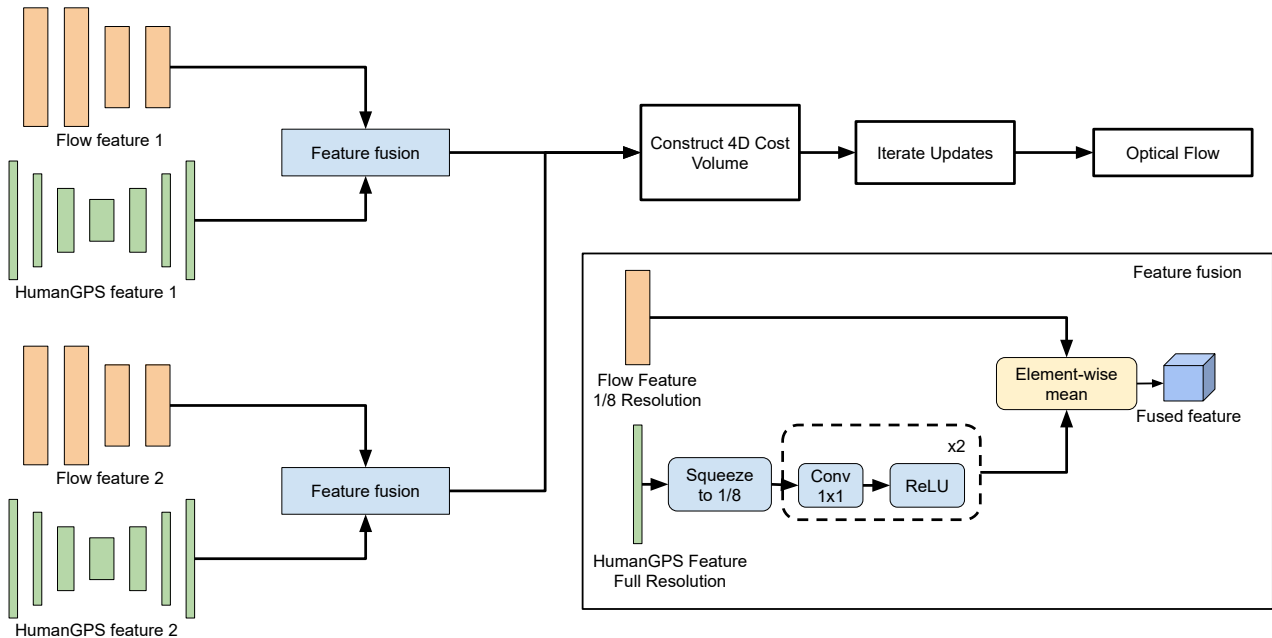
Figure D. Proposed end-to-end architecture for optical flow. We fuse our GPS feature with the original feature extractor from RAFT [26]. As showed in the main paper, this substantially improves the performance even when compared to a RAFT with a larger capacity.

our method generalizes well onto various real images in the wild. It 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.

5

| Frame 1 | Frame 2 | Wei et al. | SDC-Net | Ours | GT | Wei et al. | SDC-Net | Ours | GT Occlusion |

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.

| Frame 1 | Frame 2 | Wei et al. | SDC-Net | Ours | GT | Wei et al. | SDC-Net | Ours | GT Occlusion |

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.
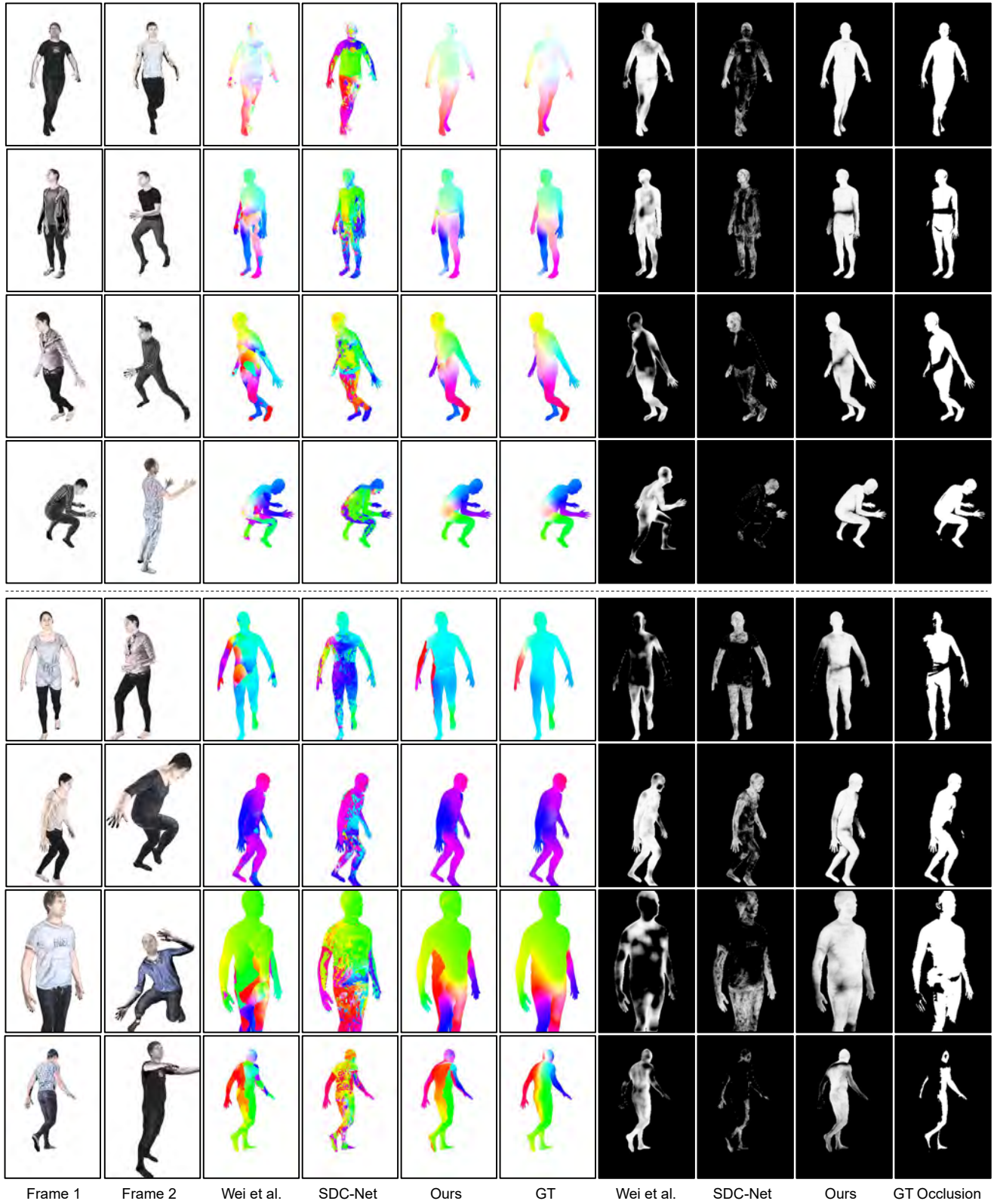
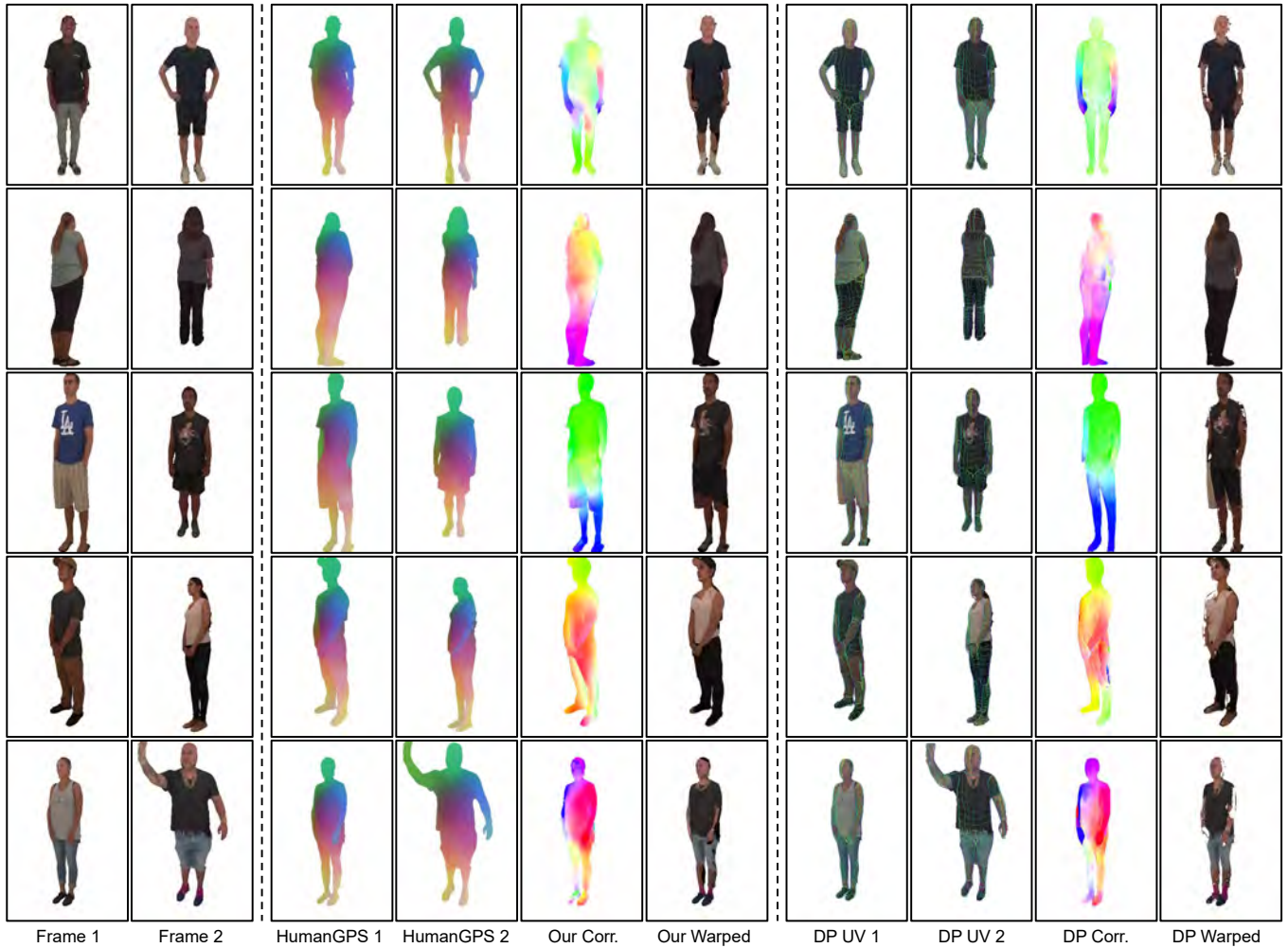| Frame 1 | Frame 2 | HumanGPS 1 | HumanGPS 2 | Our Corr. | Our Warped | DP UV 1 | DP UV 2 | DP Corr. | DP Warped |

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].

| Frame 1 | Frame 2 | GPS Feature 1 | GPS Feature 2 | Corr. | Warped Image | Frame 1 | Frame 2 | GPS Feature 1 | GPS Feature 2 | Corr. | Warped Image |

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			Inter-Subject
	SMPL [16]	The Relightables [10]	RenderPeople [3]	SMPL [16]
SDC-Net [24]	56.40	48.17	58.38	28.98
Wei *et al*. [28]	32.20	25.78	34.13	32.25
Ours	**71.20**	**56.08**	**67.65**	**69.33**
PWC-Net [25]	90.25	85.16	87.06	72.18
PWC-Net*	92.20	85.33	89.32	63.06
PWC-Net + GPS	**94.93**	**87.67**	**91.38**	**80.91**

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	DensePose [9, 19]			HumanGPS		
	5cm	10cm	20cm	5cm	10cm	20cm
ResNet-101 FCN [9]	43.05	65.23	74.17	49.09	73.12	84.51
ResNet-101 FCN* [9]	51.32	75.50	85.76	53.01	76.77	87.18
HG Stack-1 [19]	49.89	74.04	82.98	50.50	75.57	87.18
HG Stack-2 [19]	52.23	76.50	84.99	52.91	77.21	88.50
HG Stack-8 [19]	56.04	79.63	87.55	55.41	79.76	89.44

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.

Method	Finetune	AEPE
PWC-Net	No	0.2185
PWC-Net	Yes	0.2185
PWC-Net*	Yes	0.1411
PWC-Net + GPS	Yes	**0.1239**

Table C. Evaluate on Single-Human Optical Flow dataset (SHOF) [21]. Our method achieve the best performance over all. Please see the main paper for explanation of the model with *.
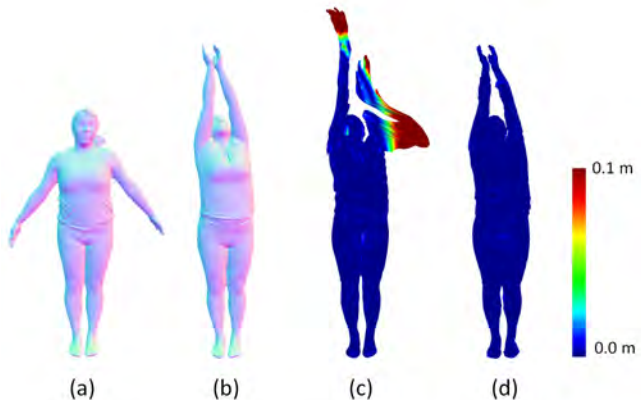

Figure I. Quantitative comparison of non-rigid tracking with learned correspondences. (a) reference geometry; (b) target geometry; non-rigid alignment without (c) and with (d) our learned correspondences. Surface errors are coded as per-pixel colors.

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-
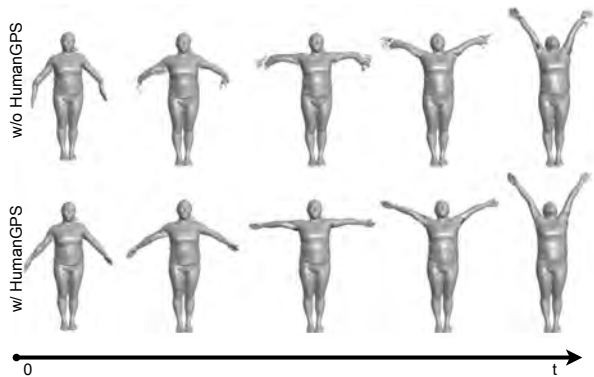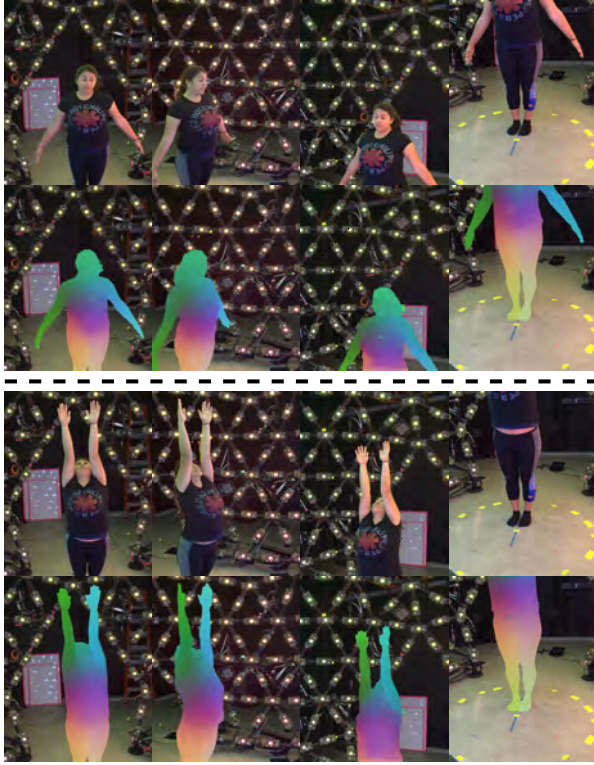
Figure J. Nonrigid Fusion Comparison. We improve the non-rigid tracking using correspondence extracted via our HumanGPS feature from the color image (top and middle). In the bottom, we show the fusion results without and with our correspondence. The standard dynamic fusion system fails quickly under fast motion, whereas successfully tracks the deformation with our correspondences.

tract reliable point to point correspondences, which are in turn used to refine the deformation in an iterative manner. Human body movements such as waving arms would easily break above requirement when performed too fast. Whereas high speed cameras [12, 11] could mitigate this behavior, here we show that HumanGPS is also an effective way to improve the results without the need of custom hardware.

The correspondences from GPS feature on color im-

ages provides additional constraints for the nonrigid tracking system, and it helps to rescue ICP failures. To demonstrate that, we provide correspondence built across 6 color images as additional initialization for the nonrigid deformation along with the ICP. The tracking algorithm takes a reference mesh and deforms it non-rigidly to align with a sequence of the target geometry. As shown in Figure I, the final target geometry (b) demonstrates large pose difference from the reference (a). Traditional non-rigid tracking algorithms (c) fail in such a case, while such large deformation can be correctly estimated benefiting from our learned correspondences (d).

In addition, non-rigid fusion algorithms such as Motion2Fusion [6] improve reconstruction accuracy with learned correspondences. This algorithm takes 6 depth images from different view point and runs non-rigid warping between canonical and live frames. The warping function is solved with additional constraints of learned correspondences from each view point 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 track the deformation with our correspondences.

Please see the supplementary webpage for video demo.

D.2. Morphing

Figure K. Morphing results based on HumanGPS. The leftmost and rightmost columns show two input images of intra- or inter-subjects. 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

[1] Carnegie-mellon mocap database. base.http://mocap.cs.cmu.edu/. 1

[2] Mixamo. https://www.mixamo.com/#/. 1

[3] Renderpeople. https://renderpeople.com/. 1, 2, 9

[4] Federica Bogo, Angjoo Kanazawa, Christoph Lassner, Peter Gehler, Javier Romero, and Michael J Black. Keep it smpl: Automatic estimation of 3d human pose and shape from a single image. In *European Conference on Computer Vision*, pages 561–578. Springer, 2016. 9

[5] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. *CoRR*, abs/1802.02611, 2018. 5

[6] 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. 9, 10

[7] 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. 9

[8] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 580–587, 2014. 5

[9] Rıza Alp Güler, Natalia Neverova, and Iasonas Kokkinos. Densepose: Dense human pose estimation in the wild. In *CVPR*, 2018. 2, 4, 8, 9

[10] 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. 1, 2, 9

[11] Kaiwen Guo, Jon Taylor, Sean Fanello, Andrea Tagliasacchi, Mingsong Dou, Philip Davidson, Adarsh Kowdle, and Shahram Izadi. Twinfusion: High framerate non-rigid fusion through fast correspondence tracking. In *3DV*, 2018. 10

[12] Adarsh Kowdle, Christoph Rhemann, Sean Fanello, Andrea Tagliasacchi, Jon Taylor, Philip Davidson, Mingsong Dou, Kaiwen Guo, Cem Keskin, Sameh Khamis, David Kim, Danhang Tang, Vladimir Tankovich, Julien Valentin, and

[13] 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. 11

[14] 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. 11

[15] Matthew Loper, Naureen Mahmood, and Michael J Black. Mosh: Motion and shape capture from sparse markers. *ACM Transactions on Graphics (TOG)*, 33(6):1–13, 2014. 1

[16] 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. 1, 9

[17] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J. Black. SMPL: A skinned multi-person linear model. *SIGGRAPH Asia*, 2015. 2

[18] Joseph SB Mitchell, David M Mount, and Christos H Papadimitriou. The discrete geodesic problem. *SIAM Journal on Computing*, 16(4):647–668, 1987. 1

[19] 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. 4, 9

[20] 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. 9

[21] Anurag Ranjan, David T Hoffmann, Dimitrios Tzionas, Siyu Tang, Javier Romero, and Michael J Black. Learning multi-human optical flow. *International Journal of Computer Vision*, pages 1–18, 2020. 1, 5, 9

[22] Anurag Ranjan, Javier Romero, and Michael J Black. Learning human optical flow. *arXiv preprint arXiv:1806.05666*, 2018. 1, 5

[23] Kathleen M Robinette, Sherri Blackwell, Hein Daanen, Mark Boehmer, and Scott Fleming. Civilian american and european surface anthropometry resource (caesar), final report. volume 1. summary. Technical report, SYTRONICS INC DAYTON OH, 2002. 1

[24] 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. 4, 5, 6, 7, 9

[25] 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. 4, 5, 9

[26] Zachary Teed and Jia Deng. Raft: Recurrent all-pairs field transforms for optical flow. *arXiv preprint arXiv:2003.12039*, 2020. 5

Shahram Izadi. The need 4 speed in real-time dense visual tracking. *SIGGRAPH Asia*, 2018. 10

[27] Gül Varol, Javier Romero, Xavier Martin, Naureen Mahmood, Michael J. Black, Ivan Laptev, and Cordelia Schmid. Learning from synthetic humans. In *CVPR*, 2017. 1

[28] 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. 4, 5, 6, 7, 9