DAViD: Data-efficient and Accurate Vision Models from Synthetic Data
Supplementary Materials

Fatemeh Saleh Sadegh Aliakbarian Charlie Hewitt Lohit Petikam Xiao-Xian
Antonio Criminisi Thomas J. Cashman Tadas Baltrušaitis

Microsoft, Cambridge, UK

In this supplementary material, we provide additional details on the data rendering and implementation of our method. We also provide additional qualitative and quantitative results. We encourage the readers to watch the supplementary video that contains additional results.

1. Synthetic Data

As described in the main paper, we use the data generation pipeline of Hewitt et al. [2], incorporating the updated face model of Petikam et al. [8], to create SynthHuman. We extend this data generation pipeline for dense prediction tasks. Specifically, we make two main changes: re-defining the hair surface normals as well as re-defining the ground-truth depth and surface normals for transparent surfaces. Below, we delve into details of these changes.

Beyond these additional output streams, in SynthHuman we update the sampling procedure to increase the number unique identities and incorporate more diverse poses, lighting, and camera views. Specifically, we sample face/body shape (from training sources and a library of 3572 scans), expression and pose (from AMASS [7], MANO [10], and more), texture (from high-res face scans with expression-based dynamic wrinkle maps blended in), hair (548 strand-level 3D hair, each with 100K+ strands), accessories (36 glasses, 57 headwear), 50 clothing tops, and environment (a mix of HDRIs and 3D environments).

1.1. Hair Surface Normals

In scan-based synthetic data, e.g., RenderPeople[1], ground-truth (GT) hair surface normals are obtained by renderings of scanned 3D human models. These scans represent hair with a coarse surface mesh. In our synthetic data we explicitly represent hair as hundreds of thousands of individual 3D strands, enabling generation of GT depth, normals, alpha, etc. with strand-level granularity. While dense strand-based 3D hair is a high-fidelity representation, when rendered from a portrait view they produce extremely high-frequency surface normals that appear noisy due to aliasing (See Fig. 1a). For generating our ground-truth surface normals, we redefine our hair strand normals to align closer to the coarse hair mesh surface normals of THuman2.1 [12] and Renderpeople [1], in which the hair normals better represent the coarse shapes of hair clumps and volumes rather than individual strands.

We wish to generate hair surface normal images with the interpretablility of Sapiens [4] hair normal training data, but without reducing the fidelity of our strand-based hair representation. We first generate a voxel-grid volume with density based on the strand geometry that occupies the voxel. Using marching cubes we convert the volume to a coarse proxy mesh that approximates the combined hair strands (Fig. 1b) with interpretable normal vectors. The proxy mesh does not capture fine-scale fly-away hair strand detail so we only use it to sample normal vectors. For a point on a strand of our synthetic hair (head hair, facial hair, eyebrows, and eyelashes), we render the normal vector of the nearest proxy mesh surface which is smooth across the pixel grid, rather than the strand normals themselves which are noisy between pixels. We render all hair strands this way to preserve the fidelity of our synthetic hair representation while generating normals representing the coarse shapes of the hair style (Fig. 1c).

1.2. Ground-truth depth and normals of transparent surfaces.

The predictions we show throughout this paper ignore the depth and normals of translucent surfaces like the lenses of glasses, instead predicting the depth and normals of the opaque surface visible behind the translucent media. For different applications we can control this behavior by choosing either to render the depth and normals of translucent surfaces or ignore them when generating our synthetic training images, as shown in Fig. 2.

(a) Naïve hair normals rendering (b) Generated hair proxy mesh (c) Our hair normals transferred from proxy mesh

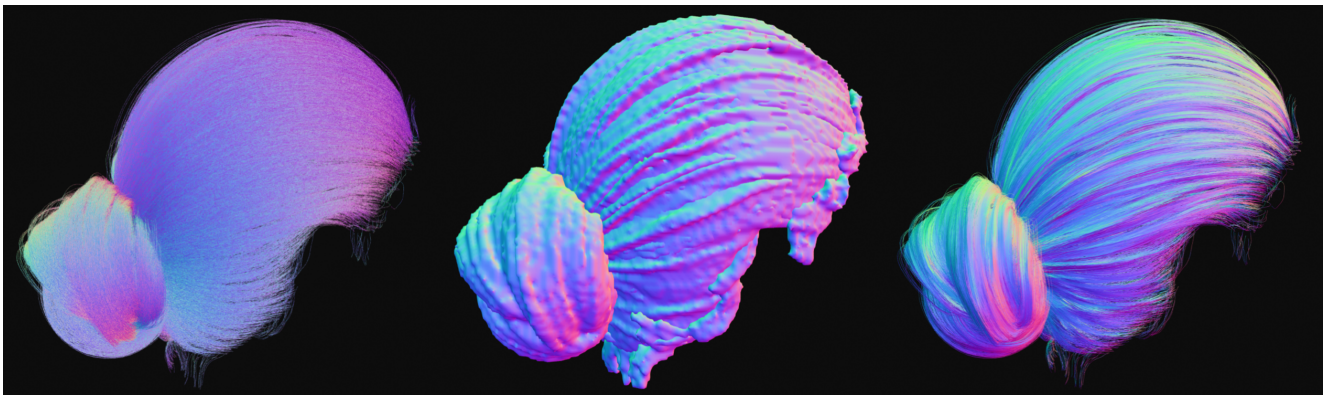

Figure 1. We generate interpretable strand-level synthetic hair normal GT training images by sampling normal directions from a proxy mesh representing the shape of the hair.

(a) RGB training image (b) GT including translucent surfaces (c) GT ignoring translucent surfaces

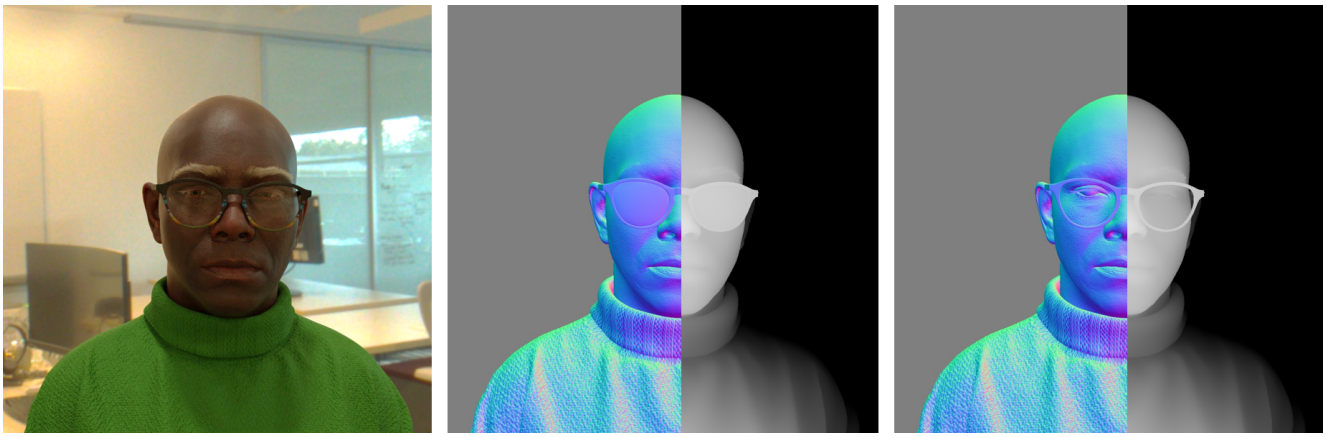

Figure 2. For different applications, we control how translucent surfaces are depicted in our generated normal and depth training images.

2. Experiments

2.1. Surface normal ground truth

Creating accurate surface normal annotations is very challenging for real data. Most approaches rely on photogrammetry or reconstruction of relatively coarse surface meshes. Both of the above approaches struggle with reconstructing thin or high frequency structures such as hair or folds in clothing. They also struggle reconstructing the area around the eyes both due to thin structures (eyelashes), poor lighting due to self shadowing, and reflective surface of the eyeball. This makes evaluating approaches that can capture such subtle details challenging as we may be seeing ceiling effect in results.

To demonstrate this we perform an experiment with taking the output of our surface normals models and blurring it using Gaussian Blur to reduce the fidelity of the output, rather than degrading the results this improves them on all metrics on the Goliath dataset. This indicates that the ability to evaluate our models is hindered by quality of the annotations.

2.2. Additional results for soft foreground segmentation.

In Tab. 2 we additionally show our soft foreground segmentation results on the two validation sets of the P3M dataset [5]. While trained solely on synthetic data, our model achieves high accuracy on this challenging dataset. However, discrepancies arise due to differences in how the ground-truth alpha is obtained in our synthetic data compared to the P3M dataset, as well as variations in defining the most dominant human subjects in the scene, objects in hand, and other factors. This makes a fair comparison with methods trained on the P3M training set difficult. To ensure a fair comparison,

Table 1. Surface normal estimation using base model and blurring the output. Note that blurring results of our model leads to an increase in accuracy across all metrics, while blurring the output of Sapiens-0.3B makes little difference.

Method	Goliath-Face			Goliath-UpperBody			Goliath-FullBody		
	Angular Error (°) ↓		% Within t° ↑	Angular Error (°) ↓		% Within t° ↑	Angular Error (°) ↓		% Within t° ↑
	Mean	Median	11.25° / 22.5° / 30°	Mean	Median	11.25° / 22.5° / 30°	Mean	Median	11.25° / 22.5° / 30°
Ours-Large	17.15	12.19	48.4 / 76.3 / 84.7	13.96	11.23	50.7 / 84.2 / 92.1	14.60	11.66	48.7 / 82.2 / 90.8
Ours with blur	**17.12**	**12.16**	**48.5 / 76.4 / 84.7**	**13.88**	**11.19**	**50.9 / 84.4 / 92.2**	**14.52**	**11.61**	**49.0 / 82.3 / 90.9**
Sapiens-0.3B	18.86	14.47	42.6 / 71.2 / 81.3	12.54	10.42	56.2 / 88.0 / 94.6	15.72	13.03	43.1 / 79.2 / 89.4
Sapiens-0.3B with blur	**18.84**	14.47	42.6 / 71.2 / 81.3	**12.51**	**10.40**	**56.3** / 88.0 / 94.6	**15.69**	13.03	43.1 / 79.2 / 89.4

Table 2. Evaluating soft foreground segmentation. Methods indicated by (*) are trained on the P3M training set.

Method	P3M-500-NP			P3M-500-P		
	SAD	SAD-T	Conn	SAD	SAD-T	Conn
Zhong et al.* [13]	10.60	**6.83**	9.77	10.04	**6.44**	9.41
BGMv2* [6]	15.66	7.72	14.65	13.90	7.23	13.13
P3M-Net* [5]	11.23	7.65	12.51	8.73	6.89	13.88
MODNet [3]	20.20	12.48	18.41	30.08	12.22	28.61
Ours (trained on SynthHuman)	14.83	10.23	14.76	12.65	9.19	12.47
Ours* (trained on P3M-train)	12.30	9.46	12.14	11.48	8.29	11.35
Ours* (trained on SynthHuman + by finetuned on P3M-train)	**9.12**	8.01	**8.94**	**8.05**	7.04	**7.90**

Table 3. Evaluating depth estimation on THuman2.1 dataset. The results for Sapiens models indicated by (*) are re-evaluated on our rendered THuman2.1 evaluation subset, using exactly the same settings as in [4], except for the HDRIs, which may differ.

Method	TH2.0-Face			TH2.0-UprBody			TH2.0-FullBody		
	RMSE	AbsRel	δ_1	RMSE	AbsRel	δ_1	RMSE	AbsRel	δ_1
MiDaS-L [9]	0.114	0.097	0.925	0.398	0.271	0.868	0.701	0.689	0.782
MiDaS-Swin2 [9]	0.050	0.036	0.995	0.122	0.081	0.948	0.292	0.171	0.862
DepthAny-B[11]	0.039	0.026	0.999	0.048	0.028	0.999	0.061	0.030	0.999
DepthAny-L[11]	0.039	0.027	0.999	0.048	0.027	0.999	0.060	0.030	0.999
Sapiens-0.3B[4]	0.012	0.008	1.000	0.015	0.009	1.000	0.021	0.010	1.000
Sapiens-2B [4]	0.008	0.005	1.000	0.010	0.006	1.000	0.016	0.008	1.000
Sapiens-0.3B*	0.008	0.005	1.000	0.011	0.006	1.000	0.016	0.007	1.000
Sapiens-2B*	0.007	0.004	1.000	0.009	0.005	1.000	0.014	0.007	1.000
Ours (trained on SynthHuman)	0.014	0.009	1.000	0.017	0.010	1.000	0.024	0.011	1.000
Ours (trained on Thuman2.1)	0.010	0.006	1.000	0.013	0.007	1.000	0.022	0.010	1.000
Ours (trained on SynthHuman + by finetuned on Thuman2.1)	0.008	0.005	1.000	0.012	0.006	1.000	0.018	0.008	1.000

we conduct additional experiments. First, instead of training on SynthHuman, we train our model on P3M training subset. This shows that training on a dataset wherein ground-truth definitions match the test scenario is effective. In another experiment, we fine-tune our model, initially trained on SynthHuman, on the P3M training subset. By starting from a good initial weights (from our synthetic data), we show that fine-tuning on P3M and fixing the mismatches in the definition of foreground region is more effective, leading to the state-of-the-art results on most metrics.

2.3. Additional results for depth estimation.

Tab. 3 summarizes our results on the THuman2.1 dataset [12]. Following [4], this synthetic dataset is rendered by placing THuman2.1 scans in HDRI environments. While we argue such synthetic data can act as a good resource for training, we do not consider them an ideal test benchmark. However, for completeness, we report our results on this dataset. Following [4], we select 526 human scans from the THuman2.1 dataset and render 1,578 images to form our evaluation set. We observe that Sapiens [4] achieves particularly strong results on this dataset, likely due to the close resemblance between THuman2.1 and RenderPeople which is used for their finetuning step. Our model, trained solely on SynthHuman dataset, also performs reasonably well on THuman2.1. However, we identify a significant difference between the quality of the rendered RGB images and depth ground-truth of THuman2.1 and those of SynthHuman. Particularly, as illustrated in Fig. 4 of the main paper, coarse and noisy scans of THuman2.1 lead to unrealistic RGB images and noisy ground-truth. To further analyze this, we utilize the remaining THuman2.1 scans to create a training set (∼100k samples), rendered by placing a virtual camera around the scans placed in HDRI environments. Fine-tuning our depth model (initially trained on SynthHuman) on this additional data for only 25 epochs allows us to achieve on-par results with Sapiens. This shows that the difference in performance is primarily due to domain adaptation rather than inherent model capability.

2.4. Remark on Resizer.

In our method, we use the Resizer module to handle any resolution while running the ViT encoder on the fixed-size version of the image (384×384). While we use the resolution of 512×512 (with 512 pixels being the height of SynthHuman images) for all the experiments in this paper, Resizer module allows us to make predictions at higher resolution. In Fig. 3, we show the output of the model when tested with input images of size 512×512 versus 1024×1024 (after padding to make square, if needed). We noticed that while still performing very fast, larger input resolution provides the model with far more details for all tasks.

2.5. Applications of Dense Prediction Tasks

In this section, we provide potential downstream applications for the dense prediction tasks we addressed in this paper. Particularly, we use our surface normal estimation model for a simple relighting. We demonstrate how we can use our

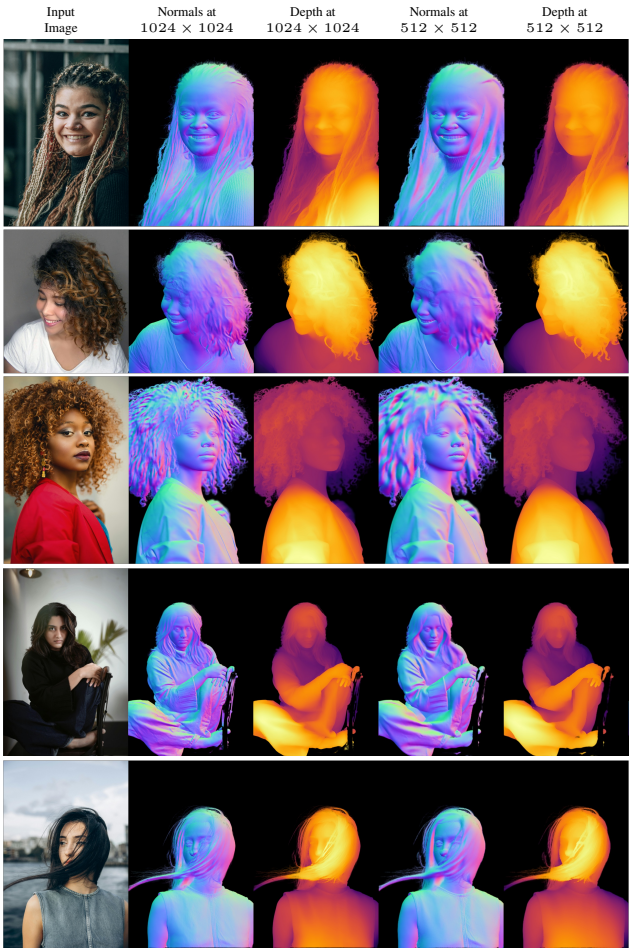

Figure 3. The Resizer module allows us to use arbitrary input size at test time. Higher resolution input provides more details to the model, thus it can capture more details in the depth and surface normals predictions.

depth estimation model to generate a 2.5D representation from a single image. And finally, we show that our soft foreground prediction model can be used for background replacement (e.g., in video conferencing).

Simple relighting from Normals. As a potential downstream application, we use our normal estimation model in a relighting pipeline to re-render images under novel lighting conditions. To this end, we first predict a surface normal map for an input image. This predicted normals, which capture fine geometric details, serve as the foundation for our relighting process. For a given image, we compute per-pixel shading based on a Lambertian reflectance model where the intensity is modulated by the cosine of the angle between each predicted normal and an externally specified light direction. To further enhance realism, we incorporate an ambient term, ensuring that areas not directly illuminated still receive a baseline level of light. As illustrated in Fig. 4, this re-rendering approach produces a visually plausible approximation of how the scene would appear under different lighting conditions.

2.5D representation from depth. We further demonstrate that our relative depth estimation model is capable of estimating the 2.5D representation of a given image. For a given image, the estimated depth map is then unnormalized using a reasonable guess of a range, which we use to generate a 3D point cloud of the visible scene. By rendering this point cloud from multiple novel viewpoints, as illustrated in Fig. 5, we demonstrate that our model captures challenging depth relations with remarkable fidelity. For example, the reconstructed geometry preserves correct facial proportions, clearly positions a hand in front of the body, and accurately depicts the shape of a hat on the head. These results illustrate that our relative depth model reliably encodes fine-grained depth cues, enabling effective 2.5D reconstruction from a single image.

Background replacement from segmentation. In addition to its primary role in supporting dense prediction tasks, our soft foreground segmentation model serves as a robust standalone solution for applications that require precise subject extraction. For example, as shown in Fig. 6, our approach enables reliable background replacement, which is particularly valuable for video conferencing. By accurately separating the human subject and preserving fine details such as hair strands, our model ensures high-quality background substitution, demonstrating its effectiveness in real-world scenarios.

2.6. Implementation Details

During training, we apply various augmentations to enhance model robustness. For geometric transformations, we use random scaling to simulate zooming in or out of the image and its corresponding ground truth. Additionally, random shift augmentation is applied to simulate the shifting of ROI in both the image and GT. For appearance augmentations, we apply random blurring to the image, with the blur strength proportional to the image size, simulating lenses with poor modulation transfer function (MTF).We adjust image brightness by adding a constant offset within a specified range and adjust the contrast using the formula:

$$\text{img} = (\text{img} - 0.5)(1 + \text{contrast}) + 0.5$$

Additionally, we randomly alter the hue and saturation, apply JPEG compression, and occasionally convert the image from BGR to greyscale. These appearance augmentations are applied with a specified probability. Following Hewitt et al. [2], we also introduce random ISO noise, inspired by

Figure 4. Examples of simple relighting using surface normals predicted by our model on in-the-wild data.

Figure 5. Results of our depth prediction model on in-the-wild images rendered as a point cloud from different viewpoints.

real camera noise, to enhance training. This noise is a combination of image intensity-dependent Poissonian noise and intensity-independent Gaussian noise.

2.7. Goliath Test Set

Tab. 4 gives the frame and camera indices which are used for selecting and rendering ground truth for the evaluation set used in our work. We render the normal and depth images at 667×1024 resolution using Blender.

2.8. Additional Qualitative Results

In this section, we provide additional qualitative results of our approach and compare them with Sapiens-2B models in Fig. 7.

References

[1] Renderpeople GmbH. Renderpeople. 1

[2] Charlie Hewitt, Fatemeh Saleh, Sadegh Aliakbarian, Lohit Petikam, Shideh Rezaeifar, Louis Florentin, Zafiirah Hose-

Figure 6. Background replacement demonstrated using results from our matting model on in-the-wild images.

Subset	Camera IDs	Subject	Frame IDs
Face	*401650*, 401645, 401655, 401894, *401962*, 402601, 402792, 402807, 402871, 402875, 402980, 403072	AXE977	02858, 13148, 23438, 28085, 29114, 34733, 49044, 62745, 75355, 87055, 99319, 110328, 121299, 132449, 139288, 140317
		QZX685	03339, 13089, 22839, 28404, 29379, 30354, 46874, 62806, 74953, 85733, 97027, 107481, 119069, 131633, 132608, 133583
		XKT970	03178, 12868, 22558, 28225, 29194, 30163, 37300, 53338, 66207, 77489, 88184, 98424, 108787, 119398, 124264, 125233
		QVC422	03280, 13990, 24730, 28636, 29707, 30778, 31849, 33856, 52762, 69555, 82046, 93762, 105020, 116706, 123621, 124692
Upper Body	401541, 400874, 400883, 400894, 400895, 400898, 400926, 400929, 400933, 400934, 400936, 401534	AXE977	00202, 02944, 05686, 08428, 11170, 13261, 14175, 22719, 25761, 28654, 31695, 34739, 37780, 40673, 43714, 46757
		QZX685	00227, 02981, 05735, 08489, 11243, 13544, 14462, 22813, 25868, 28773, 31825, 34881, 37935, 40838, 43890, 46944
		XKT970	00313, 03049, 05785, 08521, 11257, 13358, 14270, 22906, 25941, 28827, 31863, 34900, 37936, 40822, 43857, 46892
		QVC422	00207, 02913, 05619, 08325, 11031, 13150, 14052, 22493, 25498, 28354, 31362, 34368, 37373, 40229, 43236, 46242
Full Body	401156, 401150, 401185, 401191, 402359, 402401, 402432, 402435, 402547, 402551, 402636, 402689	AXE977	00202, 02944, 05686, 08428, 11170, 13261, 14175, 22719, 25761, 28654, 31695, 34739, 37780, 40673, 43714, 46757
		QZX685	00227, 02981, 05735, 08489, 11243, 13544, 14462, 22813, 25868, 28773, 31825, 34881, 37935, 40838, 43890, 46944
		XKT970	00313, 03049, 05785, 08521, 11257, 13358, 14270, 22906, 25941, 28827, 31863, 34900, 37936, 40822, 43857, 46892
		QVC422	00207, 02913, 05619, 08325, 11031, 13150, 14052, 22493, 25498, 28354, 31362, 34368, 37373, 40229, 43236, 46242

Table 4. Goliath evaluation set camera and frame selection. There are 12 cameras per subset and 16 frames per camera. Note 401650 is missing for calibration for subject XKT970 and 401962 is missing calibration for subject QZX685, so in total there are 2272 images.

nie, Thomas J Cashman, Julien Valentin, Darren Cosker, and Tadas Baltrušaitis. Look ma, no markers: holistic performance capture without the hassle. *ACM Transactions on Graphics (TOG)*, 36(6), 2024. 1, 4

[3] Zhanghan Ke, Jiayu Sun, Kaican Li, Qiong Yan, and Rynson WH Lau. Modnet: Real-time trimap-free portrait matting via objective decomposition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 1140–1147, 2022. 3

[4] Rawal Khirodkar, Timur Bagautdinov, Julieta Martinez, Su Zhaoen, Austin James, Peter Selednik, Stuart Anderson, and Shunsuke Saito. Sapiens: Foundation for human vision models. In *European Conference on Computer Vision*, pages 206–228. Springer, 2024. 1, 3, 7

[5] Jizhizi Li, Sihan Ma, Jing Zhang, and Dacheng Tao. Privacy-preserving portrait matting. In *Proceedings of the 29th ACM international conference on multimedia*, pages 3501–3509, 2021. 2, 3

[6] Shanchuan Lin, Andrey Ryabtsev, Soumyadip Sengupta, Brian L Curless, Steven M Seitz, and Ira Kemelmacher-Shlizerman. Real-time high-resolution background matting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8762–8771, 2021. 3

[7] Naureen Mahmood, Nima Ghorbani, Nikolaus F Troje, Ger-

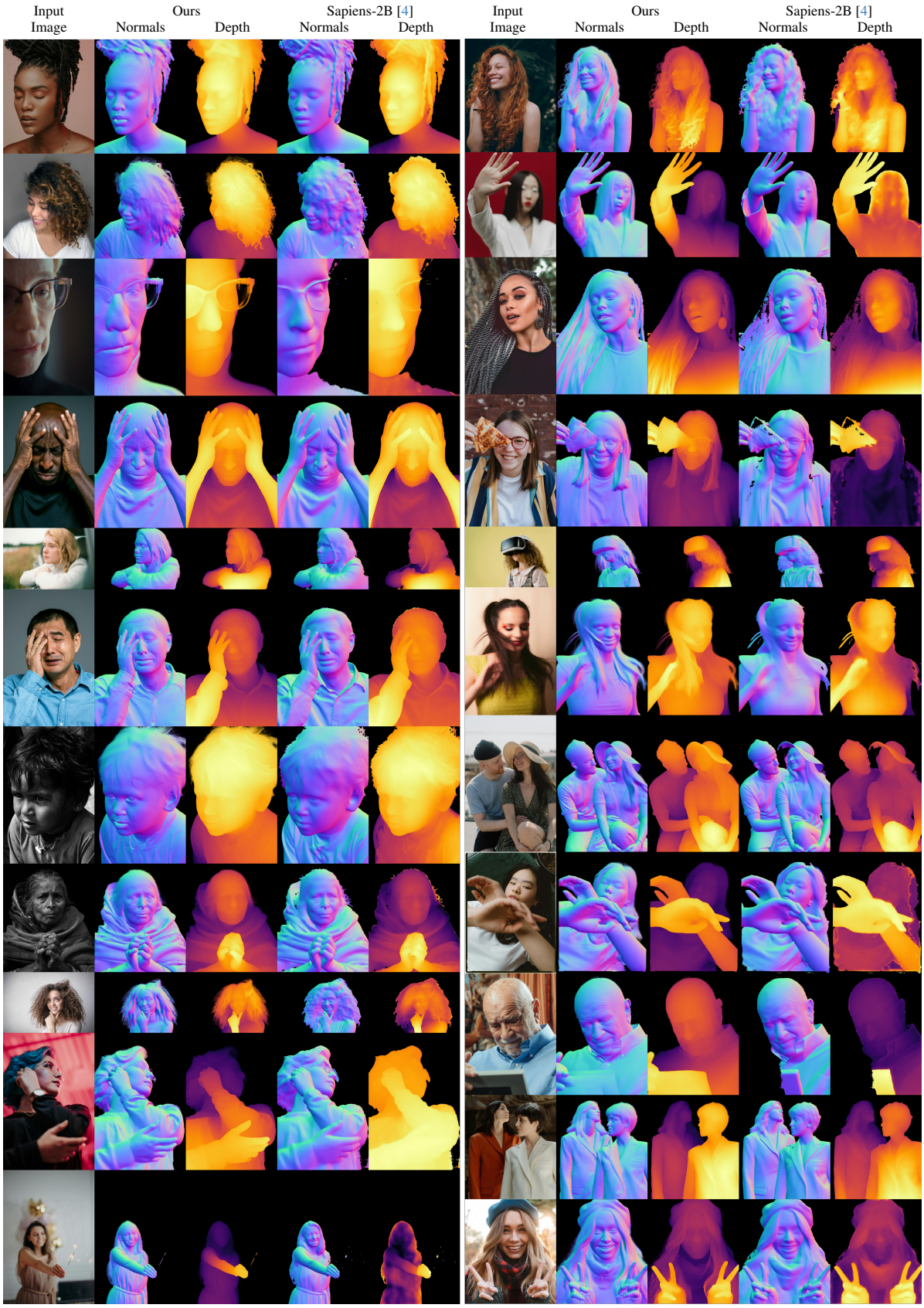

Figure 7. Additional qualitative comparisons.

ard Pons-Moll, and Michael J Black. Amass: Archive of motion capture as surface shapes. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 5442–5451, 2019. 1

[8] Lohit Petikam, Charlie Hewitt, Fatemeh Saleh, and Tadas Baltrušaitis. Eyelid fold consistency in facial modeling. In *SIGGRAPH Asia 2024 Technical Communications*, pages 1–4. Association for Computing Machinery, 2024. 1

[9] René Ranftl, Katrin Lasinger, David Hafner, Konrad Schindler, and Vladlen Koltun. Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer. *IEEE transactions on pattern analysis and machine intelligence*, 44(3):1623–1637, 2020. 3

[10] Javier Romero, Dimitrios Tzionas, and Michael J Black. Embodied hands: modeling and capturing hands and bodies together. *ACM Transactions on Graphics (TOG)*, 36(6):1–17, 2017. 1

[11] Lihe Yang, Bingyi Kang, Zilong Huang, Zhen Zhao, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything v2. *Advances in Neural Information Processing Systems*, 37: 21875–21911, 2025. 3

[12] Tao Yu, Zerong Zheng, Kaiwen Guo, Pengpeng Liu, Qionghai Dai, and Yebin Liu. Function4d: Real-time human volumetric capture from very sparse consumer rgbd sensors. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR2021)*, 2021. 1, 3

[13] Yatao Zhong and Ilya Zharkov. Lightweight portrait matting via regional attention and refinement. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 4158–4167, 2024. 3