This ICCV paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

DAViD: Data-efficient and Accurate Vision Models from Synthetic Data*

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

Microsoft, Cambridge, UK



Figure 1. Given a single, real image of a person, our human-centric models, trained entirely on synthetic data, predict accurate relative depth, surface normals, and soft foreground segmentation. Please zoom in to see details such as hair strands, eye glasses and clothes folds.

Abstract

The state of the art in human-centric computer vision achieves high accuracy and robustness across a diverse range of tasks. The most effective models in this domain have billions of parameters, thus requiring extremely large datasets, expensive training regimes, and compute-intensive inference. In this paper, we demonstrate that it is possible to train models on much smaller but high-fidelity synthetic datasets, with no loss in accuracy and higher efficiency. Using synthetic training *data provides us with excellent levels of detail and perfect labels, while providing strong guarantees for data provenance, usage rights, and user consent. Procedural data synthesis also provides us with explicit control on data diversity, that we can use to address unfairness in the models we train. Extensive quantitative assessment on* real *input images demonstrates accuracy of our models on three dense prediction tasks: depth estimation, surface normal estimation, and soft foreground segmentation. Our models require only a fraction of the cost of training and inference when compared with foundational models of similar accuracy. Our human-centric synthetic dataset and trained models are available at* https://aka.ms/DAViD.

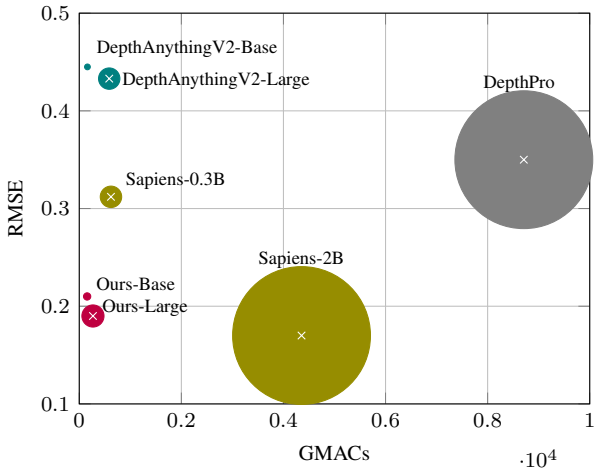


Figure 2. Compute cost vs error, comparing our method with state-of-the-art depth estimation models. Compute cost and error are measured with giga-multiply-accumulate count (GMACs) and root-mean-squared error (RMSE), respectively, on the combination of Goliath [25] and Hi4D [47] datasets. The radius of each marker is proportional to the number of model parameters. The most efficient and accurate models are in the lower-left corner.

1. Introduction

Progress in human-centric computer vision has been driven in large part by advances in data. This is both due to the

*DAViD also references Michelangelo's David—an iconic symbol of anatomical precision—and the David vs. Goliath story, reflecting our small yet powerful dataset and models.

scale and diversity of available data [15, 17, 20, 45] and the quality of annotations [14, 35]. Some types of ground-truth labels can be annotated by humans (e.g., landmarks [14, 35], coarse semantic classes [21, 40] and bounding boxes [45]). However, labels such as per-pixel depth, normals, or dense landmarks are significantly more challenging or even impossible for humans to annotate. Gathering such annotations often relies on complex camera rigs [3, 4, 24, 25, 38], or specialist sensors [26]. This leads to imperfect ground truth annotations, as they are derived from photogrammetry or noisy sensors. Further, in-lab captures significantly limit diversity of subjects and environments, as it is extremely challenging to capture truly in-the-wild data for such tasks. Training only on such datasets leads to models that produce coarse or inaccurate predictions, and which struggle to generalize outside the domain of the collected data [1, 17, 46].

In order to satisfy requirements for scale, diversity and high fidelity of annotations, recent approaches rely on large quantities of diverse data and a smaller amount of annotated data [2, 17, 46]. These techniques typically follow a two-stage approach: first, large-scale pretraining on real data with no or lower-quality ground truth, followed by fine-tuning on data with high-quality ground-truth annotations. These methods show good accuracy, but come at a considerable computational model training cost, and require complex multi-stage training. Finally, the accuracy of such methods is limited by the quality of the data used for fine-tuning. For example, Sapiens [17] relies on coarse synthetic data, and struggles to capture fine details such as facial wrinkles, eyelids, or subtle texture variations in clothing (see Fig. 4 for ground truth quality and Fig. 6 for qualitative results).

Instead, we propose to tackle both diversity and fidelity of training data through the use of procedurally-generated synthetic data [11]. We demonstrate that a single high-fidelity dataset is sufficient to tackle multiple dense prediction tasks and achieve state-of-the-art accuracy. Our approach requires a fraction of the data size, model size, computational complexity, and training time of competing approaches, all without sacrificing model accuracy on challenging cross-dataset evaluations (see Fig. 2). We demonstrate this on three challenging dense prediction tasks: relative depth estimation, surface normal estimation, and soft foreground segmentation, with our models capturing subtle details, handling thin structures, and maintaining accurate human proportions.

Our approach is different from techniques such as Depth-Pro [2], DepthAnything-v2 [44], and Sapiens [17], which either develop large, task-specific models, employ complex training regimes, or rely on large-scale data collections. We use a single architecture and a single dataset to tackle all three tasks. Importantly, training on synthetic data alone allows us to verify compliance with privacy, copyright, licensing, consent and diversity requirements, which would be more challenging to achieve with large datasets of real images. The core contribution of this paper is to demonstrate a fundamentally more *efficient* paradigm for human-centric vision. Our work demonstrates that it is possible to train performant and state-of-the-art human-centric models in a fraction of the time and on a fraction of data by relying solely on high-quality synthetic data. Details of how to access the SynthHuman dataset and trained models are available on the project website: `https://aka.ms/DAViD`.

2. Related Work

Human vision data. The availability of high-quality training data has boosted accuracy of recent computer vision models [2, 6, 27, 31], with no exception for human-centric tasks [1, 21, 45]. This is especially true for face detection [45], pose estimation [1], landmark localization [35], and semantic segmentation [21], where manual annotation is feasible with current tools and methodologies [14, 20, 36]. However, obtaining pixel-wise annotations manually for tasks such as matting, depth and surface normals is much harder [2, 5, 16, 44]. To alleviate this, some approaches have relied on curated multi-view real-image datasets to reconstruct human meshes [3, 25, 47]. While providing rich annotations, these datasets are limited in subject and environment diversity, due to the high costs of data collection. Further, as they rely on model-fitting or photogrammetry, they struggle with very thin structures like hair, reflective or semi-transparent surfaces like glasses and eyes, and are not able to capture high-frequency details (see Fig. 4). Our procedural synthetic data generation pipeline allows us to create data that is both diverse and has pixel-perfect labels.

Synthetic training datasets. Synthetic data has emerged as an alternative to overcome the annotation bottleneck in human-centric vision tasks. Early efforts focused on rendering pre-defined 3D human meshes acquired through photogrammetry [10, 28, 48]. While allowing for automatic dense annotations, the resulting data is limited by lack of reflective objects (e.g., glasses) and the quality of meshes, which are often low-fidelity, especially around hair, eyes, and digits. Procedural synthetic data can provide improved fidelity and diversity. For example, Wood et al. [41] demonstrated how a procedural synthetic data pipeline can be used to train facial landmark detection and face parsing models. BEDLAM [1] offers a full-body synthetic pipeline, featuring clothed subjects captured in diverse lighting environments. Built on the SMPL-X body model [29], BEDLAM introduces variability in body shape and pose, however it lacks high-fidelity faces, hair, and mesh-based environments. Our work builds upon the synthetic data pipeline of Hewitt et al. [11], and allows for high-fidelity expressive bodies and faces. Further, it benefits from artist-created accessories, clothing, and environments to increase the diversity of generated data. This allows the models trained on our dataset to exhibit high accuracy and to better generalize to unseen scenarios.

Training on Synthetic Data. To address data diversity and quality issues, hybrid data strategies have been proposed. Depth Anything v2 [44], uses a robust teacher model (DINOv2-G) which is trained exclusively on 595K synthetic images. This model then generates precise pseudo ground truth for a large collection of 62M unlabeled real images, which are subsequently used to train a student model. DepthPro [2] follows a two-stage training curriculum. In the first stage, the model is trained on a mix of multiple real datasets with noisy ground truth, utilizing carefully selected loss functions to improve convergence. In the second stage, the model is trained on synthetic datasets with perfect ground truth.

More recently, Sapiens [17] propose pre-training a large model on 300M real images using self-supervised learning and fine-tuning it on 500K high-resolution synthetic images for depth and surface normal estimation. While achieving promising results, it comes at a significant computational cost. Pre-training the largest variant required 18 days on 1,024 A100 GPUs*. In contrast, our work simplifies the training strategy and eliminates the need for data mixing by using a single small-scale and high-fidelity dataset.

3. Method

3.1. SynthHuman: Human-centric Synthetic Data

To train our models, we use exclusively synthetic data. To this end, a common choice is to use scan-based synthetic data generation [9, 48]. However, their quality is often limited by the 3D scanning technology used and the 3D mesh representation (see Fig. 4 for the comparison of the ground truth quality). Recently, higher fidelity synthetic data, following the practices of games and visual effects, has been demonstrated to be more effective for certain tasks such as landmark prediction and 3D reconstruction [1, 11, 41]. In this work, we extend the use of such high-fidelity synthetic data to *dense* prediction tasks where realism and annotation quality are even more critical, and for which annotations on real data are often impossible. Specifically, we use the data generation pipeline of Hewitt et al. [11], incorporating the updated face model of Petikam et al. [30], to create a human-centric synthetic dataset with a high degree of realism, as well as high-fidelity ground-truth annotations. Our *SynthHuman*, dataset contains 300K images of resolution 384×512, covering examples of faces, upper body, and full body scenarios equally. Along with the RGB rendered image, each sample includes soft foreground mask, surface normals, and depth ground-truth annotations, used to train our models. We design SynthHuman such that it is diverse in terms of poses, environments, lighting, and appearances, and not tailored to any specific evaluation set. This allows us to train models that generalize across a range of benchmark datasets, as well as on in-the-wild data. Examples of our training data are shown in Fig. 3. Rendering the dataset took 72 hours on a cluster of 300 machines with M60 GPUs*.

Our results demonstrate that using this high-quality data enables very accurate results with smaller models and less data, leading to a far more economical training and inference.

*The authors did not discuss the computational costs of fine-tuning.



Figure 3. Random samples of our synthetic training images for the face, upper and fully body.

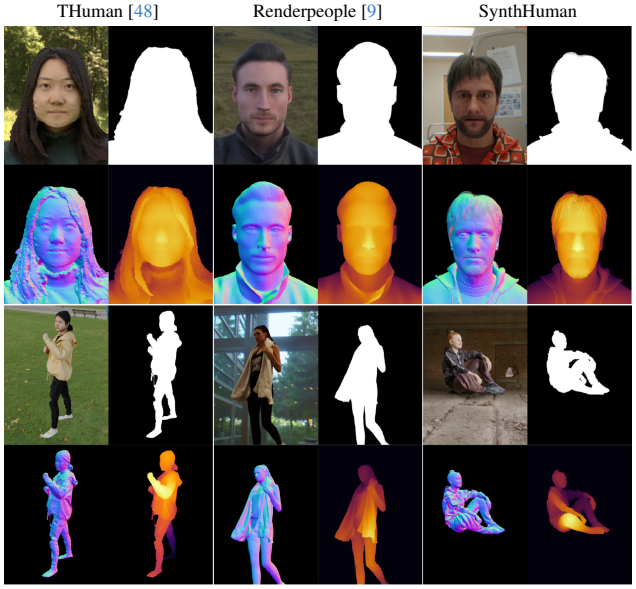


Figure 4. Ground-truth annotations for depth, surface normals and soft foreground segmentation for our synthetic data in comparison to synthetic data used in other work. Note the significantly higher fidelity annotations, particularly for hair and clothing, in our data. Our data is also free of scanning artifacts common in THuman data.

3.2. Model Architecture

We use a single model architecture (with varying number of output channels) to tackle the three dense prediction tasks. We adapt the dense prediction transformer (DPT) [34] to

*The cost is equivalent to 2 weeks of an A100 machine with 4 GPUs.

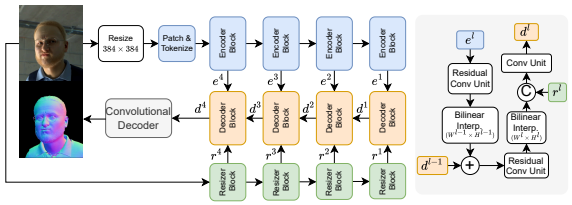


Figure 5. (Left) Overview of the model architecture, with an example of surface normal prediction. (Right) Our decoder block.

handle variable input resolutions efficiently. As illustrated in Fig. 5, our architecture has three main components: encoder blocks, resizer blocks, and decoder blocks.

Encoder. We use the ViT [7] architecture as our image encoder backbone. The encoder design follows DPT's encoder with $Read_{proj}$ as the read operation (see Ranftl et al. [34]).

$$e^l = \text{mlp}(\text{cat}(\texttt{CLS}^l, t_i^l)) \tag{1}$$

wherein e^l is the sequence of updated visual tokens at layer l after projection, $\texttt{CLS}^l$ is the optimized $\texttt{CLS}$ token at layer l, and t_i^l is the i^{th} visual token at layer l.

Resizer. While we keep a fixed resolution for the input to the ViT encoder (specifically 384×384), we utilize another light-weight fully convolutional image encoder to carry information at any resolution. These features are computed on the original image size and are used in the decoder blocks, described below. Each image resizer block is a convolutional module, defined as g. Particularly, $r^l = g^l(r^{l-1})$ at layer l computes new features at half the resolution of its input tensor. To form the full resizer module, we stack four resizer blocks, similar to the number of encoder blocks we use to extract intermediate features. Note that this is to alleviate the need for running the ViT encoder on a potentially higher-resolution image, which comes at a much higher computational cost due to the quadratic nature of self-attention.

Decoder. The decoder aims at generating feature representations that the convolutional head (described below) can generate the output from. Each decoder block in the decoder module, as depicted in Fig. 5 (right), works with 3 inputs: (1) The output from previous decoder block, $d^{\cdot}$, if available. (2) Corresponding feature from the encoder, $e^{\cdot}$. (3) Corresponding feature from image resizer, $r^{\cdot}$.

$$\begin{aligned} d^l_{\text{int}} &= \text{RConv}(d^{l-1} + \text{Interp}(\text{RConv}(e^l))) \\ d^l &= \text{Conv}([r^l, \text{Interp}(d^l_{int})]) \end{aligned} \tag{2}$$

where d^l_{int} is an intermediate feature used for internal computations in the decoder block, Interp is the bilinear interpolation, Conv is the convolutional unit and RConv is the residual convolutional unit. In particular, the decoder block first fuses the output of previous decoder block with the corresponding encoder features by first upsampling a learned residual from the encoder feature and adding it to previously decoded features. The resulting representation is then transformed into another feature map via a residual convolutional unit, followed by upsampling to the resolution of the corresponding image resizer features. The results are then concatenated with the image resizer features, producing the output after going through a convolutional unit.

Convolutional Head. The convolutional head for each task also follows the design of DPT, with different number of output channels for different tasks: 1 for portrait matting, 1 for relative depth, and 3 for surface normals.

Remark on Resizer. We explicitly use a fixed-size input to the ViT for constant inference cost of the encoder and handle variable resolutions with the Resizer and the modified decoder (Fig. 5 (right)). This is a more efficient alternative to increasing the number of visual tokens (as done in, e.g., Sapiens [17])* if the input image has higher resolution. We empirically observed that not only is this faster, it also yields compelling results capturing fine-grained details (see supplementary material for results).

3.3. Loss Functions

Having presented the model architecture, next we present the training losses used to address our three prediction tasks.

Soft Foreground Segmentation. For this task, the model only predicts a soft alpha mask, $\hat{\alpha}$, without learning the composition. To train the model, we use a loss function as

$$\mathcal{L}_\alpha = \mathcal{L}_{\text{BCE}} + \mathcal{L}_{L1} + \mathcal{L}_{\text{dice}} + \omega_{\text{lap}}\mathcal{L}_{\text{lap}} \tag{3}$$

wherein $\mathcal{L}_{\text{BCE}}$ is the binary cross-entropy loss, $\mathcal{L}_{L1}$ is the $L1$ loss, $\mathcal{L}_{\text{dice}}$ is the dice loss [39], and finally $\mathcal{L}_{\text{lap}}$ is the $L1$ reconstruction loss between the Laplacian pyramid representation [12] of the ground truth soft mask, α, and $\hat{\alpha}$. All terms except for $\mathcal{L}_{\text{lap}}$ are weighted equally. We observed that $\omega_{\text{lap}} < 1$ leads to better accuracy.

Surface Normal Estimation. The model predicts the per-pixel xyz components of the normal vector, forming a 3-channel output, $\hat{\eta}$, at the same resolution as the input image. Our model is trained to maximize the alignment between the predicted normalized and ground truth surface normal maps, $\hat{\eta}$ and η, respectively, using cosine similarity, $\mathcal{L}_\eta = 1 - \eta.\hat{\eta}$, computed on the foreground region.

Monocular Relative Depth Estimation. For relative depth estimation, we first normalize the ground-truth metric depth, d^*, by $d = \frac{d^* - \min(d^*)}{\max(d^*) - \min(d^*)}$. The model estimates a relative depth, $\hat{d}$, that closely matches the normalized ground-truth depth. We use a shift-and-scale-invariant loss [32]. To encourage sharper boundaries, we supervise gradients of the predictions [2]

$$\mathcal{L}_d = \mathcal{L}_{\text{MSE}}(s\hat{d} + t, d) + \omega_{\text{grad}}\mathcal{L}_{\text{grad}}(s\hat{d} + t, d) \tag{4}$$

*Additionally, it allows us to evaluate any resolution as opposed to constraining on a multiplier of p.

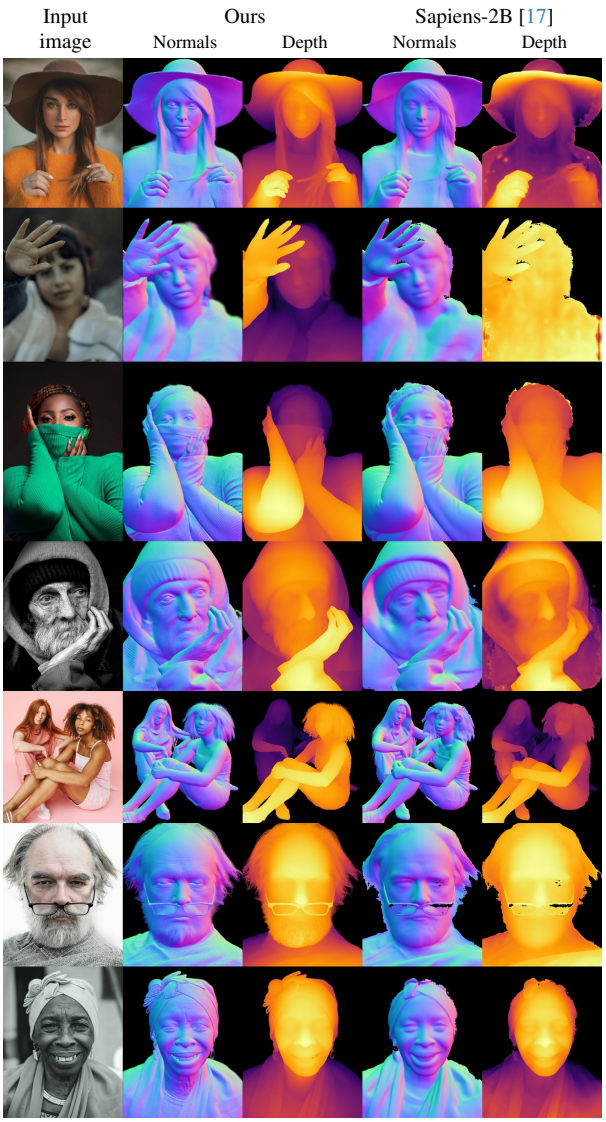


Figure 6. Qualitative comparison between our method and Sapiens [17] on challenging in-the-wild images. We use the publicly released Sapiens-1B segmentation model for Sapiens' foreground segmentation. See supplementary material for more results.

where s and t are scale and shift scalars, computed using the method of Ranftl et al. [32]. We compute the depth loss on the foreground region only.

4. Experiments

4.1. Implementation details

As mentioned in Sec. 3.1, we train our models on 300,000 Synthetic images from our SynthHuman dataset. For each task, we train the model for 100 epochs, with an AdamW optimizer [23] with a starting learning rate of $1e^{-5}$ decreasing following a cosine annealing scheduler [22]. We use a batch size of 24 on each GPU of a A100×4 compute node. The training images are rendered at the resolution of 384×512, with an aspect ratio of 3:4, which aligns with human-centric test sets. Since the original DPT encoder requires a square image, we pad around the image (by replicating the sides) to 512×512. We provide more details about augmentation during training in the supplementary materials.

4.2. Evaluation protocol

Evaluation Datasets. We evaluate our approach on multiple challenging real benchmark datasets. We use the Goliath [25] and Hi4D [47] datasets to evaluate our depth and surface normal estimation models*. Goliath contains data from four subjects, for which we use the head and fully-clothed captures to create three subsets of face, upper body, and full body. Each subset uses 12 cameras and 16 frames (minus two missing cameras) resulting in total of 2,272 test samples (generation details in the supplementary material). The Hi4D dataset provides captures of subject pairs interacting. Following the evaluation protocol in Sapiens [17], we selected the same sequences from pairs 28, 32, and 37, which include 6 unique subjects recorded by camera 4. This selection results in total of 1,195 multi-human real images for testing. For soft foreground segmentation, we report our results on the PhotoMatte85 [19] and PPM-100 [16] datasets. We also provide more results on the P3M [18] dataset in the supplementary materials.

Evaluation Metrics. To evaluate depth estimation models, we report the mean absolute value of the relative depth (AbsRel) and the root mean square error (RMSE), following standard practice [17, 33, 44]. To evaluate surface normal estimation models, we report the standard metrics [8, 17] of mean and median angular error, as well as the percentage of pixels within t° error for $t \in \{11.25, 22.5, 30\}$. For soft foreground segmentation, we report common metrics following Li et al. [18] including sum of absolute differences (SAD), mean squared error (MSE), mean absolute difference (MAD), and Connectivity (Conn.).

4.3. Comparison to the state of the art

Unlike prior approaches, which rely on distinct models and/or datasets for each task and often require task-specific tuning, architectural modifications, or additional processing modules (e.g., refiner networks or guided filters for matting, multi-resolution decoders or focal length prediction for depth estimation), our method employs a unified architecture trained on a single dataset. The only variations are the loss functions and the number of output channels. Despite its significantly smaller model size, our approach achieves competitive performance across multiple benchmarks, underscoring the crucial role of high-fidelity training data.

*See supplementary material for additional experiments on the synthetically generated dataset from THuman2.1 [48], following Sapiens [17].

Relative Depth Estimation. We evaluate our approach on two challenging real datasets Goliath and Hi4D. Tab. 1 compares our method with the existing state-of-the-art approaches. The variants of our method (Base and Large) demonstrate remarkable performance for such comparably efficient models. Our Large variant, with 0.3B parameters, performs on-par with foundation model of Sapiens [17] which has over 2B parameters, while running ∼16x faster (measured in MACs). Our model also outperforms DepthPro [44], which is trained specifically for sharp depth estimation from high resolution images. This shows the quality of the training data plays a major role in the accuracy of the model, and enables training of very simple models without custom designs*. Not only are our models accurate, they are much smaller and faster than the competing approaches, running at ∼48 FPS on NVIDIA A100. On the Hi4D dataset, we noticed that our replication of existing baselines lead to negligibly different (better) accuracy for depth estimation task*, so we re-evaluated all methods in Tab. 1 for a fair comparison. Fig. 6 illustrates the robustness of our approach when tested on the long tail (i.e., the less common, more specialized cases) of the distribution of human-centric images.

Surface Normal Estimation. We compare our surface normal prediction model on Goliath and Hi4D and report the results in Tab. 2. Similar to our depth model, our surface normal prediction model achieves very competitive performance while requiring far fewer parameters. Specifically, our model outperforms baselines of similar size, e.g., Sapiens-0.3B and performs on par with largest models, e.g., Sapiens-2B. Although Goliath and Hi4D provide good source of real data for testing, the ground-truth annotations are very noisy. Fig. 7 demonstrates that a large source of error in the metric is the lack of detail in the ground truth, indicating that we may be observing ceiling effects when evaluating our models. Specifically, detail is lacking for the mouth interior, wrinkles in clothing, and there are incorrect connected regions (attached fingers or arm-body attachment). As illustrated, our model captures far more detail such as wrinkles, evident in Fig. 7 and in challenging in-the-wild data depicted in Fig. 6. See supplementary material for further discussion on annotation quality of surface normals in the test data.

Soft Foreground Segmentation. For human-centric dense prediction tasks, such as the ones we tackle in this work, we need to separate the foreground human from the background. We do this by predicting the soft foreground mask using the Large version of our model. The closest task to this is foreground matting* for which we provide the results in Tab. 3. Our approach generalizes well, evident by the performance on PhotoMatte85 and PPM-100. Note that some prior works listed in Tab. 3, e.g., Zhong and Zharkov [49], are highly optimized for real-time portrait matting, making them considerably more efficient than ours. We prioritize maintaining a unified architecture across without task-specific modifications, achieving superior accuracy and seamlessly integrating with the other two tasks.

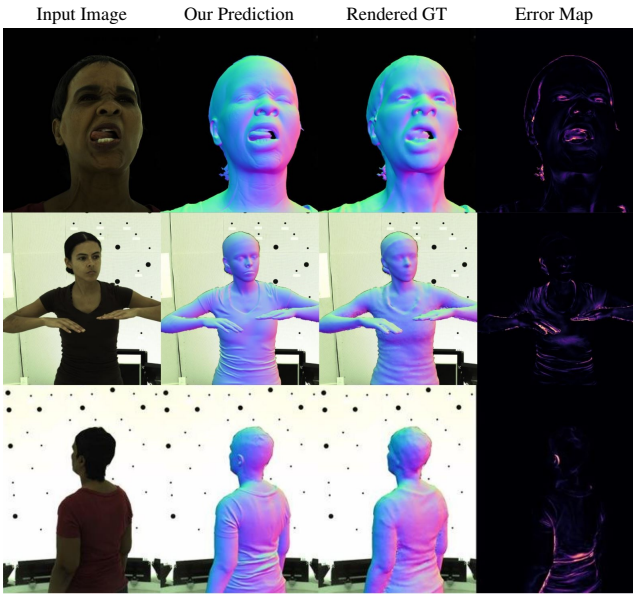


Figure 7. Qualitative results on Goliath dataset. As shown in the last column (the error map between our prediction and the ground truth), the main source of error is in high frequency details. While our approach captures very fine details, we observed that the ground truth is very coarse, lacking fine-grained details, such as details of mouth interior, face wrinkles as a result of expression, detailed wrinkles in clothing, and separation of body-arm and fingers.

4.4. Ablation Studies

In this section, we evaluate our design choices, including the impact of the synthetic data source, training data size, and model size. We also demonstrate that, since we use a single training dataset across multiple tasks to train a single model architecture, it is feasible to train a single model to perform all three tasks. We then demonstrate the accuracy of that model compared to three separately trained models specializing on each task. Unless otherwise stated, we use the depth estimation task for the ablation studies.

Impact of data source. To compare the impact of the dataset quality, we render synthetic datasets from RenderPeople and THuman2.0 of similar size to our SynthHuman dataset and use these to train a Large depth estimation model with the same hyper-parameters as our model. In Tab. 4a we see that the fidelity of the ground truth and the diversity of samples play a key role in achieving the best results. While the coarse depth estimated by each model is roughly the same,

*We acknowledge that DepthPro is trained on generic datasets, including human-centric ones, e.g., Bedlam [1]. Similarly, DepthAnythingV2 and MiDaS are also trained for depth estimation in any scenario.

*We suspect this is due to differences in ground truth rendering.

*Our approach does not tackle the full matting problem, however, for the lack of better benchmark, we evaluate our approach on matting datasets.

Table 1. Depth estimation on Goliath and Hi4D dataset.

Method	GFLOPS	Params	Goliath-Face		Goliath-UpperBody		Goliath-FullBody		Hi4D		Averaged over all	
			RMSE ↓	AbsRel ↓	RMSE ↓	AbsRel ↓	RMSE ↓	AbsRel ↓	RMSE ↓	AbsRel ↓	RMSE ↓	AbsRel ↓
MiDaS-DPT_L [33]	-	0.34B	0.224	0.016	0.553	0.015	0.973	0.027	0.148	0.042	0.437	0.027
DepthAnythingV2-L [44]	1827	0.34B	0.229	0.017	0.492	0.014	1.039	0.029	0.130	0.034	0.433	0.025
Sapiens-0.3B [17]	1242	0.34B	0.179	0.012	0.368	0.010	0.690	0.019	0.116	0.035	0.312	0.021
Sapiens-2B [17]	8709	2.16B	0.158	**0.009**	**0.204**	**0.005**	**0.266**	**0.007**	0.095	0.030	**0.170**	0.015
Depth-Pro [2]	4370	0.50B	0.295	0.020	0.442	0.010	0.723	0.016	0.084	**0.018**	0.350	0.016
Ours-Base	344	0.12B	0.142	**0.009**	0.316	0.009	0.376	0.010	0.085	0.024	0.212	0.014
Ours-Large	663	0.34B	**0.140**	**0.009**	0.283	0.008	0.334	0.009	**0.072**	0.019	0.191	**0.012**

Table 2. Surface normal estimation results on Goliath and Hi4D. All results on the Hi4D dataset are taken from [17].

Method	Goliath-Face			Goliath-UpperBody			Goliath-FullBody			Hi4D		
	Angular Error (°) ↓		% Within t° ↑	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°	Mean	Median	11.25° / 22.5° / 30°
PIFuHD [37]	-	-	-	-	-	-	-	-	-	22.39	19.26	23.0 / 60.1 / 77.0
HDNet [13]	-	-	-	-	-	-	-	-	-	28.60	26.85	19.1 / 57.9 / 70.1
ICON [42]	-	-	-	-	-	-	-	-	-	20.18	17.52	26.8 / 66.3 / 82.7
ECON [43]	-	-	-	-	-	-	-	-	-	18.46	16.47	29.3 / 68.1 / 84.9
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	15.04	12.22	47.1 / 81.5 / 90.7
Sapiens-2B	**16.04**	**11.66**	**51.7 / 78.3 / 86.3**	**10.65**	**8.67**	**65.5 / 92.5 / 96.7**	**11.49**	**9.07**	**62.3 / 90.2 / 95.4**	**12.14**	**9.62**	**60.2 / 89.1 / 94.7**
Ours-Base	17.33	12.36	47.7 / 75.9 / 84.5	14.10	11.32	50.3 / 83.9 / 91.8	14.60	11.79	48.1 / 82.3 / 91.1	15.72	12.95	43.2 / 78.7 / 89.2
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	15.37	12.51	45.1 / 79.7 / 89.6

Table 3. Cross dataset evaluation for soft foreground segmentation.

Method	PhotoMatte85			PPM-100	
	SAD ↓	MSE ↓	Conn ↓	SAD ↓	Conn ↓
Zhong et al. [49]	-	-	-	90.28	84.09
BGMv2 [19]	-	-	-	159.44	149.79
P3M-Net [18]	20.05	0.007	19.76	142.74	139.89
MODNet [16]	13.94	0.003	11.18	104.35	96.45
Ours	**5.85**	**0.0009**	**5.60**	**78.17**	**74.72**

the model trained on SynthHuman is capable of capturing far more detail, shown in Fig. 8.

Impact of training data size. In Tab. 4b, we show the effect of the size of the training data. While even a small but high-fidelity dataset, as small as 60K, leads to reasonable accuracy for the relative depth estimation task, the model achieves better performance as we increase the training data size. This highlights that the diversity and fidelity of our dataset is considerable and the trainings do not saturate on a portion of dataset. Comparing this result with the last row of Tab. 4c also highlights that our synthetic data contributes positively as we scale on both the data and model size.

Impact of model size. Another aspect of training on relatively small datasets is interaction with model size. To ensure that our training data serves models of multiple sizes we train models with ViT variants of small, base, and large, with results reported in Tab. 4c. As expected, increasing model size lead to increase in the performance of the model.

Multi-task model. Using a single dataset and a single model architecture allows us to easily train a single model with three convolution heads to perform multiple task learning. This is particularly important to combine soft foreground segmentation with depth and normal estimation, as for human-centric tasks it is needed to separate the human from the background. We observe that using three separate Large models yields slightly better results than a single Large multi-task model with one-third of the total parameters (3×0.34B vs 0.35B), see Tab. 4d. Jointly training all three tasks in a multi-task model, however, performs better than three separate models with similar combined number of parameters (3×0.12B for three Base models vs 0.35B for the Large multi-task one).

5. Potential societal impact

As for all human-centric computer vision, the models we train and demonstrate in this work could have lower accuracy for some demographic groups. We find that our use of synthetic data helps in addressing any lack of fairness we discover in model evaluations, given the precise control we have over the training data distribution. Nevertheless, there are aspects of human diversity that are not yet represented by our datasets (see Sec. 6), and there may also be lack of fairness that we have not yet discovered in evaluations.

A negative impact of the trend towards huge real datasets is difficulty in ensuring informed consent for training AI models, both from the rights holders and the people appearing in the images. By demonstrating that models trained only on synthetic data can be as accurate as large foundational models, we hope to show that state-of-the-art human understanding need not be in tension with user privacy.

Another negative societal impact comes from the environmental cost of training and running inference on models that are larger than necessary. By showing that human-centric vi-

Table 4. Ablation study: the effect of training data source, training data size, backbone size, and multi-task learning for depth estimation task evaluated on Goliath and Hi4D datasets.

(a) Impact of training data.

Source	Goliath		Hi4D	
	RMSE ↓	AbsRel ↓	RMSE ↓	AbsRel ↓
THuman2.0	0.495	0.017	0.137	0.040
RenderPeople	0.278	0.011	0.076	0.021
Ours	0.253	0.008	0.072	0.019

(b) Impact of training data size.

Dataset size	Goliath		Hi4D	
	RMSE ↓	AbsRel ↓	RMSE ↓	AbsRel ↓
ViT-Base [60K]	0.324	0.011	0.101	0.028
ViT-Base [150K]	0.305	0.010	0.085	0.022
ViT-Base [300K]	0.278	0.009	0.085	0.024

(c) Impact of model size.

Arch.	Goliath		Hi4D	
	RMSE ↓	AbsRel ↓	RMSE ↓	AbsRel ↓
ViT-Small	0.310	0.010	0.089	0.025
ViT-Base	0.278	0.009	0.085	0.024
ViT-Large	0.253	0.008	0.072	0.019

(d) Impact of multi-task training compared to single-task training.

		Depth				Surface Normal				Matting		
		Goliath		Hi4D		Goliath		Hi4D		PPM-100	PhotoMatte85	
Setting	Params	RMSE ↓	AbsRel ↓	RMSE ↓	AbsRel ↓	MAE(°) ↓	% W 30° ↑	MAE(°) ↓	% W 30° ↑	SAD ↓	SAD ↓	MSE ↓
Single-task-Large	3×0.34B	0.253	0.008	0.072	0.019	15.24	89.19	15.37	89.56	78.17	5.85	0.0009
Single-task-Base	3×0.12B	0.278	0.009	0.085	0.024	15.34	89.13	15.72	89.18	90.86	7.97	0.0017
Multi-task-Large	1×0.35B	0.270	0.009	0.078	0.021	15.27	89.12	15.61	89.48	66.08	5.40	0.0008

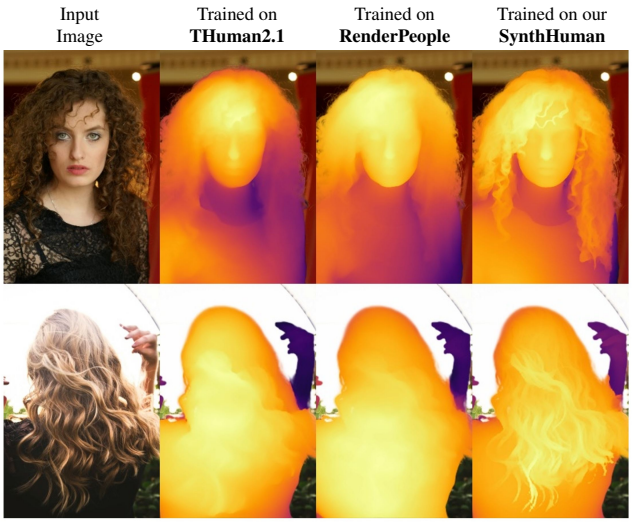


Figure 8. Comparing the accuracy of models trained on Thuman2.1, RenderPeople and our SynthHuman dataset. Our dataset contains details (e.g., hair curls) that scan-based datasets struggle to capture. Note that the only change here is the training data (see Tab. 4a).

sion models can achieve state-of-the-art accuracy at smaller model sizes, we hope to show that these techniques can be cost-effective and responsible in the use of compute resources, while sacrificing nothing in accuracy or robustness.

6. Limitations

Despite the strong generalizability of our trained models to real-world images, certain challenging scenarios still lead to failure cases, as illustrated in Fig. 9. For instance, extreme lighting conditions can introduce inaccuracies in defining surfaces. Our surface normal prediction model may misinterpret printed patterns on clothing or tattoos as distinct geometric structures instead of recognizing the underlying surface as continuous. Our relative depth estimation model



Figure 9. Failures of our models in the presence of tattoos, extreme lighting, uncommon scale variations, and challenging clothing.

struggles with rare scale variations. For example, when a baby is held in an adult's hand, the model incorrectly perceives the large hand as significantly closer to the camera than the baby's face. Many of these failure cases could be mitigated by enhancing our synthetic dataset with more diverse assets and scene variations, thereby improving the model's robustness to such real-world diversity.

7. Conclusion

We have demonstrated that it is possible to train accurate human-centric vision models without the need for large models, huge datasets, and complex methodologies. This was achieved through procedural synthetic data that allows us to have both diverse and well annotated data. Given the smaller dataset, we can train comparatively compact models in a fraction of the time (we can train ∼800 models with the compute used to train a single Sapiens-2B [17] model), while achieving results that are on par with or surpasses existing state-of-the-art methods. We release our datasets and models to encourage further research in this space.

References

[1] Michael J. Black, Priyanka Patel, Joachim Tesch, and Jinlong Yang. BEDLAM: A synthetic dataset of bodies exhibiting detailed lifelike animated motion. In *Proceedings IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 8726–8737, 2023. 2, 3, 6

[2] Aleksei Bochkovskii, Amaël Delaunoy, Hugo Germain, Marcel Santos, Yichao Zhou, Stephan R. Richter, and Vladlen Koltun. Depth pro: Sharp monocular metric depth in less than a second. In *International Conference on Learning Representations*, 2025. 2, 3, 4, 7

[3] Oliver Boyne, James Charles, and Roberto Cipolla. Find: An unsupervised implicit 3d model of articulated human feet. In *British Machine Vision Conference (BMVC)*, 2022. 2

[4] G. J. Brostow, C. Hernández, G. Vogiatzis, B. Stenger, and R. Cipolla. Video normals from colored lights. *TPAMI*, 33(10): 2104–2114, 2011. 2

[5] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3213–3223, 2016. 2

[6] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pages 248–255. IEEE, 2009. 2

[7] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, G Heigold, S Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2020. 4

[8] Ainaz Eftekhar, Alexander Sax, Jitendra Malik, and Amir Zamir. Omnidata: A scalable pipeline for making multi-task mid-level vision datasets from 3d scans. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10786–10796, 2021. 5

[9] Renderpeople GmbH. Renderpeople. 3

[10] Sang-Hun Han, Min-Gyu Park, Ju Hong Yoon, Ju-Mi Kang, Young-Jae Park, and Hae-Gon Jeon. High-fidelity 3d human digitization from single 2k resolution images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 2

[11] Charlie Hewitt, Fatemeh Saleh, Sadegh Aliakbarian, Lohit Petikam, Shideh Rezaeifar, Louis Florentin, Zafiirah Hosenie, 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. 2, 3

[12] Qiqi Hou and Feng Liu. Context-aware image matting for simultaneous foreground and alpha estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4130–4139, 2019. 4

[13] Yasamin Jafarian and Hyun Soo Park. Learning high fidelity depths of dressed humans by watching social media dance videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12753–12762, 2021. 7

[14] Sheng Jin, Lumin Xu, Jin Xu, Can Wang, Wentao Liu, Chen Qian, Wanli Ouyang, and Ping Luo. Whole-body human pose estimation in the wild. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2020. 2

[15] Kimmo Karkkainen and Jungseock Joo. Fairface: Face attribute dataset for balanced race, gender, and age for bias measurement and mitigation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1548–1558, 2021. 2

[16] 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. 2, 5, 7

[17] 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. 2, 3, 4, 5, 6, 7, 8

[18] 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. 5, 7

[19] 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. 5, 7

[20] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. *European conference on computer vision*, pages 740–755, 2014. 2

[21] Yinglu Liu, Hailin Shi, Hao Shen, Yue Si, Xiaobo Wang, and Tao Mei. A new dataset and boundary-attention semantic segmentation for face parsing. In *AAAI*, pages 11637–11644, 2020. 2

[22] Ilya Loshchilov and Frank Hutter. Sgdr: Stochastic gradient descent with warm restarts. *arXiv preprint arXiv:1608.03983*, 2016. 5

[23] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 5

[24] Wan-Chun Ma, Tim Hawkins, Pieter Peers, Charles-Felix Chabert, Malte Weiss, and Paul Debevec. Rapid acquisition of specular and diffuse normal maps from polarized spherical gradient illumination. In *Proceedings of the 18th Eurographics Conference on Rendering Techniques*, page 183–194, Goslar, DEU, 2007. Eurographics Association. 2

[25] Julieta Martinez, Emily Kim, Javier Romero, Timur Bagautdinov, Shunsuke Saito, Shoou-I Yu, Stuart Anderson, Michael Zollhöfer, Te-Li Wang, Shaojie Bai, Chenghui Li, Shih-En Wei, Rohan Joshi, Wyatt Borsos, Tomas Simon, Jason Saragih, Paul Theodosis, Alexander Greene, Anjani Josyula, Silvio Mano Maeta, Andrew I. Jewett, Simon Venshtain,

Christopher Heilman, Yueh-Tung Chen, Sidi Fu, Mohamed Ezzeldin A. Elshaer, Tingfang Du, Longhua Wu, Shen-Chi Chen, Kai Kang, Michael Wu, Youssef Emad, Steven Longay, Ashley Brewer, Hitesh Shah, James Booth, Taylor Koska, Kayla Haidle, Matt Andromalos, Joanna Hsu, Thomas Dauer, Peter Selednik, Tim Godisart, Scott Ardisson, Matthew Cipperly, Ben Humberston, Lon Farr, Bob Hansen, Peihong Guo, Dave Braun, Steven Krenn, He Wen, Lucas Evans, Natalia Fadeeva, Matthew Stewart, Gabriel Schwartz, Divam Gupta, Gyeongsik Moon, Kaiwen Guo, Yuan Dong, Yichen Xu, Takaaki Shiratori, Fabian Prada, Bernardo R. Pires, Bo Peng, Julia Buffalini, Autumn Trimble, Kevyn McPhail, Melissa Schoeller, and Yaser Sheikh. Codec Avatar Studio: Paired Human Captures for Complete, Driveable, and Generalizable Avatars. *NeurIPS Track on Datasets and Benchmarks*, 2024. 1, 2, 5

[26] Rui Min, Neslihan Kose, and Jean-Luc Dugelay. Kinectfacedb: A kinect database for face recognition. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, 44(11): 1534–1548, 2014. 2

[27] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel HAZIZA, Francisco Massa, Alaaeldin El-Nouby, Mido Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Herve Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. DINOv2: Learning robust visual features without supervision. *Transactions on Machine Learning Research*, 2024. Featured Certification. 2

[28] Priyanka Patel, Chun-Hao P. Huang, Joachim Tesch, David T. Hoffmann, Shashank Tripathi, and Michael J. Black. AGORA: Avatars in geography optimized for regression analysis. In *Proceedings IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2021. 2

[29] Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed A. A. Osman, Dimitrios Tzionas, and Michael J. Black. Expressive body capture: 3d hands, face, and body from a single image. In *Proceedings IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2019. 2

[30] 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. 3

[31] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR, 2021. 2

[32] 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. 4, 5

[33] 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. 5, 7

[34] René Ranftl, Alexey Bochkovskiy, and Vladlen Koltun. Vision transformers for dense prediction. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 12179–12188, 2021. 3, 4

[35] Christos Sagonas, Epameinondas Antonakos, Georgios Tzimiropoulos, Stefanos Zafeiriou, and Maja Pantic. 300 faces in-the-wild challenge: database and results. *Image and Vision Computing*, 47:3–18, 2016. 300-W, the First Automatic Facial Landmark Detection in-the-Wild Challenge. 2

[36] Christos Sagonas, Epameinondas Antonakos, Georgios Tzimiropoulos, Stefanos Zafeiriou, and Maja Pantic. 300 faces in-the-wild challenge: Database and results. *Image and vision computing*, 47:3–18, 2016. 2

[37] 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/CVF conference on computer vision and pattern recognition*, pages 84–93, 2020. 7

[38] Soubhik Sanyal, Timo Bolkart, Haiwen Feng, and Michael Black. Learning to regress 3D face shape and expression from an image without 3D supervision. In *Proceedings IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 7763–7772, 2019. 2

[39] Carole H Sudre, Wenqi Li, Tom Vercauteren, Sebastien Ourselin, and M Jorge Cardoso. Generalised dice overlap as a deep learning loss function for highly unbalanced segmentations. In *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support: Third International Workshop, DLMIA 2017, and 7th International Workshop, ML-CDS 2017, Held in Conjunction with MICCAI 2017, Québec City, QC, Canada, September 14, Proceedings 3*, pages 240–248. Springer, 2017. 4

[40] Yujiang Wang, Bingnan Luo, Jie Shen, and Maja Pantic. Face mask extraction in video sequence. *International Journal of Computer Vision*, 127(6):625–641, 2019. 2

[41] Erroll Wood, Tadas Baltrušaitis, Charlie Hewitt, Sebastian Dziadzio, Thomas J. Cashman, and Jamie Shotton. Fake it till you make it: face analysis in the wild using synthetic data alone. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 3681–3691, 2021. 2, 3

[42] Yuliang Xiu, Jinlong Yang, Dimitrios Tzionas, and Michael J Black. Icon: Implicit clothed humans obtained from normals. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13286–13296. IEEE, 2022. 7

[43] Yuliang Xiu, Jinlong Yang, Xu Cao, Dimitrios Tzionas, and Michael J Black. Econ: Explicit clothed humans optimized via normal integration. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 512–523, 2023. 7

[44] 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. 2, 3, 5, 6, 7

[45] Shuo Yang, Ping Luo, Chen Change Loy, and Xiaoou Tang. Wider face: A face detection benchmark. *2016 IEEE Confer-*

ence on Computer Vision and Pattern Recognition (CVPR), pages 5525–5533, 2015. 2

[46] Zhitao Yang, Zhongang Cai, Haiyi Mei, Shuai Liu, Zhaoxi Chen, Weiye Xiao, Yukun Wei, Zhongfei Qing, Chen Wei, Bo Dai, Wayne Wu, Chen Qian, Dahua Lin, Ziwei Liu, and Lei Yang. Synbody: Synthetic dataset with layered human models for 3d human perception and modeling. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 20282–20292, 2023. 2

[47] Yifei Yin, Chen Guo, Manuel Kaufmann, Juan Zarate, Jie Song, and Otmar Hilliges. Hi4d: 4d instance segmentation of close human interaction. In *Computer Vision and Pattern Recognition (CVPR)*, 2023. 1, 2, 5

[48] 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. 2, 3, 5

[49] 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. 6, 7