Sapiens: Foundation for Human Vision Models

Rawal Khirodkar, Timur Bagautdinov, Julieta Martinez, Su Zhaoen, Austin James, Peter Selednik, Stuart Anderson, Shunsuke Saito

Codec Avatars Lab, Meta

Abstract. We present *Sapiens*, a family of models for four fundamental human-centric vision tasks – 2D pose estimation, body-part segmentation, depth estimation, and surface normal prediction. Our models natively support 1K high-resolution inference and are extremely easy to adapt for individual tasks by simply fine-tuning foundation models pretrained on over 300 million in-the-wild human images. We observe that, given the same computational budget, self-supervised pretraining on a curated dataset of human images significantly boosts the performance for a diverse set of human-centric tasks. The resulting models exhibit remarkable generalization to in-the-wild data, even when labeled data is scarce or entirely synthetic. Our simple model design also brings scalability – model performance across tasks significantly improves as we scale the number of parameters from 0.3 to 2 billion. Sapiens consistently surpasses existing complex baselines across various human-centric benchmarks. Specifically, we achieve significant improvements over the prior state-of-the-art on Humans-5K (pose) by 7.6 mAP, Humans-2K (part-seg) by 17.1 mIoU, Hi4D (depth) by 22.4% relative RMSE, and THuman2 (normal) by 53.5% relative angular error.

Keywords: Human-centric, Large-scale pretraining, In-the-wild

"*Sapiens*—pertaining to, or resembling modern humans."

1 Introduction

Recent years have witnessed remarkable strides towards generating photorealistic humans in 2D [17, 28, 50, 118] and 3D [69, 89, 102, 109]. The success of these methods is greatly attributed to the robust estimation of various assets such as 2D keypoints [14,67], fine-grained body-part segmentation [119], depth [113], and surface normals [89,108]. However, robust and accurate estimation of these assets is still an active research area, and complicated systems to boost performance for individual tasks often hinder wider adoption. Moreover, obtaining accurate ground-truth annotation in-the-wild is notoriously difficult to scale. Our goal is to provide a unified framework and models to infer these assets in-the-wild to unlock a wide range of human-centric applications for everybody.

We argue that such human-centric models should satisfy three criteria: generalization, broad applicability, and high fidelity. Generalization ensures robustness to unseen conditions, enabling the model to perform consistently across varied environments. Broad applicability indicates the versatility of the model, making it suitable for a wide range of tasks with minimal modifications. High fidelity denotes the ability of the model to produce precise, high-resolution outputs,

Fig. 1: *Sapiens* models are finetuned for four human tasks - 2D pose estimation, body-part segmentation, depth prediction and normal prediction. Our models generalize across a variety of in-the-wild face, upper-body, full-body and multi-person images.

essential for faithful human generation tasks such as 2D to 3D lifting. This paper details the development of models that embody these attributes, collectively referred to as *Sapiens*.

Following the insights from [34, 79, 91], leveraging large datasets and scalable model architectures is key for generalization. For broader applicability, we adopt the pretrain-then-finetune approach, enabling post-pretraining adaptation to specific tasks with minimal adjustments. This approach raises a critical question: *What type of data is most effective for pretraining?* Given computational limits, should the emphasis be on collecting as many human images as possible, or is it preferable to pretrain on a less curated set to better reflect real-world variability? Existing methods often overlook the pretraining data distribution in the context of downstream tasks. To study the influence of pretraining data distribution on human-specific tasks, we collect the Humans-300M dataset, featuring 300 million diverse human images. These unlabelled images are used to pretrain a family of vision transformers [27] from scratch, with parameter counts ranging from 300M to 2B.

Among various self-supervision methods for learning general-purpose visual features from large datasets [5,19,34,45,46,121], we choose the masked-autoencoder (MAE) approach [45] for its simplicity and efficiency in pretraining. MAE, having a single-pass inference model compared to contrastive or multi-inference strategies, allows processing a larger volume of images with the same computational resources. For higher-fidelity, in contrast to prior methods, we increase the native input resolution of our pretraining to 1024 pixels, resulting in a $\sim 4\times$ increase in FLOPs compared to the largest existing vision backbone [91]. Each model is pretrained on 1.2 trillion tokens. Table 1 outlines a comparison with earlier approaches. For finetuning on human-centric tasks [15, 101, 113, 119], we use a consistent encoder-decoder architecture. The encoder is initialized with weights from pretraining, while the decoder, a lightweight and task-specific head, is initialized randomly. Both components are then finetuned in an end-to-end manner. We focus on four key tasks - 2D pose estimation, body-part segmentation, depth, and normal estimation, as shown in Fig. 1.

Consistently with prior studies [56,122], we affirm the critical impact of label quality on the model's in-the-wild performance. Public benchmarks [23,41,55] often contain noisy labels, providing inconsistent supervisory signals during model fine-tuning. At the same time, it is important to utilize fine-grained and precise annotations to align closely with our primary goal of 3D human digitization. To this end, we propose a substantially denser set of 2D whole body keypoints for pose estimation and a detailed class vocabulary for body part segmentation, surpassing the scope of previous datasets (please refer to Fig. 1). Specifically, we introduce a comprehensive collection of 308 keypoints encompassing the body, hands, feet, surface, and face. Additionally, we expand the segmentation class vocabulary to 28 classes, covering body parts such as the hair, tongue, teeth, upper/lower lip, and torso. To guarantee the quality and consistency of annotations and a high degree of automation, we utilize a multi-view capture setup to collect pose and segmentation annotations. We also utilize human-centric synthetic data for depth and normal estimation, leveraging 600 detailed scans from RenderPeople [82] to generate high-resolution depth maps and surface normals.

We show that the combination of domain-specific large-scale pretraining with limited, yet high-quality annotations leads to robust in-the-wild generalization. Overall, our method demonstrates an effective strategy for developing highly precise discriminative models capable of performing in real-world scenarios without the need for collecting a costly and diverse set of annotations.

Method	Dataset	#Params	GFLOPs	Image size	Domain
DINO [16]	ImageNet1k	86 M	17.6	224	General
iBOT [121]	ImageNet21k	307 M	61.6	224	General
DINOv2 [79]	LVD-142M	1 B	291.0	224	General
ViT-6.5B [91]	IG-3B	6.5 B	1657.0	224	General
AIM [34]	DFN-2B	6.5 B	1657.0	224	General
Sapiens (Ours)	Humans-300M	2 B	8709.0	1024	Human

Table 1: Comparison of state-of-the-art pretrained vision models. Sapiens adopts a higher resolution backbone on a large dataset of in-the-wild human images.

Our contributions are summarized as follows.

- We introduce *Sapiens*, a family of vision transformers pretrained on a large-scale, curated dataset of diverse human images.
- This study shows that simple data curation and large-scale pretraining significantly boost the model's performance with the same computational budget.
- Our models, fine-tuned with high-quality annotations or even synthetic data, demonstrate robust in-the-wild generalization.
- The first 1K high-resolution model that natively supports high-fidelity inference essential for human-centric tasks, achieving state-of-the-art performance on benchmarks for 2D pose estimation, body-part segmentation, depth, and normal estimation.

2 Related Work

Our work explores the limits of training large architectures on a large number of in-the-wild human images. We build on prior work from different areas: pretraining at scale, human vision tasks, and large vision transformers.

Pretraining at Scale. The remarkable success of large-scale pretraining [26,95] followed by task-specific finetuning for language modeling [2, 13, 53, 96, 99, 100] has established this approach as a standard practice. Similarly, computer vision methods [1, 4, 33, 34, 42, 79, 83, 85, 87, 120] are progressively embracing extensive data scales for pretraining. The emergence of large datasets, such as LAION-5B [90], Instagram-3.5B [77], JFT-300M [92], LVD-142M [79], Visual Genome [60], and YFCC100M [97], has enabled the exploration of a data corpus well beyond the scope of traditional benchmarks [61, 67, 86]. Salient work in this domain includes DINOv2 [79], MAWS [91], and AIM [34]. DINOv2 achieves state-of-the-art performance in generating self-supervised features by scaling the contrastive iBot [121] method on the LDV-142M dataset [79]. MAWS [91] studies the scaling of masked-autoencoders (MAE) [45] on billion images. AIM [34] explores the scalability of autoregressive visual pretraining similar to BERT [26] for vision transformers [27]. In contrast to these methods which mainly focus on general image pretraining or zero-shot image classification, we take a distinctly human-centric approach: our models leverage a vast collection of human images for pretraining, subsequently fine-tuning for a range of human-related tasks.

Human Vision Tasks. The pursuit of large-scale 3D human digitization [8, 44, 64, 74] remains a pivotal goal in computer vision [12]. Significant progress has been made within controlled or studio environments [3, 59, 63, 69, 70, 76, 89], yet challenges persist in extending these methods to unconstrained environments [29]. To address these challenges, developing versatile models capable of multiple fundamental tasks such as keypoint estimation [21, 36, 47, 51, 57, 78, 80, 93, 106], body-part segmentation [35, 40, 40, 41, 75, 104, 105], depth estimation [9, 10, 32, 43, 52, 66, 84, 113], and surface normal prediction [6, 7, 31, 39, 62, 88, 101, 108] from images in natural settings is crucial. In this work, we aim to develop models for these essential human vision tasks which generalize to in-the-wild settings.

Scaling Architectures. Currently, the largest publicly-accessible language models contain upwards of 100B parameters [49], while the more commonly used language models [94, 100] contain around 7B parameters. In contrast, Vision Transformers (ViT) [27], despite sharing a similar architecture, have not been scaled to this extent successfully. While there are notable endeavors in this direction, including the development of a dense ViT-4B [20] trained on both text and images, and the formulation of techniques for the stable training of a ViT-22B [25], commonly utilized vision backbones still range between 300M to 600M parameters [24, 38, 48, 68] and are primarily pretrained at an image resolution of about 224 pixels. Similarly, existing transformer-based image generation models, such as DiT [81] use less than 700M parameters, and operate on a highly compressed latent space. To address this gap, we introduce Sapiens - a collection of large, high-resolution ViT models that are pretrained natively at a 1024 pixel image resolution on millions of human images.

3 Method

3.1 Humans-300M Dataset

We utilize a large proprietary dataset for pretraining of approximately 1 billion in-the-wild images, focusing exclusively on human images. The preprocessing involves discarding images with watermarks, text, artistic depictions, or unnatural elements. Subsequently, we use an off-the-shelf person bounding-box detector [103] to filter images, retaining those with a detection score above 0.9 and bounding box dimensions exceeding 300 pixels. Fig. 2 provides an overview of the distribution of the number of people per image in our dataset, noting that over 248 million images contain multiple subjects.

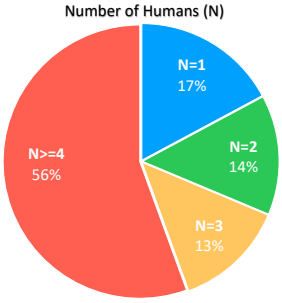

Fig. 2: Overview of number of humans per image in the Humans-300M dataset.

3.2 Pretraining

We follow the masked-autoencoder [45] (MAE) approach for pretraining. Our model is trained to reconstruct the original human image given its partial observation. Like all autoencoders, our model has an encoder that maps the visible image to a latent representation and a decoder that reconstructs the original image from this latent representation. Our pretraining dataset consists of both single and multi-human images; each image is resized to a fixed size with a square aspect ratio. Similar to ViT [27], we divide an image into regular non-overlapping patches with a fixed patch size. A subset of these patches is randomly selected and masked, leaving the rest visible. The proportion of masked patches to visible ones is defined as the masking ratio, which remains fixed throughout training. We refer to MAE [45] for more details. Fig. 3 (*Top*) shows the reconstruction of our pretrained model on unseen human images. Our models exhibit generalization across a variety of image characteristics including scales, crops, the age and ethnicity of subjects, and number of subjects. Each patch token in our model accounts for 0.02% of the image area compared to 0.4% in standard ViTs, a 16×

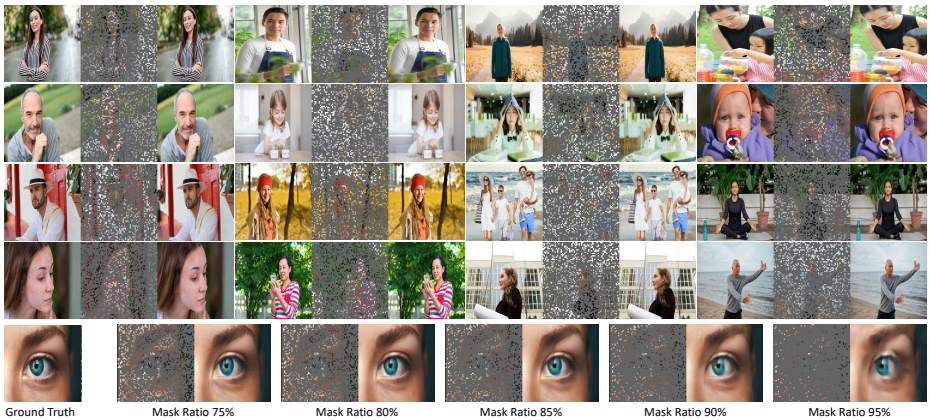

Fig. 3: In-the-wild generalization of *Sapiens* on unseen images. *Top*: Each triplet illustrates the ground truth (left), the masked image (center), and the MAE reconstruction (right), with a masking ratio of 75%, a patch size of 16, and an image size of 1024. *Bottom*: Varying the mask ratio between [0.75, 0.95] during inference reveals a minimal reduction in quality, underscoring the model's robust understanding of human images.

reduction - this provides a fine-grained inter-token reasoning for our models. Fig.3 (*Bottom*) shows that even with an increased mask ratio of 95%, our model achieves a plausible reconstruction of human anatomy on held-out samples.

3.3 2D Pose Estimation

We follow the top-down 2D pose estimation paradigm, which aims to detect the locations of K keypoints from an input image $\mathbf{I} \in \mathbb{R}^{H \times W \times 3}$. Most methods pose this problem as heatmap prediction, where each of K heatmaps represents the probability of the corresponding keypoint being at any spatial location. Similar to [111], we define a pose estimation transformer, $\mathcal{P}$, for keypoint detection. The bounding box at training and inference is scaled to $H \times W$ and is provided as an input to $\mathcal{P}$. Let $\mathbf{y} \in \mathbb{R}^{H \times W \times K}$ denote the K heatmaps corresponding to the ground truth keypoints for a given input $\mathbf{I}$. The pose estimator transforms input $\mathbf{I}$ to a set of predicted heatmaps, $\hat{\mathbf{y}} \in \mathbb{R}^{H \times W \times K}$, such that $\hat{\mathbf{y}} = \mathcal{P}(\mathbf{I})$. $\mathcal{P}$ is trained to minimize the mean squared loss $\mathcal{L}_{\text{pose}} = \texttt{MSE}(\mathbf{y}, \hat{\mathbf{y}})$. During finetuning, the encoder of $\mathcal{P}$ is initialized with the weights from pretaining, and the decoder is initialized randomly. The aspect ratio $H : W$ is set to be $4 : 3$, with the pretrained positional embedding being interpolated accordingly [58]. We use lightweight decoders with deconvolution and convolution operations.

We finetune both the encoder and the decoder in $\mathcal{P}$ across multiple skeleton formats, including $K = 17$ [67], $K = 133$ [55] and a new highly-detailed skeleton format, with $K = 308$, as shown in Fig. 4 (*Left*). Compared to existing formats with at most 68 facial keypoints, our annotations consist of 243 facial keypoints, including representative points around the eyes, eyebrows, lips, nose, and ears. This design is tailored to meticulously capture the nuanced details of facial expressions in the real world. With these keypoints, we manually annotated 1 million images at $4K$ resolution from an indoor capture setup.

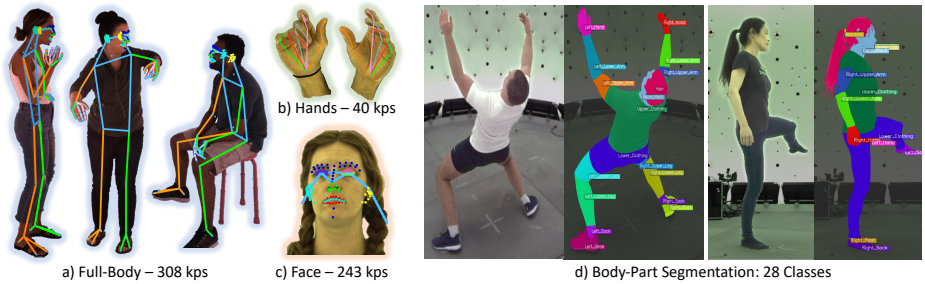

Fig. 4: Ground-truth annotations for 2D pose estimation and body-part segmentation.

3.4 Body-Part Segmentation

Commonly referred to as human parsing, body-part segmentation aims to classify pixels in the input image $\mathbf{I}$ into C classes. Most methods [41] transform this problem to estimating per-pixel class probabilities to create a probability map $\hat{\mathbf{p}} \in \mathbb{R}^{H \times W \times C}$ such that $\hat{\mathbf{p}} = \mathcal{S}(\mathbf{I})$, where $\mathcal{S}$ is the segmentation model. As outlined previously, we adopt the same encoder-decoder architecture and initialization scheme for $\mathcal{S}$. $\mathcal{S}$ is finetuned to minimize the weighted cross-entropy loss between the actual $\mathbf{p}$ and predicted $\hat{\mathbf{p}}$ probability maps, $\mathcal{L}_{\text{seg}} = \texttt{WeightedCE}(\mathbf{p}, \hat{\mathbf{p}})$.

We finetune $\mathcal{S}$ across two part-segmentation vocabularies: a standard set with $C = 20$ [41] and a new larger vocabulary with $C = 28$, as illustrated in Fig.4 (*Right*). Our proposed vocabulary goes beyond previous datasets in important ways. It distinguishes between the upper and lower halves of limbs and incorporates more detailed classifications such as upper/lower lips, teeth, and tongue. To this end, we manually annotate $100K$ images at $4K$ resolution with this vocabulary.

3.5 Depth Estimation

For depth estimation, we adopt the architecture used for segmentation, with the modification that the decoder output channel is set to 1 for regression. We denote the ground-truth depth map of image $\mathbf{I}$ by $\mathbf{d} \in \mathbb{R}^{H \times W}$, the depth estimator by $\mathcal{D}$, where $\hat{\mathbf{d}} = \mathcal{D}(\mathbf{I})$, and M as the number of human pixels in the image. For the relative depth estimation, we normalize $\mathbf{d}$ to the range $[0, 1]$ using max and min depths in the image. The $\mathcal{L}_{\text{depth}}$ loss [32] for $\mathcal{D}$ is defined as follows:

$$\Delta\mathbf{d} = \log(\mathbf{d}) - \log(\hat{\mathbf{d}}), \tag{1}$$

$$\overline{\Delta\mathbf{d}} = \frac{1}{M}\sum_{i=1}^{M}\Delta\mathbf{d}_i, \qquad \overline{(\Delta\mathbf{d})^2} = \frac{1}{M}\sum_{i=1}^{M}(\Delta\mathbf{d}_i)^2, \tag{2}$$

$$\mathcal{L}_{\text{depth}} = \sqrt{\overline{(\Delta\mathbf{d})^2} - \frac{1}{2}\overline{(\Delta\mathbf{d})}^2}. \tag{3}$$

We render $500,000$ synthetic images using 600 high-resolution photogrammetry human scans as shown in Fig. 5 to obtain a robust monocular depth

estimation model with high-fidelity. A random background is selected from a 100 HDRI environment map collection. We place a virtual camera within the scene, randomly adjusting its focal length, rotation, and translation to capture images and their associated ground-truth depth maps at 4K resolution.

Fig. 5: Ground-truth synthetic annotations for depth and surface normal estimation.

3.6 Surface Normal Estimation

Similar to previous tasks, we set the decoder output channels of the normal estimator $\mathcal{N}$ to be 3, corresponding to the xyz components of the normal vector at each pixel. The generated synthetic data is also used as supervision for surface normal estimation. Let $\mathbf{n}$ be the ground-truth normal map for image $\mathbf{I}$ and $\hat{\mathbf{n}} = \mathcal{N}(\mathbf{I})$. Similar to depth, the loss $\mathcal{L}_{\text{normal}}$ is only computed for human pixels in the image and is defined as follows:

$$\mathcal{L}_{\text{normal}} = ||\mathbf{n} - \hat{\mathbf{n}}||_1 + (1 - \mathbf{n} \cdot \hat{\mathbf{n}}) \tag{4}$$

4 Experiments

In this section, we initially provide an overview of the implementation details. Subsequently, we conduct comprehensive benchmarking across four tasks: pose estimation, part segmentation, depth estimation, and normal estimation.

4.1 Implementation Details

Our largest model, Sapiens-2B, is pretrained using 1024 A100 GPUs for 18 days using PyTorch. We use the AdamW [73] optimizer for all our experiments. The learning schedule includes a brief linear warm-up, followed by cosine annealing [72] for pretraining and linear decay [65] for finetuning. All models are pretrained from scratch at a resolution of 1024×1024 with a patch size of 16. For finetuning, the input image is resized to a 4:3 ratio, *i.e.* 1024×768. We use standard augmentations like cropping, scaling, flipping, and photometric distortions. A random background from non-human COCO [67] images is added for

Model	#Params	FLOPs	Hidden size	Layers	Heads	Batch size
Sapiens-0.3B	0.336 B	1.242 T	1024	24	16	98,304
Sapiens-0.6B	0.664 B	2.583 T	1280	32	16	65,536
Sapiens-1B	1.169 B	4.647 T	1536	40	24	40,960
Sapiens-2B	2.163 B	8.709 T	1920	48	32	20,480

Table 2: Sapiens encoder specifications for pretraining on Human-300M dataset.

segmentation, depth, and normal prediction tasks. Importantly, we use differential learning rates [115] to preserve generalization *i.e.* lower learning rates for initial layers and progressively higher rates for subsequent layers. The layer-wise learning rate decay is set to 0.85 with a weight decay of 0.1 for the encoder. We detail the design specifications of Sapiens in Table. 2. Following [34, 100], we prioritize scaling models by width rather than depth. Note that the Sapiens-0.3B model, while architecturally similar to the traditional ViT-Large, consists of twentyfold more FLOPs due to its higher resolution.

4.2 2D Pose Estimation

We finetune Sapiens for face, body, feet, and hand ($K = 308$) pose estimation on our high-fidelity annotations. For training, we use the `train` set with $1M$ images and for evaluation, we use the `test` set, named Humans-5K, with $5K$ images. Our evaluation is top-down [111] *i.e.* we use an off-the-shelf detector [37] for bounding-box and conduct single human pose inference. Table 3 shows a comparison of our models with existing methods for whole-body pose estimation. We evaluate all methods on 114 common keypoints between our 308 keypoint vocabulary and the 133 keypoint vocabulary from COCO-WholeBody [55]. Sapiens-0.6B surpasses the current state-of-the-art, DWPose-l [114] by +2.8 AP. Contrary to DWPose [114], which utilizes a complex student-teacher framework with feature distillation tailored for the task, Sapiens adopts a general encoder-decoder architecture with large human-centric pretraining.

Interestingly, even with the same parameter count, our models demonstrate superior performance compared to their counterparts. For instance, Sapiens-0.3B exceeds VitPose+-L by +5.6 AP, and Sapiens-0.6B outperforms VitPose+-H by +7.9 AP. Within the Sapiens family, our results indicate a direct correlation between model size and performance. Sapiens-2B sets a state-of-the-art with 61.1 AP, a significant improvement of **+7.6 AP** to the prior art. Despite fine-tuning

Model	Input Size	Body		Foot		Face		Hand		Whole-body	
		AP	AR	AP	AR	AP	AR	AP	AR	AP	AR
DeepPose [98]	384 × 288	32.1	43.5	25.3	41.2	37.8	53.9	15.7	31.6	23.9	37.2
SimpleBaseline [106]	384 × 288	52.3	60.1	49.8	62.5	59.6	67.3	41.4	51.8	44.6	53.7
HRNet [93]	384 × 288	55.8	62.6	45.2	55.4	58.9	64.5	39.3	47.6	45.7	53.9
ZoomNAS [110]	384 × 288	59.7	66.3	48.1	57.9	74.5	79.2	49.8	60.6	52.1	60.7
ViTPose+-L [112]	256 × 192	61.0	66.8	62.4	68.2	50.1	55.7	41.5	47.3	47.8	53.6
ViTPose+-H [112]	256 × 192	61.6	67.4	63.2	69.0	50.7	56.3	42.0	47.8	48.3	54.1
RTMPose-x [54]	384 × 288	57.1	63.7	55.3	66.8	74.4	78.5	46.3	55.0	51.9	59.6
DWPose-m [114]	256 × 192	54.2	61.4	49.9	63.0	68.5	74.2	40.1	50.0	47.7	55.8
DWPose-l [114]	384 × 288	57.9	64.2	56.5	67.4	74.3	78.4	49.3	57.4	53.1	60.6
Sapiens-0.3B (Ours)	1024 × 768	58.1	64.5	56.8	67.7	74.5	78.6	49.6	57.7	53.4 (+0.3)	60.9 (+0.3)
Sapiens-0.6B (Ours)	1024 × 768	59.8	65.5	64.7	72.3	75.2	79.0	52.1	60.3	56.2 (+2.8)	62.4 (+2.1)
Sapiens-1B (Ours)	1024 × 768	62.9	68.2	68.3	75.1	76.4	79.7	55.9	63.4	59.4 (+5.9)	65.3 (+5.1)
Sapiens-2B (Ours)	1024 × 768	**64.7**	**69.9**	**69.4**	**76.2**	**76.9**	**79.9**	**57.1**	**64.4**	**61.1**(+7.6)	**67.1**(+7.0)

Table 3: Pose estimation results on Humans-5K `test` set. Flip test is used.

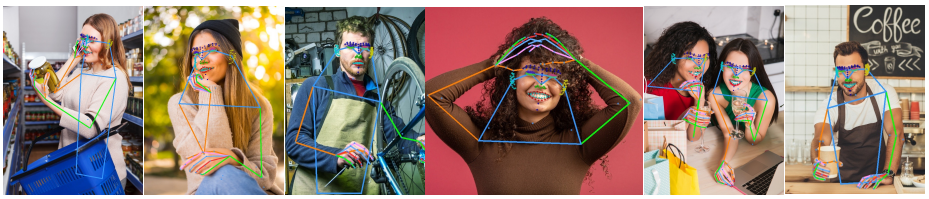

Fig. 6: Pose estimation with Sapiens-1B for 308 keypoints on in-the-wild images.

with annotations from a indoor capture studio, Sapiens demonstrate robust generalization to real-world, as shown in Fig. 6.

4.3 Body-Part Segmentation

We fine-tune and evaluate our annotations with a segmentation vocabulary of 28 classes. Our `train` set consists of $100K$ images, and the `test` set, Humans-2K, consists of $2K$ images. We compare Sapiens with existing body-part segmentation methods fine-tuned on our `train` set. Importantly, we use suggested pretrained checkpoints by each method as initialization. Similar to pose, we observe generalization to segmentation as shown in Table 4.

Model	mIoU(%)	mAcc(%)
FCN* [71]	48.2	57.6
SegFormer* [107]	53.5	62.9
Mask2Former* [22]	58.7	68.3
DeepLabV3+* [18]	64.1	74.8
Sapiens-0.3B (Ours)	76.7	86.1
Sapiens-0.6B (Ours)	77.8	86.3
Sapiens-1B (Ours)	79.9	89.1
Sapiens-2B (Ours)	**81.2**	**89.4**

Table 4: We report mIoU and mAcc on Humans-2K `test` set. Methods with * are trained by us.

Interestingly, our smallest model, Sapiens-0.3B outperforms existing state-of-the-art segmentation methods like Mask2Former [22] and DeepLabV3+ [18] by 12.6 mIoU due to its higher resolution and large human-centric pretraining. Furthermore, increasing the model size improves segmentation performance. Sapiens-2B achieves the best performance of 81.2 mIoU and 89.4 mAcc on the `test` set. Fig. 7 shows the qualitative results of our models.

4.4 Depth Estimation

We evaluate our models on THuman2.0 [117] and Hi4D [116] datasets for depth estimation. THuman2.0 consists of 526 high-quality human scans, from which we derive three sets of images for testing: a) face, b) upper body, and c) full body using a virtual camera. THuman2.0 with 1578 images thus enables the evaluation of our models' performance on single-human images across multiple scales. Conversely, the Hi4D dataset focuses on multi-human scenarios, with each sequence showcasing two subjects engaged in activities involving human-human

Fig. 7: Human segmentation with Sapiens-1B for 28 categories on in-the-wild images.

Method	TH2.0-Face			TH2.0-UprBody			TH2.0-FullBody			Hi4D		
	RMSE ↓	AbsRel ↓	δ_1 ↑	RMSE	AbsRel	δ_1	RMSE	AbsRel	δ_1	RMSE	AbsRel	δ_1
MiDaS-L [11]	0.114	0.097	0.925	0.398	0.271	0.868	0.701	0.689	0.782	0.261	0.082	0.975
MiDaS-Swin2 [11]	0.050	0.036	0.995	0.122	0.081	0.948	0.292	0.171	0.862	0.209	0.063	0.997
DepthAny-B [113]	0.039	0.026	0.999	0.048	0.028	0.999	0.061	0.030	0.999	0.143	**0.034**	0.997
DepthAny-L [113]	0.039	0.027	0.999	0.048	0.027	0.999	0.060	0.030	0.999	0.147	0.035	0.997
Sapiens-0.3B (Ours)	0.012	0.008	1.000	0.015	0.009	1.000	0.021	0.010	1.000	0.148	0.046	1.000
Sapiens-0.6B (Ours)	0.011	0.008	1.000	0.015	0.009	1.000	0.021	0.010	1.000	0.142	0.044	1.000
Sapiens-1B (Ours)	0.009	0.006	1.000	0.012	0.007	1.000	0.019	0.009	1.000	0.125	0.039	1.000
Sapiens-2B (Ours)	**0.008**	**0.005**	**1.000**	**0.010**	**0.006**	**1.000**	**0.016**	**0.008**	**1.000**	**0.114**	0.036	**1.000**

Table 5: Comparison of Sapiens for monocular depth estimation on human images.

interactions. We select sequences from pair 28, pair 32, and pair 37, featuring 6 unique subjects from camera 4, totaling 1195 multi-human real images for testing. We follow the relative-depth evaluation protocols established by MiDaS-v3.1 [11], reporting standard metrics such as AbsRel and δ_1. In addition, we also report RMSE as our primary metric since δ_1 does not effectively reflect performance in human scenes characterized by subtle depth variations.

Table 5 compares our models with existing state-of-the-art monocular depth estimators. Sapiens-2B, finetuned solely on synthetic data, remarkably outperforms prior art across all single-human scales and multi-human scenarios. We observe a 20% RMSE reduction compared to the top-performing Depth-Anything model on Hi4D images. It is important to highlight that while baseline models are trained on a variety of scenes, Sapiens specializes in human-centric depth estimation. Fig. 8 presents a qualitative comparison of depth estimation between Sapiens-1B and DepthAnything-L. To ensure a fair comparison, the predicted depth is renormalized using the human mask in the baseline visualizations.

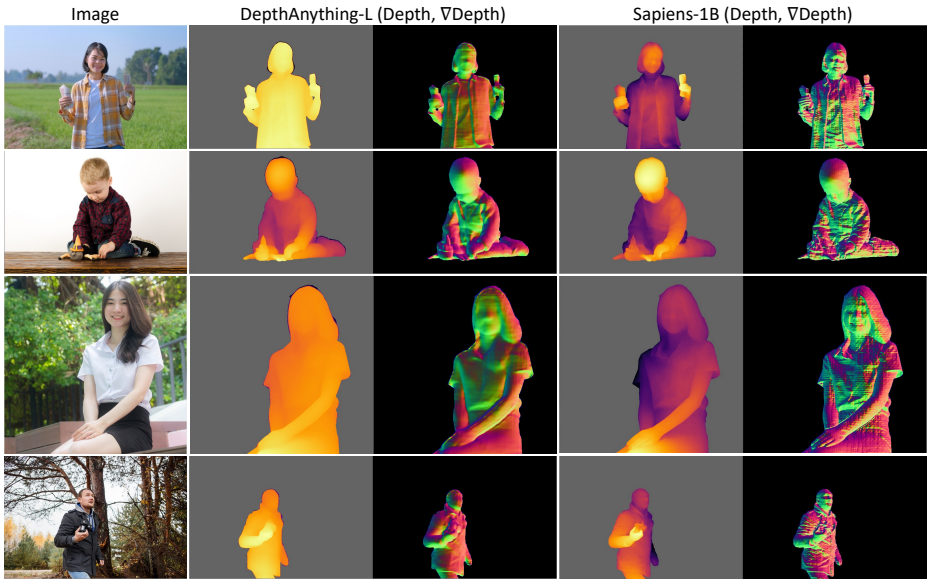

Fig. 8: We compare our depth prediction with DepthAnything [113]. To showcase the consistency of predicted depth, we also visualize the ∇depth as pseudo surface normals.

4.5 Surface Normal Estimation

The datasets for surface normal evaluation are identical to those used for depth estimation. Following [30], we report the mean and median angular error, along with the percentage of pixels within $t°$ error for $t \in \{11.25°, 22.5°, 30°\}$. Table 6 compares our models with existing human-specific surface normal estimators. All our models outperform existing methods by a significant margin. Sapiens-2B achieves a mean error of around 12° on the THuman2.0 (single-human) and Hi4D (multi-human) datasets. We qualitatively compare Sapiens-1B with PIFuHD [89] and ECON [108] for surface normal estimation in Figure 9. Note that PIFuHD [89] is trained with the identical set of 3D scans as ours, and ECON [108] is trained with 4000 scans that are a super set of our 3D scan data.

Method	THuman2.0 [117]					Hi4D [116]				
	Angular Error°		% Within $t°$			Angular Error°		% Within $t°$		
	Mean	Median	11.25°	22.5°	30°	Mean	Median	11.25°	22.5°	30°
PIFuHD [89]	30.51	27.13	15.81	42.97	58.86	22.39	19.26	22.98	60.14	77.02
HDNet [52]	34.82	30.60	17.44	39.26	54.51	28.60	26.85	19.08	57.93	70.14
ICON [109]	28.74	25.52	22.81	47.83	63.73	20.18	17.52	26.81	66.34	82.73
ECON [108]	25.45	23.67	32.95	55.86	69.03	18.46	16.47	29.35	68.12	84.88
Sapiens-0.3B	13.02	10.33	57.37	86.20	92.7	15.04	12.22	47.07	81.49	90.70
Sapiens-0.6B	12.86	10.23	57.85	86.68	93.30	14.06	11.47	50.59	84.37	92.54
Sapiens-1B	12.11	9.40	61.97	88.03	93.84	12.18	**9.59**	**60.36**	88.62	94.44
Sapiens-2B	**11.84**	**9.16**	**63.16**	**88.60**	**94.18**	**12.14**	9.62	60.22	**89.08**	**94.74**

Table 6: Comparison of Sapiens for surface normal estimation on human images.

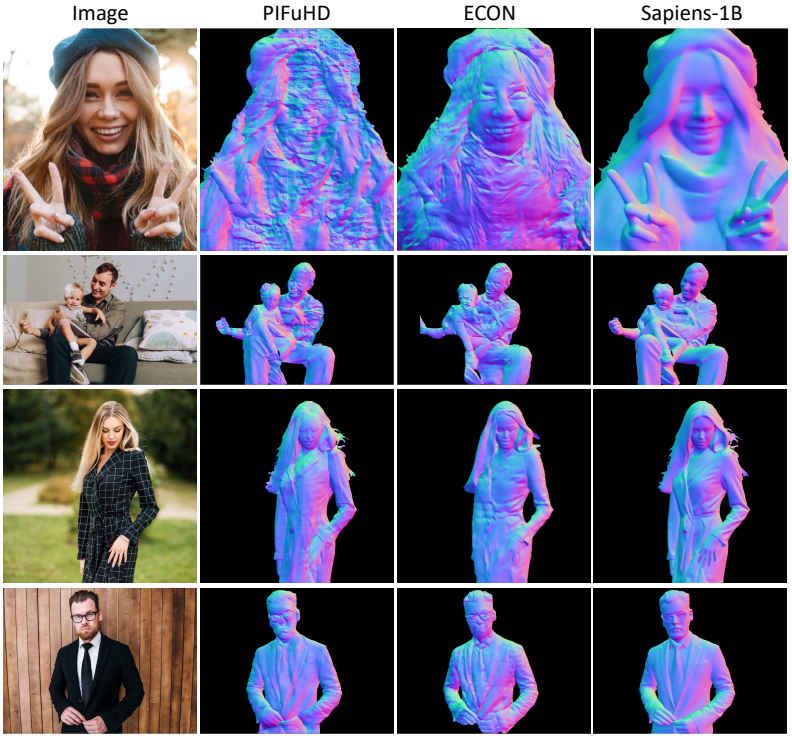

Fig. 9: Qualitative comparison of Sapiens-1B with PIFuHD [89] and ECON [108] for monocular surface normal estimation on challenging in-the-wild images.

4.6 Discussion

Importance of Pretraining Data Source. The feature quality is closely linked to the pretraining data quality. We assess the importance of pretraining on various data sources for human-centric tasks by pretraining Sapiens-0.3B on each dataset under identical training schedules and number of iterations. We fine-tune the model on each task and select early checkpoints for evaluation, reasoning that early-stage fine-tuning better reflects the model's generalization capability. We investigate the impact of pretraining at scale on general images (which may include humans) versus exclusively human images using Sapiens. We randomly select 100 million and 300 million general images from our 1 billion image corpus to create the General-100M and General-300M datasets, respectively. Table 7 showcases the comparison of pretraining outcomes. We report mAP for pose on Humans-5K, mIoU for segmentation on Humans-2K, RMSE for depth on THuman2.0, and mean angular error in degrees for normal estimation on Hi4D. Aligned with findings from [112], our results show that pretraining with Human300M leads to superior performance across all metrics, highlighting the benefits of human-centric pretraining within a fixed computational budget.

We also study the effect of number of unique human images seen during pretraining with normal estimation performance. We report % within 30°. Again, we maintain identical conditions for Sapiens-0.3B pretraining and finetuning. Fig.10 shows a steady improvement in performance as the pretraining data size increases without saturation. In summary, the diversity of human images during pretraining directly correlates with improved generalization to down-stream tasks.

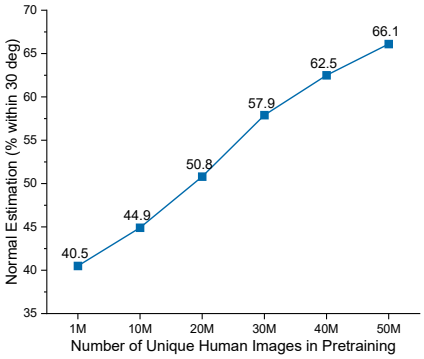

Fig. 10: Sapiens-0.3B's normal estimation performance with unique human images seen during pretraining.

Zero-Shot Generalization. Our models exhibit broad generalization to a variety of settings. For instance, in segmentation, Sapiens are finetuned on single-human images with limited subject diversity, minimal background variation, and solely third-person views (see Fig. 4). Nevertheless, our large-scale pretraining enables generalization across number of subjects, varying ages, and egocentric views, as shown in Fig. 11. These observations similarly hold for other tasks.

Pretraining Source	#Images	Pose (↑)	Seg(↑)	Depth(↓)	Normal(↓)
Random Initialization	-	30.2	40.3	0.720	35.4
General-100M	100M	35.7	50.1	0.351	27.5
General-300M	300M	37.3	52.8	0.347	26.8
Humans-100M	100M	43.6	61.2	0.316	24.0
Humans-300M (Full)	300M	**47.0**	**66.5**	**0.288**	**21.8**

Table 7: Comparison of Sapiens-0.3B pretrained on various data sources. A domain-specific pretraining yields superior results compared to general data sources.

Fig. 11: *Sapiens* achieve broad generalization via large human-centric pretraining.

Applications. Accurately predicting pose, body-part segmentation, depth and surface-normals for in-the-wild human images enables a wide range of applications, such as 3D human scans and the generation of controllable images. We adapt ControlNet [118] to integrate keypoints for faces and hands predicted by Sapiens, and train this modified model on a subset of the Humans-300M dataset, as shown in Fig. 12. This grants more precise control over image generation, producing anatomically consistent human images.

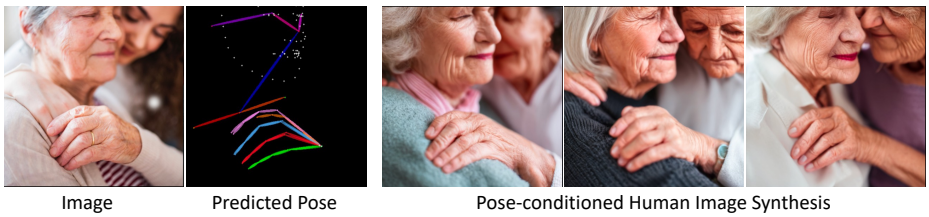

Fig. 12: Image synthesis using ControlNet [118] trained on Sapiens pose predictions.

Limitations. While our models generally perform well, they are not perfect. Human images with complex/rare poses, crowding, and severe occlusion are challenging (see supplemental for details). Although aggressive data augmentation and a detect-and-crop strategy could mitigate these issues, we envision our models as a tool for acquiring large-scale, real-world supervision with human-in-the-loop to develop the next generations of human vision models.

5 Conclusion

Sapiens represents a significant step toward elevating human-centric vision models into the realm of foundation models. Our models demonstrate strong generalization capabilities on a variety of human-centric tasks. We attribute the state-of-the-art performance of our models to: (i) large-scale pretraining on a large curated dataset, which is specifically tailored to understanding humans, (ii) scaled high-resolution and high-capacity vision transformer backbones, and (iii) high-quality annotations on augmented studio and synthetic data. We believe that these models can become a key building block for a multitude of downstream tasks, and provide access to high-quality vision backbones to a significantly wider part of the community. A potential direction for future work would be extending *Sapiens* to 3D and multi-modal datasets.

References

1. Abnar, S., Dehghani, M., Neyshabur, B., Sedghi, H.: Exploring the limits of large scale pre-training. arXiv preprint arXiv:2110.02095 (2021) 4
2. Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F.L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., et al.: Gpt-4 technical report. arXiv preprint arXiv:2303.08774 (2023) 4
3. Alldieck, T., Zanfir, M., Sminchisescu, C.: Photorealistic monocular 3d reconstruction of humans wearing clothing. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1506–1515 (2022) 4
4. Bai, Y., Geng, X., Mangalam, K., Bar, A., Yuille, A., Darrell, T., Malik, J., Efros, A.A.: Sequential modeling enables scalable learning for large vision models. arXiv preprint arXiv:2312.00785 (2023) 4
5. Bao, H., Dong, L., Piao, S., Wei, F.: Beit: Bert pre-training of image transformers. arXiv preprint arXiv:2106.08254 (2021) 3
6. Barron, J.T., Malik, J.: Intrinsic scene properties from a single rgb-d image. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 17–24 (2013) 4
7. Barron, J.T., Malik, J.: Shape, illumination, and reflectance from shading. IEEE transactions on pattern analysis and machine intelligence **37**(8), 1670–1687 (2014) 4
8. Bartol, K., Bojanić, D., Petković, T., Pribanić, T.: A review of body measurement using 3d scanning. Ieee Access **9**, 67281–67301 (2021) 4
9. Bhat, S.F., Alhashim, I., Wonka, P.: Adabins: Depth estimation using adaptive bins. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4009–4018 (2021) 4
10. Bhat, S.F., Birkl, R., Wofk, D., Wonka, P., Müller, M.: Zoedepth: Zero-shot transfer by combining relative and metric depth. arXiv preprint arXiv:2302.12288 (2023) 4
11. Birkl, R., Wofk, D., Müller, M.: Midas v3. 1–a model zoo for robust monocular relative depth estimation. arXiv preprint arXiv:2307.14460 (2023) 11
12. Bojic, L.: Metaverse through the prism of power and addiction: what will happen when the virtual world becomes more attractive than reality? European Journal of Futures Research **10**(1), 1–24 (2022) 4
13. Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J.D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al.: Language models are few-shot learners. Advances in neural information processing systems **33**, 1877–1901 (2020) 4
14. Cao, Z., Simon, T., Wei, S., Sheikh, Y.: Realtime multi-person 2d pose estimation using part affinity fields. corr abs/1611.08050. arXiv preprint arXiv:1611.08050 (2016) 1
15. Cao, Z., Hidalgo, G., Simon, T., Wei, S.E., Sheikh, Y.: Openpose: realtime multi-person 2d pose estimation using part affinity fields. arXiv preprint arXiv:1812.08008 (2018) 3
16. Caron, M., Touvron, H., Misra, I., Jégou, H., Mairal, J., Bojanowski, P., Joulin, A.: Emerging properties in self-supervised vision transformers. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 9650–9660 (2021) 3
17. Chan, C., Ginosar, S., Zhou, T., Efros, A.A.: Everybody dance now. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 5933–5942 (2019) 1

18. Chen, L.C., Zhu, Y., Papandreou, G., Schroff, F., Adam, H.: Encoder-decoder with atrous separable convolution for semantic image segmentation. In: Proceedings of the European conference on computer vision (ECCV). pp. 801–818 (2018) 10
19. Chen, T., Kornblith, S., Norouzi, M., Hinton, G.: A simple framework for contrastive learning of visual representations. In: International conference on machine learning. pp. 1597–1607. PMLR (2020) 3
20. Chen, X., Wang, X., Changpinyo, S., Piergiovanni, A., Padlewski, P., Salz, D., Goodman, S., Grycner, A., Mustafa, B., Beyer, L., et al.: Pali: A jointly-scaled multilingual language-image model. arXiv preprint arXiv:2209.06794 (2022) 5
21. Chen, Y., Wang, Z., Peng, Y., Zhang, Z., Yu, G., Sun, J.: Cascaded pyramid network for multi-person pose estimation. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 7103–7112 (2018) 4
22. Cheng, B., Misra, I., Schwing, A., Kirillov, A., Girdhar, R.: Masked-attention mask transformer for universal image segmentation. arxiv 2022. arXiv preprint arXiv:2112.01527 (2021) 10
23. Couprie, C., Farabet, C., Najman, L., LeCun, Y.: Indoor semantic segmentation using depth information. arXiv preprint arXiv:1301.3572 (2013) 3
24. Dai, Z., Liu, H., Le, Q.V., Tan, M.: Coatnet: Marrying convolution and attention for all data sizes. Advances in neural information processing systems **34**, 3965–3977 (2021) 5
25. Dehghani, M., Djolonga, J., Mustafa, B., Padlewski, P., Heek, J., Gilmer, J., Steiner, A.P., Caron, M., Geirhos, R., Alabdulmohsin, I., et al.: Scaling vision transformers to 22 billion parameters. In: International Conference on Machine Learning. pp. 7480–7512. PMLR (2023) 5
26. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805 (2018) 4
27. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020) 2, 4, 5
28. Drobyshev, N., Chelishev, J., Khakhulin, T., Ivakhnenko, A., Lempitsky, V., Zakharov, E.: Megaportraits: One-shot megapixel neural head avatars. In: Proceedings of the 30th ACM International Conference on Multimedia. pp. 2663–2671 (2022) 1
29. Du, Y., Kips, R., Pumarola, A., Starke, S., Thabet, A., Sanakoyeu, A.: Avatars grow legs: Generating smooth human motion from sparse tracking inputs with diffusion model. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 481–490 (2023) 4
30. Eftekhar, A., Sax, A., Malik, J., Zamir, A.: 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. pp. 10786–10796 (2021) 12
31. Eigen, D., Fergus, R.: Predicting depth, surface normals and semantic labels with a common multi-scale convolutional architecture. In: Proceedings of the IEEE international conference on computer vision. pp. 2650–2658 (2015) 4
32. Eigen, D., Puhrsch, C., Fergus, R.: Depth map prediction from a single image using a multi-scale deep network. Advances in neural information processing systems **27** (2014) 4, 7

33. El-Nouby, A., Izacard, G., Touvron, H., Laptev, I., Jegou, H., Grave, E.: Are large-scale datasets necessary for self-supervised pre-training? arXiv preprint arXiv:2112.10740 (2021) 4
34. El-Nouby, A., Klein, M., Zhai, S., Bautista, M.A., Toshev, A., Shankar, V., Susskind, J.M., Joulin, A.: Scalable pre-training of large autoregressive image models. arXiv preprint arXiv:2401.08541 (2024) 2, 3, 4, 9
35. Fang, H.S., Lu, G., Fang, X., Xie, J., Tai, Y.W., Lu, C.: Weakly and semi supervised human body part parsing via pose-guided knowledge transfer. arXiv preprint arXiv:1805.04310 (2018) 4
36. Fang, H.S., Xie, S., Tai, Y.W., Lu, C.: Rmpe: Regional multi-person pose estimation. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 2334–2343 (2017) 4
37. Fang, Y., Sun, Q., Wang, X., Huang, T., Wang, X., Cao, Y.: Eva-02: A visual representation for neon genesis. arXiv preprint arXiv:2303.11331 (2023) 9
38. Fang, Y., Wang, W., Xie, B., Sun, Q., Wu, L., Wang, X., Huang, T., Wang, X., Cao, Y.: Eva: Exploring the limits of masked visual representation learning at scale. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 19358–19369 (2023) 5
39. Fouhey, D.F., Gupta, A., Hebert, M.: Data-driven 3d primitives for single image understanding. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 3392–3399 (2013) 4
40. Gong, K., Liang, X., Li, Y., Chen, Y., Yang, M., Lin, L.: Instance-level human parsing via part grouping network. In: Proceedings of the European conference on computer vision (ECCV). pp. 770–785 (2018) 4
41. Gong, K., Liang, X., Zhang, D., Shen, X., Lin, L.: Look into person: Self-supervised structure-sensitive learning and a new benchmark for human parsing. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 932–940 (2017) 3, 4, 7
42. Goyal, P., Caron, M., Lefaudeux, B., Xu, M., Wang, P., Pai, V., Singh, M., Liptchinsky, V., Misra, I., Joulin, A., et al.: Self-supervised pretraining of visual features in the wild. arXiv preprint arXiv:2103.01988 (2021) 4
43. Guizilini, V., Vasiljevic, I., Chen, D., Ambruș, R., Gaidon, A.: Towards zero-shot scale-aware monocular depth estimation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 9233–9243 (2023) 4
44. Halstead, M.A., Barsky, B.A., Klein, S.A., Mandell, R.B.: Reconstructing curved surfaces from specular reflection patterns using spline surface fitting of normals. In: Proceedings of the 23rd annual conference on Computer graphics and interactive techniques. pp. 335–342 (1996) 4
45. He, K., Chen, X., Xie, S., Li, Y., Dollár, P., Girshick, R.: Masked autoencoders are scalable vision learners. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 16000–16009 (2022) 3, 4, 5
46. He, K., Fan, H., Wu, Y., Xie, S., Girshick, R.: Momentum contrast for unsupervised visual representation learning. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 9729–9738 (2020) 3
47. He, K., Gkioxari, G., Dollár, P., Girshick, R.B.: Mask r-cnn. corr abs/1703.06870 (2017). arXiv preprint arXiv:1703.06870 (2017) 4
48. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016) 5

49. Hoffmann, J., Borgeaud, S., Mensch, A., Buchatskaya, E., Cai, T., Rutherford, E., Casas, D.d.L., Hendricks, L.A., Welbl, J., Clark, A., et al.: Training compute-optimal large language models. arXiv preprint arXiv:2203.15556 (2022) 5
50. Hu, L., Gao, X., Zhang, P., Sun, K., Zhang, B., Bo, L.: Animate anyone: Consistent and controllable image-to-video synthesis for character animation. arXiv preprint arXiv:2311.17117 (2023) 1
51. Huang, S., Gong, M., Tao, D.: A coarse-fine network for keypoint localization. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 3028–3037 (2017) 4
52. Jafarian, Y., Park, H.S.: 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. pp. 12753–12762 (2021) 4, 12
53. Jiang, A.Q., Sablayrolles, A., Mensch, A., Bamford, C., Chaplot, D.S., Casas, D.d.l., Bressand, F., Lengyel, G., Lample, G., Saulnier, L., et al.: Mistral 7b. arXiv preprint arXiv:2310.06825 (2023) 4
54. Jiang, T., Lu, P., Zhang, L., Ma, N., Han, R., Lyu, C., Li, Y., Chen, K.: Rtmpose: Real-time multi-person pose estimation based on mmpose. arXiv preprint arXiv:2303.07399 (2023) 9
55. Jin, S., Xu, L., Xu, J., Wang, C., Liu, W., Qian, C., Ouyang, W., Luo, P.: Whole-body human pose estimation in the wild. In: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part IX 16. pp. 196–214. Springer (2020) 3, 6, 9
56. Kato, N., Li, T., Nishino, K., Uchida, Y.: Improving multi-person pose estimation using label correction. arXiv preprint arXiv:1811.03331 (2018) 3
57. Khirodkar, R., Chari, V., Agrawal, A., Tyagi, A.: Multi-instance pose networks: Rethinking top-down pose estimation. In: Proceedings of the IEEE/CVF International conference on computer vision. pp. 3122–3131 (2021) 4
58. Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A.C., Lo, W.Y., et al.: Segment anything. arXiv preprint arXiv:2304.02643 (2023) 6
59. Kocabas, M., Chang, J.H.R., Gabriel, J., Tuzel, O., Ranjan, A.: Hugs: Human gaussian splats. arXiv preprint arXiv:2311.17910 (2023) 4
60. Krishna, R., Zhu, Y., Groth, O., Johnson, J., Hata, K., Kravitz, J., Chen, S., Kalantidis, Y., Li, L.J., Shamma, D.A., et al.: Visual genome: Connecting language and vision using crowdsourced dense image annotations. International journal of computer vision **123**, 32–73 (2017) 4
61. Krizhevsky, A., Hinton, G., et al.: Learning multiple layers of features from tiny images. arXiv preprint (2009) 4
62. Ladickỳ, L., Zeisl, B., Pollefeys, M.: Discriminatively trained dense surface normal estimation. In: Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13. pp. 468–484. Springer (2014) 4
63. Lawrence, J., Goldman, D.B., Achar, S., Blascovich, G.M., Desloge, J.G., Fortes, T., Gomez, E.M., Häberling, S., Hoppe, H., Huibers, A., et al.: Project starline: A high-fidelity telepresence system. arXiv preprint (2021) 4
64. Levoy, M., Pulli, K., Curless, B., Rusinkiewicz, S., Koller, D., Pereira, L., Ginzton, M., Anderson, S., Davis, J., Ginsberg, J., et al.: The digital michelangelo project: 3d scanning of large statues. In: Proceedings of the 27th annual conference on Computer graphics and interactive techniques. pp. 131–144 (2000) 4
65. Lewkowycz, A.: How to decay your learning rate. arXiv preprint arXiv:2103.12682 (2021) 8

66. Li, Z., Wang, X., Liu, X., Jiang, J.: Binsformer: Revisiting adaptive bins for monocular depth estimation. arXiv preprint arXiv:2204.00987 (2022) 4
67. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: European conference on computer vision. pp. 740–755. Springer (2014) 1, 4, 6, 8
68. Liu, Z., Hu, H., Lin, Y., Yao, Z., Xie, Z., Wei, Y., Ning, J., Cao, Y., Zhang, Z., Dong, L., et al.: Swin transformer v2: Scaling up capacity and resolution. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 12009–12019 (2022) 5
69. Lombardi, S., Simon, T., Saragih, J., Schwartz, G., Lehrmann, A., Sheikh, Y.: Neural volumes: Learning dynamic renderable volumes from images. arXiv preprint arXiv:1906.07751 (2019) 1, 4
70. Lombardi, S., Simon, T., Schwartz, G., Zollhoefer, M., Sheikh, Y., Saragih, J.: Mixture of volumetric primitives for efficient neural rendering. ACM Transactions on Graphics (ToG) **40**(4), 1–13 (2021) 4
71. Long, J., Shelhamer, E., Darrell, T., Berkeley, U.: Fully convolutional networks for semantic segmentation. arxiv 2015. arXiv preprint arXiv:1411.4038 (2014) 10
72. Loshchilov, I., Hutter, F.: Sgdr: Stochastic gradient descent with warm restarts. arXiv preprint arXiv:1608.03983 (2016) 8
73. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. arXiv preprint arXiv:1711.05101 (2017) 8
74. Lowe, D.G.: Three-dimensional object recognition from single two-dimensional images. Artificial intelligence **31**(3), 355–395 (1987) 4
75. Luo, Y., Zheng, Z., Zheng, L., Guan, T., Yu, J., Yang, Y.: Macro-micro adversarial network for human parsing. In: Proceedings of the European conference on computer vision (ECCV). pp. 418–434 (2018) 4
76. Ma, S., Simon, T., Saragih, J., Wang, D., Li, Y., De La Torre, F., Sheikh, Y.: Pixel codec avatars. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 64–73 (2021) 4
77. Mahajan, D., Girshick, R., Ramanathan, V., He, K., Paluri, M., Li, Y., Bharambe, A., Van Der Maaten, L.: Exploring the limits of weakly supervised pretraining. In: Proceedings of the European conference on computer vision (ECCV). pp. 181–196 (2018) 4
78. Newell, A., Yang, K., Deng, J.: Stacked hourglass networks for human pose estimation. In: European conference on computer vision. pp. 483–499. Springer (2016) 4
79. Oquab, M., Darcet, T., Moutakanni, T., Vo, H., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., et al.: Dinov2: Learning robust visual features without supervision. arXiv preprint arXiv:2304.07193 (2023) 2, 3, 4
80. Papandreou, G., Zhu, T., Kanazawa, N., Toshev, A., Tompson, J., Bregler, C., Murphy, K.: Towards accurate multi-person pose estimation in the wild. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 4903–4911 (2017) 4
81. Peebles, W., Xie, S.: Scalable diffusion models with transformers. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 4195–4205 (2023) 5
82. render people: 3d people for architectural visualization | renderpeople, https://renderpeople.com/3d-people, accessed: 2024-02-22 3

83. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International conference on machine learning. pp. 8748–8763. PMLR (2021) 4
84. Ranftl, R., Lasinger, K., Hafner, D., Schindler, K., Koltun, V.: 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
85. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 10684–10695 (2022) 4
86. Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., et al.: Imagenet large scale visual recognition challenge. International journal of computer vision **115**, 211–252 (2015) 4
87. Saharia, C., Chan, W., Saxena, S., Li, L., Whang, J., Denton, E.L., Ghasemipour, K., Gontijo Lopes, R., Karagol Ayan, B., Salimans, T., et al.: Photorealistic text-to-image diffusion models with deep language understanding. Advances in Neural Information Processing Systems **35**, 36479–36494 (2022) 4
88. Saito, S., Huang, Z., Natsume, R., Morishima, S., Kanazawa, A., Li, H.: Pifu: Pixel-aligned implicit function for high-resolution clothed human digitization. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 2304–2314 (2019) 4
89. Saito, S., Simon, T., Saragih, J., Joo, H.: 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. pp. 84–93 (2020) 1, 4, 12
90. Schuhmann, C., Beaumont, R., Vencu, R., Gordon, C., Wightman, R., Cherti, M., Coombes, T., Katta, A., Mullis, C., Wortsman, M., et al.: Laion-5b: An open large-scale dataset for training next generation image-text models. Advances in Neural Information Processing Systems **35**, 25278–25294 (2022) 4
91. Singh, M., Duval, Q., Alwala, K.V., Fan, H., Aggarwal, V., Adcock, A., Joulin, A., Dollár, P., Feichtenhofer, C., Girshick, R., et al.: The effectiveness of mae pre-pretraining for billion-scale pretraining. arXiv preprint arXiv:2303.13496 (2023) 2, 3, 4
92. Sun, C., Shrivastava, A., Singh, S., Gupta, A.: Revisiting unreasonable effectiveness of data in deep learning era. In: Proceedings of the IEEE international conference on computer vision. pp. 843–852 (2017) 4
93. Sun, K., Xiao, B., Liu, D., Wang, J.: Deep high-resolution representation learning for human pose estimation. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 5693–5703 (2019) 4, 9
94. Taori, R., Gulrajani, I., Zhang, T., Dubois, Y., Li, X., Guestrin, C., Liang, P., Hashimoto, T.B.: Alpaca: A strong, replicable instruction-following model. Stanford Center for Research on Foundation Models. https://crfm. stanford. edu/2023/03/13/alpaca. html **3**(6), 7 (2023) 5
95. Tay, Y., Dehghani, M., Rao, J., Fedus, W., Abnar, S., Chung, H.W., Narang, S., Yogatama, D., Vaswani, A., Metzler, D.: Scale efficiently: Insights from pre-training and fine-tuning transformers. arXiv preprint arXiv:2109.10686 (2021) 4

96. Team, G., Anil, R., Borgeaud, S., Wu, Y., Alayrac, J.B., Yu, J., Soricut, R., Schalkwyk, J., Dai, A.M., Hauth, A., et al.: Gemini: a family of highly capable multimodal models. arXiv preprint arXiv:2312.11805 (2023) 4
97. Thomee, B., Shamma, D.A., Friedland, G., Elizalde, B., Ni, K., Poland, D., Borth, D., Li, L.J.: Yfcc100m: The new data in multimedia research. Communications of the ACM **59**(2), 64–73 (2016) 4
98. Toshev, A., Szegedy, C.: Deeppose: Human pose estimation via deep neural networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1653–1660 (2014) 9
99. Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., et al.: Llama: Open and efficient foundation language models. arXiv preprint arXiv:2302.13971 (2023) 4
100. Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., Bhosale, S., et al.: Llama 2: Open foundation and fine-tuned chat models. arXiv preprint arXiv:2307.09288 (2023) 4, 5, 9
101. Wang, X., Fouhey, D., Gupta, A.: Designing deep networks for surface normal estimation. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 539–547 (2015) 3, 4
102. Weng, C.Y., Curless, B., Srinivasan, P.P., Barron, J.T., Kemelmacher-Shlizerman, I.: Humannerf: Free-viewpoint rendering of moving people from monocular video. In: Proceedings of the IEEE/CVF conference on computer vision and pattern Recognition. pp. 16210–16220 (2022) 1
103. Wu, Y., Kirillov, A., Massa, F., Lo, W.Y., Girshick, R.: Detectron2. https://github.com/facebookresearch/detectron2 (2019) 5
104. Xia, F., Wang, P., Chen, L.C., Yuille, A.L.: Zoom better to see clearer: Human and object parsing with hierarchical auto-zoom net. In: Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part V 14. pp. 648–663. Springer (2016) 4
105. Xia, F., Wang, P., Chen, X., Yuille, A.L.: Joint multi-person pose estimation and semantic part segmentation. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 6769–6778 (2017) 4
106. Xiao, B., Wu, H., Wei, Y.: Simple baselines for human pose estimation and tracking. In: Proceedings of the European conference on computer vision (ECCV). pp. 466–481 (2018) 4, 9
107. Xie, E., Wang, W., Yu, Z., Anandkumar, A., Alvarez, J.M., Luo, P.: Segformer: Simple and efficient design for semantic segmentation with transformers. Advances in neural information processing systems **34**, 12077–12090 (2021) 10
108. Xiu, Y., Yang, J., Cao, X., Tzionas, D., Black, M.J.: Econ: Explicit clothed humans optimized via normal integration. arXiv preprint arXiv:2212.07422 (2022) 1, 4, 12
109. Xiu, Y., Yang, J., Tzionas, D., Black, M.J.: Icon: Implicit clothed humans obtained from normals. In: 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 13286–13296. IEEE (2022) 1, 12
110. Xu, L., Jin, S., Liu, W., Qian, C., Ouyang, W., Luo, P., Wang, X.: Zoomnas: searching for whole-body human pose estimation in the wild. IEEE Transactions on Pattern Analysis and Machine Intelligence **45**(4), 5296–5313 (2022) 9
111. Xu, Y., Zhang, J., Zhang, Q., Tao, D.: Vitpose: Simple vision transformer baselines for human pose estimation. Advances in Neural Information Processing Systems **35**, 38571–38584 (2022) 6, 9

112. Xu, Y., Zhang, J., Zhang, Q., Tao, D.: Vitpose+: Vision transformer foundation model for generic body pose estimation. arXiv preprint arXiv:2212.04246 (2022) 9, 13
113. Yang, L., Kang, B., Huang, Z., Xu, X., Feng, J., Zhao, H.: Depth anything: Unleashing the power of large-scale unlabeled data. arXiv preprint arXiv:2401.10891 (2024) 1, 3, 4, 11
114. Yang, Z., Zeng, A., Yuan, C., Li, Y.: Effective whole-body pose estimation with two-stages distillation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 4210–4220 (2023) 9
115. Yang, Z., Dai, Z., Yang, Y., Carbonell, J., Salakhutdinov, R.R., Le, Q.V.: Xlnet: Generalized autoregressive pretraining for language understanding. Advances in neural information processing systems **32** (2019) 9
116. Yin, Y., Guo, C., Kaufmann, M., Zarate, J.J., Song, J., Hilliges, O.: Hi4d: 4d instance segmentation of close human interaction. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 17016–17027 (2023) 10, 12
117. Yu, T., Zheng, Z., Guo, K., Liu, P., Dai, Q., Liu, Y.: Function4d: Real-time human volumetric capture from very sparse consumer rgbd sensors. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 5746–5756 (2021) 10, 12
118. Zhang, L., Rao, A., Agrawala, M.: Adding conditional control to text-to-image diffusion models. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 3836–3847 (2023) 1, 14
119. Zhang, S.H., Li, R., Dong, X., Rosin, P., Cai, Z., Han, X., Yang, D., Huang, H., Hu, S.M.: Pose2seg: Detection free human instance segmentation. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 889–898 (2019) 1, 3
120. Zhang, X., Lu, Y., Wang, W., Yan, A., Yan, J., Qin, L., Wang, H., Yan, X., Wang, W.Y., Petzold, L.R.: Gpt-4v (ision) as a generalist evaluator for vision-language tasks. arXiv preprint arXiv:2311.01361 (2023) 4
121. Zhou, J., Wei, C., Wang, H., Shen, W., Xie, C., Yuille, A., Kong, T.: ibot: Image bert pre-training with online tokenizer. arXiv preprint arXiv:2111.07832 (2021) 3, 4
122. Zlateski, A., Jaroensri, R., Sharma, P., Durand, F.: On the importance of label quality for semantic segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 1479–1487 (2018) 3