

Sapiens: Foundation for Human Vision Models

Rawal Khirodkar, Timur Bagautdinov, Julieta Martinez, Su Zhaoen,
 Austin James, Peter Selednik, Stuart Anderson, Shunsuke Saito
<https://about.meta.com/realitylabs/codecavatars/sapiens>



Figure 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.

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 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 improves as we scale the number of parameters from 0.3 to 2 billion. *Sapiens* consistently surpasses existing baselines across various human-centric benchmarks. 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.

“*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, essential for faithful human generation tasks. 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, 47, 48, 121], we choose the masked-autoencoder (MAE) approach [48] 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 end-to-end. 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, 40, 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 [84] 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 dataset of 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 or even synthetic labels, demonstrate in-the-wild generalization.
- The first 1K resolution model that natively supports high-fidelity inference for human-centric tasks, achieving state-of-the-art performance on benchmarks for 2D pose, 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, 82, 85, 87, 120] are progressively embracing extensive data scales for pre-training. 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) [48] 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 finetuning 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, 35, 46, 51, 57, 78, 80, 93, 106], body-part segmentation [36, 40, 41, 41, 75, 104, 105], depth estima-

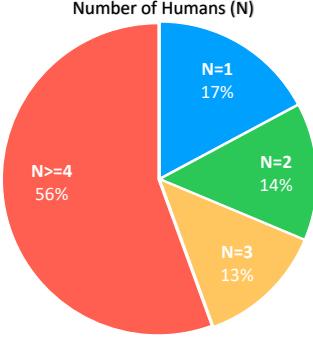


Figure 2. Overview of number of humans per image in the Humans-300M dataset.

tion [9, 10, 32, 43, 52, 66, 83, 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, 45, 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.

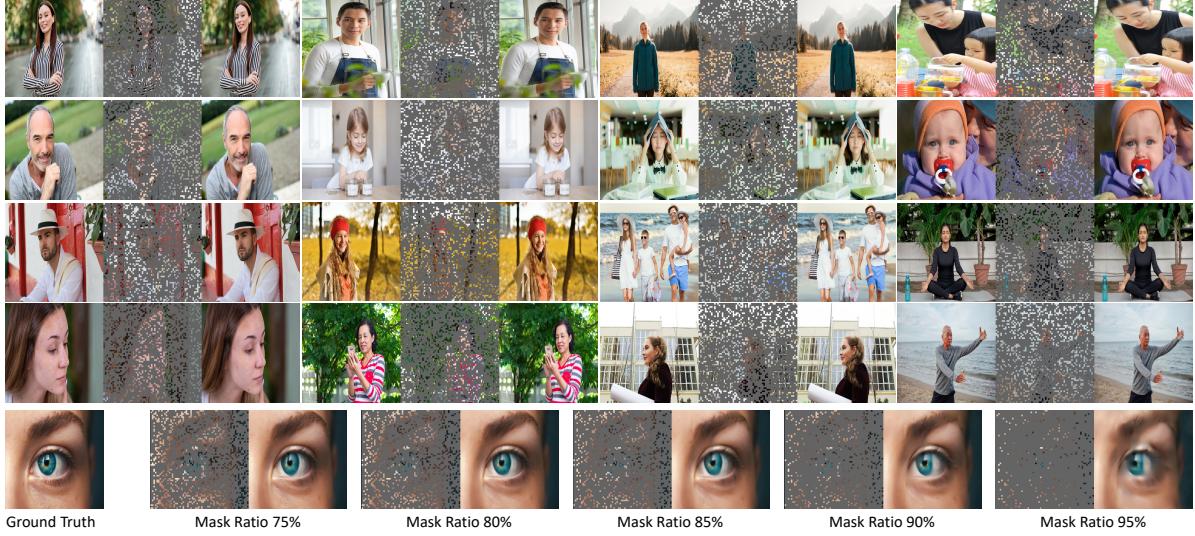


Figure 3. *Sapiens* reconstruction on unseen images. *Top*: Each triplet contains the ground truth (left), the masked image (center), and the MAE reconstruction (right), with a masking ratio of 75%, a patch size of 16, and an image size of 1024. *Bottom*: Varying the mask ratio between [0.75, 0.95] during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

3.2. Pretraining

We follow the masked-autoencoder [48] (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 [48] 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 $16\times$ 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 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 prob-

ability 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 $\mathcal{L}_{\text{pose}} = \text{MSE}(\mathbf{y}, \hat{\mathbf{y}})$. During finetuning, the encoder of \mathcal{P} is initialized with the weights from pretraining, 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 the encoder and decoder in \mathcal{P} across multiple skeletons, including $K = 17$ [67], $K = 133$ [55] and a new highly-detailed skeleton, 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, 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.

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 [40] 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

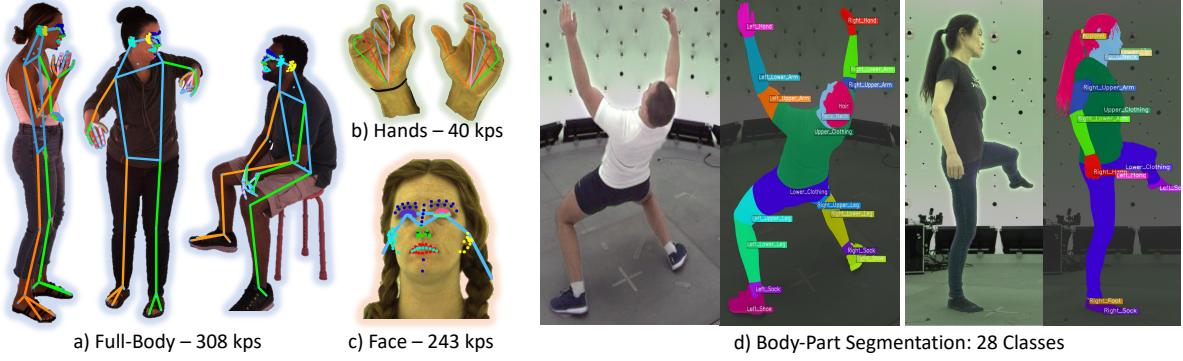


Figure 4. Ground-truth annotations for 2D pose estimation and body-part segmentation.

\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, $\mathcal{L}_{\text{seg}} = \text{WeightedCE}(\mathbf{p}, \hat{\mathbf{p}})$.

We finetune \mathcal{S} across two part-segmentation vocabularies: a standard set with $C = 20$ [40] 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:

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

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

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

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.



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

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.

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:

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

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 segmentation, depth, and normal prediction tasks. Importantly, we use differential learning rates [114] 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, DW-Pose-1 [115] by +2.8 AP. Contrary to DW-Pose [115], 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 with annotations from a indoor capture studio, Sapiens demonstrate robust generalization to real-world, as shown in Fig. 6.



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

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 [115]	256 × 192	54.2	61.4	49.9	63.0	68.5	74.2	40.1	50.0	47.7	55.8
DWPose-l [115]	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.

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.

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 pre-training. 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.

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.



Figure 7. Body-part segmentation with Sapiens-1B for 28 categories on single and multi-human images.

Method	TH2.0-Face			TH2.0-UprBody			TH2.0-FullBody			Hi4D		
	RMSE ↓	AbsRel ↓	$\delta_1 \uparrow$	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.

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 interactions. We select sequences from pair 28, 32, and 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.

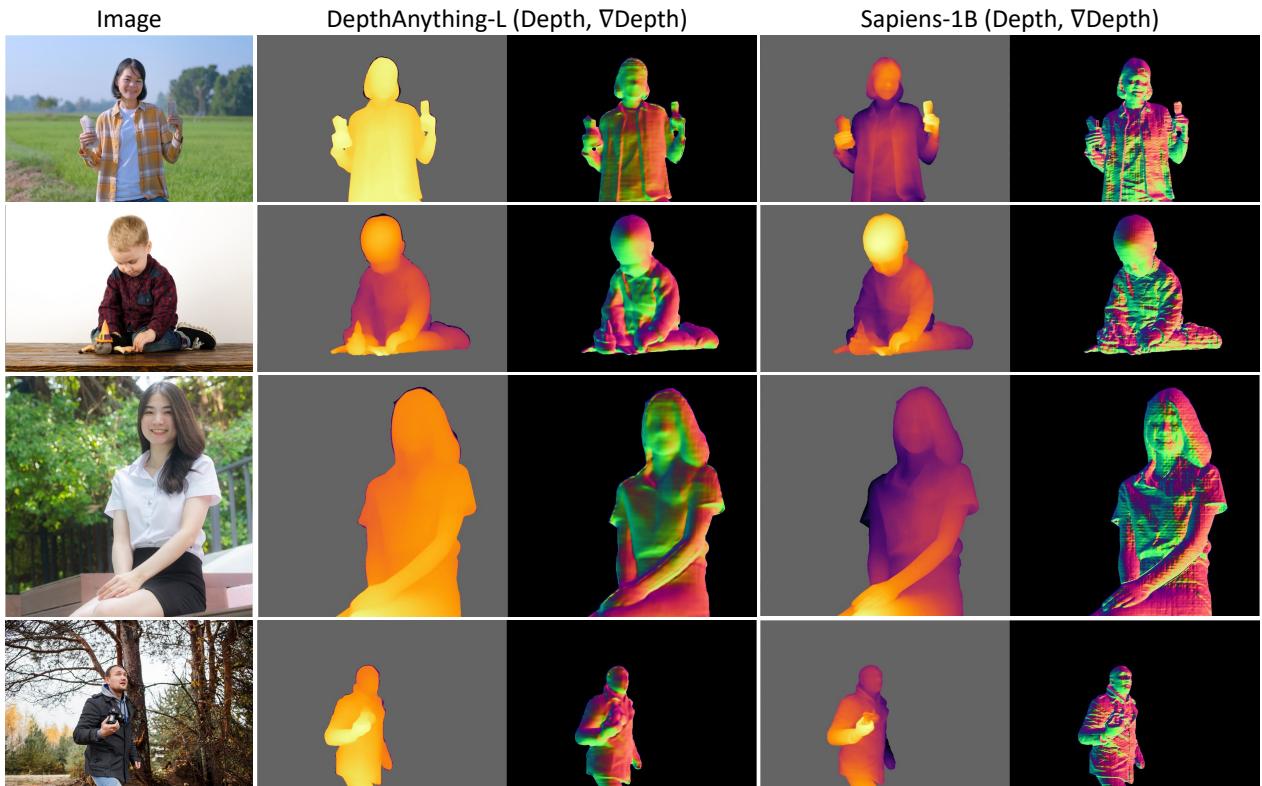


Figure 8. We compare our depth prediction with DepthAnything [113]. To showcase the consistency of predicted depth, we also visualize the ∇ depth as pseudo surface normals.

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.

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 $t \in \{11.25^\circ, 22.5^\circ, 30^\circ\}$. 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 PI-

FuHD [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.

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

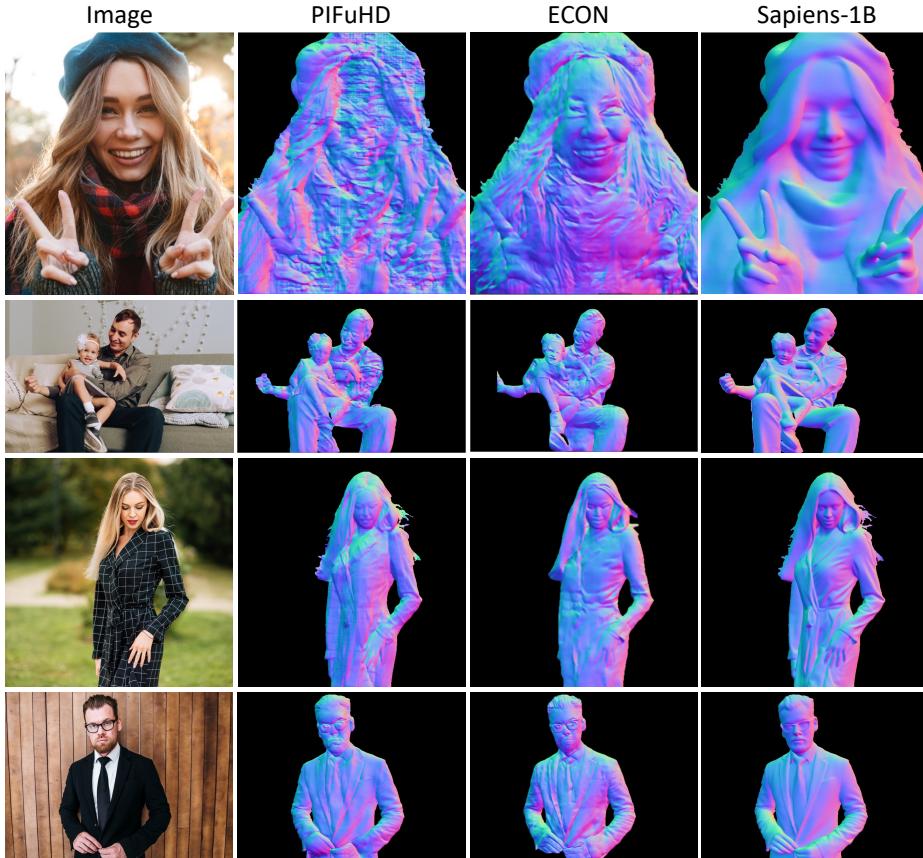


Figure 9. Qualitative comparison of Sapiens-1B with PIFuHD [89] and ECON [108] for surface normal estimation on in-the-wild images.

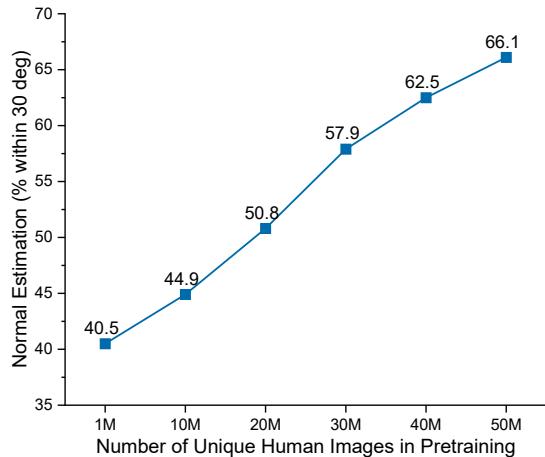


Figure 10. Sapiens-0.3B’s normal estimation performance with unique human images seen during pretraining.

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 fine-

Pretraining Source	#Images	Pose (\uparrow)	Seg(\uparrow)	Depth(\downarrow)	Normal(\downarrow)
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.



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

tuning. Fig.10 shows a steady improvement in performance as the pretraining data size increases without saturation. In summary, the diversity of human images observed during pretraining directly correlates with improved generalization to down-stream tasks.

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.

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.

Acknowledgements: We would like to acknowledge He Wen and Srivathsan Govindarajan for their contributions with training, and optimizing Sapiens.

References

- [1] Samira Abnar, Mostafa Dehghani, Behnam Neyshabur, and Hanie Sedghi. Exploring the limits of large scale pre-training. *arXiv preprint arXiv:2110.02095*, 2021. 3
- [2] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023. 3
- [3] Thiendo Alldieck, Mihai Zanfir, and Cristian Sminchisescu. Photorealistic monocular 3d reconstruction of humans wearing clothing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1506–1515, 2022. 3
- [4] Yutong Bai, Xinyang Geng, Karttikeya Mangalam, Amir Bar, Alan Yuille, Trevor Darrell, Jitendra Malik, and Alexei A Efros. Sequential modeling enables scalable learning for large vision models. *arXiv preprint arXiv:2312.00785*, 2023. 3
- [5] Hangbo Bao, Li Dong, Songhao Piao, and Furu Wei. Beit: Bert pre-training of image transformers. *arXiv preprint arXiv:2106.08254*, 2021. 2
- [6] Jonathan T Barron and Jitendra Malik. Intrinsic scene properties from a single rgb-d image. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 17–24, 2013. 3
- [7] Jonathan T Barron and Jitendra Malik. Shape, illumination, and reflectance from shading. *IEEE transactions on pattern analysis and machine intelligence*, 37(8):1670–1687, 2014. 3
- [8] Kristijan Bartol, David Bojanic, Tomislav Petković, and Tomislav Pribanic. A review of body measurement using 3d scanning. *Ieee Access*, 9:67281–67301, 2021. 3
- [9] Shariq Farooq Bhat, Ibraheem Alhashim, and Peter Wonka. Adabins: Depth estimation using adaptive bins. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4009–4018, 2021. 3
- [10] Shariq Farooq Bhat, Reiner Birk, Diana Wofk, Peter Wonka, and Matthias Müller. Zoedepth: Zero-shot transfer by combining relative and metric depth. *arXiv preprint arXiv:2302.12288*, 2023. 3
- [11] Reiner Birk, Diana Wofk, and Matthias Müller. Midas v3. 1—a model zoo for robust monocular relative depth estimation. *arXiv preprint arXiv:2307.14460*, 2023. 8
- [12] Ljubisa Bojic. 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. 3
- [13] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020. 3
- [14] Z Cao, T Simon, S Wei, and Y Sheikh. Realtime multi-person 2d pose estimation using part affinity fields. corr abs/1611.08050. *arXiv preprint arXiv:1611.08050*, 2016. 2
- [15] Zhe Cao, Gines Hidalgo, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Openpose: realtime multi-person 2d pose estimation using part affinity fields. *arXiv preprint arXiv:1812.08008*, 2018. 2
- [16] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9650–9660, 2021. 2
- [17] Caroline Chan, Shiry Ginosar, Tinghui Zhou, and Alexei A Efros. Everybody dance now. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 5933–5942, 2019. 2
- [18] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florentin Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *Proceedings of the European conference on computer vision (ECCV)*, pages 801–818, 2018. 7
- [19] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR, 2020. 2
- [20] Xi Chen, Xiao Wang, Soravit Changpinyo, AJ Piergiovanni, Piotr Padlewski, Daniel Salz, Sebastian Goodman, Adam Grycner, Basil Mustafa, Lucas Beyer, et al. Pali: A jointly-scaled multilingual language-image model. *arXiv preprint arXiv:2209.06794*, 2022. 3
- [21] Yilun Chen, Zhicheng Wang, Yuxiang Peng, Zhiqiang Zhang, Gang Yu, and Jian Sun. Cascaded pyramid network for multi-person pose estimation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7103–7112, 2018. 3
- [22] B Cheng, I Misra, AG Schwing, A Kirillov, and R Girdhar. Masked-attention mask transformer for universal image segmentation. arxiv 2022. *arXiv preprint arXiv:2112.01527*, 2021. 7
- [23] Camille Couprise, Clément Farabet, Laurent Najman, and Yann LeCun. Indoor semantic segmentation using depth information. *arXiv preprint arXiv:1301.3572*, 2013. 2
- [24] Zihang Dai, Hanxiao Liu, Quoc V Le, and Mingxing Tan. Coatnet: Marrying convolution and attention for all data sizes. *Advances in neural information processing systems*, 34:3965–3977, 2021. 3
- [25] Mostafa Dehghani, Josip Djolonga, Basil Mustafa, Piotr Padlewski, Jonathan Heek, Justin Gilmer, Andreas Peter Steiner, Mathilde Caron, Robert Geirhos, Ibrahim Alabdulmohsin, et al. Scaling vision transformers to 22 billion parameters. In *International Conference on Machine Learning*, pages 7480–7512. PMLR, 2023. 3
- [26] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018. 3
- [27] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner,

- Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. [2](#), [3](#), [4](#)
- [28] Nikita Drobyshev, Jenya Chelishev, Taras Khakhulin, Aleksei Ivakhnenko, Victor Lempitsky, and Egor Zakharov. Megaportraits: One-shot megapixel neural head avatars. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 2663–2671, 2022. [2](#)
- [29] Yuming Du, Robin Kips, Albert Pumarola, Sebastian Starke, Ali Thabet, and Artsiom Sanakoyeu. 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*, pages 481–490, 2023. [3](#)
- [30] Ainaz Eftekhari, 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. [9](#)
- [31] David Eigen and Rob Fergus. Predicting depth, surface normals and semantic labels with a common multi-scale convolutional architecture. In *Proceedings of the IEEE international conference on computer vision*, pages 2650–2658, 2015. [3](#)
- [32] David Eigen, Christian Puhrsch, and Rob Fergus. Depth map prediction from a single image using a multi-scale deep network. *Advances in neural information processing systems*, 27, 2014. [3](#), [5](#)
- [33] Alaaeldin El-Nouby, Gautier Izacard, Hugo Touvron, Ivan Laptev, Hervé Jegou, and Edouard Grave. Are large-scale datasets necessary for self-supervised pre-training? *arXiv preprint arXiv:2112.10740*, 2021. [3](#)
- [34] Alaaeldin El-Nouby, Michal Klein, Shuangfei Zhai, Miguel Angel Bautista, Alexander Toshev, Vaishaal Shankar, Joshua M Susskind, and Armand Joulin. Scalable pre-training of large autoregressive image models. *arXiv preprint arXiv:2401.08541*, 2024. [2](#), [3](#), [6](#)
- [35] Hao-Shu Fang, Shuqin Xie, Yu-Wing Tai, and Cewu Lu. Rmpe: Regional multi-person pose estimation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2334–2343, 2017. [3](#)
- [36] Hao-Shu Fang, Guansong Lu, Xiaolin Fang, Jianwen Xie, Yu-Wing Tai, and Cewu Lu. Weakly and semi supervised human body part parsing via pose-guided knowledge transfer. *arXiv preprint arXiv:1805.04310*, 2018. [3](#)
- [37] Yuxin Fang, Quan Sun, Xinggang Wang, Tiejun Huang, Xinlong Wang, and Yue Cao. Eva-02: A visual representation for neon genesis. *arXiv preprint arXiv:2303.11331*, 2023. [6](#)
- [38] Yuxin Fang, Wen Wang, Binhui Xie, Quan Sun, Ledell Wu, Xinggang Wang, Tiejun Huang, Xinlong Wang, and Yue Cao. Eva: Exploring the limits of masked visual representation learning at scale. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19358–19369, 2023. [3](#)
- [39] David F Fouhey, Abhinav Gupta, and Martial Hebert. Data-driven 3d primitives for single image understanding. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3392–3399, 2013. [3](#)
- [40] Ke Gong, Xiaodan Liang, Dongyu Zhang, Xiaohui Shen, and Liang Lin. 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*, pages 932–940, 2017. [2](#), [3](#), [4](#), [5](#)
- [41] Ke Gong, Xiaodan Liang, Yicheng Li, Yimin Chen, Ming Yang, and Liang Lin. Instance-level human parsing via part grouping network. In *Proceedings of the European conference on computer vision (ECCV)*, pages 770–785, 2018. [3](#)
- [42] Priya Goyal, Mathilde Caron, Benjamin Lefauveaux, Min Xu, Pengchao Wang, Vivek Pai, Mannat Singh, Vitaliy Liptchinsky, Ishan Misra, Armand Joulin, et al. Self-supervised pretraining of visual features in the wild. *arXiv preprint arXiv:2103.01988*, 2021. [3](#)
- [43] Vitor Guizilini, Igor Vasiljevic, Dian Chen, Rareş Ambrus, and Adrien Gaidon. Towards zero-shot scale-aware monocular depth estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9233–9243, 2023. [3](#)
- [44] Mark A Halstead, Brian A Barsky, Stanley A Klein, and Robert B Mandell. 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*, pages 335–342, 1996. [3](#)
- [45] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. [3](#)
- [46] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross B Girshick. Mask r-cnn. corr abs/1703.06870 (2017). *arXiv preprint arXiv:1703.06870*, 2017. [3](#)
- [47] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9729–9738, 2020. [2](#)
- [48] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16000–16009, 2022. [2](#), [3](#), [4](#)
- [49] Jordan Hoffmann, Sébastien Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, et al. Training compute-optimal large language models. *arXiv preprint arXiv:2203.15556*, 2022. [3](#)
- [50] Li Hu, Xin Gao, Peng Zhang, Ke Sun, Bang Zhang, and Liefeng Bo. Animate anyone: Consistent and controllable image-to-video synthesis for character animation. *arXiv preprint arXiv:2311.17117*, 2023. [2](#)
- [51] Shaoli Huang, Mingming Gong, and Dacheng Tao. A coarse-fine network for keypoint localization. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3028–3037, 2017. [3](#)

- [52] 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. 3, 9
- [53] Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lamplé, Lucile Saulnier, et al. Mistral 7b. *arXiv preprint arXiv:2310.06825*, 2023. 3
- [54] Tao Jiang, Peng Lu, Li Zhang, Ningsheng Ma, Rui Han, Chengqi Lyu, Yining Li, and Kai Chen. Rtmpose: Real-time multi-person pose estimation based on mmpose. *arXiv preprint arXiv:2303.07399*, 2023. 7
- [55] 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 *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part IX 16*, pages 196–214. Springer, 2020. 2, 4, 6
- [56] Naoki Kato, Tianqi Li, Kohei Nishino, and Yusuke Uchida. Improving multi-person pose estimation using label correction. *arXiv preprint arXiv:1811.03331*, 2018. 2
- [57] Rawal Khirodkar, Visesh Chari, Amit Agrawal, and Ambrish Tyagi. Multi-instance pose networks: Rethinking top-down pose estimation. In *Proceedings of the IEEE/CVF International conference on computer vision*, pages 3122–3131, 2021. 3
- [58] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. *arXiv preprint arXiv:2304.02643*, 2023. 4
- [59] Muhammed Kocabas, Jen-Hao Rick Chang, James Gabriel, Oncel Tuzel, and Anurag Ranjan. Hugs: Human gaussian splats. *arXiv preprint arXiv:2311.17910*, 2023. 3
- [60] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International journal of computer vision*, 123:32–73, 2017. 3
- [61] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. *arXiv preprint*, 2009. 3
- [62] L’ubor Ladický, Bernhard Zeisl, and Marc Pollefeys. Discriminatively trained dense surface normal estimation. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6–12, 2014, Proceedings, Part V 13*, pages 468–484. Springer, 2014. 3
- [63] Jason Lawrence, Dan B Goldman, Supreeth Achar, Gregory Major Blascovich, Joseph G Desloge, Tommy Fortes, Eric M Gomez, Sascha Häberling, Hugues Hoppe, Andy Huibers, et al. Project starline: A high-fidelity telepresence system. *arXiv preprint*, 2021. 3
- [64] Marc Levoy, Kari Pulli, Brian Curless, Szymon Rusinkiewicz, David Koller, Lucas Pereira, Matt Ginzton, Sean Anderson, James Davis, Jeremy Ginsberg, et al. The digital michelangelo project: 3d scanning of large statues. In *Proceedings of the 27th annual conference on Computer graphics and interactive techniques*, pages 131–144, 2000. 3
- [65] Aitor Lewkowycz. How to decay your learning rate. *arXiv preprint arXiv:2103.12682*, 2021. 6
- [66] Zhenyu Li, Xuyang Wang, Xianming Liu, and Junjun Jiang. Binsformer: Revisiting adaptive bins for monocular depth estimation. *arXiv preprint arXiv:2204.00987*, 2022. 3
- [67] 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. In *European conference on computer vision*, pages 740–755. Springer, 2014. 2, 3, 4, 6
- [68] Ze Liu, Han Hu, Yutong Lin, Zhuliang Yao, Zhenda Xie, Yixuan Wei, Jia Ning, Yue Cao, Zheng Zhang, Li Dong, et al. Swin transformer v2: Scaling up capacity and resolution. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12009–12019, 2022. 3
- [69] Stephen Lombardi, Tomas Simon, Jason Saragih, Gabriel Schwartz, Andreas Lehrmann, and Yaser Sheikh. Neural volumes: Learning dynamic renderable volumes from images. *arXiv preprint arXiv:1906.07751*, 2019. 2, 3
- [70] Stephen Lombardi, Tomas Simon, Gabriel Schwartz, Michael Zollhoefer, Yaser Sheikh, and Jason Saragih. Mixture of volumetric primitives for efficient neural rendering. *ACM Transactions on Graphics (ToG)*, 40(4):1–13, 2021. 3
- [71] Jonathan Long, Evan Shelhamer, Trevor Darrell, and UC Berkeley. Fully convolutional networks for semantic segmentation. *arxiv* 2015. *arXiv preprint arXiv:1411.4038*, 2014. 7
- [72] Ilya Loshchilov and Frank Hutter. Sgdr: Stochastic gradient descent with warm restarts. *arXiv preprint arXiv:1608.03983*, 2016. 6
- [73] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 6
- [74] David G Lowe. Three-dimensional object recognition from single two-dimensional images. *Artificial intelligence*, 31(3):355–395, 1987. 3
- [75] Yawei Luo, Zhedong Zheng, Liang Zheng, Tao Guan, Junqing Yu, and Yi Yang. Macro-micro adversarial network for human parsing. In *Proceedings of the European conference on computer vision (ECCV)*, pages 418–434, 2018. 3
- [76] Shugao Ma, Tomas Simon, Jason Saragih, Dawei Wang, Yuecheng Li, Fernando De La Torre, and Yaser Sheikh. Pixel codec avatars. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 64–73, 2021. 3
- [77] Dhruv Mahajan, Ross Girshick, Vignesh Ramanathan, Kaiming He, Manohar Paluri, Yixuan Li, Ashwin Bharambe, and Laurens Van Der Maaten. Exploring the limits of weakly supervised pretraining. In *Proceedings of the European conference on computer vision (ECCV)*, pages 181–196, 2018. 3
- [78] Alejandro Newell, Kaiyu Yang, and Jia Deng. Stacked hourglass networks for human pose estimation. In *European conference on computer vision*, pages 483–499. Springer, 2016. 3

- [79] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023. 2, 3
- [80] George Papandreou, Tyler Zhu, Nori Kanazawa, Alexander Toshev, Jonathan Tompson, Chris Bregler, and Kevin Murphy. Towards accurate multi-person pose estimation in the wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4903–4911, 2017. 3
- [81] William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4195–4205, 2023. 3
- [82] 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. 3
- [83] René Ranftl, Katrin Lasinger, David Hafner, Konrad Schindler, and Vladlen Koltun. Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer. *IEEE transactions on pattern analysis and machine intelligence*, 44(3):1623–1637, 2020. 3
- [84] render people. 3d people for architectural visualization — renderpeople. Accessed: 2024-02-22. 2
- [85] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022. 3
- [86] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115:211–252, 2015. 3
- [87] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in Neural Information Processing Systems*, 35:36479–36494, 2022. 3
- [88] Shunsuke Saito, Zeng Huang, Ryota Natsume, Shigeo Morishima, Angjoo Kanazawa, and Hao Li. Pifu: Pixel-aligned implicit function for high-resolution clothed human digitization. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 2304–2314, 2019. 3
- [89] 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. 2, 3, 9
- [90] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, 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. 3
- [91] Mannat Singh, Quentin Duval, Kalyan Vasudev Alwala, Haoqi Fan, Vaibhav Aggarwal, Aaron Adcock, Armand Joulin, Piotr Dollár, Christoph Feichtenhofer, Ross Girshick, et al. The effectiveness of mae pre-training for billion-scale pretraining. *arXiv preprint arXiv:2303.13496*, 2023. 2, 3
- [92] Chen Sun, Abhinav Shrivastava, Saurabh Singh, and Abhinav Gupta. Revisiting unreasonable effectiveness of data in deep learning era. In *Proceedings of the IEEE international conference on computer vision*, pages 843–852, 2017. 3
- [93] Ke Sun, Bin Xiao, Dong Liu, and Jingdong Wang. Deep high-resolution representation learning for human pose estimation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5693–5703, 2019. 3, 7
- [94] Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B Hashimoto. 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. 3
- [95] Yi Tay, Mostafa Dehghani, Jinfeng Rao, William Fedus, Samira Abnar, Hyung Won Chung, Sharan Narang, Dani Yogatama, Ashish Vaswani, and Donald Metzler. Scale efficiently: Insights from pre-training and fine-tuning transformers. *arXiv preprint arXiv:2109.10686*, 2021. 3
- [96] Gemini Team, Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023. 3
- [97] Bart Thomee, David A Shamma, Gerald Friedland, Benjamin Elizalde, Karl Ni, Douglas Poland, Damian Borth, and Li-Jia Li. Yfcc100m: The new data in multimedia research. *Communications of the ACM*, 59(2):64–73, 2016. 3
- [98] Alexander Toshev and Christian Szegedy. Deeppose: Human pose estimation via deep neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1653–1660, 2014. 7
- [99] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023. 3
- [100] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaee, Nikolay Bashlykov, Soumya Batra, Prajwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023. 3, 6
- [101] Xiaolong Wang, David Fouhey, and Abhinav Gupta. Designing deep networks for surface normal estimation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 539–547, 2015. 2, 3
- [102] Chung-Yi Weng, Brian Curless, Pratul P Srinivasan, Jonathan T Barron, and Ira Kemelmacher-Shlizerman. Hu-

- mannerf: Free-viewpoint rendering of moving people from monocular video. In *Proceedings of the IEEE/CVF conference on computer vision and pattern Recognition*, pages 16210–16220, 2022. 2
- [103] Yuxin Wu, Alexander Kirillov, Francisco Massa, Wan-Yen Lo, and Ross Girshick. Detectron2. <https://github.com/facebookresearch/detectron2>, 2019. 3
- [104] Fangting Xia, Peng Wang, Liang-Chieh Chen, and Alan L Yuille. 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*, pages 648–663. Springer, 2016. 3
- [105] Fangting Xia, Peng Wang, Xianjie Chen, and Alan L Yuille. Joint multi-person pose estimation and semantic part segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6769–6778, 2017. 3
- [106] Bin Xiao, Haiping Wu, and Yichen Wei. Simple baselines for human pose estimation and tracking. In *Proceedings of the European conference on computer vision (ECCV)*, pages 466–481, 2018. 3, 7
- [107] Enze Xie, Wenhui Wang, Zhiding Yu, Anima Anandkumar, Jose M Alvarez, and Ping Luo. Segformer: Simple and efficient design for semantic segmentation with transformers. *Advances in neural information processing systems*, 34:12077–12090, 2021. 7
- [108] Yuliang Xiu, Jinlong Yang, Xu Cao, Dimitrios Tzionas, and Michael J Black. Econ: Explicit clothed humans optimized via normal integration. *arXiv preprint arXiv:2212.07422*, 2022. 2, 3, 9
- [109] 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. 2, 9
- [110] Lumin Xu, Sheng Jin, Wentao Liu, Chen Qian, Wanli Ouyang, Ping Luo, and Xiaogang Wang. Zoomnas: searching for whole-body human pose estimation in the wild. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(4):5296–5313, 2022. 7
- [111] Yufei Xu, Jing Zhang, Qiming Zhang, and Dacheng Tao. Vitpose: Simple vision transformer baselines for human pose estimation. *Advances in Neural Information Processing Systems*, 35:38571–38584, 2022. 4, 6
- [112] Yufei Xu, Jing Zhang, Qiming Zhang, and Dacheng Tao. Vitpose+: Vision transformer foundation model for generic body pose estimation. *arXiv preprint arXiv:2212.04246*, 2022. 7, 10
- [113] Lihe Yang, Bingyi Kang, Zilong Huang, Xiaogang Xu, Jiasi Feng, and Hengshuang Zhao. Depth anything: Unleashing the power of large-scale unlabeled data. *arXiv preprint arXiv:2401.10891*, 2024. 2, 3, 8
- [114] Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. Xlnet: Generalized autoregressive pretraining for language understanding. *Advances in neural information processing systems*, 32, 2019. 6
- [115] Zhendong Yang, Ailing Zeng, Chun Yuan, and Yu Li. Effective whole-body pose estimation with two-stages distillation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4210–4220, 2023. 6, 7
- [116] Yifei Yin, Chen Guo, Manuel Kaufmann, Juan Jose Zarate, Jie Song, and Otmar Hilliges. Hi4d: 4d instance segmentation of close human interaction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17016–17027, 2023. 8, 9
- [117] 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 *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5746–5756, 2021. 8, 9
- [118] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3836–3847, 2023. 2
- [119] Song-Hai Zhang, Ruilong Li, Xin Dong, Paul Rosin, Zixi Cai, Xi Han, Dingcheng Yang, Haozhi Huang, and Shi-Min Hu. Pose2seg: Detection free human instance segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 889–898, 2019. 2
- [120] Xinlu Zhang, Yujie Lu, Weizhi Wang, An Yan, Jun Yan, Lianke Qin, Heng Wang, Xifeng Yan, William Yang Wang, and Linda Ruth Petzold. Gpt-4v (ision) as a generalist evaluator for vision-language tasks. *arXiv preprint arXiv:2311.01361*, 2023. 3
- [121] Jinghao Zhou, Chen Wei, Huiyu Wang, Wei Shen, Ci-hang Xie, Alan Yuille, and Tao Kong. ibot: Image bert pre-training with online tokenizer. *arXiv preprint arXiv:2111.07832*, 2021. 2, 3
- [122] Aleksandar Zlateski, Ronnachai Jaroensri, Prafull Sharma, and Frédéric Durand. On the importance of label quality for semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1479–1487, 2018. 2