# Sapiens: Foundation for Human Vision Models

Xiantong Xiang

Shandong University

Tuesday, December 3rd, 2024

Khirodkar R, Bagautdinov T, Martinez J, et al. Sapiens: Foundation for Human Vision Models [A]. 2024: arXiv:2408.12569. arXiv: 2408.12569.

# Contents

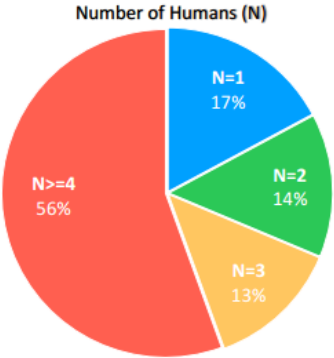
# Background

- leveraging large datasets and scalable model architectures is key for generalization
- What type of data is most effective for pretraining?
- the critical impact of label quality on the model’ s in-the-wild performance

# Dataset



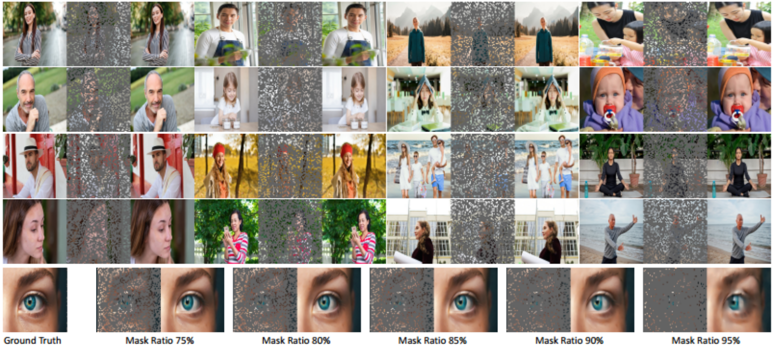
- Humans-300M
- approximately 1 billion in-the-wild images
- Figure provides an overview of the distribution of the number of people per image in our dataset, noting that over 248 million images contain multiple subjects



# Pretraining



- Masked-Autoencoder (MAE)
  - **Encoder**: maps the visible image to a latent representation
  - **Decoder**: reconstructs the original image from this latent representation
- an image ⇒ regular non-overlapping patches (fixed patch size)

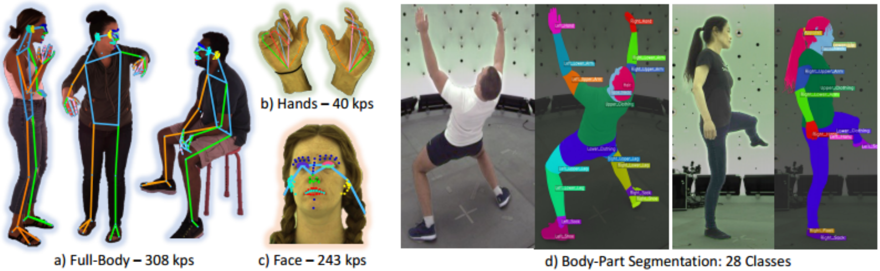


# 2D Pose Estimation

- **Heatmap Prediction**:
  - detect the locations of $K$ keypoints from an input image $I \in \mathbb{R}^{H \times W \times 3}$
  - each of K heatmaps represents the probability of the corresponding keypoint being at any spatial location
- **Pose Estimation Transformer $\mathcal{P}$**:
  - **Input**: $I \in \mathbb{R}^{H \times W \times 3}$
  - **Output**: $y \in \mathbb{R}^{H \times W \times K}$
    K heatmaps corresponding to the ground truth keypoints
  - Minimize the mean squared loss:

$$\mathcal{L}_{\text{pose}} = \text{MSE}(y, \hat{y})$$

# 2D Pose Estimation




- Compared to existing formats with **at most 68 facial keypoints**, their annotations consist of **243 facial keypoints**, including representative points around the eyes, lips, nose, and ears.
- annotated **1 million** images at **4K** resolution

# Body-Part Segmentation



- **Method**: adopt the same encoder-decoder architecture and initialization scheme
- minimize the **weighted cross-entropy loss**

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

- finetune $S$ across two part-segmentation vocabularies
  - a standard set with C = 20
  - a new larger vocabulary with C = 28
- annotate **100K** images at **4K** resolution

# Depth Estimation



- **Method**
  - adopt the architecture used for segmentation
  - the decoder outputchannel is set to 1 for regression
- The $\mathcal{L}_{\text{depth}}$ loss for $\mathcal{D}$ is defined as follows:

$$\Delta d = \log(d) - \log\left(\hat{d}\right)$$

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

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

groundtruth depth map $d \in \mathbb{R}^{H \times W}$, $\hat{d} = D(I)$, the number of human pixels $M$

# Depth Estimation



- **the relative depth estimation**:
  - normalize $d$ to the range $[0, 1]$ using max and min depths
- render $\mathbf{500,000}$ synthetic images using 600 highresolution photogrammetry human scans
- **4K** resolution

# Surface Normal Estimation



- **Method**
  - adopt the same architecture
  - the decoder outputchannel is set 3
- the loss $\mathcal{L}_{\text{normal}}$ is only computed for human pixels

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

# 2D Pose Estimation



- **Pretrain**: 1024 A100 GPUs for 18 days
- **AdamW optimizer**
- **The learning schedule**
  - a brief linear warm-up
  - cosine annealing for pretraining
  - linear decay for finetuning
- **differential learning rates**
  - lower learning rates for initial layers
  - progressively higher rates for subsequent layers

# 2D Pose Estimation



| 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) |

- Sapiens-0.3B exceeds VitPose+-L by +5.6 AP
- Sapiens-0.6B outperforms VitPose+-H by +7.9 AP
- Sapiens-2B, a significant improvement of +7.6 AP to the prior art

# Conclusion



- **largescale pretraining** on a large curated dataset, which is specifically tailored to understanding humans
- scaled highresolution and high-capacity **vision transformer** backbones
- **high-quality annotations** on augmented studio and synthetic data