Sapiens2: Unified Human-Centric Pretraining for Semantics and Pixel Fidelity

Rawal Khirodkar, He Wen, Julieta Martinez, Yuan Dong, Su Zhaoen, Shunsuke Saito

ICLR

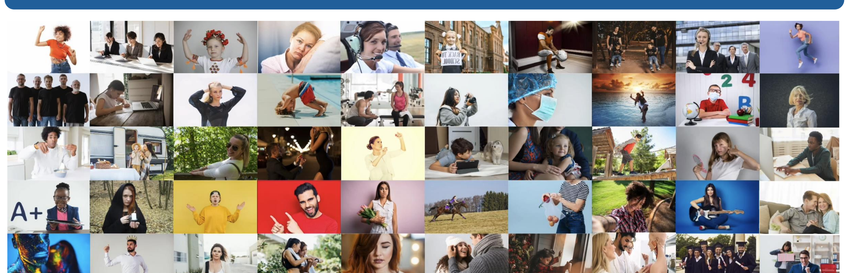
OVERVIEW

Image | Attention Map | Tasks: Seg, Depth, Normal

Contributions:

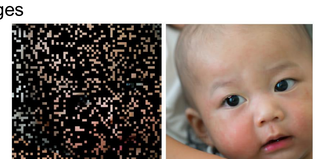
- Vision transformers (0.4B – 5B) pretrained on **1 billion human images**.
- Joint pretraining objective: masked reconstruction + contrastive learning.
- 4K resolution backbones. Learns features that support both high-level semantic reasoning and low-level dense prediction.

HUMAN-CENTRIC PRETRAINING



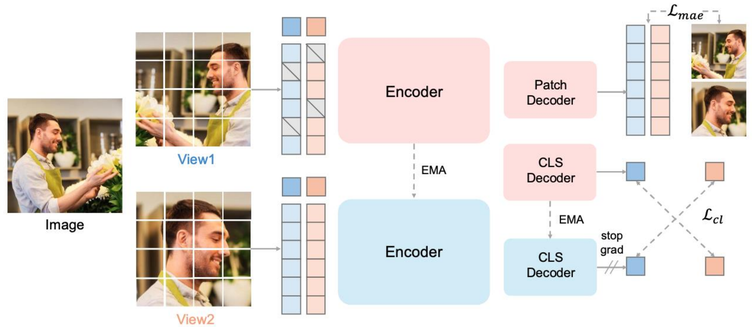
In-the-Wild 1 Billion Human Images

- **Sapiens** (ECCV 2024) based on masked autoencoders.
- Feature encode low-level image details like color.
- Useful for high-resolution dense predictions.
- Lower semantic reasoning compared to contrastive methods like CLIP, SigLIP, DINOv3.

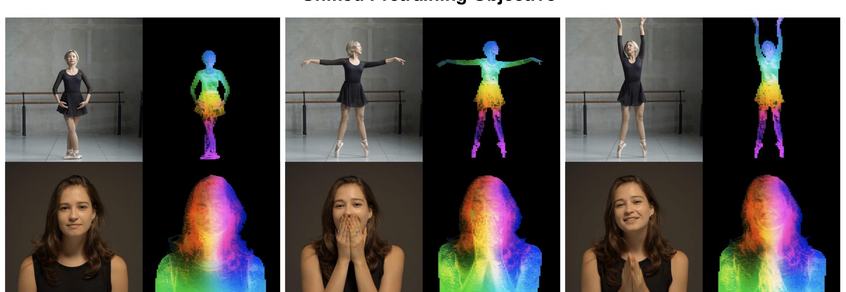


Input (75% Mask) | Output

MASKED MODELING + CONTRASTIVE LEARNING



Unified Pretraining Objective



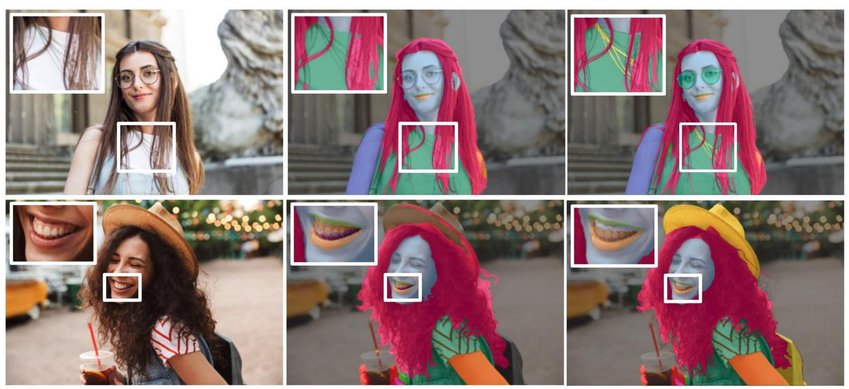
Sapiens2 learn human semantics at scale. Colorized PCA visualization of last layer features.

Query | Retrieval Capabilities: CLS-Token kNN using Sapiens vs Sapiens2 (k=3).

RESULTS

Our models generalize across human tasks with **very few or synthetic labels.**

State-of-the-art performance for part-segmentation, normal and point-map estimation.



Sapiens, ECCV 2024 | Sapiens2 (Ours)

DAViD, ICCV 2025 | Sapiens2 (Ours)

MoGe, CVPR 2025 | Sapiens2 (Ours)