Novelty Assessment Report

Paper: Sapiens2

PDF URL: https://openreview.net/pdf?id=IVAlYCqdvW

Venue: ICLR 2026 Conference Submission

Year: 2026

Report Generated: 2025-12-27

Abstract

We present Sapiens2, a model family of high-resolution transformers for human-centric vision focused on generalization, versatility, and high-fidelity outputs. Our model sizes range from 0.4 to 5 billion parameters, with native 1K resolution and hierarchical variants that support 4K. Sapiens2 substantially improves over its predecessor in both pretraining and post-training. First, to learn features that capture low-level details (for dense prediction) and high-level semantics (for zero-shot or few-label settings), we combine masked image reconstruction with self-distilled contrastive objectives. Our evaluations show that this unified pretraining objective is better suited for a wider range of downstream tasks. Second, along the data axis, we pretrain on a curated dataset of 1 billion high-quality human images and improve the quality and quantity of task annotations. Third, architecturally, we incorporate advances from frontier models that enable longer training schedules with improved stability. Our 4K models adopt windowed attention to reason over longer spatial context and are pretrained with 2K output resolution. Sapiens2 sets a new state-of-the-art and improves over the first generation on pose (+4 mAP), body-part segmentation (+22.3 mIoU), normal estimation (+29.2 rel-angular error) and extends to new tasks such as pointmap and albedo estimation.

Disclaimer

This report is **AI-GENERATED** using Large Language Models and WisPaper (a scholar search engine). It analyzes academic papers' tasks and contributions against retrieved prior work. While this system identifies **POTENTIAL** overlaps and novel directions, **ITS COVERAGE IS NOT EXHAUSTIVE AND JUDGMENTS ARE APPROXIMATE**. These results are intended to assist human reviewers and **SHOULD NOT** be relied upon as a definitive verdict on novelty.

Note that some papers exist in multiple, slightly different versions (e.g., with different titles or URLs). The system may retrieve several versions of the same underlying work. The current automated pipeline does not reliably align or distinguish these cases, so human reviewers will need to disambiguate them manually.

If you have any questions, please contact: mingzhang23@m.fudan.edu.cn

Core Task Landscape

This paper addresses: **human-centric vision with high-resolution transformers**

A total of **50 papers** were analyzed and organized into a taxonomy with **18 categories**.

Taxonomy Overview

The research landscape has been organized into the following main categories:

- **Dense Prediction Tasks for Human Body Analysis**
- **High-Resolution Image Generation and Synthesis**
- **Scene Understanding and Environmental Analysis**
- **Efficient Transformer Architectures and Training**
- **Domain-Specific Applications and Specialized Tasks**

Complete Taxonomy Tree

- human-centric vision with high-resolution transformers Survey Taxonomy
- Dense Prediction Tasks for Human Body Analysis
  - Multi-Task Human-Centric Foundation Models ★ (3 papers)
  - [0] Sapiens2 (Anon et al., 2026) View paper
  - [3] Sapiens: Foundation for Human Vision Models (Rawal Khirodkar, 2024) View paper
  - [7] Unihcp: A unified model for human-centric perceptions (Yuanzheng Ci, 2023) View paper
  - Human Pose Estimation Architectures
  - Efficient Sparse Representation Methods (3 papers)
    - [2] Not All Tokens Are Equal: Human-centric Visual Analysis via Token Clustering Transformer (Wang Zeng, 2022) View paper
    - [4] Sharpose: Sparse high-resolution representation for human pose estimation (An XiaoQi, 2024) View paper
    - [23] Potter: Pooling attention transformer for efficient human mesh recovery (Ce Zheng, 2023) View paper
  - High-Resolution Parallel Architectures (4 papers)
    - [20] MTPose: Human pose estimation with high-resolution multi-scale transformers (Rui Wang, 2022) View paper
    - [22] HRPVT: High-Resolution Pyramid Vision Transformer for medium and small-scale human pose estimation (Zhoujie Xu, 2024) View paper
    - [33] HRPoseFormer: High-Resolution Transformer for Human Pose Estimation via Multi-Scale Token Aggregation (Xiao-Wei Yu, 2022) View paper
    - [46] High-Resolution Representation Learning for Human Pose Estimation based on Transformer (Dengyu Fu, 2022) View paper
  - Plain Vision Transformer Baselines (2 papers)
    - [6] ViTPose: Simple Vision Transformer Baselines for Human Pose Estimation (Xu, 2022) View paper
    - [31] AggPose: Deep Aggregation Vision Transformer for Infant Pose Estimation (Cao Xu, 2022) View paper
  - Specialized Pose Estimation Contexts (5 papers)
    - [17] Efficient High-Resolution Visual Representation Learning with State Space Model for Human Pose Estimation (Zhang Hao, 2024) View paper
    - [21] Rethinking the Sparse End-to-End Multiperson Pose Estimation (Xixia Xu, 2025) View paper
    - [25] Adaptive Vision Transformer for Event-Based Human Pose Estimation (Nannan Yu, 2024) View paper
    - [48] UViT: Efficient and lightweight U-shaped hybrid vision transformer for human pose estimation (Biao Li, 2024) View paper
    - [50] HEViTPose: High-Efficiency Vision Transformer for Human Pose Estimation (Wu ChengPeng, 2023) View paper
  - Body Surface Reconstruction and Mesh Recovery (2 papers)

  - [24] Capturing and Inferring Dense Full-Body Human-Scene Contact (Chun-Hao P. Huang, 2022) View paper
  - [30] Quality transformer for human parsing (Yao Guo, 2025) View paper
- High-Resolution Image Generation and Synthesis
  - Human-Centric Image and Video Generation (4 papers)
  - [10] UnitedHuman: Harnessing Multi-Source Data for High-Resolution Human Generation (Fu, 2023) View paper
  - [14] BeyondScene: Higher-Resolution Human-Centric Scene Generation With Pretrained Diffusion (Gwang-Hyun Kim, 2024) View paper
  - [16] OpenHumanVid: A Large-Scale High-Quality Dataset for Enhancing Human-Centric Video Generation (Hui Li, 2024) View paper
  - [45] Scalable High-Resolution Pixel-Space Image Synthesis with Hourglass Diffusion Transformers (Crowson, 2024) View paper
  - Super-Resolution with Vision Transformers (4 papers)
  - [5] Hybrid Vision Transformer and Convolutional Neural Network for Super-Resolution Image Quality Assessment (X Li, 2025) View paper
  - [13] LMLT: Low-to-high Multi-Level Vision Transformer for Lightweight Image Super-Resolution (J Kim, 2025) View paper
  - [26] Lmlt: Low-to-high multi-level vision transformer for image super-resolution (Kim, 2024) View paper
  - [35] SVTSR: image super-resolution using scattering vision transformer (Jiabao Liang, 2024) View paper
  - High-Resolution Scene and Image Synthesis (3 papers)
  - [8] Design of an Integrated Model Combining CycleGAN, PPO, and Vision Transformer for Adaptive Scene Rendering in the Metaverse (Durga Prasad Kavadi, 2025) View paper
  - [27] Swin-UNIT: Transformer-based GAN for High-resolution Unpaired Image Translation (Yifan Li, 2023) View paper
  - [36] StyleSwin: Transformer-based GAN for High-resolution Image Generation (Zhang Bo-wen, 2022) View paper
- Scene Understanding and Environmental Analysis
  - Urban and Environmental Perception (3 papers)
  - [1] Integrating spatiotemporal vision transformer into digital twins for high-resolution heat stress forecasting in campus environments (Gong Wenjing, 2025) View paper
  - [9] Interpretable Multimodal Framework for Human-Centered Street Assessment: Integrating Visual-Language Models for Perceptual Urban Diagnostics (Lan, 2025) View paper
  - [39] Pan-Arctic Permafrost Landform and Human-built Infrastructure Feature Detection with Vision Transformers and Location Embeddings (FernÃ¡ndez, 2025) View paper
  - Ultra-High Resolution Scene Segmentation (3 papers)
  - [11] Ultra-High Resolution Segmentation via Boundary-Enhanced Patch-Merging Transformer (Haopeng Sun, 2024) View paper
  - [28] Adaptive Patching for High-resolution Image Segmentation with Transformers (Zhang, 2024) View paper
  - [32] Recurrent Multi-scale Transformer for High-Resolution Salient Object Detection (Xinhao Deng, 2023) View paper
  - Human-Scene Interaction and Referring Expression (2 papers)
  - [19] A large-scale human-centric benchmark for referring expression comprehension in the LMM era (Fangyun Wei, 2024) View paper
  - [34] Human-Centered Visual Segmentation (Qin, 2025) View paper
- Efficient Transformer Architectures and Training
  - Neural Architecture Search for Transformers (2 papers)
  - [12] Hr-nas: Searching efficient high-resolution neural architectures with lightweight transformers (Ding, 2021) View paper
  - [15] Searching efficient neural architecture with multi-resolution fusion transformer for appearance-based gaze estimation (Vikrant Nagpure, 2023) View paper
  - Efficient Attention Mechanisms (3 papers)
  - [38] Aggregated Contextual Transformations for High-Resolution Image Inpainting. (Zeng Yan-hong, 2023) View paper
  - [40] Glance-and-Gaze Vision Transformer (Yu, 2021) View paper
  - [41] Restormer: Efficient Transformer for High-Resolution Image Restoration (Zamir, 2022) View paper
  - Resolution Adaptation and Scalability (1 papers)
  - [49] ViTAR: Vision Transformer with Any Resolution (Fan, 2024) View paper
- Domain-Specific Applications and Specialized Tasks
  - Medical and Biomedical Imaging (4 papers)
  - [29] Swallowing Assessment using High-Resolution Cervical Auscultations and Transformer-based Neural Networks. (Ayman Anwar, 2025) View paper
  - [42] Toward high resolution scoring system for HER2 expression in breast cancer based on pathological images via deep learning (Yao Li, 2024) View paper
  - [43] Abstract LB087: Inferring single-cell spatial gene expression with tissue morphology via explainable deep learning (Yue Zhao, 2025) View paper
  - [47] Improved Detection of Urolithiasis Using High-Resolution Computed Tomography Images by a Vision Transformer Model (Hyoung Sun Choi, 2023) View paper
  - Multimodal and Cross-Domain Learning (2 papers)
  - [18] Disentangling the Factors of Convergence between Brains and Computer Vision Models (Szafraniec, 2025) View paper
  - [37] LM-MCVT: A Lightweight Multi-modal Multi-view Convolutional-Vision Transformer Approach for 3D Object Recognition (Xiong Song-song, 2025) View paper
  - Theoretical and Methodological Foundations (1 papers)
  - [44] Towards Efficient Deep Learning for Human-Centric Visual Understanding and Generation (Ma, 2024) View paper

Narrative

Core task: human-centric vision with high-resolution transformers. This field centers on leveraging transformer architectures to process high-resolution imagery for tasks involving human subjects, ranging from pose estimation and body parsing to generation and synthesis. The taxonomy reveals five main branches: Dense Prediction Tasks for Human Body Analysis focuses on pixel-level understanding of human anatomy and attributes, often employing multi-task frameworks like ViTPose[6] and unified models such as Unihcp[7]; High-Resolution Image Generation and Synthesis addresses the creation and manipulation of detailed human imagery, including super-resolution methods like Hybrid Vision SuperResolution[5] and style-based approaches such as StyleSwin[36]; Scene Understanding and Environmental Analysis extends beyond isolated humans to contextual reasoning in complex environments; Efficient Transformer Architectures and Training explores computational optimizations like token clustering (Token Clustering Transformer[2]) and neural architecture search (Hr-nas[12]); and Domain-Specific Applications targets specialized settings from medical imaging to metaverse avatars (CycleGAN PPO Metaverse[8]).

A particularly active line of work involves multi-task human-centric foundation models that unify diverse prediction tasks under a single architecture, balancing generalization with task-specific performance. Sapiens2[0] exemplifies this direction by extending the capabilities of its predecessor Sapiens Foundation[3], aiming to handle multiple human analysis tasks at high resolution within a shared framework. Compared to Sapiens Foundation[3], which established foundational multi-task learning for human understanding, Sapiens2[0] pushes toward broader task coverage and improved scalability. Meanwhile, Unihcp[7] represents a parallel effort in unified human-centric prediction, emphasizing modular design for flexible task composition. These works collectively address the trade-off between model complexity and the ability to capture fine-grained human details, a challenge that remains central as the field moves toward more comprehensive and efficient foundation models for human-centric vision.

Related Works in Same Category

The following **2 sibling papers** share the same taxonomy leaf node with the original paper:

1. Sapiens: Foundation for Human Vision Models

Authors: Rawal Khirodkar, Timur Bagautdinov, Julieta MartÃnez, Su Zhaoen, Austin James, et al. (8 authors total) | **Year/Venue**: 2024 • European Conference on Computer Vision | **URL**: View paper

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 si...

⚠ Similarity Notice

These papers appear to be successive versions of the same research project. Both present Sapiens models for human-centric vision tasks with high-resolution transformers, share nearly identical task scopes (pose estimation, body-part segmentation, depth/normal estimation), and follow the same architectural philosophy of self-supervised pretraining on large-scale human image datasets. The candidate paper (Sapiens) is likely the first generation, while the original paper (Sapiens2) represents an evolved version with expanded scale and methodology.

2. Unihcp: A unified model for human-centric perceptions

Authors: Yuanzheng Ci, Yizhou Wang, Meilin Chen, Tang Shixiang, Lei Bai, et al. (10 authors total) | **Year/Venue**: 2023 | **URL**: View paper

Abstract

Human-centric perceptions (e.g., pose estimation, human parsing, pedestrian detection, person re-identification, etc.) play a key role in industrial applications of visual models. While specific human-centric tasks have their own relevant semantic aspect to focus on, they also share the same underlying semantic structure of the human body. However, few works have attempted to exploit such homogeneity and design a general-propose model for human-centric tasks. In this work, we revisit a broad ran...

Relationship Analysis

Both papers belong to the Multi-Task Human-Centric Foundation Models category, focusing on unified models pretrained on large-scale human image datasets for multiple dense prediction tasks. They overlap in addressing pose estimation, body-part segmentation, and depth/normal estimation using transformer architectures with large-scale pretraining. However, Sapiens2 emphasizes high-resolution (1K-4K) processing with masked reconstruction combined with contrastive learning on 750M images, while UniHCP focuses on unifying five distinct human-centric tasks (including ReID and attribute prediction) through task-specific queries and a shared interpreter with 99.97% parameter sharing across 33 datasets.

Contributions Analysis

Overall novelty summary. Sapiens2 contributes a family of high-resolution transformers (0.4–5B parameters) for multi-task human-centric vision, combining masked reconstruction with self-distilled contrastive pretraining and scaling to 4K resolution via windowed attention. The paper resides in the 'Multi-Task Human-Centric Foundation Models' leaf, which contains only three papers including Sapiens2 itself. This is a relatively sparse research direction within the broader taxonomy of 50 papers across 18 leaf nodes, suggesting the work targets an emerging but not yet crowded subfield focused on unified architectures for diverse human analysis tasks.

The taxonomy reveals neighboring leaves addressing related but distinct challenges: 'Human Pose Estimation Architectures' explores specialized keypoint detection methods (e.g., token clustering, high-resolution parallel branches), while 'Body Surface Reconstruction and Mesh Recovery' focuses on 3D geometry. Sapiens2 diverges by pursuing a unified multi-task framework rather than task-specific architectures. The 'High-Resolution Image Generation and Synthesis' branch addresses generative modeling, whereas Sapiens2 emphasizes discriminative dense prediction. The scope_note for its leaf explicitly excludes single-task specialists, positioning Sapiens2 as a generalist foundation model rather than a narrow solution.

Among 30 candidates examined, the unified pretraining objective (Contribution A) shows 2 refutable candidates from 10 examined, indicating some prior work on combining reconstruction and contrastive learning for human-centric tasks. The Humans-750M dataset (Contribution B) found no refutations across 10 candidates, suggesting novelty in dataset curation scale or composition. The 4K hierarchical architecture (Contribution C) encountered 3 refutable candidates from 10, reflecting existing exploration of windowed attention or multi-resolution strategies. The limited search scope means these statistics capture top-30 semantic matches, not exhaustive coverage of all relevant prior work.

Given the sparse taxonomy leaf and the scale of literature examined, Sapiens2 appears to advance an emerging research direction where unified multi-task human-centric models remain relatively underexplored. The contribution-level statistics suggest incremental architectural and pretraining innovations rather than entirely unprecedented techniques, though the dataset and integration choices may offer practical value. This assessment is constrained by the top-30 candidate scope and does not account for concurrent or unpublished work in this rapidly evolving area.

This paper presents **3 main contributions**, each analyzed against relevant prior work:

Contribution 1: Sapiens2 model family with unified pretraining objective

Description: The authors introduce Sapiens2, a family of vision transformers ranging from 0.4B to 5B parameters that support native 1K and hierarchical 4K resolution. The models combine masked image reconstruction with self-distilled contrastive objectives to learn features capturing both low-level details for dense prediction and high-level semantics for zero-shot or few-label settings.

This contribution was assessed against **10 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

1. MMCLIP: Cross-modal Attention Masked Modelling for Medical Language-Image Pre-Training

URL: View paper

Brief Assessment

MMCLIP[58] focuses on medical vision-language pretraining with cross-modal attention masking for medical images and reports, not general human-centric vision with unified masked reconstruction and contrastive objectives for vision transformers.

2. A Theoretical Analysis of Self-Supervised Learning for Vision Transformers

URL: View paper

Brief Assessment

Self-Supervised Vision Theory[55] provides theoretical analysis of MAE and contrastive learning differences for vision transformers, but does not propose a unified pretraining objective or model family. The candidate focuses on theoretical understanding rather than practical model development.

3. Contrastive feature masking open-vocabulary vision transformer

URL: View paper

Brief Assessment

Contrastive Feature Masking[53] focuses on open-vocabulary object detection with masked feature reconstruction in joint image-text embedding space, while Sapiens2 targets human-centric dense prediction tasks combining masked image reconstruction with self-distilled contrastive objectives. The technical approaches and application domains differ substantially.

4. Bringing masked autoencoders explicit contrastive properties for point cloud self-supervised learning

URL: View paper

Brief Assessment

Masked Autoencoders Point[60] focuses on point cloud self-supervised learning with dual-masking strategies for 3D data, not human-centric vision with image transformers combining masked reconstruction and contrastive objectives for dense prediction tasks.

5. Masked contrastive reconstruction for cross-modal medical image-report retrieval

URL: View paper

Brief Assessment

Masked Contrastive Reconstruction[57] focuses on cross-modal medical image-report retrieval using masked data for both contrastive learning and reconstruction. The original paper addresses human-centric vision with vision transformers at high resolution, which is a fundamentally different domain and application.

6. Unveiling the power of audio-visual early fusion transformers with dense interactions through masked modeling

URL: View paper

Brief Assessment

Audio-Visual Masked Modeling[59] focuses on audio-visual early fusion transformers for multimodal learning, not vision-only transformers for human-centric tasks. The candidate combines masked reconstruction with contrastive learning for audio-visual data, while the original applies this to human-centric vision models.

7. Cross-Modal Contrastive Masked AutoEncoder for Compressed Video Pre-Training

URL: View paper

Brief Assessment

Cross-Modal Compressed Video[54] focuses on compressed video understanding using motion vectors and residuals from video codecs, not general-purpose vision transformers for human-centric tasks. The technical domains and applications are fundamentally different.

8. X-Former: Unifying Contrastive and Reconstruction Learning for MLLMs

URL: View paper

Brief Assessment

X-Former[52] focuses on combining CL and MIM encoders for multimodal language models (MLLMs) with vision-language understanding, while Sapiens2 targets human-centric vision tasks with native high-resolution processing and dense prediction capabilities.

9. Mimco: Masked image modeling pre-training with contrastive teacher

URL: View paper

Prior Art Analysis

Mimco[56] demonstrates that combining masked image reconstruction with contrastive objectives was already proposed and implemented prior to Sapiens2. The candidate paper explicitly describes a two-stage framework where masked image modeling is combined with contrastive learning from a pre-trained teacher model, using both patch-level and image-level reconstruction losses. This directly challenges the novelty claim of Sapiens2's unified pretraining objective that combines masked reconstruction with self-distilled contrastive objectives.

Evidence

Evidence 1 - **Rationale**: Both papers describe combining masked image modeling with contrastive learning objectives. Mimco[56] explicitly proposes this combination before Sapiens2, using a teacher-student framework with reconstruction and contrastive losses. - **Original**: to learn features that capture low-level details (for dense prediction) and high-level semantics (for zero-shot or few-label settings), we combine masked image reconstruction with self-distilled contrastive objectives. our evaluations show that this unified pretraining objective is better suited for... - **Candidate**: in this work, we propose a novel and flexible pre-training framework, named mimco, which combines mim and contrastive learning through two-stage pre-training. specifically, mimco takes a pre-trained contrastive learning model as the teacher model and is pre-trained with two types of learning targets...

Evidence 2 - **Rationale**: Both papers address the same fundamental problem of combining reconstruction with contrastive objectives to improve representation quality. Mimco[56] explicitly tackles this combination to improve linear separability and transfer performance. - **Original**: s apiens 2 addresses these limitations by coupling a reconstruction objective with contrastive objectives, anchoring features in pixel space (huang et al., 2023) while organizing them semantically. the result is a general-purpose representation that transfers across zero-shot, fewshot (song et al., ... - **Candidate**: this inspires us to think whether the linear separability of mim pretrained representation can be further improved, thereby improving the pre-training performance. since mim and contrastive learning tend to utilize different data augmentations and training strategies, combining these two pretext tas...

Evidence 3 - **Rationale**: Both papers describe using reconstruction losses combined with contrastive objectives. Mimco[56] proposes patch-level and image-level reconstruction losses that work with contrastive teacher features, demonstrating prior work on this unified objective approach. - **Original**: we use masked reconstruction with contrastive objectives to learn features that generalize in zero-shot settings on human tasks while preserving fine details in dense predictions. - **Candidate**: to take full advantage of the contrastive teacher model, we further propose two types of reconstruction losses. the first is the patch-level reconstruction loss. for the masked patches, we take the corresponding features of the contrastive teacher model as reconstruction targets. the second is the i...

10. Contrastive masked autoencoders are stronger vision learners

URL: View paper

Prior Art Analysis

Contrastive Masked Autoencoders[51] demonstrates that combining masked image reconstruction with contrastive objectives was already proposed and implemented prior to Sapiens2. The candidate paper explicitly describes a unified pretraining framework that combines masked image modeling (MIM) with contrastive learning (CL) to learn representations capturing both low-level details and high-level semantics. The paper states 'we propose contrastive masked autoencoders (cmae), a new self-supervised pre-training method for learning more comprehensive and capable vision representations. by elaboratively unifying contrastive learning (cl) and masked image model (mim) through novel designs, cmae leverages their respective advantages and learns representations with both strong instance discriminability and local perceptibility.' This directly refutes the novelty claim of Sapiens2's unified pretraining objective, as CMAE was published earlier and addresses the same technical challenge of combining reconstruction and contrastive objectives.

Evidence

Evidence 1 - **Rationale**: Both papers describe combining masked reconstruction with contrastive objectives to learn representations that capture both low-level details and high-level semantics. CMAE explicitly states it unifies CL and MIM, which is the same core contribution claimed by Sapiens2. - **Original**: to learn features that capture low-level details (for dense prediction) and high-level semantics (for zero-shot or few-label settings), we combine masked image reconstruction with self-distilled contrastive objectives. our evaluations show that this unified pretraining objective is better suited for... - **Candidate**: we propose contrastive masked autoencoders (cmae), a new self-supervised pre-training method for learning more comprehensive and capable vision representations. by elaboratively unifying contrastive learning (cl) and masked image model (mim) through novel designs, cmae leverages their respective adv...

Evidence 2 - **Rationale**: Both papers identify the same limitation of MIM methods (lack of discriminative features) and propose the same solution (adding contrastive learning). CMAE explicitly asks whether MIM can benefit from contrastive learning, which is the exact motivation behind Sapiens2's unified objective. - **Original**: s apiens 2 addresses these limitations by coupling a reconstruction objective with contrastive objectives, anchoring features in pixel space (huang et al., 2023) while organizing them semantically. the result is a general-purpose representation that transfers across zero-shot, fewshot (song et al., ... - **Candidate**: in contrast, mim focuses more on learning local relations in input image for fulfilling the reconstruction task, instead of modeling the relation among different images [35]. therefore, it is suspected that mim is less efficient in learning discriminative representations. this issue has been manifes...

Evidence 3 - **Rationale**: Both papers describe the same technical approach: using reconstruction to preserve fine details while using contrastive learning to enhance discriminability. The architectural implementation differs slightly, but the core unified objective is identical. - **Original**: we use masked reconstruction with contrastive objectives to learn features that generalize in zero-shot settings on human tasks while preserving fine details in dense predictions. - **Candidate**: the online encoder reconstructs original images from latent representations of masked images to learn holistic features. the momentum encoder, fed with the full images, enhances the feature discriminability via contrastive learning with its online counterpart.

Evidence 4 - **Rationale**: Both papers describe nearly identical technical implementations of the unified objective, including tokenization, encoder-decoder architecture, and dual objectives for reconstruction and contrastive learning. - **Original**: let i denote the training set. we sample an image x ∼ iand draw v random augmentations to obtain views {xi}v i=1. each view is patchified into n tokens indexed by p = {1, . . . , n}, i.e., xi = {xp i }p∈p. let {ep pos}p∈p be positional embeddings (dosovitskiy et al., 2020) and φenc, φdec, φcls be ou... - **Candidate**: let us denote the input image is to the online encoder as being tokenized into a sequence of n image patch tokens {xs i }n i=1, where n is the total number of image patches. for the masked version of is, the set of visible tokens is represented as {xv}. similarly, the input image it is tokenized int...

Contribution 2: Humans-750M pretraining dataset

Description: The authors curate and introduce a large-scale dataset of 750 million high-quality human images from a web-scale corpus through multi-stage filtering. The dataset spans diverse ages, ethnicities, backgrounds, and real-world conditions with no task-specific labels or human-specific priors injected during pretraining.

This contribution was assessed against **10 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

1. Large-scale reinforcement learning for diffusion models

URL: View paper

Brief Assessment

Reinforcement Learning Diffusion[73] focuses on diffusion models for sequential data (video, fluid dynamics) and does not discuss human image datasets or human-centric vision pretraining.

2. Wukong: A 100 million large-scale chinese cross-modal pre-training benchmark

URL: View paper

Brief Assessment

Wukong[72] focuses on Chinese cross-modal (image-text) pre-training with 100M pairs, not human-centric vision. The dataset construction targets general web images with Chinese text, not specifically human images or human-centric tasks.

3. Personvit: large-scale self-supervised vision transformer for person re-identification

URL: View paper

Brief Assessment

Personvit[79] uses the LUPerson dataset (4.18 million images) for pretraining, which is a different dataset than the Humans-750M corpus. The candidate focuses on person re-identification rather than general human-centric vision tasks.

4. Scaling up vision-language pre-training for image captioning

URL: View paper

Brief Assessment

Vision-Language Captioning Scaling[74] focuses on image-text pairs for vision-language captioning tasks, not human-centric vision datasets. Their dataset contains alt-text descriptions for general images, while the original paper curates human-specific images for dense prediction tasks.

5. Multimodal c4: An open, billion-scale corpus of images interleaved with text

URL: View paper

Brief Assessment

Multimodal c4[76] focuses on creating an interleaved image-text corpus from web documents for multimodal language models, not specifically human-centric vision datasets. The datasets serve fundamentally different purposes and domains.

6. Laion-5b: An open large-scale dataset for training next generation image-text models

URL: View paper

Brief Assessment

Laion-5b[71] focuses on general image-text pairs from web crawling (5.85 billion pairs across all domains), not specifically human-centric images. The original paper curates 750M high-quality human images with human-specific filtering criteria, which is a distinct contribution.

7. The Neglected Tails in Vision-Language Models

URL: View paper

Brief Assessment

Neglected Tails Vision[78] focuses on analyzing long-tailed concept distributions in vision-language models' pretraining data (e.g., LAION) and does not present a new human-centric image dataset for pretraining.

8. Disco: Disentangled control for realistic human dance generation

URL: View paper

Brief Assessment

Disco[75] focuses on human dance generation using diffusion models, not on pretraining datasets for vision transformers. The candidate paper does not discuss large-scale human image dataset curation or pretraining methodologies.

9. Image representations learned with unsupervised pre-training contain human-like biases

URL: View paper

Brief Assessment

Unsupervised Human Biases[77] focuses on analyzing social biases in ImageNet-trained models, not on curating large-scale human image datasets for pretraining vision models.

10. Sapiens: Foundation for Human Vision Models

URL: View paper

Brief Assessment

Sapiens Foundation[3] uses 300 million human images for pretraining, which is a smaller-scale dataset than the 750M claimed in the original paper. The candidate does not demonstrate that a 750M-scale human image dataset existed prior to the original work.

Contribution 3: Hierarchical 4K architecture with windowed attention

Description: The authors introduce a hierarchical architecture design for 4K resolution processing that uses windowed self-attention in early layers to capture local structure, followed by spatial downsampling and global attention layers. This design enables high-resolution dense prediction while maintaining computational tractability and compatibility with masked pretraining.

This contribution was assessed against **10 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

1. AutoHFormer: Efficient Hierarchical Autoregressive Transformer for Time Series Prediction

URL: View paper

Brief Assessment

AutoHFormer[62] focuses on time series forecasting with hierarchical temporal modeling for sequential data, not high-resolution image processing. The windowed attention serves different purposes: temporal causality in time series versus spatial locality in vision tasks.

2. Pedestrian Trajectory Prediction via Window Attention and Spatial Graph Interaction Network

URL: View paper

Brief Assessment

Pedestrian Trajectory Prediction[70] uses windowed attention for temporal modeling in trajectory prediction tasks, not for high-resolution dense prediction in vision transformers. The architectural goals and application domains are fundamentally different.

3. A Pyramid Fusion MLP for Dense Prediction

URL: View paper

Brief Assessment

Pyramid Fusion MLP[65] uses multi-scale pooling and fully connected layers for feature pyramids, not windowed self-attention. It focuses on MLP architectures for dense prediction rather than transformer-based hierarchical designs with windowed attention for 4K resolution processing.

4. Hierarchical multi-scale attention for semantic segmentation

URL: View paper

Brief Assessment

Hierarchical Multiscale Attention[61] focuses on combining multi-scale predictions for semantic segmentation using attention mechanisms between different image scales (0.5x, 1.0x, 2.0x), not on hierarchical architectures with windowed attention for processing high-resolution inputs like 4K images. The candidate's hierarchical structure refers to learning relative attention between adjacent inference scales, while the original paper's contribution concerns a spatial hierarchical architecture with windowed self-attention layers followed by spatial downsampling for 4K resolution processing.

5. Hrformer: High-resolution vision transformer for dense predict

URL: View paper

Prior Art Analysis

Hrformer Dense Predict[63] demonstrates that hierarchical architectures with windowed attention for high-resolution processing were established prior to the original paper. The candidate explicitly describes a multi-resolution parallel design where early layers use local-window self-attention over non-overlapping windows to capture local structure, followed by spatial downsampling and global attention layers. This design enables high-resolution dense prediction while maintaining computational efficiency, which directly parallels the original paper's claimed contribution. The candidate was published at NeurIPS 2021, predating the original submission.

Evidence

Evidence 1 - **Rationale**: Both papers emphasize that windowed attention enables computational tractability and is compatible with masked pretraining approaches. The candidate demonstrates this design principle was established in 2021. - **Original**: this layout is naturally compatible with mae-style pretraining: after the local stage, masked tokens can be dropped so that information does not flow across masked regions, avoiding the leakage that convolutional backbones typically require masked convolutions to prevent - **Candidate**: at each resolution, the local-window self-attention mechanism is adopted to reduce the memory and computation complexity. we partition the representation maps into a set of non-overlapping small image windows and perform self-attention in each image window separately. this reduces the memory and com...

Evidence 2 - **Rationale**: The candidate paper explicitly describes using windowed attention for high-resolution dense prediction tasks, establishing this approach before the original paper's submission. - **Original**: our 4k models adopt windowed attention to reason over longer spatial context and are pretrained with 2k output resolution. - **Candidate**: we present a high-resolution transformer (hrformer) that learns high-resolution representations for dense prediction tasks, in contrast to the original vision transformer that produces low-resolution representations and has high memory and computational cost. we take advantage of the multi-resolutio...

6. Attention guided multi-level feature aggregation network for camouflaged object detection

URL: View paper

Brief Assessment

Camouflaged Object Detection[68] focuses on detecting camouflaged objects using spatial pyramid attention mechanisms, not on hierarchical architectures with windowed attention for high-resolution dense prediction tasks.

7. Attention receptive pyramid network for ship detection in SAR images

URL: View paper

Brief Assessment

Ship Detection Pyramid[69] focuses on SAR ship detection using receptive field blocks and attention modules for feature pyramids, not on hierarchical architectures with windowed self-attention for high-resolution dense prediction tasks.

8. Swin Transformer: Hierarchical Vision Transformer using Shifted Windows

URL: View paper

Brief Assessment

[Final Audit Failure] The model insisted on a refutation claim but failed to provide verifiable evidence after multiple retries. Marked as cannot_refute for safety. Please manually verify the candidate text.

9. HRFormer: High-Resolution Transformer for Dense Prediction

URL: View paper

Prior Art Analysis

HRFormer[64] demonstrates prior work on hierarchical architectures with windowed attention for high-resolution dense prediction. Both papers employ windowed self-attention in early layers to capture local structure, followed by spatial processing for global context. The candidate paper explicitly describes using 'local-window self-attention that performs self-attention over small non-overlapping image windows' for dense prediction tasks, which directly parallels the original paper's approach of 'windowed self-attention layers operates locally to capture texture and fine boundaries' followed by global attention. This establishes that the hierarchical windowed attention design for high-resolution processing was not first introduced by the original paper.

Evidence

Evidence 1 - **Rationale**: Both papers describe using windowed self-attention in early layers for local processing in high-resolution architectures. The candidate explicitly uses 'local-window self-attention' over 'small non-overlapping image windows' for dense prediction, which is the same core design principle as the original's 'windowed self-attention layers operates locally' approach. - **Original**: to make 4k tractable, we adopt a hierarchical design (li et al., 2022): an initial stack of windowed self-attention layers operates locally to capture texture and fine boundaries, from each window we pool a summary token and then apply global self-attention-mirroring our 1k models-to fuse long-range... - **Candidate**: we take advantage of the multi-resolution parallel design introduced in high-resolution convolutional networks (hrnet), along with local-window self-attention that performs self-attention over small non-overlapping image windows, for improving the memory and computation efficiency.

Evidence 2 - **Rationale**: Both papers address the same problem of making high-resolution processing tractable for dense prediction tasks through hierarchical transformer designs. The candidate's HRFormer is explicitly designed for 'dense prediction tasks' with 'high-resolution representations', matching the original's goal of '4k backbone' for 'dense prediction'. - **Original**: we introduce a 4k backbone pretrained and post-trained for dense prediction, with task heads that decode to 2k resolution across tasks. to make 4k tractable, we adopt a hierarchical design (li et al., 2022): an initial stack of windowed self-attention layers operates locally to capture texture and f... - **Candidate**: we present a high-resolution transformer (hrformer) that learns high-resolution representations for dense prediction tasks, in contrast to the original vision transformer that produces low-resolution representations and has high memory and computational cost.

10. Fastervit: Fast vision transformers with hierarchical attention

URL: View paper

Prior Art Analysis

Fastervit[67] demonstrates that hierarchical architectures with windowed attention for high-resolution processing were already established prior to the ORIGINAL paper. Fastervit[67] explicitly describes a hierarchical design where early layers use windowed self-attention to capture local structure, followed by spatial downsampling and global attention layers - the exact same architectural pattern claimed as novel in the ORIGINAL paper. Both papers use windowed attention in early stages for local feature extraction, then transition to global attention after downsampling, and both are compatible with masked pretraining objectives. The architectural similarities are substantial and directly overlap with the claimed contribution.

Evidence

Evidence 1 - **Rationale**: Both papers describe hierarchical architectures where early stages use local/windowed attention for high-resolution processing, followed by global attention in later stages. This demonstrates prior work on the same architectural pattern. - **Original**: to make 4k tractable, we adopt a hierarchical design (li et al., 2022): an initial stack of windowed self-attention layers operates locally to capture texture and fine boundaries, from each window we pool a summary token and then apply global self-attention-mirroring our 1k models-to fuse long-range... - **Candidate**: we propose to leverage residual convolutional blocks in the high-resolution stages of the architecture (i.e., stage 1, 2), while employing transformer-blocks in later stages (i.e., stage 3, 4). this strategy allows for fast generation of high-level tokens which can be further processed with the tran...

Evidence 2 - **Rationale**: Both papers describe architectures compatible with masked pretraining, though Fastervit[67] uses a patch-based approach while the ORIGINAL paper explicitly mentions MAE-style pretraining compatibility. - **Original**: this layout is naturally compatible with mae-style pretraining: after the local stage, masked tokens can be dropped so that information does not flow across masked regions, avoiding the leakage that convolutional backbones typically require masked convolutions to prevent (gao et al., 2022). - **Candidate**: given an input image x prhˆwˆ3 is converted into overlapping patches by two consecutive3ˆ3 convolutional layers, each with a stride of 2, which project them into a d-dimensional embedding. the embedded tokens are further batch-normalized (ioffe & szegedy, 2015) and use the relu activation function a...

Evidence 3 - **Rationale**: Both papers use summary tokens from local windows followed by global attention mechanisms. Fastervit[67]'s carrier tokens serve the same purpose as the ORIGINAL paper's pooled summary tokens, demonstrating prior work on this specific design pattern. - **Original**: we adopt a hierarchical design (li et al., 2022): an initial stack of windowed self-attention layers operates locally to capture texture and fine boundaries, from each window we pool a summary token and then apply global self-attention - **Candidate**: specifically, our proposed hierarchical attention (see fig. 2) learns carrier tokens as a summary of each local window and efficiently models the cross-interaction between these regions. the computational complexity of the hierarchical attention grows almost linearly with input image resolution, as ...

Evidence 4 - **Rationale**: Both papers demonstrate their hierarchical architectures on dense prediction tasks, showing that the architectural approach for high-resolution dense prediction was already established in Fastervit[67]. - **Original**: beyond standard 1k backbones (khirodkar et al., 2024), we introduce a 4k backbone pretrained and post-trained for dense prediction, with task heads that decode to 2k resolution across tasks. - **Candidate**: we have extensively validated the effectiveness of the proposed fastervit model on various image tasks and datasets such as imagenet-1k for image classification, ms coco for object detection and instance segmentation and ade20k dataset for semantic segmentation. fastervit achieves state-of-the-art p...

Appendix: Text Similarity Detection

Textual similarity detection checked 31 papers and found 1 similarity segment(s) across 1 paper(s).

The following **1 paper(s)** were detected to have high textual similarity with the original paper. These may represent different versions of the same work, duplicate submissions, or papers with substantial textual overlap. Readers are advised to verify these relationships independently.

1. Personvit: large-scale self-supervised vision transformer for person re-identification

Detected in: Contribution: contribution_2

⚠ **Note**: This paper shows substantial textual similarity with the original paper. It may be a different version, a duplicate submission, or contain significant overlapping content. Please review carefully to determine the nature of the relationship.

References

- [0] Sapiens2 View paper
- [1] Integrating spatiotemporal vision transformer into digital twins for high-resolution heat stress forecasting in campus environments View paper
- [2] Not All Tokens Are Equal: Human-centric Visual Analysis via Token Clustering Transformer View paper
- [3] Sapiens: Foundation for Human Vision Models View paper
- [4] Sharpose: Sparse high-resolution representation for human pose estimation View paper
- [5] Hybrid Vision Transformer and Convolutional Neural Network for Super-Resolution Image Quality Assessment View paper
- [6] ViTPose: Simple Vision Transformer Baselines for Human Pose Estimation View paper
- [7] Unihcp: A unified model for human-centric perceptions View paper
- [8] Design of an Integrated Model Combining CycleGAN, PPO, and Vision Transformer for Adaptive Scene Rendering in the Metaverse View paper
- [9] Interpretable Multimodal Framework for Human-Centered Street Assessment: Integrating Visual-Language Models for Perceptual Urban Diagnostics View paper
- [10] UnitedHuman: Harnessing Multi-Source Data for High-Resolution Human Generation View paper
- [11] Ultra-High Resolution Segmentation via Boundary-Enhanced Patch-Merging Transformer View paper
- [12] Hr-nas: Searching efficient high-resolution neural architectures with lightweight transformers View paper
- [13] LMLT: Low-to-high Multi-Level Vision Transformer for Lightweight Image Super-Resolution View paper
- [14] BeyondScene: Higher-Resolution Human-Centric Scene Generation With Pretrained Diffusion View paper
- [15] Searching efficient neural architecture with multi-resolution fusion transformer for appearance-based gaze estimation View paper
- [16] OpenHumanVid: A Large-Scale High-Quality Dataset for Enhancing Human-Centric Video Generation View paper
- [17] Efficient High-Resolution Visual Representation Learning with State Space Model for Human Pose Estimation View paper
- [18] Disentangling the Factors of Convergence between Brains and Computer Vision Models View paper
- [19] A large-scale human-centric benchmark for referring expression comprehension in the LMM era View paper
- [20] MTPose: Human pose estimation with high-resolution multi-scale transformers View paper
- [21] Rethinking the Sparse End-to-End Multiperson Pose Estimation View paper
- [22] HRPVT: High-Resolution Pyramid Vision Transformer for medium and small-scale human pose estimation View paper
- [23] Potter: Pooling attention transformer for efficient human mesh recovery View paper
- [24] Capturing and Inferring Dense Full-Body Human-Scene Contact View paper
- [25] Adaptive Vision Transformer for Event-Based Human Pose Estimation View paper
- [26] Lmlt: Low-to-high multi-level vision transformer for image super-resolution View paper
- [27] Swin-UNIT: Transformer-based GAN for High-resolution Unpaired Image Translation View paper
- [28] Adaptive Patching for High-resolution Image Segmentation with Transformers View paper
- [29] Swallowing Assessment using High-Resolution Cervical Auscultations and Transformer-based Neural Networks. View paper
- [30] Quality transformer for human parsing View paper
- [31] AggPose: Deep Aggregation Vision Transformer for Infant Pose Estimation View paper
- [32] Recurrent Multi-scale Transformer for High-Resolution Salient Object Detection View paper
- [33] HRPoseFormer: High-Resolution Transformer for Human Pose Estimation via Multi-Scale Token Aggregation View paper
- [34] Human-Centered Visual Segmentation View paper
- [35] SVTSR: image super-resolution using scattering vision transformer View paper
- [36] StyleSwin: Transformer-based GAN for High-resolution Image Generation View paper
- [37] LM-MCVT: A Lightweight Multi-modal Multi-view Convolutional-Vision Transformer Approach for 3D Object Recognition View paper

- [38] Aggregated Contextual Transformations for High-Resolution Image Inpainting. View paper
- [39] Pan-Arctic Permafrost Landform and Human-built Infrastructure Feature Detection with Vision Transformers and Location Embeddings View paper
- [40] Glance-and-Gaze Vision Transformer View paper
- [41] Restormer: Efficient Transformer for High-Resolution Image Restoration View paper
- [42] Toward high resolution scoring system for HER2 expression in breast cancer based on pathological images via deep learning View paper
- [43] Abstract LB087: Inferring single-cell spatial gene expression with tissue morphology via explainable deep learning View paper
- [44] Towards Efficient Deep Learning for Human-Centric Visual Understanding and Generation View paper
- [45] Scalable High-Resolution Pixel-Space Image Synthesis with Hourglass Diffusion Transformers View paper
- [46] High-Resolution Representation Learning for Human Pose Estimation based on Transformer View paper
- [47] Improved Detection of Urolithiasis Using High-Resolution Computed Tomography Images by a Vision Transformer Model View paper
- [48] UViT: Efficient and lightweight U-shaped hybrid vision transformer for human pose estimation View paper
- [49] ViTAR: Vision Transformer with Any Resolution View paper
- [50] HEViTPose: High-Efficiency Vision Transformer for Human Pose Estimation View paper
- [51] Contrastive masked autoencoders are stronger vision learners View paper
- [52] X-Former: Unifying Contrastive and Reconstruction Learning for MLLMs View paper
- [53] Contrastive feature masking open-vocabulary vision transformer View paper
- [54] Cross-Modal Contrastive Masked AutoEncoder for Compressed Video Pre-Training View paper
- [55] A Theoretical Analysis of Self-Supervised Learning for Vision Transformers View paper
- [56] Mimco: Masked image modeling pre-training with contrastive teacher View paper
- [57] Masked contrastive reconstruction for cross-modal medical image-report retrieval View paper
- [58] MMCLIP: Cross-modal Attention Masked Modelling for Medical Language-Image Pre-Training View paper
- [59] Unveiling the power of audio-visual early fusion transformers with dense interactions through masked modeling View paper
- [60] Bringing masked autoencoders explicit contrastive properties for point cloud self-supervised learning View paper
- [61] Hierarchical multi-scale attention for semantic segmentation View paper
- [62] AutoHFormer: Efficient Hierarchical Autoregressive Transformer for Time Series Prediction View paper
- [63] Hrformer: High-resolution vision transformer for dense predict View paper
- [64] HRFormer: High-Resolution Transformer for Dense Prediction View paper
- [65] A Pyramid Fusion MLP for Dense Prediction View paper
- [66] Swin Transformer: Hierarchical Vision Transformer using Shifted Windows View paper
- [67] Fastervit: Fast vision transformers with hierarchical attention View paper
- [68] Attention guided multi-level feature aggregation network for camouflaged object detection View paper
- [69] Attention receptive pyramid network for ship detection in SAR images View paper
- [70] Pedestrian Trajectory Prediction via Window Attention and Spatial Graph Interaction Network View paper
- [71] Laion-5b: An open large-scale dataset for training next generation image-text models View paper
- [72] Wukong: A 100 million large-scale chinese cross-modal pre-training benchmark View paper
- [73] Large-scale reinforcement learning for diffusion models View paper
- [74] Scaling up vision-language pre-training for image captioning View paper
- [75] Disco: Disentangled control for realistic human dance generation View paper
- [76] Multimodal c4: An open, billion-scale corpus of images interleaved with text View paper
- [77] Image representations learned with unsupervised pre-training contain human-like biases View paper
- [78] The Neglected Tails in Vision-Language Models View paper
- [79] Personvit: large-scale self-supervised vision transformer for person re-identification View paper