arXiv:2312.05525v1 [cs.CV] 9 Dec 2023

You Only Learn One Query: Learning Unified Human Query for Single-Stage Multi-Person Multi-Task Human-Centric Perception

Sheng Jin[1,2*] Shuhuai Li[2*] Tong Li[2] Wentao Liu[2,3] Chen Qian[2] Ping Luo[1,3]
[1] The University of Hong Kong [2] SenseTime Research and Tetras.AI [3] Shanghai AI Laboratory
{jinsheng, liuwentao, qianchen}@tetras.ai {lishuhuai, litong}@sensetime.com pluo@cs.hku.hk

Abstract

*Human-centric perception (*e.g. *pedetrian detection, segmentation, pose estimation, and attribute analysis) is a long-standing problem for computer vision. This paper introduces a unified and versatile framework (HQNet) for single-stage multi-person multi-task human-centric perception (HCP). Our approach centers on learning a unified human query representation, denoted as Human Query, which captures intricate instance-level features for individual persons and disentangles complex multi-person scenarios. Although different HCP tasks have been well-studied individually, single-stage multi-task learning of HCP tasks has not been fully exploited in the literature due to the absence of a comprehensive benchmark dataset. To address this gap, we propose COCO-UniHuman benchmark dataset to enable model development and comprehensive evaluation. Experimental results demonstrate the proposed method's state-of-the-art performance among multi-task HCP models and its competitive performance compared to task-specific HCP models. Moreover, our experiments underscore Human Query's adaptability to new HCP tasks, thus demonstrating its robust generalization capability. Codes and data will be publicly accessible.*

1. Introduction

Human-centric visual perception (*e.g.* pedestrian detection, pose estimation[1], human segmentation and human attribute recognition have attracted increasing research attention owing to their widespread industrial applications such as sports analysis, virtual reality, and augmented reality.

The task of single-stage multi-person multi-task human-centric perception (HCP) has not been fully exploited in the literature due to the absence of a representative benchmark dataset. Consequently, previous studies [12] resorted to training models on various datasets for each HCP task, which can introduce certain limitations. Firstly, there is inherent scale variance across different datasets. For example, human detection datasets [46] consist of scene images with multiple interacting people, while attribute recognition datasets [52] typically contain images with a single cropped person. This hampers the development of single-stage multi-task algorithms that can comprehensively address various HCP tasks as a unified problem. Secondly, single-task datasets are often designed for specific application scenarios, resulting in strong dataset biases across different datasets. For example, some datasets [27] are captured in controlled lab environments, while some [52] are captured from a surveillance viewpoint. Naively training models on a combination of these datasets inevitably introduces dataset biases and hinders performance in real-world, unconstrained scenarios. Although there are separate benchmarks for individual HCP tasks, a comprehensive benchmark to simultaneously evaluate multiple HCP tasks is still lacking. To address this problem, we introduce a large-scale benchmark dataset called COCO-UniHuman, specifically designed for unified human-centric perceptions.

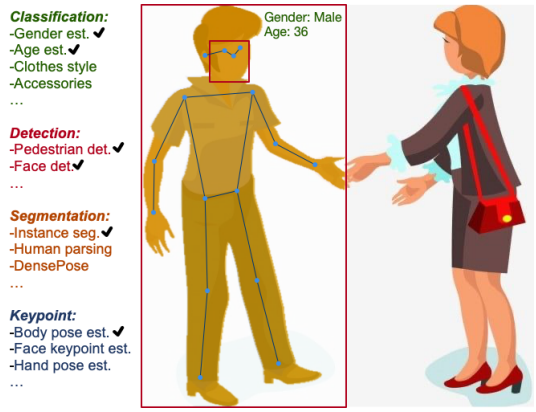

Figure 1. Multi-person human-centric perception tasks can be categorized into 4 groups: classification, detection, segmentation and keypoint localization.

*Equal contribution.

[1]In this paper, we use "keypoint localization" and "pose estimation" interchangeably.

As shown in Figure 1, most popular HCP tasks[2] can be grouped into four fundamental categories: classification, detection, segmentation, and keypoint localization. The COCO-UniHuman dataset extends COCO dataset by extensively annotating gender and age labels for each person instance. It encompasses all these four categories, covering 6 diverse HCP tasks (marked with check marks in Figure 1).

Prior works on multi-person multi-task HCP have predominantly employed a multi-stage approach. These approaches typically involve employing a human detector to detect human instances, followed by task-specific models for each individual human perception task such as pose estimation and instance segmentation. However, these approaches exhibit three significant drawbacks. Firstly, they suffer from the issue of early commitment: the performance of the whole pipeline highly relies on body detection, and there is no recourse to recovery if the body detector fails. Secondly, the run-time is proportional to the number of individuals present in the image, making them computationally expensive for real-time applications. In contrast, single-stage methods estimate all required properties for human-centric analysis in a single pass, resulting in improved efficiency. Thirdly, these approaches overlook the potential inter-task synergy. Different HCP tasks are highly correlated as they share a common understanding of human body structure. In this work, we develop a simple, straightforward and versatile baseline framework, called HQNet, for single-stage multi-task HCP. It unifies various distinct human-centric tasks, including pedetrian detection, human segmentation, human pose estimation, and human attribute analysis (specifically gender and age).

Different HCP tasks have their own relevant features of diverse granularity to focus on. For instance, pedestrian detection emphasizes global semantic features; attribute recognition necessitates both global and local semantic cues; person segmentation relies on fine-grained semantic features; and pose estimation require fine-grained semantic and localization information. In this paper, we propose to learn unified all-in-one query representations, termed Human Query, to encode instance-specific features of diverse granularity from multiple perspectives. Our work is inspired by DETR-based methods [8, 40, 51, 96, 106], which employ learnable query embeddings to represent objects and infer the relations of the objects and the image features. This study expands upon these works by learning versatile instance-level query representations for general human-centric perceptions. Despite the absence of intricate model designs, HQNet achieves state-of-the-art results across various benchmarks. Furthermore, we highlight several noteworthy characteristics of HQNet, including its flexibility, scalability, and transferability. (1) **Flexibility:** HQNet can readily integrate with diverse backbone networks, such as ResNet [23], Swin [54] and ViT [17]. (2) **Scalibility:** the weight-sharing backbone, transformer encoder, and decoder in HQNet enables seamless integration with multiple tasks, with minimal overhead from each task-specific head, thus demonstrating remarkable scalability. (3) **Transferability:** Experiments demonstrate strong transferability of the learned Human Query to novel HCP tasks, such as face detection and multi-object tracking. It should be noted that this work does not claim the algorithmic superiority but rather establishes a solid baseline with superior performance in human-centric perception.

Our work makes the following key contributions: (1) We introduce the COCO-UniHuman benchmark, a large-scale dataset that comprehensively covers all representative HCP tasks, *i.e*. classification (gender and age estimation), detection (body and face detection), segmentation, and keypoint localization. (2) We develop a simple yet effective baseline called HQNet, unifying multiple distinctive HCP tasks in a single-stage multi-task manner. The key idea is to learn unified all-in-one query representations, termed Human Query, which encode instance-specific features of diverse granularity from various perspectives. Our approach achieves state-of-the-art results on different HCP tasks, demonstrating the strong representation capability of the learnt Human Query. Furthermore, experiments show the strong transferability of the learned Human Query to novel HCP tasks, such as face detection and multi-object tracking. We hope our work can shed light on future research on developing single-stage multi-person multi-task HCP algorithms.

2. Related Works

2.1. Human-centric perception tasks and datasets

Human-centric perception (HCP) tasks including pose estimation [7, 29, 78, 81, 87, 91], segmentation [22, 24, 69], and attribute recognition [28, 38, 68], have been extensively studied in computer vision. Approaches to multi-person HCP can be categorized into top-down, bottom-up, and single-stage methods. **Top-down methods** follow a detect-then-analyze approach. They first localize human instances, and then perform single person analysis. Top-down approaches can be divided into two types: those using separate pre-trained detectors and task-specific perception models [69], and those jointly learning detection and perception modules [24]. **Bottom-up methods** learn instance-agnostic keypoints/masks and cluster them using integer linear programming [26, 30, 36], heuristic greedy parsing [6, 65], embedding clustering [31, 37, 60], or learnable clustering [32]. **Single-stage methods** directly predict keypoints or masks for each individual, with different representations for pose estimation (coordinate-based [63, 76, 83, 93], heatmap-based [71, 79], or hybrid [20, 57, 105]) and segmentation (contour-based [88] or mask-based [5]). While most exist-

[2]In this work, we focus on 2D human-centric perception tasks.

Table 1. Overview of representative HCP datasets. "# Img", "# Inst", and "# ID" mean the number of total images, instances and identities respectively. "Crop" indicates whether the images are cropped for "face" or "body". "*" means head box annotation. "group:n" means age classification with n groups, "real" means real age estimation, and "appa" means apparent age estimation.

Dataset	# Img	# Inst	# ID	Crop	Body Box	Face Box	Body Kpt	Body Mask	Gender	Age
Caltech [16]	250K	350K	2.3K	✗	✓	✗	✗	✗	✗	✗
CityPersons [99]	5K	32K	32K	✗	✓	✗	✗	✗	✗	✗
CrowdHuman [70]	24K	552K	552K	✗	✓	*	✗	✗	✗	✗
MPII [3]	25K	40K	-	✗	✓	*	✓	✗	✗	✗
PoseTrack [4]	23K	153K	-	✗	✓	*	✓	✗	✗	✗
CIHP [22]	38K	129K	129K	✗	✓	✗	✗	✓	✗	✗
MHP [42]	5K	15K	15K	✗	✓	✗	✗	✓	✗	✗
CelebA [53]	200K	200K	10K	face	✗	✗	✗	✗	✓	group:4
APPA-REAL [2]	7.5K	7.5K	7.5K	face	✗	✗	✗	✗	✓	appa & real
MegaAge [102]	40K	40K	40K	face	✗	✗	✗	✗	✓	real
WIDER-Attr [44]	13K	57K	57K	✗	✓	✗	✗	✗	✓	group:6
PETA [14]	19K	19K	8.7K	body	✗	✗	✗	✗	✓	group:4
PA-100K [52]	100K	100K	-	body	✗	✗	✗	✗	✓	group:3
OCHuman [100]	5K	13K	13K	✗	✓	✗	✓	✓	✗	✗
COCO [46]	200K	273K	273K	✗	✓	✗	✓	✓	✗	✗
COCO-WholeBody [33, 90]	200K	273K	273K	✗	✓	✓	✓	✗	✗	✗
COCO-UniHuman	200K	273K	273K	✗	✓	✓	✓	✓	✓	appa

ing approaches focus on individual HCP tasks, we aim to unify human-centric perception by learning a single model that handles multiple tasks simultaneously, enabling a comprehensive understanding of humans.

As shown in Table 1, there are many task-specific datasets separately annotated for different HCP tasks, including pedestrian detection [16, 70, 99], pose estimation [3, 4, 18, 85], human segmentation [22, 42], and human attribute recognition [52, 102]. Datasets for multiple HCP tasks also exist. Widely used COCO [33, 46, 90] offers thorough annotations: body box, keypoints, and segmentation mask. Our COCO-UniHuman dataset further extends COCO featuring extensive gender and age annotations.

2.2. Unified methods for HCP

General network architecture for different HCP tasks. Some works design general network backbones, including CNN-based [80] and Transformer-based backbones [95]. Others unify HCP tasks with novel perception heads, such as UniHead [45]. Unlike these methods, which employ separate task-specific models, we consolidate diverse HCP tasks within a single network. **Pre-training on HCP tasks.** There are works [10, 25, 74] on pre-training on diverse human-centric tasks with large-scale data. UniHCP [12] presents a unified vision transformer model to perform multitask pre-training at scale. It employs task-specific queries for attending to relevant features, but tackles one task at a time. Unlike ours, our approach simultaneously solves multiple HCP tasks in a single forward pass. Our approach contrasts with these pre-training based methods by avoiding pre-training, minimizing fine-tuning, and circumventing resource-intensive multi-dataset training. Unlike them, we handle multiple HCP tasks concurrently in a single-stage, multi-task manner, diverging from their single-person focus. **Co-learning on HCP tasks.** Many works have investigated the correlations between pairs of HCP tasks [50, 61, 62, 75, 97]. We propose a single-stage model that learns a general unified representation to handle all representative HCP tasks simultaneously.

2.3. Object-centric representation learning

DETR [8] pioneers learnable object queries to represent objects and interact with image features. Deformable DETR [106] introduces deformable attention modules to focus on key sampling points, enhancing convergence speed. DAB-DETR [51] treats each positional query as a dynamic 4D anchor box, updated across decoder layers. DN-DETR [40] employs denoising training for faster convergence. Recently, DINO [96] amalgamates these techniques, introducing a mixed query selection and look-forward-twice strategy to expedite and stabilize training. Our work is inspired by DETR-based methods. Especially we build upon DINO and extend it to develop a versatile framework for single-stage multi-task HCP, unifying multiple distinct human-centric tasks.

3. COCO-UniHuman Dataset

COCO-UniHuman is the first large-scale dataset, which provides annotations for all four representative HCP tasks in multi-person scenarios. Building upon COCO [33, 46], we have enriched the annotations by including gender and age information for each individual.

Uniqueness. The newly introduced dataset possesses several noteworthy properties in comparison to existing HCP datasets. **(1) Comprehensiveness:** This is the first large-scale multi-person HCP dataset that encompasses all four basic HCP tasks, *i.e.* classification, detection, segmentation, and keypoint localization in multi-person scenar-

ios. It facilitates the development and evaluation of single-stage multi-person multi-task HCP algorithms. **(2) Large scale and high diversity:** With over 200,000 images and 273,000 identities, this dataset exhibits significant variations in terms of lighting conditions, image resolutions, human poses, and indoor/outdoor environments. **(3) Multi-person attribute recognition:** Unlike most existing human attribute recognition datasets that solely provide single-person center cropped images, our proposed dataset offers a valuable benchmark for multi-person attribute recognition in challenging scenarios. **(4) Body-based apparent age estimation:** While previous research has primarily focused on predicting a person's age based on facial images, our dataset emphasizes the utilization of richer visual cues derived from whole-body images. Incorporating body-based visual cues such as skin elasticity, body posture, and body height proves beneficial for estimating a person's age, particularly in situations where the facial image lacks clarity (*e.g*. captured from a distance). Notably, existing large-scale pedestrian attribute datasets [14] typically only offer coarse age group annotations, while facial attribute datasets [2] often provide fine-grained apparent or real age annotations. Our proposed dataset bridges this gap and serves as the pioneering large-scale dataset for body-based apparent age estimation in the wild. **(5) Enhanced human representation:** The extended attribute labels provide additional descriptive information about individuals beyond the existing labels. By leveraging these attribute labels, models can learn improved representations of humans, consequently enhancing the performance of other HCP tasks. Furthermore, the inclusion of gender and age labels can be valuable in downstream applications (*e.g*. gender/age-specific SMPL[55] model selection).

3.1. Data Annotation

To ensure accurate annotations, we employ trained annotators to manually label the gender and apparent age for each human instance in the dataset. We discard images full of non-human objects, and exclude all *Small* category persons that are hardly attribute-recognizable. **Gender annotation.** For each valid human instance, we adopt a body-based annotation approach. Using the provided human bounding boxes, we crop the body images and request annotators to label the gender. To maintain data quality, we conduct quality inspections and manual corrections throughout the labeling process. **Age annotation.** To enhance the quality of annotation, we employ a two-stage strategy based on body-based annotation. Similar to gender annotation, age annotation is also performed on cropped body images. We implement a coarse-to-fine two-stage annotation strategy, considering age group annotation to be comparatively easier than apparent age annotation [2]. In the first stage, age groups are annotated. Following [44], we divide the age ranges into six groups, *i.e*. "baby", "kid", "teen", "young", "middle aged", and "elderly". For each cropped person image, we request a group of 10 annotators to independently and repeatedly label the age groups (6-category classification task). We take the mode of the 10 votes as the ground-truth age group. In the second stage, the apparent age is annotated. Given the age group as a prior, a group of 10 annotators independently and repeatedly annotate the apparent age. Consequently, we obtain 10 votes for each human instance. We remove the outliers and take the average as the final ground-truth apparent age. As a summary, the dataset contains over 1M apparent votes. Experiments validate the effectiveness of the body-based annotation strategy and the two-stage annotation strategy (see Sec. A2).

4. Method

4.1. Overview

This study endeavors to develop a single-stage framework that effectively supports a wide range of human-centric perception (HCP) tasks. The key is to learn a comprehensive human representation, which can be universally employed across various HCP tasks. To achieve this, we employ a query-based methodology and investigate the feasibility of representing each human instance as a single shared query.

Our framework is characterized by simplicity, flexibility, and scalability. Unlike previous task-specific HCP models that may incorporate specialized designs tailored to specific tasks (*e.g*. "mask-enhanced anchor box initialization" in Mask DINO [41]), our approach aims to handle various human-centric analysis tasks in a unified manner. To maximize knowledge sharing among various HCP tasks, we attempt to share most weights across different HCP tasks.

As illustrated in Figure 2, our framework consists of four key components: a backbone network, a Transformer encoder, a task-shared Transformer decoder and task-specific heads. The backbone network, such as ResNet [23], takes an image as input and produces multi-scale features. These features, along with corresponding positional embeddings, are then passed through the Transformer encoder to enhance the feature representation. We use the mixed query selection technique to select initial anchor boxes as positional queries for the Transformer decoder. Following DINO [96], we only initialize the positional queries but do not initialize content queries. Unlike previous approaches that employ task-specific Transformer decoders, we propose to use a task-shared decoder for all HCP tasks. The Transformer decoder incorporates the deformable attention [106] to refine the queries across decoder layers. We refer to the refined content queries as "Human Query" as they encode diverse information pertaining to human instances. Finally, the Human Queries are fed into each light-weight task-specific head for final prediction.

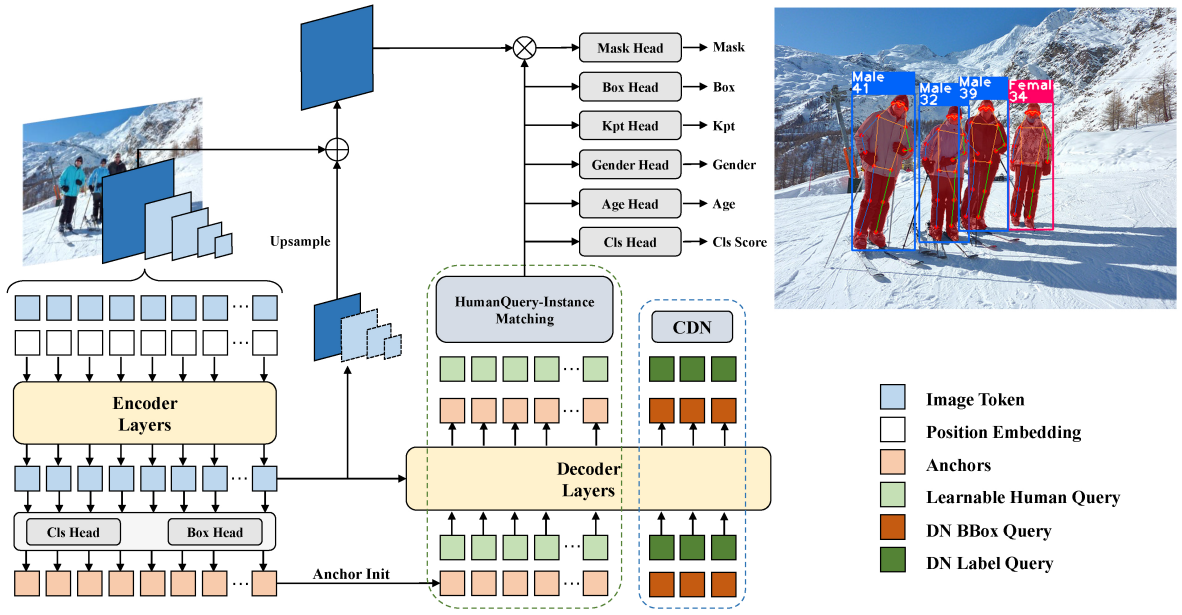

Figure 2. Overview of HQNet. HQNet unifies various representative HCP tasks in a single network by learning shared Human Query.

4.2. Task-shared Transformer Decoder

Queries in DETR-like models are formed by two parts: positional queries and content queries. Each positional query is formulated as a 4D anchor box, encoding the center x-y coordinates, width and height of the box, respectively. Our content query, denoted as Human Query, encapsulates various features (local and global appearance features, as well as coarse- and fine-grained localization features) specific to each instance.

Contrastive DeNoising (CDN). To enhance training stability and acceleration, we employ Contrastive DeNoising (CDN) as introduced in DINO [96]. Notably, we observe that incorporating auxiliary DeNoise losses for other tasks (*e.g.* segmentation and pose) does not yield significant improvements. Consequently, we only apply DN losses for human detection.

HumanQuery-Instance Matching. To ensure consistent and unique predictions for each ground-truth instance across all HCP tasks, including classification (Cls.), detection (Det.), keypoint (Kpt.), and segmentation (Seg.), we employ HumanQuery-Instance (HQ-Ins) Matching. $\lambda_{cls}L_{cls} + \lambda_{det}L_{det} + \lambda_{seg}L_{seg} + \lambda_{kpt}L_{kpt}$, where λ are the corresponding loss weights (see Sec. A4 for details).

4.3. Task-specific Heads

To ensure scalability, we categorize HCP tasks into three groups and design specific implementation paradigms for each category. **Coordinate prediction** tasks (*e.g.* object detection and pose estimation) share common reference points with bounding box prediction and directly regress the normalized offsets of each point. **Dense prediction** tasks (*e.g.* instance segmentation and human parsing) follow the design of Mask DINO [41], which involves constructing a high-resolution pixel embedding map by integrating features from both the backbone and the Transformer encoder. By performing a dot-product operation between the content query embedding and the pixel embedding map, an instance-aware pixel embedding map is generated, facilitating pixel-level classification. **Classification** tasks (*e.g.* determining if an instance is human, gender and age estimation) directly map the Human Query to the classification prediction results, as the Human Query inherently encodes the positional information.

To minimize the overhead of incorporating new tasks, we employ lightweight task-specific heads comprising simple linear layers. The details of our four task-specific heads are as follows: **Human detection head.** A 3-layer multi-layer perceptron (MLP) with a hidden dimension of d is utilized to predict the normalized center x-y coordinates, height, and width of the bounding box w.r.t. the input image. Additionally, a linear projection layer (FC) is employed to predict the class label (human or non-human). **Pose estimation head.** Following the coordinate prediction paradigm, the learned Human Query is fed into a pose regression head (MLP) to regress the relative pose offsets w.r.t. the shared reference points of the detection head. A confidence prediction head (FC) is used to predict confidence score of having visible keypoints. Following PETR [72], joint decoder layers are employed to refine body poses by leveraging structured relations between body keypoints. An auxiliary heatmap branch is used to aid training and discarded during testing. **Human instance segmentation head.** A 3-layer MLP is

Table 2. **Comparisons with task-specific and multi-task models** on the COCO-UniHuman `val` set. We report AP for the "Person" category without *Small* category person. * denotes models trained to handle general 80 classes. † denotes flip testing. We compare with ♢ top-down, ♡ bottom-up, ★ one-stage approaches.

Model	Backbone	Det.			Seg.			Kpt.			Cls. (Gender)			Cls. (Age)		
		AP	AP^M	AP^L	AP	AP^M	AP^L	AP	AP^M	AP^L	AP	AP^M	AP^L	AP	AP^M	AP^L
Faster R-CNN [66]	R-50	65.3	61.5	71.2	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗
DINO [96]	R-50	73.3	68.1	79.9	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗
♢ Mask R-CNN [24]	R-50-FPN	66.7	62.3	73.1	58.4	51.8	66.2	✗	✗	✗	✗	✗	✗	✗	✗	✗
★ PolarMask [88]	R-50-FPN	✗	✗	✗	45.1	38.5	57.1	✗	✗	✗	✗	✗	✗	✗	✗	✗
★ YOLACT [5]	R-50-FPN	✗	✗	✗	47.4	40.1	61.2	✗	✗	✗	✗	✗	✗	✗	✗	✗
★ MEInst [98]	R-50-FPN	✗	✗	✗	49.3	42.3	57.6	✗	✗	✗	✗	✗	✗	✗	✗	✗
★ CondInst [77]	R-50-FPN	✗	✗	✗	54.8	43.3	69.0	✗	✗	✗	✗	✗	✗	✗	✗	✗
★ Mask DINO [41]	R-50	72.3	66.5	79.5	64.8	57.3	73.4	✗	✗	✗	✗	✗	✗	✗	✗	✗
♢ SBL† [87]	R-50	✗	✗	✗	✗	✗	✗	70.4	67.1	77.2	✗	✗	✗	✗	✗	✗
♢ Swin† [54]	Swin-L	✗	✗	✗	✗	✗	✗	74.3	70.6	81.2	✗	✗	✗	✗	✗	✗
♢ HRNet† [73]	HRNet-32	✗	✗	✗	✗	✗	✗	74.4	70.8	81.0	✗	✗	✗	✗	✗	✗
♢ PRTR† [43]	R-50	✗	✗	✗	✗	✗	✗	68.2	63.2	76.2	✗	✗	✗	✗	✗	✗
♡ HrHRNet† [11]	HRNet-w32	✗	✗	✗	✗	✗	✗	67.1	61.5	76.1	✗	✗	✗	✗	✗	✗
♡ DEKR† [20]	HRNet-w32	✗	✗	✗	✗	✗	✗	68.0	62.1	77.7	✗	✗	✗	✗	✗	✗
♡ SWAHR† [56]	HRNet-w32	✗	✗	✗	✗	✗	✗	68.9	63.0	77.4	✗	✗	✗	✗	✗	✗
★ CID† [79]	R-50-FPN	✗	✗	✗	✗	✗	✗	52.0	48.6	58.0	✗	✗	✗	✗	✗	✗
★ CID† [79]	HRNet-w32	✗	✗	✗	✗	✗	✗	69.8	64.0	78.9	✗	✗	✗	✗	✗	✗
★ FCPose [57]	R-50	✗	✗	✗	✗	✗	✗	63.0	59.1	70.3	✗	✗	✗	✗	✗	✗
★ InsPose [71]	R-50	✗	✗	✗	✗	✗	✗	65.2	60.6	72.2	✗	✗	✗	✗	✗	✗
★ PETR [72]	R-50	✗	✗	✗	✗	✗	✗	68.8	62.7	77.7	✗	✗	✗	✗	✗	✗
♢ StrongBL [28]	R-50	-	-	-	✗	✗	✗	✗	✗	✗	46.4	35.2	53.2	✗	✗	✗
♢ Mask R-CNN [24]	R-50	66.3	61.9	72.8	✗	✗	✗	✗	✗	✗	46.7	36.3	52.8	✗	✗	✗
♢ StrongBL [28]	R-50	-	-	-	✗	✗	✗	✗	✗	✗	✗	✗	✗	42.3	31.9	48.3
♢ Mask R-CNN [24]	R-50	66.3	62.1	72.5	✗	✗	✗	✗	✗	✗	✗	✗	✗	37.4	27.9	43.3
♢ Pose2Seg [100]	R-50-FPN	✗	✗	✗	55.5	49.8	67.0	59.9	-	-	✗	✗	✗	✗	✗	✗
♡ MultiPoseNet [1]	R-50	-	58.0	68.1	-	-	-	62.3	57.7	70.4	✗	✗	✗	✗	✗	✗
♡ PersonLab [65]	R-152	✗	✗	✗	-	48.3	59.5	66.5	62.3	73.2	✗	✗	✗	✗	✗	✗
★ CenterNet [105]	Hourglass	-	-	-	✗	✗	✗	64.0	59.4	72.1	✗	✗	✗	✗	✗	✗
★ LSNet-5 [101]	DLA-34	✗	✗	✗	56.2	44.2	71.0	-	-	-	✗	✗	✗	✗	✗	✗
★ UniHead* [45]	R-50-FPN	67.3	62.6	74.4	38.6	37.2	42.2	57.5	55.3	61.9	✗	✗	✗	✗	✗	✗
★ HQNet (Ours)	R-50	74.9	70.4	80.7	65.8	58.7	73.9	69.3	63.8	77.3	56.0	42.5	63.3	53.8	39.7	61.2
★ HQNet (Ours)	Swin-L	77.3	73.3	82.7	68.1	60.9	75.9	72.6	67.4	80.1	57.9	43.1	65.8	56.2	41.5	63.9
★ HQNet (Ours)	ViT-L	78.0	73.6	83.7	68.6	61.4	76.5	75.3	69.8	83.5	58.0	44.7	65.0	58.0	40.9	66.7

used to process the instance-aware pixel embedding map and output a one-channel mask, which is then upsampled to match the original input image size. **Human attribute head.** The gender estimation head and the age estimation head operate in parallel. Both heads consist of two-layer MLPs. Gender estimation involves binary classification, while age estimation is formulated as an 85-class ([1, 85]) classification with softmax expected value [68] estimation.

5. Experiments

5.1. Dataset and Evaluation Metric

COCO-UniHuman Dataset. Our model training exclusively employs COCO-UniHuman train data (in addition to ImageNet pre-training). We follow DINO [96] for augmentation and adopt the 100-epoch training schedule. Model evaluation takes place on COCO-UniHuman `val` set (5K images). Due to limitations of the COCO `test-dev` evaluation server, which lacks support for "Person" category evaluation and attribute recognition, we mainly report results on the `val` set. The evaluation is based on the standard COCO metrics including Average Precision (AP), AP^M for medium-sized persons and AP^L for large-sized persons. Following [100, 101], we exclude the *Small* category persons during evaluation due to the lack of annotations in COCO. For attribute recognition, we also use AP with Age-10 metric for evaluation, where the age estimation is considered correct if the prediction error is no larger than 10.

OCHuman Dataset. OCHuman dataset is a large benchmark dataset that focuses on heavily occluded humans. It contains no training samples and is intended solely for evaluation purposes. Following [100], we train models on the COCO `train` set and evaluate models on OCHuman `val` set (4731 images) and `test` set (8110 images).

Table 3. Comparison with state-of-the-art models on the OCHuman dataset. † denotes flip testing. We compare with ♢ top-down, ♡ bottom-up, ★ one-stage approaches.

Model	Backbone	OCHuman Val			OCHuman Test		
		Det.	Seg.	Kpt.	Det.	Seg.	Kpt.
♢ Mask R-CNN	R-50-FPN	-	16.3	✗	-	16.9	✗
♢ SBL†	R-50	✗	✗	37.8	✗	✗	30.4
♢ Pose2Seg	R-50-FPN	✗	22.2	28.5	✗	23.8	30.3
♡ DEKR†	HRNet-w32	✗	✗	37.9	✗	✗	36.5
♡ HrHRNet†	HRNet-w32	✗	✗	40.0	✗	✗	39.4
★ YOLACT	R-101-FPN	✗	13.2	✗	✗	13.5	✗
★ CondInst	R-50-FPN	✗	20.3	✗	✗	20.1	✗
★ LSNet-5	DLA-34	✗	25.0	✗	✗	24.9	✗
★ LOGO-CAP†	HRNet-w32	✗	✗	39.0	✗	✗	38.1
★ CID†	R-50-FPN	✗	✗	29.2	✗	✗	28.3
★ CID†	HRNet-w32	✗	✗	44.9	✗	✗	44.0
★ HQNet (Ours)	R-50	30.6	31.5	40.3	29.5	31.1	40.0
★ HQNet (Ours)	ViT-L	36.9	39.9	46.8	35.8	38.8	45.6

5.2. Results on COCO dataset

We compare our method to task-specific and multi-task HCP models on the COCO-UniHuman dataset in Table 2. It outperforms multi-task HCP models and achieves very competitive results against task-specific HCP models. More details about the baselines can be found in Sec. A5.

Comparison with task-specific HCP models. For human detection, we compare two baseline approaches, *i.e*. Faster-RCNN [66] and DINO [96]. For human instance segmentation, we contrast HQNet with state-of-the-art general and human-specific instance segmentation methods, including Mask R-CNN [24], PolarMask [88], MEInst [98], YOLACT [5], and CondInst [77]. For human pose estimation, we compare with several representative top-down methods (SBL [87], HRNet [73], Swin [54] and PRTR [43]), bottom-up approaches (HrHRNet [11], DEKR [20], and SWAHR [56]) and single-stage approaches (FCPose [57], InsPose [71], PETR [72] and CID [79]). For gender and age estimation, we establish baselines using StrongBL [28] and Mask R-CNN [24]. Our approach achieves very competitive performance compared to other task-specific HCP models when using the R-50 backbone. Moreover, with stronger backbones such as Swin-L and ViT-L, we establish new state-of-the-art results.

Comparison with multi-task HCP methods. Pose2Seg [100] is a two-stage human pose-based instance segmentation approach. It uses a standalone keypoint detector for pose estimation and employs human skeleton features for top-down instance segmentation guidance. MultiPoseNet [1] and PersonLab [65] follow bottom-up strategies. CenterNet [105], LSNet [101], and UniHead [45][3] are single-stage alternatives. Our R-50 model achieves superior detection, keypoint estimation, and segmentation performance in multi-task HCP, without bells and whistles.

[3]UniHead trains separate models for different HCP tasks.

5.3. Results on the OCHuman dataset

To verify the performance of HQNet in more challenging crowded scenarios, we compare it with recent works on the OCHuman benchmark, which is a popular crowded scene benchmark for human detection, segmentation, and pose estimation. We show that our model outperforms previous methods under the same ResNet50 backbone network by a large margin. For instance, it outperforms SBL by 9.6 keypoint AP and CondInst by 11.0 segmentation AP on `test` set. It even achieves superior performance than HrHRNet (40.3 vs 40.0) and LOGO-CAP (40.3 vs 39.0) even with a much smaller backbone (ResNet-50 vs. HRNet-w32). With a stronger backbone, *i.e*. ViT-L, our HQNet sets new state-of-the-art results on detection (35.8 AP), segmentation (38.8 AP), and pose estimation (45.6 AP).

5.4. Generalize to new HCP tasks

Finetuning evaluation. Similar to linear probing in image classification, we freeze our backbone and transformer encoder (from Table 2) and finetune other parts to evaluate the generalization ability of HQNet on a new HCP task, *i.e*. face detection. In Table 4, we compare our approach with Faster R-CNN [66] and ZoomNet [33]. Our HQNet can not only better exploit the inherent multi-level structure of the human body, but also preserve the efficiency of single-stage detection. It outperforms Faster R-CNN (68.4 AP vs 43.9 AP) and ZoomNet (68.4 AP vs 58.2 AP) by a large margin.

Unseen-task (zero-shot) generalization. We evaluate the generalization ability of our approach through an unseen task evaluation, specifically multiple object tracking (MOT) on the PoseTrack21 dataset [15]. Our models are trained solely on the COCO-UniHuman image-based dataset without explicit tuning for MOT. We hypothesize that our learned human query embeddings, which encode instance-specific features of diverse granularity, can serve as strong cues for distinguishing different objects. We utilized DeepSORT [84] and used the learned Human Query as re-identification features for association. In Table 5, we compare our results with two state-of-the-art single-network MOT methods that were pretrained on COCO and fine-tuned on PoseTrack21. Despite not explicitly being trained for MOT, our HQNet (R-50) achieves highly competitive results (64.6 IDF1 and 51.1 MOTA). This demonstrates the generalization ability of our learned Human Query. HQNet-Det refers to HQNet trained solely on the detection task, and we observed that co-training on multiple HCP tasks improved the quality of the query embeddings (64.6 IDF1 vs. 62.4 IDF1). Furthermore, by employing a stronger ViT backbone, our approach achieves state-of-the-art performance in the "zero-shot" MOT evaluation.

Table 4. **Finetuning evaluation on novel face detection tasks.** Face detection results are reported on COCO-UniHuman `val` dataset.

Method	Face detection	
	AP	AR
Faster RCNN [66]	43.9	71.2
ZoomNet [33]	58.2	72.8
HQNet (R-50)	**68.4**	**83.2**

Table 5. **Unseen-task (MOT) evaluation on PoseTrack21 [15].** 'FT' means fine-tuning on PoseTrack21. Our models are evaluated without training on MOT.

Method	FT	IDF1	MOTA
TRMOT [82]	✓	57.3	47.2
FairMOT [103]	✓	63.2	56.3
HQNet-Det (R-50)	✗	62.4	48.6
HQNet (R-50)	✗	64.6	51.1
HQNet (ViT-L)	✗	**69.1**	**57.0**

Table 6. **Robustness to domain shift.** All models are evaluated on Human-Art [34] `val` set without training on Human-Art.

Method	Det.	Kpt.
Faster R-CNN [66] + HRNet [73]	12.0	22.2
YOLOX [19] + ViTPose [92]	14.4	28.7
HigherHRNet [11]	-	34.6
ED-Pose [94]	-	37.5
HQNet (Swin-L)	15.8	43.0
HQNet (ViT-L)	**18.7**	**52.2**

Table 7. Analysis of multi-task co-learning on COCO-UniHuman `val` set. We report AP for the "Person" category without *Small* category person. All models use ResNet-50 backbone with 100-epoch training setting for fair comparisons.

Model	Params	Det.			Seg.			Kpt.			Cls. (Gender)			Cls. (Age)		
		AP	AP^M	AP^L	AP	AP^M	AP^L	AP	AP^M	AP^L	AP	AP^M	AP^L	AP	AP^M	AP^L
Det.	47.29 M	73.3	68.1	79.8	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗
Det.+Seg.	48.30 M	73.0	68.0	79.4	63.6	57.6	72.1	✗	✗	✗	✗	✗	✗	✗	✗	✗
Det.+Kpt.+Seg.	51.99 M	74.5	70.3	80.1	65.7	**58.7**	73.8	**69.5**	**64.4**	77.0	✗	✗	✗	✗	✗	✗
Det.+Kpt.+Seg.+Cls.	52.14 M	**74.9**	**70.4**	**80.7**	**65.8**	**58.7**	**73.9**	69.3	63.8	**77.3**	**53.8**	**39.7**	**61.2**	**56.0**	**42.5**	**63.3**

5.5. Robustness to domain shift

HumanArt [34] contains images from both natural and artificial (*e.g.* cartoon and painting) scenarios, which can be used for evaluating the robustness to domain shift. In Table 6, we conduct a system-level cross-domain evaluation by directly evaluating all models on Human-Art `val` set without any finetuning. We observe that all models, particularly two-stage models, experienced a decline in performance when a domain gap was present. However, our approach maintained competitive performance, showcasing its resilience to the domain gap.

5.6. More Analysis

Multi-task co-learning. In Table 7, we observed that co-learning with multiple human-centric tasks leads to improved overall performance. This enhancement can be attributed to the inter-task synergy that arises from jointly training different HCP tasks.

Computational cost analysis. In Table 7, we follow [72] to analyze the computational cost. Specifically, we use ResNet-50 backbone with the input size of 800×800. In HQNet, multiple tasks share the computation cost of the backbone, transformer encoder and decoder. The overhead of each task-specific head is negligible, showing good scalability of HQNet in terms of increasing the number of tasks.

Effect of HQ-Ins Matching. As shown in Figure 3 (left), with box only matching [96], there may be some erroneous cases when one person's pose is matched to another person. HQ-Ins Matching avoids such errors by comprehensively considering multiple tasks as a whole. Quantitative evaluation can be found in Sec. A3.2.

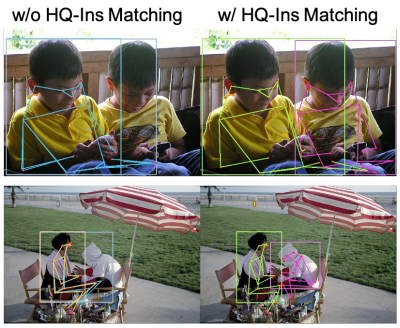

Figure 3. Effect of HumanQuery-Instance (HQ-Ins) Matching. **w/o HQ-Ins Matching**: Using detection loss only for bipartite matching [96]. **w/ HQ-Ins Matching**: Comprehensively using detection, keypoint, and segmentation loss for bipartite matching.

6. Conclusion

In this work, we present a unified solution towards single-stage multi-task human-centric perception, called HQNet. The core idea is to learn a unified query representation (Human Query) that encodes different features (local and global appearance features, coarse and fine-grained localization features) for each instance. To facilitate model training and evaluation, we introduce a large-scale benchmark, termed COCO-UniHuman benchmark, to unify different representative HCP tasks (including classification, detection, segmentation, and keypoint localization). We extensively compare our proposed method with several state-of-the-art task-specific and multi-task approaches, and show the effectiveness of our proposed method.

Limitations. While we focus on 2D tasks, tasks about 3D [58, 59, 86] or sequential data [21, 89] also hold significant potential. We encourage future research to explore more comprehensive multi-task human-centric perception.

Acknowledgement. We thank Wang Zeng and Ruijie Yao for valuable discussions. This paper is partially supported by the National Key R&D Program of China No.2022ZD0161000 and the General Research Fund of Hong Kong No.17200622.

References

[1] Abrar H Abdulnabi, Gang Wang, Jiwen Lu, and Kui Jia. Multi-task cnn model for attribute prediction. *IEEE Trans. Multimedia*, 17(11):1949–1959, 2015. 6, 7, 19

[2] Eirikur Agustsson, Radu Timofte, Sergio Escalera, Xavier Baro, Isabelle Guyon, and Rasmus Rothe. Apparent and real age estimation in still images with deep residual regressors on appa-real database. In *IEEE Int. Conf. Auto. Face & Gesture Recog.*, pages 87–94, 2017. 3, 4, 19

[3] Mykhaylo Andriluka, Leonid Pishchulin, Peter Gehler, and Bernt Schiele. 2d human pose estimation: New benchmark and state of the art analysis. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2014. 3

[4] Mykhaylo Andriluka, Umar Iqbal, Eldar Insafutdinov, Leonid Pishchulin, Anton Milan, Juergen Gall, and Bernt Schiele. Posetrack: A benchmark for human pose estimation and tracking. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2018. 3

[5] Daniel Bolya, Chong Zhou, Fanyi Xiao, and Yong Jae Lee. Yolact: Real-time instance segmentation. In *Int. Conf. Comput. Vis.*, pages 9157–9166, 2019. 2, 6, 7, 18

[6] Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Realtime multi-person 2d pose estimation using part affinity fields. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2017. 2

[7] Zhe Cao, Gines Hidalgo, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Openpose: realtime multi-person 2d pose estimation using part affinity fields. *IEEE Trans. Pattern Anal. Mach. Intell.*, 2018. 2

[8] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *Eur. Conf. Comput. Vis.*, pages 213–229, 2020. 2, 3

[9] Kai Chen, Jiaqi Wang, Jiangmiao Pang, Yuhang Cao, Yu Xiong, Xiaoxiao Li, Shuyang Sun, Wansen Feng, Ziwei Liu, Jiarui Xu, et al. Mmdetection: Open mmlab detection toolbox and benchmark. *arXiv preprint arXiv:1906.07155*, 2019. 14, 16

[10] Weihua Chen, Xianzhe Xu, Jian Jia, Hao Luo, Yaohua Wang, Fan Wang, Rong Jin, and Xiuyu Sun. Beyond appearance: a semantic controllable self-supervised learning framework for human-centric visual tasks. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 15050–15061, 2023. 3, 18

[11] Bowen Cheng, Bin Xiao, Jingdong Wang, Honghui Shi, Thomas S Huang, and Lei Zhang. Higherhrnet: Scale-aware representation learning for bottom-up human pose estimation. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 5386–5395, 2020. 6, 7, 8, 17

[12] Yuanzheng Ci, Yizhou Wang, Meilin Chen, Shixiang Tang, Lei Bai, Feng Zhu, Rui Zhao, Fengwei Yu, Donglian Qi, and Wanli Ouyang. Unihcp: A unified model for human-centric perceptions. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 17840–17852, 2023. 1, 3, 18

[13] MMPose Contributors. Openmmlab pose estimation toolbox and benchmark. https://github.com/open-mmlab/mmpose, 2020. 18

[14] Yubin Deng, Ping Luo, Chen Change Loy, and Xiaoou Tang. Pedestrian attribute recognition at far distance. In *ACM Int. Conf. Multimedia*, pages 789–792, 2014. 3, 4, 19

[15] Andreas Doering, Di Chen, Shanshan Zhang, Bernt Schiele, and Juergen Gall. Posetrack21: A dataset for person search, multi-object tracking and multi-person pose tracking. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 20963–20972, 2022. 7, 8, 14

[16] Piotr Dollár, Christian Wojek, Bernt Schiele, and Pietro Perona. Pedestrian detection: A benchmark. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 304–311, 2009. 3

[17] 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. *Int. Conf. Learn. Represent.*, 2021. 2

[18] Haodong Duan, Kwan-Yee Lin, Sheng Jin, Wentao Liu, Chen Qian, and Wanli Ouyang. Trb: a novel triplet representation for understanding 2d human body. In *Int. Conf. Comput. Vis.*, pages 9479–9488, 2019. 3

[19] Zheng Ge, Songtao Liu, Feng Wang, Zeming Li, and Jian Sun. Yolox: Exceeding yolo series in 2021. *arXiv preprint arXiv:2107.08430*, 2021. 8

[20] Zigang Geng, Ke Sun, Bin Xiao, Zhaoxiang Zhang, and Jingdong Wang. Bottom-up human pose estimation via disentangled keypoint regression. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 14676–14686, 2021. 2, 6, 7, 17

[21] Rohit Girdhar, Georgia Gkioxari, Lorenzo Torresani, Manohar Paluri, and Du Tran. Detect-and-track: Efficient pose estimation in videos. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 350–359, 2018. 8

[22] Ke Gong, Xiaodan Liang, Yicheng Li, Yimin Chen, Ming Yang, and Liang Lin. Instance-level human parsing via part grouping network. In *Eur. Conf. Comput. Vis.*, pages 770–785, 2018. 2, 3

[23] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2016. 2, 4, 14, 18

[24] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Int. Conf. Comput. Vis.*, pages 2961–2969, 2017. 2, 6, 7, 18, 19

[25] Fangzhou Hong, Liang Pan, Zhongang Cai, and Ziwei Liu. Versatile multi-modal pre-training for human-centric perception. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 16156–16166, 2022. 3, 18

[26] Eldar Insafutdinov, Leonid Pishchulin, Bjoern Andres, Mykhaylo Andriluka, and Bernt Schiele. Deepercut: A deeper, stronger, and faster multi-person pose estimation model. In *Eur. Conf. Comput. Vis.*, pages 34–50, 2016. 2

[27] Catalin Ionescu, Dragos Papava, Vlad Olaru, and Cristian Sminchisescu. Human3.6m: Large scale datasets and predictive methods for 3d human sensing in natural environments. *IEEE Trans. Pattern Anal. Mach. Intell.*, 36(7): 1325–1339, 2013. 1

[28] Jian Jia, Houjing Huang, Wenjie Yang, Xiaotang Chen, and Kaiqi Huang. Rethinking of pedestrian attribute recognition: Realistic datasets with efficient method. *arXiv preprint arXiv:2005.11909*, 2020. 2, 6, 7, 18, 19

[29] Wentao Jiang, Sheng Jin, Wentao Liu, Chen Qian, Ping Luo, and Si Liu. Posetrans: A simple yet effective pose transformation augmentation for human pose estimation. In *Eur. Conf. Comput. Vis.*, pages 643–659, 2022. 2

[30] Sheng Jin, Xujie Ma, Zhipeng Han, Yue Wu, Wei Yang, Wentao Liu, Chen Qian, and Wanli Ouyang. Towards multi-person pose tracking: Bottom-up and top-down methods. In *Int. Conf. Comput. Vis. Worksh.*, page 7, 2017. 2

[31] Sheng Jin, Wentao Liu, Wanli Ouyang, and Chen Qian. Multi-person articulated tracking with spatial and temporal embeddings. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 5664–5673, 2019. 2

[32] Sheng Jin, Wentao Liu, Enze Xie, Wenhai Wang, Chen Qian, Wanli Ouyang, and Ping Luo. Differentiable hierarchical graph grouping for multi-person pose estimation. In *Eur. Conf. Comput. Vis.*, pages 718–734, 2020. 2

[33] 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 *Eur. Conf. Comput. Vis.*, 2020. 3, 7, 8

[34] Xuan Ju, Ailing Zeng, Jianan Wang, Qiang Xu, and Lei Zhang. Human-art: A versatile human-centric dataset bridging natural and artificial scenes. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 618–629, 2023. 8

[35] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 15

[36] Alexander Kirillov, Evgeny Levinkov, Bjoern Andres, Bogdan Savchynskyy, and Carsten Rother. Instancecut: from edges to instances with multicut. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 5008–5017, 2017. 2

[37] Shu Kong and Charless C Fowlkes. Recurrent pixel embedding for instance grouping. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 9018–9028, 2018. 2

[38] Gil Levi and Tal Hassner. Age and gender classification using convolutional neural networks. In *IEEE Conf. Comput. Vis. Pattern Recog. Worksh.*, pages 34–42, 2015. 2

[39] Dangwei Li, Zhang Zhang, Xiaotang Chen, and Kaiqi Huang. A richly annotated pedestrian dataset for person retrieval in real surveillance scenarios. *IEEE Trans. Image Process.*, 28(4):1575–1590, 2019. 19

[40] Feng Li, Hao Zhang, Shilong Liu, Jian Guo, Lionel M Ni, and Lei Zhang. Dn-detr: Accelerate detr training by introducing query denoising. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 13619–13627, 2022. 2, 3, 16

[41] Feng Li, Hao Zhang, Huaizhe Xu, Shilong Liu, Lei Zhang, Lionel M Ni, and Heung-Yeung Shum. Mask dino: Towards a unified transformer-based framework for object detection and segmentation. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 3041–3050, 2023. 4, 5, 6

[42] Jianshu Li, Jian Zhao, Yunchao Wei, Congyan Lang, Yidong Li, Terence Sim, Shuicheng Yan, and Jiashi Feng. Multiple-human parsing in the wild. *arXiv preprint arXiv:1705.07206*, 2017. 3

[43] Ke Li, Shijie Wang, Xiang Zhang, Yifan Xu, Weijian Xu, and Zhuowen Tu. Pose recognition with cascade transformers. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 1944–1953, 2021. 6, 7, 17

[44] Yining Li, Chen Huang, Chen Change Loy, and Xiaoou Tang. Human attribute recognition by deep hierarchical contexts. In *Eur. Conf. Comput. Vis.*, 2016. 3, 4, 19

[45] Jianming Liang, Guanglu Song, Biao Leng, and Yu Liu. Unifying visual perception by dispersible points learning. In *Eur. Conf. Comput. Vis.*, pages 439–456, 2022. 3, 6, 7, 18

[46] 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 *Eur. Conf. Comput. Vis.*, 2014. 1, 3

[47] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection, 2017. 14

[48] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Int. Conf. Comput. Vis.*, pages 2980–2988, 2017. 15

[49] Yutian Lin, Liang Zheng, Zhedong Zheng, Yu Wu, Zhilan Hu, Chenggang Yan, and Yi Yang. Improving person re-identification by attribute and identity learning. *Pattern Recognition*, 2019. 19

[50] Yiming Lin, Jie Shen, Yujiang Wang, and Maja Pantic. Fp-age: Leveraging face parsing attention for facial age estimation in the wild. *IEEE Trans. Image Process.*, 2022. 3, 19

[51] Shilong Liu, Feng Li, Hao Zhang, Xiao Yang, Xianbiao Qi, Hang Su, Jun Zhu, and Lei Zhang. Dab-detr: Dynamic anchor boxes are better queries for detr. *Int. Conf. Learn. Represent.*, 2022. 2, 3

[52] Xihui Liu, Haiyu Zhao, Maoqing Tian, Lu Sheng, Jing Shao, Junjie Yan, and Xiaogang Wang. Hydraplus-net: Attentive deep features for pedestrian analysis. In *Int. Conf. Comput. Vis.*, pages 1–9, 2017. 1, 3, 19

[53] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Int. Conf. Comput. Vis.*, 2015. 3

[54] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Int. Conf. Comput. Vis.*, pages 10012–10022, 2021. 2, 6, 7, 17

[55] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J. Black. SMPL: A skinned multi-person linear model. *ACM Trans. Graphics (Proc. SIGGRAPH Asia)*, 34(6):248:1–248:16, 2015. 4

[56] Zhengxiong Luo, Zhicheng Wang, Yan Huang, Liang Wang, Tieniu Tan, and Erjin Zhou. Rethinking the heatmap

regression for bottom-up human pose estimation. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 13264–13273, 2021. 6, 7, 17

[57] Weian Mao, Zhi Tian, Xinlong Wang, and Chunhua Shen. Fcpose: Fully convolutional multi-person pose estimation with dynamic instance-aware convolutions. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 9034–9043, 2021. 2, 6, 7, 17

[58] Julieta Martinez, Rayat Hossain, Javier Romero, and James J Little. A simple yet effective baseline for 3d human pose estimation. In *Int. Conf. Comput. Vis.*, pages 2640–2649, 2017. 8

[59] Hao Meng, Sheng Jin, Wentao Liu, Chen Qian, Mengxiang Lin, Wanli Ouyang, and Ping Luo. 3d interacting hand pose estimation by hand de-occlusion and removal. In *Eur. Conf. Comput. Vis.*, pages 380–397, 2022. 8

[60] Alejandro Newell, Zhiao Huang, and Jia Deng. Associative embedding: End-to-end learning for joint detection and grouping. In *Adv. Neural Inform. Process. Syst.*, 2017. 2

[61] Xuecheng Nie, Jiashi Feng, and Shuicheng Yan. Mutual learning to adapt for joint human parsing and pose estimation. In *Eur. Conf. Comput. Vis.*, pages 502–517, 2018. 3, 19

[62] Xuecheng Nie, Jiashi Feng, Yiming Zuo, and Shuicheng Yan. Human pose estimation with parsing induced learner. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 2100–2108, 2018. 3, 19

[63] Xuecheng Nie, Jiashi Feng, Jianfeng Zhang, and Shuicheng Yan. Single-stage multi-person pose machines. In *Int. Conf. Comput. Vis.*, pages 6951–6960, 2019. 2

[64] Hongyu Pan, Hu Han, Shiguang Shan, and Xilin Chen. Mean-variance loss for deep age estimation from a face. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 5285–5294, 2018. 15

[65] George Papandreou, Tyler Zhu, Liang-Chieh Chen, Spyros Gidaris, Jonathan Tompson, and Kevin Murphy. Personlab: Person pose estimation and instance segmentation with a bottom-up, part-based, geometric embedding model. In *Eur. Conf. Comput. Vis.*, pages 269–286, 2018. 2, 6, 7, 19

[66] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster R-CNN: Towards real-time object detection with region proposal networks. In *Adv. Neural Inform. Process. Syst.*, 2015. 6, 7, 8, 16, 18

[67] Hamid Rezatofighi, Nathan Tsoi, JunYoung Gwak, Amir Sadeghian, Ian Reid, and Silvio Savarese. Generalized intersection over union: A metric and a loss for bounding box regression. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 658–666, 2019. 15

[68] Rasmus Rothe, Radu Timofte, and Luc Van Gool. Dex: Deep expectation of apparent age from a single image. In *Int. Conf. Comput. Vis. Worksh.*, pages 10–15, 2015. 2, 6

[69] Tao Ruan, Ting Liu, Zilong Huang, Yunchao Wei, Shikui Wei, and Yao Zhao. Devil in the details: Towards accurate single and multiple human parsing. In *AAAI Conf. Artif. Intell.*, pages 4814–4821, 2019. 2

[70] Shuai Shao, Zijian Zhao, Boxun Li, Tete Xiao, Gang Yu, Xiangyu Zhang, and Jian Sun. Crowdhuman: A benchmark for detecting human in a crowd. *arXiv preprint arXiv:1805.00123*, 2018. 3

[71] Dahu Shi, Xing Wei, Xiaodong Yu, Wenming Tan, Ye Ren, and Shiliang Pu. Inspose: instance-aware networks for single-stage multi-person pose estimation. In *ACM Int. Conf. Multimedia*, pages 3079–3087, 2021. 2, 6, 7, 17

[72] Dahu Shi, Xing Wei, Liangqi Li, Ye Ren, and Wenming Tan. End-to-end multi-person pose estimation with transformers. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 11069–11078, 2022. 5, 6, 7, 8, 15, 17

[73] Ke Sun, Bin Xiao, Dong Liu, and Jingdong Wang. Deep high-resolution representation learning for human pose estimation. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 5693–5703, 2019. 6, 7, 8, 17, 18

[74] Shixiang Tang, Cheng Chen, Qingsong Xie, Meilin Chen, Yizhou Wang, Yuanzheng Ci, Lei Bai, Feng Zhu, Haiyang Yang, Li Yi, et al. Humanbench: Towards general human-centric perception with projector assisted pretraining. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 21970–21982, 2023. 3, 16, 18

[75] Yonglong Tian, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Pedestrian detection aided by deep learning semantic tasks. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 5079–5087, 2015. 3, 19

[76] Zhi Tian, Hao Chen, and Chunhua Shen. Directpose: Direct end-to-end multi-person pose estimation. *arXiv preprint arXiv:1911.07451*, 2019. 2

[77] Zhi Tian, Chunhua Shen, and Hao Chen. Conditional convolutions for instance segmentation. In *Eur. Conf. Comput. Vis.*, pages 282–298, 2020. 6, 7, 18

[78] Can Wang, Sheng Jin, Yingda Guan, Wentao Liu, Chen Qian, Ping Luo, and Wanli Ouyang. Pseudo-labeled auto-curriculum learning for semi-supervised keypoint localization. *Int. Conf. Learn. Represent.*, 2022. 2

[79] Dongkai Wang and Shiliang Zhang. Contextual instance decoupling for robust multi-person pose estimation. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 11060–11068, 2022. 2, 6, 7, 17

[80] Jingdong Wang, Ke Sun, Tianheng Cheng, Borui Jiang, Chaorui Deng, Yang Zhao, Dong Liu, Yadong Mu, Mingkui Tan, Xinggang Wang, et al. Deep high-resolution representation learning for visual recognition. *IEEE Trans. Pattern Anal. Mach. Intell.*, 2020. 3, 18

[81] Jiahang Wang, Sheng Jin, Wentao Liu, Weizhong Liu, Chen Qian, and Ping Luo. When human pose estimation meets robustness: Adversarial algorithms and benchmarks. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 11855–11864, 2021. 2

[82] Zhongdao Wang, Liang Zheng, Yixuan Liu, Yali Li, and Shengjin Wang. Towards real-time multi-object tracking. In *Eur. Conf. Comput. Vis.*, pages 107–122, 2020. 8

[83] Fangyun Wei, Xiao Sun, Hongyang Li, Jingdong Wang, and Stephen Lin. Point-set anchors for object detection, instance segmentation and pose estimation. In *Eur. Conf. Comput. Vis.*, pages 527–544, 2020. 2, 18

[84] Nicolai Wojke, Alex Bewley, and Dietrich Paulus. Simple online and realtime tracking with a deep association metric.

In *IEEE Int. Conf. Image Process.*, pages 3645–3649, 2017. 7

[85] Jiahong Wu, He Zheng, Bo Zhao, Yixin Li, Baoming Yan, Rui Liang, Wenjia Wang, Shipei Zhou, Guosen Lin, Yanwei Fu, et al. Ai challenger: a large-scale dataset for going deeper in image understanding. *arXiv preprint arXiv:1711.06475*, 2017. 3

[86] Size Wu, Sheng Jin, Wentao Liu, Lei Bai, Chen Qian, Dong Liu, and Wanli Ouyang. Graph-based 3d multi-person pose estimation using multi-view images. In *Int. Conf. Comput. Vis.*, pages 11148–11157, 2021. 8

[87] Bin Xiao, Haiping Wu, and Yichen Wei. Simple baselines for human pose estimation and tracking. In *Eur. Conf. Comput. Vis.*, 2018. 2, 6, 7, 17, 18

[88] Enze Xie, Peize Sun, Xiaoge Song, Wenhai Wang, Xuebo Liu, Ding Liang, Chunhua Shen, and Ping Luo. Polarmask: Single shot instance segmentation with polar representation. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 12193–12202, 2020. 2, 6, 7, 18

[89] Lumin Xu, Yingda Guan, Sheng Jin, Wentao Liu, Chen Qian, Ping Luo, Wanli Ouyang, and Xiaogang Wang. Vipnas: Efficient video pose estimation via neural architecture search. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 16072–16081, 2021. 8

[90] 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 Trans. Pattern Anal. Mach. Intell.*, 45(4):5296–5313, 2022. 3

[91] Lumin Xu, Sheng Jin, Wang Zeng, Wentao Liu, Chen Qian, Wanli Ouyang, Ping Luo, and Xiaogang Wang. Pose for everything: Towards category-agnostic pose estimation. In *Eur. Conf. Comput. Vis.*, pages 398–416, 2022. 2

[92] Yufei Xu, Jing Zhang, Qiming Zhang, and Dacheng Tao. Vitpose: Simple vision transformer baselines for human pose estimation. *Adv. Neural Inform. Process. Syst.*, 35: 38571–38584, 2022. 8

[93] Nan Xue, Tianfu Wu, Gui-Song Xia, and Liangpei Zhang. Learning local-global contextual adaptation for multi-person pose estimation. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 13065–13074, 2022. 2

[94] Jie Yang, Ailing Zeng, Shilong Liu, Feng Li, Ruimao Zhang, and Lei Zhang. Explicit box detection unifies end-to-end multi-person pose estimation. *arXiv preprint arXiv:2302.01593*, 2023. 8

[95] Wang Zeng, Sheng Jin, Wentao Liu, Chen Qian, Ping Luo, Wanli Ouyang, and Xiaogang Wang. Not all tokens are equal: Human-centric visual analysis via token clustering transformer. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 11101–11111, 2022. 3, 18

[96] Hao Zhang, Feng Li, Shilong Liu, Lei Zhang, Hang Su, Jun Zhu, Lionel M Ni, and Heung-Yeung Shum. Dino: Detr with improved denoising anchor boxes for end-to-end object detection. *Int. Conf. Learn. Represent.*, 2023. 2, 3, 4, 5, 6, 7, 8, 15, 16

[97] Ning Zhang, Manohar Paluri, Marc'Aurelio Ranzato, Trevor Darrell, and Lubomir Bourdev. Panda: Pose aligned networks for deep attribute modeling. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 1637–1644, 2014. 3, 19

[98] Rufeng Zhang, Zhi Tian, Chunhua Shen, Mingyu You, and Youliang Yan. Mask encoding for single shot instance segmentation. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 10226–10235, 2020. 6, 7, 18

[99] Shanshan Zhang, Rodrigo Benenson, and Bernt Schiele. Citypersons: A diverse dataset for pedestrian detection. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 3213–3221, 2017. 3

[100] 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 *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 889–898, 2019. 3, 6, 7, 19

[101] Xiangzhou Zhang, Bingpeng Ma, Hong Chang, Shiguang Shan, and Xilin Chen. Location sensitive network for human instance segmentation. *IEEE Trans. Image Process.*, 30:7649–7662, 2021. 6, 7, 18

[102] Yunxuan Zhang, Li Liu, Cheng Li, and Chen Change Loy. Quantifying facial age by posterior of age comparisons. In *Brit. Mach. Vis. Conf.*, 2017. 3, 13, 19

[103] Yifu Zhang, Chunyu Wang, Xinggang Wang, Wenjun Zeng, and Wenyu Liu. Fairmot: On the fairness of detection and re-identification in multiple object tracking. *Int. J. Comput. Vis.*, 129:3069–3087, 2021. 8

[104] Liang Zheng, Liyue Shen, Lu Tian, Shengjin Wang, Jingdong Wang, and Qi Tian. Scalable person re-identification: A benchmark. In *Int. Conf. Comput. Vis.*, 2015. 19

[105] Xingyi Zhou, Dequan Wang, and Philipp Krähenbühl. Objects as points. *arXiv preprint arXiv:1904.07850*, 2019. 2, 6, 7, 16

[106] Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable detr: Deformable transformers for end-to-end object detection. *Int. Conf. Learn. Represent.*, 2021. 2, 3, 4, 16

A1. COCO-UniHuman Dataset Statistics

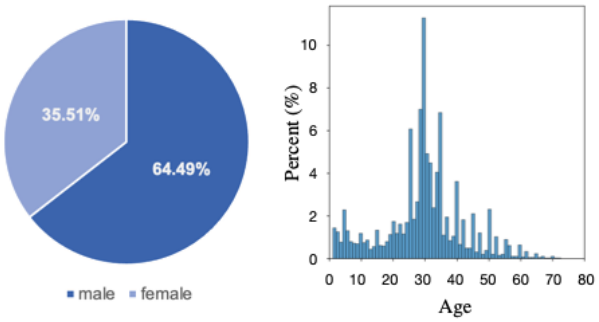

Figure A1. Statistics of COCO-UniHuman benchmark. (a) The gender distribution of COCO-UniHuman is biased towards male. (b) The age distribution ranges from [1, 84] and is biased towards young adults, since images are from public Internet repositories.

In Figure A1, we show statistics of our proposed COCO-UniHuman dataset. The plots show the distribution of the gender and the apparent age. We find gender and age biases existed in the widely used COCO dataset. The occurrence of men is significantly higher than women in COCO dataset. More specifically, male to female ratio is about 65:35. In addition, the dataset has an unbalanced age distribution. The apparent age distribution ranges from [1, 84], and it is mainly concentrated between the ages of 25 and 35. Analyzing and addressing the gender and age bias in the computer vision system can also be an important topic in the AI community. Future work could also use our benchmark dataset to comprehensively measure and analyze such biases, but it is out of the scope of this paper.

A2. COCO-UniHuman Dataset Annotation

Obtaining the reliable apparent age is challenging even for human perception. The apparent age will be influenced not only by the real age, but also by other biological and sociological factors of "aging". Therefore, there are significant variations on appearance among people of the same age. In this work, we propose the body-based and two-stage annotation strategy to improve the age annotation quality. We also conduct some experiments to show the effectiveness of the proposed age annotation strategy.

A2.1. Body-based vs face-based annotation strategies.

In this study, we design experiments to compare three different annotation strategies. (1) face-based without face alignment, where the annotation is based on the cropped face image, (2) face-based with face alignment, where face cropping and face alignment pre-processing [102] is applied before annotation, (3) body-based, where the annotation is based on the cropped body image. We randomly selected 500 sample person images, and applied the aforementioned 3 strategies to process the data individually, and obtained 3 data sets. We also randomly divided 30 well-trained annotators into three groups of 10 annotators each. Each data set was labeled by one group of annotators. Each annotator was asked to independently give votes of apparent age for the whole data set. As a result, for each body or face image, we have 10 votes. We take the average of the 10 votes as the ground-truth age annotation, and calculate the Age-5 and Age-10 consistency separately. Age-n consistency is defined as:

$$\frac{1}{K \times N} \sum_{\substack{1 \le i \le N \\ 1 \le j \le K}} \mathbb{I}\{|x_{i,j} - x_i^*| \le n\} \times 100\%, \qquad (1)$$

where $N = 500$ is the total number of images, and $K = 10$ is the number of votes for each image. $x_{i,j}$ is the j-th vote for the i-th image, while x_i^* means the ground-truth age annotation for the i-th image.

From Table A1, we find that the body-based age annotation is better than the face-based age annotation, indicating that the whole-body image contains richer visual cues for age estimation. Interestingly, we also find that face alignment will help improve the age estimation consistency even for human annotators.

Table A1. Comparisons of age annotation strategies.

Annotation Strategy	Age-5	Age-10
face w/o alignment	75.3	93.5
face w/ alignment [102]	78.2	96.5
body	**80.9**	**98.1**

Table A2. Effect of two-stage age annotation.

Annotation Strategy	Age-5	Age-10
One-stage age annotation	80.9	98.1
Two-stage age annotation	**82.1**	**98.5**

A2.2. Two-stage vs one-stage annotation strategies.

In this study, we design experiments to compare the two-stage and one-stage annotation strategies. For one-stage annotation, we directly annotate the apparent age of the subject. For two-stage annotation, we first annotate the age group and then label the apparent age based on the age group. Table A2, shows that two-stage annotation strategy improves the annotation consistency.

A3. More Experimental Analysis

Multi-task co-learning can mitigate over-fitting. From Figure A2, we observe that training task-specific models on "Person" category only will easily lead to over-fitting

Table A3. Comparison of general 80 class models and person-specific models on the COCO-UniHuman `val` set. We report AP for the "Person" category without *Small* category person. "R" is ResNet [23], and "FPN" is feature pyramid network [47]. The asterisk * denotes models trained to handle general 80 classes.

Model	Backbone	Det.			Seg.		
		AP	AP^M	AP^L	AP	AP^M	AP^L
Faster R-CNN*	R-50	63.0	59.8	68.1	✗	✗	✗
Faster R-CNN	R-50	65.3	61.5	71.2	✗	✗	✗
Mask R-CNN*	R-50-FPN	64.1	60.5	69.6	56.3	50.1	63.9
Mask R-CNN	R-50-FPN	66.7	62.3	73.1	58.4	51.8	66.2

problem, the performance decreases with increasing number of epochs. Specifically, in this experiment, we compare the common 1x, 2x, and 4x training settings for RCNN-based methods (*i.e*. Faster-RCNN, Mask RCNN), and compare 50-epoch and 100-epoch settings for DETR-based methods (*i.e*. DINO, Mask DINO, and our HQNet). The models are trained using MMDetection [9] with suggested hyper-parameters. We report the Average Precision (AP) for both human detection (solid lines) and the human instance segmentation (dashed lines) on COCO-UniHuman `val` set. Interestingly, we find that our presented multi-task co-learning (HQNet) can mitigate the over-fitting problem, and the performance consistently improves with the increasing training epochs, demonstrating good scalability.

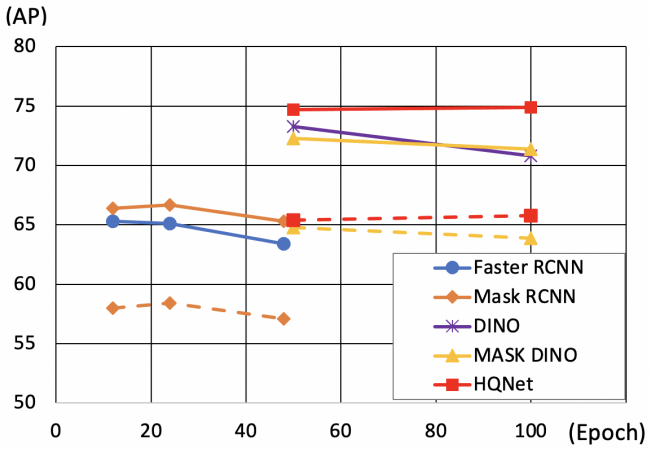

Figure A2. Results of human detection (solid lines) and segmentation (dashed lines) with different training schedules on COCO-UniHuman `val` set. For RCNN-based models, we choose 1x, 2x, and 4x training settings. For DETR-based models, we use 50-epoch and 100-epoch training settings.

A3.1. General models vs person-specific models

In Table A3, we compare general 80-class models and person-specific models on COCO-UniHuman `val` set. We find that person-specific models achieve slightly better performance than the general 80-class models for human analysis. The asterisk * denotes models trained to handle general 80 classes. All models are evaluated on "Person" category without *Small* person. As shown in previous section, training on "Person" category only may lead to over-fitting problem. In the experiments, we report the best result for these baseline models. Specifically, Faster R-CNN is trained for 1x, Mask R-CNN for 2x. More details can be found in the section of "Details about Baseline Models" below.

A3.2. Effect of HumanQuery-Instance Matching

In Table A4, we quantitatively analyze the effect of HumanQuery-Instance (HQ-Ins) Matching on COCO-UniHuman `val` set using the ResNet-50 backbone. Note that we use the standard 100-epoch training setting in the experiment. We report AP for 'Det', 'Seg', 'Kpt', 'Gener', and 'Age', which represent detection, keypoint estimation, instance segmentation, gender and age estimation respectively. We show the effectiveness of our proposed HumanQuery-Instance Matching in making the optimization of multi-task HCP learning more consistent and achieving better balance of multiple human-centric analysis tasks.

A3.3. Qualitative Results

In Figure A3 and Figure A4, we show some qualitative results of HQNet with ResNet-50 backbone. In Figure A3, we show some qualitative results on COCO-UniHuman `val` dataset for human detection, human pose estimation, human instance segmentation, and human attribute recognition. The model can recognize the gender and age of different people

In Figure A4, we visualize the results of human detection and tracking (same color for same id), human pose estimation, human instance segmentation, gender estimation and age estimation. As introduced in the section of "Unseen-task generalization" in the main paper, we directly apply our HQNet on multiple object tracking (with pose estimation and segmentation) on the challenging PoseTrack21 [15] dataset, where our models are trained only on the COCO-UniHuman image-based dataset without explicitly tuned for multi-object tracking (MOT) on video-based dataset like PoseTrack21. Note that our results are simply obtained by image-based human analysis without performing further smoothing post-process. Our learned Human Query, which encodes both spatial and visual cues, can serve as good embedding features to distinguish different human instances. Therefore, our human tracking is robust to heavy occlusion, and the id can recover from occlusions. Our HQNet makes a comprehensive all-in-one human analysis system that can achieve multiple functions: multiple object tracking with human pose estimation, human instance segmentation, and human attribute recognition.

Table A4. Effect of HumanQuery-Instance (HQ-Ins) Matching. Experiments are conducted on COCO-UniHuman `val` set using the ResNet-50 backbone with 100-epoch training setting. We report AP for the "Person" category without *Small* category person.

Matching			Det.			Seg.			Kpt.			Cls. (Gender)			Cls. (Age)		
Box	Pose	Mask	AP	AP^M	AP^L	AP	AP^M	AP^L	AP	AP^M	AP^L	AP	AP^M	AP^L	AP	AP^M	AP^L
✓			76.2	71.4	82.4	66.1	59.1	74.2	66.8	61.0	75.0	52.1	37.3	60.7	54.0	41.2	62.0
✓	✓		74.4	70.2	80.1	65.5	58.5	73.2	69.0	63.9	76.4	54.4	39.9	62.7	55.9	42.0	63.9
✓	✓	✓	74.9	70.4	80.7	65.8	58.7	73.9	69.3	63.8	77.3	53.8	39.7	61.2	56.0	42.5	63.3

A3.4. Attention Visualization

In Figure A5, we visualize the sampling locations of deformable attention for different HCP models. We show the results of the last decoder layer in HQNet-ResNet50. Each sampling point is marked as a red-filled circle. The left results are from the model trained for detection and segmentation ($M_{Det.+Seg.}$). The middle ones are from the model trained for detection and keypoint ($M_{Det.+Kpt.}$). And the right ones are from the model trained for detection, segmentation, keypoint and attribute ($M_{Det.+Kpt.+Seg.+Cls.}$). With the segmentation task, we notice that some of the sampling points of $M_{Det.+Seg.}$ are distributed near the boundary of the human body, and some are distributed in the background to capture more context information. The sampling points of $M_{Det.+Kpt.}$ have higher probability to distribute inside the human body and some of the points are located closer to the defined human body keypoints, especially the face, arms, and legs. $M_{Det.+Kpt.+Seg.+Cls.}$ combines the characteristics of $M_{Det.+Seg.}$ and $M_{Det.+Kpt.}$.

A4. More Implementation Details

A4.1. Loss Functions

In this work, we jointly train multiple human-centric perception (HCP) tasks, including human detection, human instance segmentation, human pose estimation and human attribute (gender and age) recognition.

For human detection, we follow DINO [96] to apply focal loss [48] for classification L_{cls}^{focal} and detection loss (L1 regression loss L_{det}^{reg} and GIOU loss [67] L_{det}^{giou}). For human pose estimation, we follow PETR [72] to use focal loss for classifying valid and invalid human instances L_{kpt}^{focal}, L1 keypoint regression loss L_{kpt}^{reg}, OKS loss L_{kpt}^{oks}, and auxiliary heatmap loss L_{kpt}^{hm}. For segmentation loss, we use binary cross-entropy loss L_{seg}^{bce} and dice loss L_{seg}^{dice}. For attribute recognition, we have binary cross-entropy loss for gender estimation L_{gender}^{bce} and mean-variance loss [64] for age estimation L_{age}^{mean} and L_{age}^{var}.

Formally, the overall loss function can be formulated as a linear combination of these sub-task loss functions:

$$
\begin{aligned}
L = & \lambda_{cls}^{focal} L_{cls}^{focal} + \lambda_{det}^{reg} L_{det}^{reg} + \lambda_{det}^{giou} L_{det}^{giou} \\
& + \lambda_{kpt}^{focal} L_{kpt}^{focal} + \lambda_{kpt}^{reg} L_{kpt}^{reg} + \lambda_{kpt}^{oks} L_{kpt}^{oks} + \lambda_{kpt}^{hm} L_{kpt}^{hm} \\
& + \lambda_{seg}^{bce} L_{seg}^{bce} + \lambda_{seg}^{dice} L_{seg}^{dice} \\
& + \lambda_{gender}^{bce} L_{gender}^{bce} + \lambda_{age}^{mean} L_{age}^{mean} + \lambda_{age}^{var} L_{age}^{var},
\end{aligned}
$$

where λs are corresponding loss weights. Detailed settings for the loss weights can be found in Table A5.

Table A5. Loss weights for training our models.

Detection	λ_{cls}^{focal}	1.0
	λ_{det}^{reg}	5.0
	λ_{det}^{giou}	2.0
Keypoint	λ_{kpt}^{focal}	1.0
	λ_{kpt}^{reg}	50.0
	λ_{kpt}^{oks}	1.5
	λ_{kpt}^{hm}	4.0
Segmentation	λ_{seg}^{bce}	8.0
	λ_{seg}^{dice}	5.0
Attribute	λ_{gender}^{bce}	1.0
	λ_{age}^{mean}	0.002
	λ_{age}^{var}	0.01

A4.2. Details about Training

We follow the setting of DINO [96] to augment the input image by random crop, random flip, and random resize. Specifically, we randomly resize the input image to have its shorter side between 480 and 800 pixels and its longer side less or equal to 1333.

The models are trained with AdamW optimizer [35] with base learning rate of 1×10^{-4}, momentum of 0.9 and weight decay of 1×10^{-4}. For all experiments, the models are trained for 100 epochs with a total batch size of 16 and the initial learning rate is decayed at 80th epoch by a factor of 0.1. We use 16 Tesla V100 GPUs for model training.

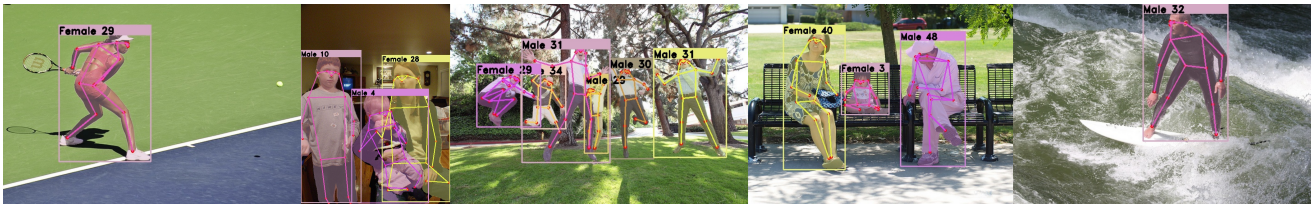

Figure A3. Qualitative results on COCO-UniHuman `val` dataset. Our HQNet achieves accurate human detection, human pose estimation, human instance segmentation, and human attribute recognition simultaneously.

Figure A4. Qualitative results on PoseTrack21 `val` set. Our HQNet makes a comprehensive human analysis system that can achieve multiple functions: multiple object tracking with pose estimation, human instance segmentation, and attribute recognition. Note that our HQNet is only trained on the COCO-UniHuman image-based dataset without finetuning on the PoseTrack21 video-based dataset.

In the experiments, we report results of three different backbones: the ResNet-50 backbone is pre-trained on ImageNet-1K dataset, Swin-L backbone is pre-trained on ImageNet-22K dataset, and ViT-L backbone whose pre-trained weights are from [74]. Unlike DINO and Mask DINO which also pre-train models on Objects365 [105], we only use COCO-UniHuman data for training without Objects365 dataset. For all backbones, we use 4 scales of feature maps feeding to the encoder and an additional high-resolution feature map for mask prediction. In contrast, DINO and MaskDINO use 5 scales for Swin-L models. Following the common practice in DETR-like models [40, 96], we use a 6-layer Transformer encoder and a 6-layer Transformer decoder and 256 as the hidden feature dimension. We use 300 queries and 100 CDN pairs for training. Following [106], we use independent auxiliary heads to refine the multi-task predictions at each decoder layer.

A4.3. Details about Inference

During inference, the input image is resized to have its shorter side being 800 and longer side at most 1333. All reported numbers are obtained without model ensemble or test-time augmentations (*e.g.* flip test and multi-scale test).

A5. Details about Baseline Models

Details about human detection baselines. For human detection, we compare two baseline approaches, *i.e.* Faster-RCNN [66] and DINO [96]. Note that these general object detectors are originally trained to handle general 80 classes (marked with * in Table A3). For fair comparisons, we use MMDetection [9] to re-train and evaluate them on "Person" category using the default experimental setting. Note that MMDetection re-implementation can be a little bit better than the original implementation.

Details about human pose estimation baselines. For human pose estimation, we compare with several represen-

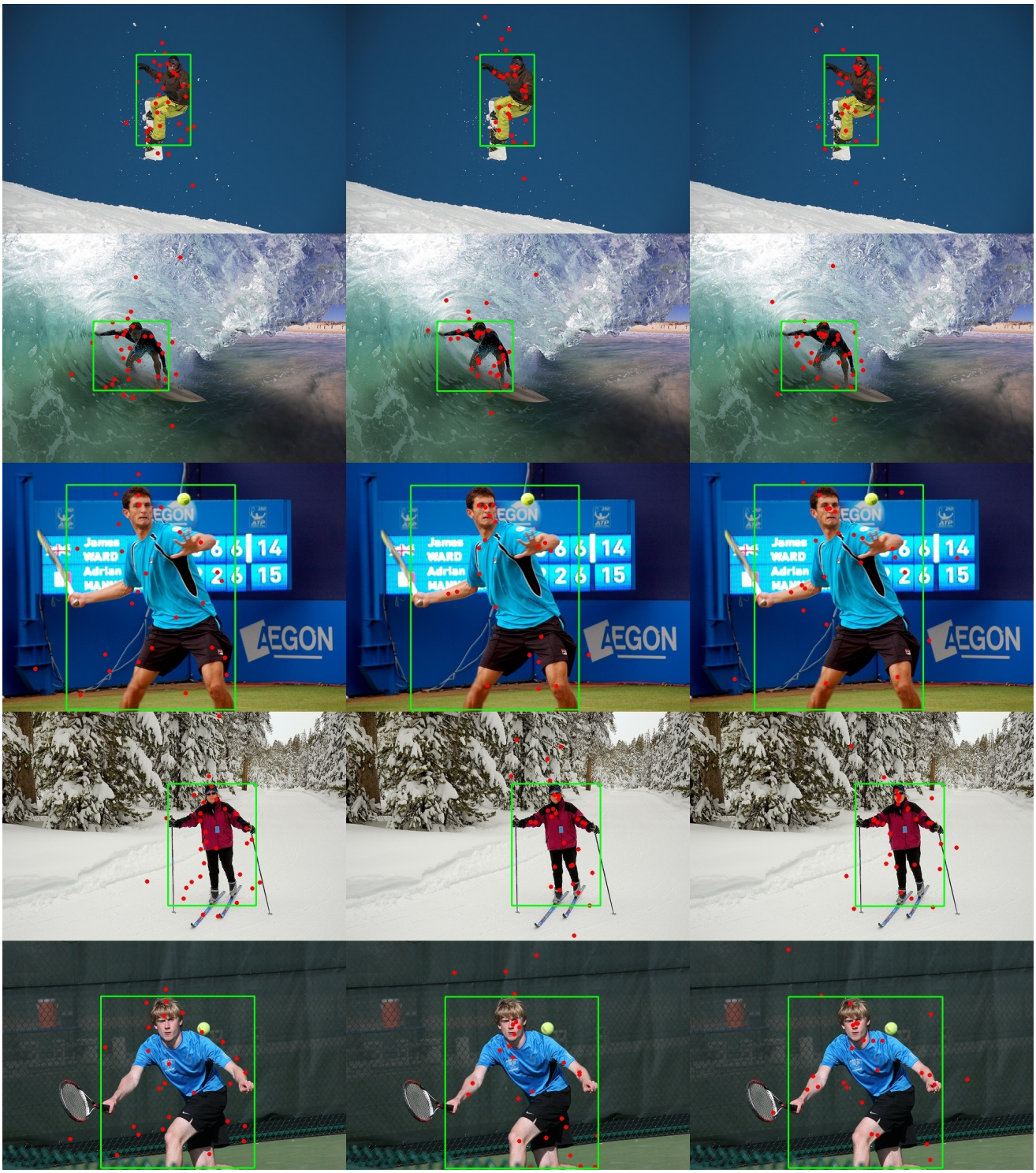

Figure A5. Visualization of deformable attention sampling points. The results are from models trained for different HCP tasks. Left: detection and segmentation. Middle: detection and keypoint. Right: detection, keypoint, segmentation and attribute.

tative top-down methods (SBL [87], HRNet [73], Swin [54] and PRTR [43]), bottom-up approaches (HrHRNet [11], DEKR [20], and SWAHR [56]) and single-stage approaches (FCPose [57], InsPose [71], PETR [72] and CID [79]). Note

that the results of Swin (Swin-L) and CID (R-50-FPN) are from MMPose [13], and other results are from their original papers. *Top-down methods* generally yield superior performance, but often rely on a separate human detector, incurring redundant computational costs. Specifically, SBL, HRNet and Swin use the same person detector provided by [87], which is a strong Faster-RCNN [66] based detector with detection AP 56.4 for the "Person" category on the COCO'2017 `val` set. PRTR applies a DETR-based person detector for human detection, which achieves 50.2 AP for the whole "Person" category on the COCO'2017 `val` set. While PRTR introduces an end-to-end variant (E2E-PRTR) optimizing detection and pose jointly, it lags behind separately trained top-down approaches. For pose estimation, the input resolution for SBL, HRNet, and Swin is set as 256×192, while the input resolution for PRTR is 384×288. *Bottom-up methods* learn instance-agnostic keypoints and then cluster them into corresponding individuals. HrHRNet, DEKR, and SWAHR adopt the strong HRNet-w32 [73] backbone network with an input resolution of 512×512. *Single-stage approaches* directly predict human body keypoints in a single stage. FCPose, InsPose and PETR adopt R-50 [23] backbone network. The input images are resized to have their shorter sides being 800 and their longer sides less or equal to 1333. For CID, we report both the results of R-50-FPN and HRNet-w32 backbones. The input resolution of CID is 512×512.

Details about human instance segmentation baselines. For human instance segmentation, we contrast HumanQuery with state-of-the-art general and human-specific instance segmentation methods. Mask R-CNN [24] is an end-to-end top-down approach that optimizes object detection and instance segmentation jointly. Given our one-stage pipeline, we also compare against one-stage methods , including PolarMask [88], MEInst [98], YOLACT [5], and CondInst [77]. Results of PolarMask, MEInst, YOLACT, CondInst are from [101], which are obtained by re-training and evaluating the models on COCO "Person" category only. PolarMask encodes the instance mask with coordinates, while MEInst encodes the mask into a compact representation vector. YOLACT and CondInst use a series of global prototypes and linear coefficients to represent instance masks. Instead we learn instance-aware Human Query to decouple each human instance.

Details about gender and age estimation baselines. Multi-person gender and age estimation remains underexplored in the literature. We establish baselines using StrongBL [28] and Mask R-CNN [24]. StrongBL is a top-down approach which requires an off-the-shelf human detector. For the detection part, we use a pre-trained Mask RCNN to produce human detection results. And for attribute part, we follow official settings to retrain StrongBL on the COCO-UniHuman dataset. The gender and age models use ResNet50 as the backbone with input resolution 256×192. Mask R-CNN is an end-to-end top-down approach, modified with gender or age branches and retrained using MMDetection with default training settings.

A6. Discussion about Unifying HCP Tasks

A6.1. General network architecture design

There are some attempts to design general network architecture for unifying human-centric perception tasks. Some works propose to design network backbones for HCP tasks. Both CNN-based (*e.g.* HRNet [80]) and Transformer-based backbone networks (*e.g.* TCFormer [95] are proposed for general human-centric visual tasks. Other works focus on designing network heads to unify different HCP tasks. For example, UniHead [45] designs a novel perception head with unified keypoint representations that can be used in different HCP tasks. Point-Set Anchors [83] designs different point-set anchors to provide task-specific initialization for different HCP tasks. Unlike these methods, which employ separate task-specific models for different HCP tasks, we consolidate diverse HCP tasks within a single network.

A6.2. Pre-training on HCP tasks

There are also works [10, 25, 74] on pre-training on diverse human-centric tasks with large-scale data. HCMoCo [25] introduces a versatile multi-modal (RGB-D) pre-training framework for single-person pose estimation and segmentation. SOLIDER [10] presents a self-supervised learning framework to learn a general human representation with more semantic information. HumanBench [74] builds a large-scale human-centric pre-training dataset and introduces the projector-assisted pre-training method with hierarchical weight sharing. More recently, UniHCP [12] presents a unified vision transformer model to perform multitask pre-training at scale. It employs task-specific queries for attending to relevant features, but tackles one task at a time. Unlike ours, our approach simultaneously solves multiple HCP tasks in a single forward pass. Our proposed method is different from this pre-training based approach. First, these methods mainly focus on the pre-training stage, and require fine-tuning for the optimal performance on specific down-stream tasks. Second, these approaches require large-scale joint training on multiple human-centric perception datasets. This makes it unfair to directly compare with models that train on one specific dataset. In addition, large-scale model training is extremely costly. For example, training of UniHCP requires more than 10,000 GPU hours. Third, these methods are designed for single-person human analysis (or top-down human analysis). In comparison, our approach solves multiple HCP tasks in a single-stage multitask manner.

A6.3. Co-learning on HCP tasks

Many works have investigated the correlations between pairs of HCP tasks [50, 61, 62, 75, 97]. For example, [75] explore to integrate fine-grained person attribute learning into the pipeline of pedestrian detection. Mask-RCNN [24] extends Faster-RCNN by adding extra keypoint localization or segmentation branch to handle pose estimation and instance segmentation respectively. Pose2Seg [100] presents a top-down approach for pose-based human instance segmentation. It uses previously generated poses as input instead of the region proposals to extract features for better alignment and performs the down-stream instance segmentation task. PersonLab [65] adopts a bottom-up scheme and solve pose estimation and instance segmentation by applying a greedy decoding process for human grouping. We propose a single-stage model that learns a general unified representation to handle all representative human-centric perception tasks simultaneously.

A7. Discussion about Human Attribute Recognition

Visual recognition of human attributes is an important research topic in computer vision. Among all the human attributes, gender and age are arguably the most popular and representative, which is also our main focus.

A7.1. Dataset

Human attribute recognition datasets can be classified into two categories, *i.e*. facial attribute recognition datasets and pedestrian attribute recognition datasets. Most existing attribute recognition datasets only provide center cropped face (facial attribute recognition) or body (for pedestrian attribute recognition) images, making it not suitable for developing and evaluating multi-person attribute recognition algorithms. In comparison, our proposed COCO-UniHuman preserves the original high resolution image and densely annotates attributes for each human instances. One exception is WIDER-Attr [44], which also provides the original images. However, the number of images is relatively small. We hope our dataset can serve as a good alternative benchmark dataset for multi-person human attribute recognition.

Age estimation datasets can be categorized into three groups, *i.e*. age group classification, real age estimation, and apparent age estimation. To our best knowledge, public large-scale pedestrian attribute datasets (*e.g*. WIDER-Attr [44], PETA [14], Market1501-Attr [49, 104], RAP-2.0 [39] and PA-100K [52]) only have coarse age group annotations. Facial attribute datasets may also have fine-grained apparent (*e.g*. APPA-REAL [2]) or real (*e.g*. MegaAge [102]) age annotations. Apparent age estimation focuses on how old a subject "looks like", instead of how old a subject "really is". It is considered to be a more practical setting for visual analysis. Our proposed dataset is the first large scale in-the-wild dataset for body-based apparent age estimation. Body-based apparent age estimation is promising especially when the facial image is not captured clear enough (*e.g*. captured in a distance). However, body-based apparent age estimation is under-explored in literature due to lack of dataset. We hope our presented COCO-UniHuman dataset can promote related research.

A7.2. Method

Human attribute recognition focuses on assigning a set of semantic attributes (*e.g*. gender and age) to each human instance. Typical approaches include global image based [1, 28], local parts based [44], and visual attention based [52] approaches. Most of them focus on single-human (or top-down) analysis without consider the relationship among different human instances. In comparison, we introduce a single-stage multi-person human attribute (*i.e*. gender and age) recognition approach.