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] ✉, Chen Qian[2], and Ping Luo[1,3] ✉

[1] The University of Hong Kong [2] SenseTime Research and Tetras.AI
[3] Shanghai AI Laboratory
js20@connect.hku.hk, lishuhuai@sensetime.com

Abstract. Human-centric perception (*e.g.* 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 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 are available at `https://github.com/lishuhuai527/COCO-UniHuman`.

Keywords: Human-Centric Perception · Unified Vision Model

1 Introduction

Human-centric perception (*e.g.* pedestrian detection, 2D keypoint estimation, 3D mesh recovery, human segmentation and 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 [11] resorted to training models on various datasets for each HCP task, which can introduce certain limitations. Firstly, there is inherent scale variance across different

⋆ Equal contribution. ✉ Corresponding authors.

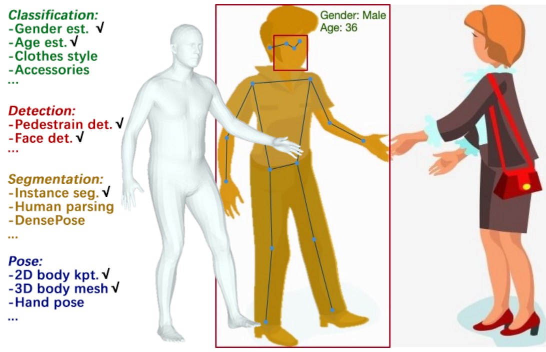

Fig. 1: Multi-person human-centric perception tasks can be categorized into 4 groups: classification, detection, segmentation and pose estimation.

datasets. For example, human detection datasets [43] consist of scene images with multiple interacting people, while attribute recognition datasets [46] 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 [23] are captured in controlled lab environments, while some [46] 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. As shown in Figure 1, most popular HCP tasks can be grouped into four fundamental categories: classification, detection, segmentation, and pose estimation. 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 7 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 keypoint 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, 2D human keypoint estimation, 3D human mesh recovery, 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, 35, 45, 82, 93], 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. In addition, we design HumanQuery-Instance Matching (HQ-Ins Matching) and Gender-aided human Model Selection (GaMS) mechanisms to further exploit the interactions among different HCP tasks and enhance the performance of multi-task HCP.

We highlight several noteworthy characteristics of HQNet. (1) **Flexibility:** HQNet can readily integrate with diverse backbone networks, such as ResNet [20], Swin [47] and ViT [15]. (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.

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 pose estimation (2D keypoint and 3D mesh recovery). (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. Additionally, we design HumanQuery-Instance (HQ-Ins) Matching and Gender-aided human Model Selection (GaMS) mechanisms to improve the performance of multi-task HCP. (3) 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.

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	BodyBox	FaceBox	BodyKpt	BodyMask	Gender	Age	Mesh
Caltech [14]	250K	350K	2.3K	✗	✓	✗	✗	✗	✗	✗	✗
CityPersons [85]	5K	32K	32K	✗	✓	✗	✗	✗	✗	✗	✗
CrowdHuman [59]	24K	552K	552K	✗	✓	*	✗	✗	✗	✗	✗
MPII [5]	25K	40K	-	✗	✓	*	✓	✗	✗	✗	✗
PoseTrack [4]	23K	153K	-	✗	✓	*	✓	✗	✗	✗	✗
CIHP [18]	38K	129K	129K	✗	✓	✗	✗	✓	✗	✗	✗
MHP [37]	5K	15K	15K	✗	✓	✗	✗	✓	✗	✗	✗
CelebA [48]	200K	200K	10K	face	✗	✗	✗	✗	✓	group:4	✗
APPA-REAL [2]	7.5K	7.5K	7.5K	face	✗	✗	✗	✗	✓	appa & real	✗
MegaAge [90]	40K	40K	40K	face	✗	✗	✗	✗	✓	real	✗
WIDER-Attr [39]	13K	57K	57K	✗	✓	✗	✗	✗	✓	group:6	✗
PETA [12]	19K	19K	8.7K	body	✗	✗	✗	✗	✓	group:4	✗
PA-100K [46]	100K	100K	-	body	✗	✗	✗	✗	✓	group:3	✗
OCHuman [86]	5K	13K	13K	✗	✓	✗	✓	✓	✗	✗	✗
COCO [43]	200K	273K	273K	✗	✓	✗	✓	✓	✗	✗	✗
COCO-WholeBody [28]	200K	273K	273K	✗	✓	✓	✓	✗	✗	✗	✗
COCO-UniHuman	200K	273K	273K	✗	✓	✓	✓	✓	✓	appa	✓

2 Related Works

2.1 Human-Centric Perception Tasks and Datasets

Approaches to multi-person human-centric perception (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 [25, 32, 42, 62, 69, 75, 81], and those jointly learning detection and perception modules [3, 19]. **Bottom-up methods** learn instance-agnostic keypoints/masks and cluster them using integer linear programming [22, 26, 33], heuristic greedy parsing [7, 56], embedding clustering [34, 51], or learnable clustering [27]. **Single-stage methods** directly predict keypoints or masks for each individual, with different representations for 2D keypoint estimation (coordinate-based [53, 66, 71, 78], heatmap-based [61, 68], or hybrid [17, 50, 92]), 3D mesh recovery [41, 63] and segmentation (contour-based [74] or mask-based [6]). While existing approaches focus on individual HCP tasks, we aim to unify HCP by learning a single model that handles multiple tasks simultaneously, enabling a comprehensive understanding of humans. As shown in Table 1, there are task-specific datasets separately annotated for different HCP tasks, including pedestrian detection [14, 59, 85, 88], keypoint estimation [4, 5], segmentation [18, 37], and attribute recognition [46, 90]. Datasets for multiple HCP tasks also exist. COCO [43] offers thorough annotations: body box, keypoints, and segmentation mask. COCO-WholeBody [28, 76] provides dense annotations of face/hand boxes and 133 whole-body keypoints. Our COCO-UniHuman dataset further extends COCO-WholeBody featuring extensive gender, age and mesh annotations.

2.2 Unified Methods for HCP

General network architecture for different HCP tasks. Some works design general network backbones, including CNN-based [69] and Transformer-based backbones [80]. Others unify HCP tasks with novel perception heads, such as UniHead [40] and UniFS [29]. 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 also works [9,21,64] on pre-training on diverse human-centric tasks with large-scale data. More recently, UniHCP [11] 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 [44, 52, 54, 65, 83]. We propose a single-stage model that learns a general unified representation to handle all representative human-centric perception tasks simultaneously.

2.3 Object-Centric Representation Learning

DETR [8] pioneers learnable object queries to represent objects and interact with image features. Deformable DETR [93] introduces deformable attention modules to focus on key sampling points, enhancing convergence speed. DAB-DETR [45] treats each positional query as a dynamic 4D anchor box, updated across decoder layers. DN-DETR [35] employs denoising training for faster convergence. Recently, DINO [82] 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 v1 dataset is the first large-scale dataset, which provides annotations for all four representative HCP tasks in multi-person scenarios. Building upon COCO [30, 43] dataset, we have enriched the annotations by including gender, age, 3D body mesh information for each individual.

3.1 Data Annotation

Human Attribute 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 [39], 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 Supplementary). **Mesh Annotation.** We follow [30] to apply Exemplar Fine-Tuning (EFT) method with the gender neutral model to generate 3D pseudo-ground truth SMPL parameters. To ensure data quality, we only retain instances with at least 12 keypoint annotations where all limbs are visible and filter out low-quality data manually.

3.2 Data 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, keypoint localization, and 3D body mesh recovery in multi-person scenarios. 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

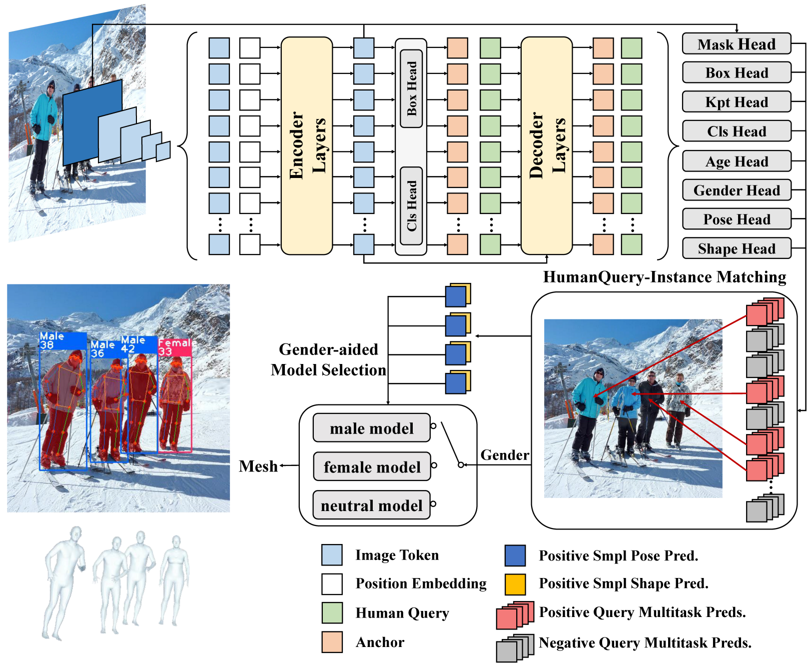

Fig. 2: Overview of HQNet. HQNet unifies various representative HCP tasks in a single network by learning shared Human Query.

in situations where the facial image lacks clarity (*e.g.* captured from a distance). Notably, existing large-scale pedestrian attribute datasets [12] 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 human attribute labels and 3D human mesh information provide additional descriptive information about individuals beyond the existing labels. By leveraging these information, models can learn improved representations of humans, consequently enhancing the performance of other HCP tasks.

4 Method

4.1 Overview

This study endeavors to develop a single-stage framework that 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.

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 [36]), 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 [20], 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 [82], 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 [93] 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.

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. To enhance training stability and acceleration, we employ Contrastive DeNoising (CDN) as introduced in DINO [82]. 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, *i.e.* classification (Cls.), detection (Det.), pose (Pose.), and segmentation (Seg.), we employ HumanQuery-Instance (HQ-Ins) Matching. $\lambda_{cls}L_{cls} + \lambda_{det}L_{det} + \lambda_{seg}L_{seg} + \lambda_{pose}L_{pose}$, where λ are loss weights. Details can be found in Supplementary.

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 keypoint 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 [36], 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. **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). **2D keypoint 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 [60], 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 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 [58] estimation. **3D mesh recovery head.** Two 3-layer MLPs with the same hidden dimension are used to predict the pose and shape parameters respectively. These parameters are then fed into a SMPL body model to generate the 3D body meshes.

Gender-aided human Model Selection (GaMS). There are three versions of SMPL models: male, female and neutral. Previous works usually use the neutral model because of the lack of gender annotation. COCO-UniHuman has both gender and 3D mesh annotations. To improve the performance of 3D mesh, we employ Gender-aided human Model Selection (GaMS) which selects different SMPL models by gender labels during the training and the inference stage.

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 [82] for augmentation and adopt the 100-epoch training schedule. Model evaluation

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.			Pose (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 [57]	R-50	65.3	61.5	71.2	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗
IterDETR [91]	R-50	71.8	66.0	78.9	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗
DINO [82]	R-50	73.3	68.1	79.9	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗
♢ Mask R-CNN [19]	R-50-FPN	66.7	62.3	73.1	58.4	51.8	66.2	✗	✗	✗	✗	✗	✗	✗	✗	✗
★ PolarMask [74]	R-50-FPN	✗	✗	✗	45.1	38.5	57.1	✗	✗	✗	✗	✗	✗	✗	✗	✗
★ YOLACT [6]	R-50-FPN	✗	✗	✗	47.4	40.1	61.2	✗	✗	✗	✗	✗	✗	✗	✗	✗
★ MEInst [84]	R-50-FPN	✗	✗	✗	49.3	42.3	57.6	✗	✗	✗	✗	✗	✗	✗	✗	✗
★ CondInst [67]	R-50-FPN	✗	✗	✗	54.8	43.3	69.0	✗	✗	✗	✗	✗	✗	✗	✗	✗
★ Mask DINO [36]	R-50	72.3	66.5	79.5	64.8	57.3	73.4	✗	✗	✗	✗	✗	✗	✗	✗	✗
♢ SBL† [73]	R-50	✗	✗	✗	✗	✗	✗	70.4	67.1	77.2	✗	✗	✗	✗	✗	✗
♢ Swin† [47]	Swin-L	✗	✗	✗	✗	✗	✗	74.3	70.6	81.2	✗	✗	✗	✗	✗	✗
♢ HRNet† [62]	HRNet-32	✗	✗	✗	✗	✗	✗	74.4	70.8	81.0	✗	✗	✗	✗	✗	✗
♢ ViTPose† [77]	ViT-L	✗	✗	✗	✗	✗	✗	78.2	74.5	85.4	✗	✗	✗	✗	✗	✗
♢ PRTR† [38]	R-50	✗	✗	✗	✗	✗	✗	68.2	63.2	76.2	✗	✗	✗	✗	✗	✗
♡ HrHRNet† [10]	HRNet-w32	✗	✗	✗	✗	✗	✗	67.1	61.5	76.1	✗	✗	✗	✗	✗	✗
♡ DEKR† [17]	HRNet-w32	✗	✗	✗	✗	✗	✗	68.0	62.1	77.7	✗	✗	✗	✗	✗	✗
♡ SWAHR† [49]	HRNet-w32	✗	✗	✗	✗	✗	✗	68.9	63.0	77.4	✗	✗	✗	✗	✗	✗
★ CID† [68]	R-50-FPN	✗	✗	✗	✗	✗	✗	52.0	48.6	58.0	✗	✗	✗	✗	✗	✗
★ CID† [68]	HRNet-w32	✗	✗	✗	✗	✗	✗	69.8	64.0	78.9	✗	✗	✗	✗	✗	✗
★ FCPose [50]	R-50	✗	✗	✗	✗	✗	✗	63.0	59.1	70.3	✗	✗	✗	✗	✗	✗
★ InsPose [61]	R-50	✗	✗	✗	✗	✗	✗	65.2	60.6	72.2	✗	✗	✗	✗	✗	✗
★ PETR [60]	R-50	✗	✗	✗	✗	✗	✗	68.8	62.7	77.7	✗	✗	✗	✗	✗	✗
♢ StrongBL [24]	R-50	-	-	-	✗	✗	✗	✗	✗	✗	46.4	35.2	53.2	✗	✗	✗
♢ Mask R-CNN [19]	R-50	66.3	61.9	72.8	✗	✗	✗	✗	✗	✗	46.7	36.3	52.8	✗	✗	✗
♢ StrongBL [24]	R-50	-	-	-	✗	✗	✗	✗	✗	✗	✗	✗	✗	42.3	31.9	48.3
♢ Mask R-CNN [19]	R-50	66.3	62.1	72.5	✗	✗	✗	✗	✗	✗	✗	✗	✗	37.4	27.9	43.3
♢ Pose2Seg [86]	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 [56]	R-152	✗	✗	✗	-	48.3	59.5	66.5	62.3	73.2	✗	✗	✗	✗	✗	✗
★ CenterNet [92]	Hourglass	-	-	-	✗	✗	✗	64.0	59.4	72.1	✗	✗	✗	✗	✗	✗
★ LSNet-5 [87]	DLA-34	✗	✗	✗	56.2	44.2	71.0	-	-	-	✗	✗	✗	✗	✗	✗
★ UniHead* [40]	R-50-FPN	67.3	62.6	74.4	38.6	37.2	42.2	57.5	55.3	61.9	✗	✗	✗	✗	✗	✗
★ HQNet (D+S)	R-50	73.0	68.0	79.4	63.6	57.6	72.1	✗	✗	✗	✗	✗	✗	✗	✗	✗
★ HQNet (D+S+P)	R-50	74.5	70.3	80.1	65.7	58.7	73.8	69.5	64.4	77.0	✗	✗	✗	✗	✗	✗
★ HQNet (D+S+P+C)	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	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	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

takes place on COCO-UniHuman `val` set (2693 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 [86, 87], 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. For human mesh recovery, we evaluate pose accuracy using MPJPE (Mean Per Joint Position Error) w.r.t. root relative poses and PA-MPJPE (Procrustes-Aligned MPJPE), which is MPJPE calculated after rigid alignment of predicted pose with the ground truth.

Table 3: Results of 3D human mesh recovery on the COCO-UniHuman `val` set. HQNets are jointly trained with all HCP tasks (D+S+P+C). ↓ means lower is better.

Model	Backbone	Bbox	Pose (Mesh.) MPJPE ↓	Pose (Mesh.) PA-MPJPE ↓
◇ HMR [32]	R-50	GT	109.62	72.03
◇ HMR+ [55]	R-50	GT	78.06	50.36
★ ROMP [63]	R-50	-	119.52	72.27
★ HQNet w/o GaSM	R-50	-	87.00	54.92
★ HQNet	R-50	-	84.74	50.80
★ HQNet	ViT-L	-	76.31	48.26

OCHuman Dataset [86] is a large benchmark that focuses on heavily occluded humans. It contains no training samples and is intended solely for evaluation purposes. Following [86], we train models on the COCO `train` set and evaluate models on OCHuman `val` set (4731 images) and `test` set (8110 images).

5.2 Results on COCO-UniHuman Dataset

We compare our method to task-specific and multi-task HCP models on the COCO-UniHuman dataset in Table 2 and Table 3. Our models outperform multi-task HCP models and achieves very competitive results against task-specific HCP models. Details about the baselines can be found in Supplementary. "D", "S", "P", "C" mean model training with Detection (Det.), Segmentation (Seg.), Pose and Classification (Cls.) task respectively.

Comparison with task-specific HCP models. For human detection, we compare three baseline approaches, *i.e.* Faster-RCNN [57], IterDETR [91] and DINO [82]. For human instance segmentation, we contrast HQNet with state-of-the-art general and human-specific instance segmentation methods, including Mask R-CNN [19], PolarMask [74], MEInst [84], YOLACT [6], and CondInst [67]. For human pose estimation, we compare with several representative top-down methods (SBL [73], HRNet [62], Swin [47], ViTPose [77] and PRTR [38]), bottom-up approaches (HrHRNet [10], DEKR [17], and SWAHR [49]) and single-stage approaches (FCPose [50], InsPose [61], PETR [60] and CID [68]). For gender and age estimation, we establish baselines using StrongBL [24] and Mask R-CNN [19]. For mesh, we compare with HMR [32], HMR+ [55] and ROMP [63]. 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 achieve SOTA among single-stage approaches.

Comparison with multi-task HCP methods. Pose2Seg [86] 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 [56] follow bottom-up strategies. CenterNet [92], LSNet [87], and UniHead [40][1]

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

Table 4: Comparison with state-of-the-art models on the OCHuman dataset. † denotes flip testing. We compare with ◇ top-down, ♡ bottom-up, ★ one-stage approaches.

		OCHuman Val			**OCHuman Test**		
Model	Backbone	Det.	Seg.	Pose (Kpt.)	Det.	Seg.	Pose (Kpt.)
◇ Mask R-CNN [19]	R-50-FPN	-	16.3	✗	-	16.9	✗
◇ SBL† [73]	R-50	✗	✗	37.8	✗	✗	30.4
◇ Pose2Seg [86]	R-50-FPN	✗	22.2	28.5	✗	23.8	30.3
♡ AE† [51]	Hourglass	✗	✗	32.1	✗	✗	29.5
♡ HGG† [27]	Hourglass	✗	✗	35.6	✗	✗	34.8
♡ DEKR† [17]	HRNet-w32	✗	✗	37.9	✗	✗	36.5
♡ HrHRNet† [10]	HRNet-w32	✗	✗	40.0	✗	✗	39.4
★ YOLACT [6]	R-101-FPN	✗	13.2	✗	✗	13.5	✗
★ CondInst [67]	R-50-FPN	✗	20.3	✗	✗	20.1	✗
★ LSNet-5 [87]	DLA-34	✗	25.0	✗	✗	24.9	✗
★ LOGO-CAP† [78]	HRNet-w32	✗	✗	39.0	✗	✗	38.1
★ CID† [68]	R-50-FPN	✗	✗	29.2	✗	✗	28.3
★ CID† [68]	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

are single-stage alternatives. Our R-50 model achieves superior performance in multi-task HCP, without bells and whistles.

Effect of multi-task co-learning. In Table 2, we also compare with different variants of HQNet for various task composition (*i.e.* D, S, P, C). 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.

5.3 Results on the OCHuman Dataset

To verify the performance of HQNet in challenging crowded scenarios, we compare it with recent works on OCHuman dataset [86], which is a crowded scene benchmark for human detection, segmentation, and pose estimation in Table 4. 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 5, we compare our approach with Faster R-CNN [57] and ZoomNet [28]. Our HQNet can not only better exploit the inherent multi-level structure of the human body, but also preserve the efficiency of single-stage

Table 5: Finetuning evaluation on novel face detection tasks. Face detection results are reported on COCO-UniHuman `val` dataset.

Method	Face detection	
	AP	AR
Faster RCNN [57]	43.9	71.2
ZoomNet [28]	58.2	72.8
HQNet (R-50)	**68.4**	**83.2**

Table 6: Unseen-task evaluation on PoseTrack21 [13]. 'FT' means fine-tuning on PoseTrack21. Our models are evaluated without training on MOT.

Method	FT	IDF1	MOTA
TRMOT [70]	✓	57.3	47.2
FairMOT [89]	✓	63.2	56.3
HQNet (D)	✗	62.4	48.6
HQNet (D+S)	✗	63.3	49.5
HQNet (R-50)	✗	64.6	51.1
HQNet (ViT-L)	✗	**69.1**	**57.0**

Table 7: Robustness to domain shift. All models are evaluated on Human-Art [31] `val` set without training on Human-Art.

Method	Det.	Kpt.
Faster R-CNN [57] + HRNet [62]	12.0	22.2
YOLOX [16] + ViTPose [77]	14.4	28.7
HigherHRNet [10]	-	34.6
ED-Pose [79]	-	37.5
HQNet (Swin-L)	15.8	43.0
HQNet (ViT-L)	**18.7**	**52.2**

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 generalization. We evaluate the generalization ability of our approach through an unseen task evaluation, specifically multiple object tracking (MOT) on the PoseTrack21 dataset [13]. 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 [72] and used the learned Human Query as re-identification features for association. In Table 6, 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 (D) and HQNet (D+S) refers to HQNet trained solely on the detection (D) and segmentation (S) tasks respectively, 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.

5.5 Robustness to Domain Shift

HumanArt [31] 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 7, 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.

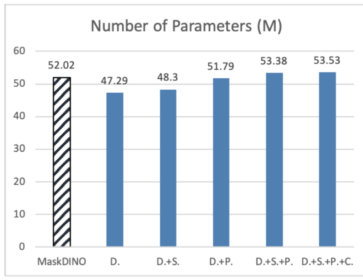

Fig. 3: Computation cost analysis validates the efficiency of HQNet.

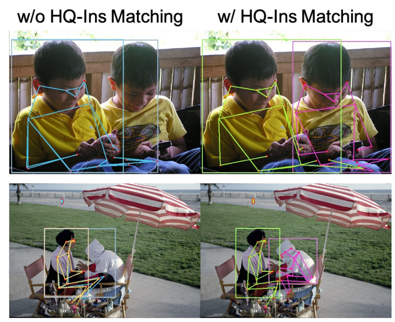

Fig. 4: Effect of HumanQuery-Instance (HQ-Ins) Matching.

5.6 More Analysis

Computation cost analysis. In Fig 3, we report the number of parameters of our Res50-based HQNet model variants of different task composition. 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. It is noteworthy that our model is efficient and its cost is comparable to the task-specific HCP models (*e.g.* MaskDINO [36]).

Effect of Gender-aided human Model Selection (GaMS) In Table 3, we analyze the effect of Gender-aided human Model Selection (GaMS). We find that incorporating the obtained gender information can assist in selecting proper 3D model of the human body, resulting in more accurate human mesh recovery.

Effect of HQ-Ins Matching. "w/o HQ-Ins Matching" means using detection loss only for bipartite matching [82]. "w/ HQ-Ins Matching" means comprehensively using detection, pose, and segmentation loss for bipartite matching. As shown in Fig. 4, with detection only matching [82], 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. More quantitative evaluation can be found in Supplementary.

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 that encodes 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. 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.

Acknowledgement. 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 and 17209324.

References

1. Abdulnabi, A.H., Wang, G., Lu, J., Jia, K.: Multi-task cnn model for attribute prediction. IEEE Trans. Multimedia **17**(11), 1949–1959 (2015)
2. Agustsson, E., Timofte, R., Escalera, S., Baro, X., Guyon, I., Rothe, R.: Apparent and real age estimation in still images with deep residual regressors on appa-real database. In: IEEE Int. Conf. Auto. Face & Gesture Recog. pp. 87–94 (2017)
3. Alp Güler, R., Neverova, N., Kokkinos, I.: Densepose: Dense human pose estimation in the wild. In: IEEE Conf. Comput. Vis. Pattern Recog. (2018)
4. Andriluka, M., Iqbal, U., Insafutdinov, E., Pishchulin, L., Milan, A., Gall, J., Schiele, B.: Posetrack: A benchmark for human pose estimation and tracking. In: IEEE Conf. Comput. Vis. Pattern Recog. (2018)
5. Andriluka, M., Pishchulin, L., Gehler, P., Schiele, B.: 2d human pose estimation: New benchmark and state of the art analysis. In: IEEE Conf. Comput. Vis. Pattern Recog. (2014)
6. Bolya, D., Zhou, C., Xiao, F., Lee, Y.J.: Yolact: Real-time instance segmentation. In: Int. Conf. Comput. Vis. pp. 9157–9166 (2019)
7. Cao, Z., Simon, T., Wei, S.E., Sheikh, Y.: Realtime multi-person 2d pose estimation using part affinity fields. In: IEEE Conf. Comput. Vis. Pattern Recog. (2017)
8. Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., Zagoruyko, S.: End-to-end object detection with transformers. In: Eur. Conf. Comput. Vis. pp. 213–229 (2020)
9. Chen, W., Xu, X., Jia, J., Luo, H., Wang, Y., Wang, F., Jin, R., Sun, X.: Beyond appearance: a semantic controllable self-supervised learning framework for human-centric visual tasks. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 15050–15061 (2023)
10. Cheng, B., Xiao, B., Wang, J., Shi, H., Huang, T.S., Zhang, L.: Higherhrnet: Scale-aware representation learning for bottom-up human pose estimation. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 5386–5395 (2020)
11. Ci, Y., Wang, Y., Chen, M., Tang, S., Bai, L., Zhu, F., Zhao, R., Yu, F., Qi, D., Ouyang, W.: Unihcp: A unified model for human-centric perceptions. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 17840–17852 (2023)
12. Deng, Y., Luo, P., Loy, C.C., Tang, X.: Pedestrian attribute recognition at far distance. In: ACM Int. Conf. Multimedia. pp. 789–792 (2014)
13. Doering, A., Chen, D., Zhang, S., Schiele, B., Gall, J.: Posetrack21: A dataset for person search, multi-object tracking and multi-person pose tracking. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 20963–20972 (2022)
14. Dollár, P., Wojek, C., Schiele, B., Perona, P.: Pedestrian detection: A benchmark. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 304–311 (2009)
15. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. Int. Conf. Learn. Represent. (2021)
16. Ge, Z., Liu, S., Wang, F., Li, Z., Sun, J.: Yolox: Exceeding yolo series in 2021. arXiv preprint arXiv:2107.08430 (2021)

17. Geng, Z., Sun, K., Xiao, B., Zhang, Z., Wang, J.: Bottom-up human pose estimation via disentangled keypoint regression. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 14676–14686 (2021)
18. Gong, K., Liang, X., Li, Y., Chen, Y., Yang, M., Lin, L.: Instance-level human parsing via part grouping network. In: Eur. Conf. Comput. Vis. pp. 770–785 (2018)
19. He, K., Gkioxari, G., Dollár, P., Girshick, R.: Mask r-cnn. In: Int. Conf. Comput. Vis. pp. 2961–2969 (2017)
20. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: IEEE Conf. Comput. Vis. Pattern Recog. (2016)
21. Hong, F., Pan, L., Cai, Z., Liu, Z.: Versatile multi-modal pre-training for human-centric perception. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 16156–16166 (2022)
22. Insafutdinov, E., Pishchulin, L., Andres, B., Andriluka, M., Schiele, B.: Deepercut: A deeper, stronger, and faster multi-person pose estimation model. In: Eur. Conf. Comput. Vis. pp. 34–50 (2016)
23. Ionescu, C., Papava, D., Olaru, V., Sminchisescu, C.: 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)
24. Jia, J., Huang, H., Yang, W., Chen, X., Huang, K.: Rethinking of pedestrian attribute recognition: Realistic datasets with efficient method. arXiv preprint arXiv:2005.11909 (2020)
25. Jiang, W., Jin, S., Liu, W., Qian, C., Luo, P., Liu, S.: Posetrans: A simple yet effective pose transformation augmentation for human pose estimation. In: Eur. Conf. Comput. Vis. pp. 643–659 (2022)
26. Jin, S., Liu, W., Ouyang, W., Qian, C.: Multi-person articulated tracking with spatial and temporal embeddings. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 5664–5673 (2019)
27. Jin, S., Liu, W., Xie, E., Wang, W., Qian, C., Ouyang, W., Luo, P.: Differentiable hierarchical graph grouping for multi-person pose estimation. In: Eur. Conf. Comput. Vis. pp. 718–734 (2020)
28. Jin, S., Xu, L., Xu, J., Wang, C., Liu, W., Qian, C., Ouyang, W., Luo, P.: Whole-body human pose estimation in the wild. In: Eur. Conf. Comput. Vis. (2020)
29. Jin, S., Yao, R., Xu, L., Liu, W., Qian, C., Wu, J., Luo, P.: Unifs: Universal few-shot instance perception with point representations. In: Eur. Conf. Comput. Vis. (2024)
30. Joo, H., Neverova, N., Vedaldi, A.: Exemplar fine-tuning for 3d human model fitting towards in-the-wild 3d human pose estimation. In: Int. Conf. 3D Vis. pp. 42–52. IEEE (2021)
31. Ju, X., Zeng, A., Wang, J., Xu, Q., Zhang, L.: Human-art: A versatile human-centric dataset bridging natural and artificial scenes. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 618–629 (2023)
32. Kanazawa, A., Black, M.J., Jacobs, D.W., Malik, J.: End-to-end recovery of human shape and pose. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 7122–7131 (2018)
33. Kirillov, A., Levinkov, E., Andres, B., Savchynskyy, B., Rother, C.: Instancecut: from edges to instances with multicut. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 5008–5017 (2017)
34. Kong, S., Fowlkes, C.C.: Recurrent pixel embedding for instance grouping. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 9018–9028 (2018)
35. Li, F., Zhang, H., Liu, S., Guo, J., Ni, L.M., Zhang, L.: Dn-detr: Accelerate detr training by introducing query denoising. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 13619–13627 (2022)

36. Li, F., Zhang, H., Xu, H., Liu, S., Zhang, L., Ni, L.M., Shum, H.Y.: Mask dino: Towards a unified transformer-based framework for object detection and segmentation. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 3041–3050 (2023)
37. Li, J., Zhao, J., Wei, Y., Lang, C., Li, Y., Sim, T., Yan, S., Feng, J.: Multiple-human parsing in the wild. arXiv preprint arXiv:1705.07206 (2017)
38. Li, K., Wang, S., Zhang, X., Xu, Y., Xu, W., Tu, Z.: Pose recognition with cascade transformers. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 1944–1953 (2021)
39. Li, Y., Huang, C., Loy, C.C., Tang, X.: Human attribute recognition by deep hierarchical contexts. In: Eur. Conf. Comput. Vis. (2016)
40. Liang, J., Song, G., Leng, B., Liu, Y.: Unifying visual perception by dispersible points learning. In: Eur. Conf. Comput. Vis. pp. 439–456 (2022)
41. Lin, J., Zeng, A., Wang, H., Zhang, L., Li, Y.: One-stage 3d whole-body mesh recovery with component aware transformer. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 21159–21168 (2023)
42. Lin, K., Wang, L., Liu, Z.: End-to-end human pose and mesh reconstruction with transformers. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 1954–1963 (2021)
43. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: Eur. Conf. Comput. Vis. (2014)
44. Lin, Y., Shen, J., Wang, Y., Pantic, M.: Fp-age: Leveraging face parsing attention for facial age estimation in the wild. IEEE Trans. Image Process. (2022)
45. Liu, S., Li, F., Zhang, H., Yang, X., Qi, X., Su, H., Zhu, J., Zhang, L.: Dab-detr: Dynamic anchor boxes are better queries for detr. Int. Conf. Learn. Represent. (2022)
46. Liu, X., Zhao, H., Tian, M., Sheng, L., Shao, J., Yan, J., Wang, X.: Hydraplus-net: Attentive deep features for pedestrian analysis. In: Int. Conf. Comput. Vis. pp. 1–9 (2017)
47. Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B.: Swin transformer: Hierarchical vision transformer using shifted windows. In: Int. Conf. Comput. Vis. pp. 10012–10022 (2021)
48. Liu, Z., Luo, P., Wang, X., Tang, X.: Deep learning face attributes in the wild. In: Int. Conf. Comput. Vis. (2015)
49. Luo, Z., Wang, Z., Huang, Y., Wang, L., Tan, T., Zhou, E.: Rethinking the heatmap regression for bottom-up human pose estimation. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 13264–13273 (2021)
50. Mao, W., Tian, Z., Wang, X., Shen, C.: Fcpose: Fully convolutional multi-person pose estimation with dynamic instance-aware convolutions. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 9034–9043 (2021)
51. Newell, A., Huang, Z., Deng, J.: Associative embedding: End-to-end learning for joint detection and grouping. In: Adv. Neural Inform. Process. Syst. (2017)
52. Nie, X., Feng, J., Yan, S.: Mutual learning to adapt for joint human parsing and pose estimation. In: Eur. Conf. Comput. Vis. pp. 502–517 (2018)
53. Nie, X., Feng, J., Zhang, J., Yan, S.: Single-stage multi-person pose machines. In: Int. Conf. Comput. Vis. pp. 6951–6960 (2019)
54. Nie, X., Feng, J., Zuo, Y., Yan, S.: Human pose estimation with parsing induced learner. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 2100–2108 (2018)
55. Pang, H.E., Cai, Z., Yang, L., Zhang, T., Liu, Z.: Benchmarking and analyzing 3d human pose and shape estimation beyond algorithms. Advances in Neural Information Processing Systems **35**, 26034–26051 (2022)

56. Papandreou, G., Zhu, T., Chen, L.C., Gidaris, S., Tompson, J., Murphy, K.: Personlab: Person pose estimation and instance segmentation with a bottom-up, part-based, geometric embedding model. In: Eur. Conf. Comput. Vis. pp. 269–286 (2018)
57. Ren, S., He, K., Girshick, R., Sun, J.: Faster R-CNN: Towards real-time object detection with region proposal networks. In: Adv. Neural Inform. Process. Syst. (2015)
58. Rothe, R., Timofte, R., Van Gool, L.: Dex: Deep expectation of apparent age from a single image. In: Int. Conf. Comput. Vis. Worksh. pp. 10–15 (2015)
59. Shao, S., Zhao, Z., Li, B., Xiao, T., Yu, G., Zhang, X., Sun, J.: Crowdhuman: A benchmark for detecting human in a crowd. arXiv preprint arXiv:1805.00123 (2018)
60. Shi, D., Wei, X., Li, L., Ren, Y., Tan, W.: End-to-end multi-person pose estimation with transformers. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 11069–11078 (2022)
61. Shi, D., Wei, X., Yu, X., Tan, W., Ren, Y., Pu, S.: Inspose: instance-aware networks for single-stage multi-person pose estimation. In: ACM Int. Conf. Multimedia. pp. 3079–3087 (2021)
62. Sun, K., Xiao, B., Liu, D., Wang, J.: Deep high-resolution representation learning for human pose estimation. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 5693–5703 (2019)
63. Sun, Y., Bao, Q., Liu, W., Fu, Y., Black, M.J., Mei, T.: Monocular, one-stage, regression of multiple 3d people. In: Int. Conf. Comput. Vis. pp. 11179–11188 (2021)
64. Tang, S., Chen, C., Xie, Q., Chen, M., Wang, Y., Ci, Y., Bai, L., Zhu, F., Yang, H., Yi, L., et al.: Humanbench: Towards general human-centric perception with projector assisted pretraining. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 21970–21982 (2023)
65. Tian, Y., Luo, P., Wang, X., Tang, X.: Pedestrian detection aided by deep learning semantic tasks. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 5079–5087 (2015)
66. Tian, Z., Chen, H., Shen, C.: Directpose: Direct end-to-end multi-person pose estimation. arXiv preprint arXiv:1911.07451 (2019)
67. Tian, Z., Shen, C., Chen, H.: Conditional convolutions for instance segmentation. In: Eur. Conf. Comput. Vis. pp. 282–298 (2020)
68. Wang, D., Zhang, S.: Contextual instance decoupling for robust multi-person pose estimation. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 11060–11068 (2022)
69. Wang, J., Sun, K., Cheng, T., Jiang, B., Deng, C., Zhao, Y., Liu, D., Mu, Y., Tan, M., Wang, X., et al.: Deep high-resolution representation learning for visual recognition. IEEE Trans. Pattern Anal. Mach. Intell. (2020)
70. Wang, Z., Zheng, L., Liu, Y., Li, Y., Wang, S.: Towards real-time multi-object tracking. In: Eur. Conf. Comput. Vis. pp. 107–122 (2020)
71. Wei, F., Sun, X., Li, H., Wang, J., Lin, S.: Point-set anchors for object detection, instance segmentation and pose estimation. In: Eur. Conf. Comput. Vis. pp. 527–544 (2020)
72. Wojke, N., Bewley, A., Paulus, D.: Simple online and realtime tracking with a deep association metric. In: IEEE Int. Conf. Image Process. pp. 3645–3649 (2017)
73. Xiao, B., Wu, H., Wei, Y.: Simple baselines for human pose estimation and tracking. In: Eur. Conf. Comput. Vis. (2018)
74. Xie, E., Sun, P., Song, X., Wang, W., Liu, X., Liang, D., Shen, C., Luo, P.: Polarmask: Single shot instance segmentation with polar representation. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 12193–12202 (2020)

75. Xu, L., Guan, Y., Jin, S., Liu, W., Qian, C., Luo, P., Ouyang, W., Wang, X.: Vipnas: Efficient video pose estimation via neural architecture search. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 16072–16081 (2021)
76. Xu, L., Jin, S., Liu, W., Qian, C., Ouyang, W., Luo, P., Wang, X.: Zoomnas: searching for whole-body human pose estimation in the wild. IEEE Trans. Pattern Anal. Mach. Intell. **45**(4), 5296–5313 (2022)
77. Xu, Y., Zhang, J., Zhang, Q., Tao, D.: Vitpose: Simple vision transformer baselines for human pose estimation. Adv. Neural Inform. Process. Syst. **35**, 38571–38584 (2022)
78. Xue, N., Wu, T., Xia, G.S., Zhang, L.: Learning local-global contextual adaptation for multi-person pose estimation. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 13065–13074 (2022)
79. Yang, J., Zeng, A., Liu, S., Li, F., Zhang, R., Zhang, L.: Explicit box detection unifies end-to-end multi-person pose estimation. Int. Conf. Learn. Represent. (2023)
80. Zeng, W., Jin, S., Liu, W., Qian, C., Luo, P., Ouyang, W., Wang, X.: Not all tokens are equal: Human-centric visual analysis via token clustering transformer. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 11101–11111 (2022)
81. Zeng, W., Ouyang, W., Luo, P., Liu, W., Wang, X.: 3d human mesh regression with dense correspondence. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 7054–7063 (2020)
82. Zhang, H., Li, F., Liu, S., Zhang, L., Su, H., Zhu, J., Ni, L.M., Shum, H.Y.: Dino: Detr with improved denoising anchor boxes for end-to-end object detection. Int. Conf. Learn. Represent. (2023)
83. Zhang, N., Paluri, M., Ranzato, M., Darrell, T., Bourdev, L.: Panda: Pose aligned networks for deep attribute modeling. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 1637–1644 (2014)
84. Zhang, R., Tian, Z., Shen, C., You, M., Yan, Y.: Mask encoding for single shot instance segmentation. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 10226–10235 (2020)
85. Zhang, S., Benenson, R., Schiele, B.: Citypersons: A diverse dataset for pedestrian detection. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 3213–3221 (2017)
86. Zhang, S.H., Li, R., Dong, X., Rosin, P., Cai, Z., Han, X., Yang, D., Huang, H., Hu, S.M.: Pose2seg: Detection free human instance segmentation. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 889–898 (2019)
87. Zhang, X., Ma, B., Chang, H., Shan, S., Chen, X.: Location sensitive network for human instance segmentation. IEEE Trans. Image Process. **30**, 7649–7662 (2021)
88. Zhang, Y., Zeng, W., Jin, S., Qian, C., Luo, P., Liu, W.: When pedestrian detection meets multi-modal learning: Generalist model and benchmark dataset. In: Eur. Conf. Comput. Vis. (2024)
89. Zhang, Y., Wang, C., Wang, X., Zeng, W., Liu, W.: Fairmot: On the fairness of detection and re-identification in multiple object tracking. Int. J. Comput. Vis. **129**, 3069–3087 (2021)
90. Zhang, Y., Liu, L., Li, C., Loy, C.C.: Quantifying facial age by posterior of age comparisons. In: Brit. Mach. Vis. Conf. (2017)
91. Zheng, A., Zhang, Y., Zhang, X., Qi, X., Sun, J.: Progressive end-to-end object detection in crowded scenes. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 857–866 (2022)
92. Zhou, X., Wang, D., Krähenbühl, P.: Objects as points. arXiv preprint arXiv:1904.07850 (2019)
93. Zhu, X., Su, W., Lu, L., Li, B., Wang, X., Dai, J.: Deformable detr: Deformable transformers for end-to-end object detection. Int. Conf. Learn. Represent. (2021)