Supplementary Material
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

S1 COCO-UniHuman Dataset Statistics

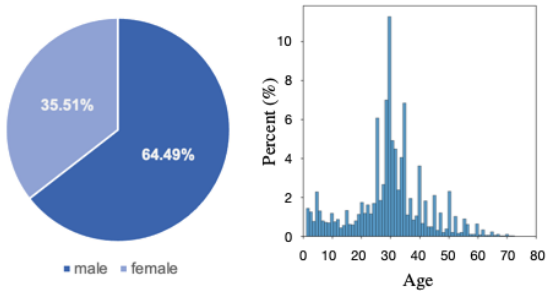

Fig. S1: Statistics of the COCO-UniHuman benchmark. (a) The gender distribution of COCO-UniHuman is biased towards male. (b) The age distribution ranges from [1, 84] and is biased towards young adults, since images are from public Internet repositories.

In Fig. S1, 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.

⋆ Equal contribution. ✉ Corresponding authors.

S2 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.

S2.1 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 [57] 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:

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

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.

Table S1: Comparisons of age annotation strategies.

Annotation Strategy	Age-5	Age-10
face w/o alignment	75.3	93.5
face w/ alignment [57]	78.2	96.5
body	**80.9**	**98.1**

From Table S1, 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.

S2.2 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 S2, shows that two-stage annotation strategy improves the annotation consistency.

Table S2: 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**

S3 More Experimental Analysis

Multi-task co-learning can mitigate over-fitting. From Fig. S2, we observe that training task-specific models on "Person" category only will easily lead to over-fitting 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 [4] 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.

S3.1 General class models vs person-specific models

In Table S3, 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.

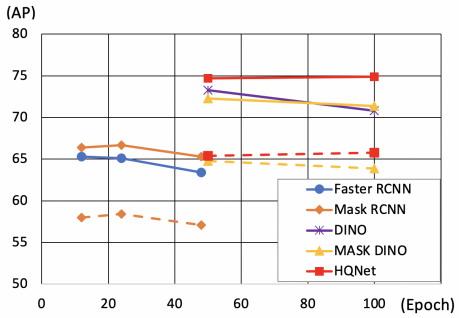

Fig. S2: Results of human detection (solid lines) and segmentation (dashed lines) with different training schedules on COCO-UniHuman `val` set. For RCNN-based models, we choose 1x, 2x, and 4x training settings. For DETR-based models, we use 50-epoch and 100-epoch training settings.

Table S3: 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 [13], and "FPN" is feature pyramid network [23]. 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

S3.2 Effect of HumanQuery-Instance Matching

In Table S4, 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.

S3.3 Qualitative Results

In Fig. S3 and Fig. S4, we show some qualitative results of HQNet with ResNet-50 backbone. In Fig. S3, we show some qualitative results on COCO-UniHuman `val` dataset for human detection, human pose estimation, human instance segmentation and human attribute recognition, and human mesh estimation. The model can recognize the gender and age of different people In Fig. S4, we visualize the results of human detection and tracking (same color for same id),

Table S4: Effect of HumanQuery-Instance 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.			Pose (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

Fig. S3: Qualitative results on COCO-UniHuman `val` dataset. Our HQNet achieves accurate human detection, human pose estimation, human instance segmentation, human attribute recognition and human mesh estimation simultaneously.

human pose estimation, human instance segmentation, gender estimation, age estimation and mesh estimation. As introduced in the section of "Unseen-task generalization" in the main paper, we directly apply our HQNet on multiple object tracking on the challenging PoseTrack21 [10] 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. 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, human attribute recognition and human mesh estimation.

S3.4 Attention Visualization

In Fig. S5, 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_{D+S}). The middle ones are from the model trained for detection and pose (M_{D+P}). And the right ones are from the model trained for detection, segmentation, pose and attribute

Fig. S4: Qualitative results on PoseTrack21 `val` video dataset. Our HQNet makes a comprehensive human analysis system that can achieve multiple functions: multiple object tracking with human pose estimation, human instance segmentation, and human attribute recognition. Our HQNet is only trained on the COCO-UniHuman image-based dataset without finetuning on the PoseTrack21 video-based dataset.

($M_{D+P+S+C}$). With the segmentation task, we notice that some of the sampling points of M_{D+S} 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_{D+P} 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. $M_{D+P+S+C}$ combines the characteristics of M_{D+S} and M_{D+P}.

S3.5 Failure Case Analysis

We have analyzed the main cases where our approach fails in the COCO-UniHuman `val` set. Fig. S6 shows an overview of common failure cases. In highly crowded and occluded scenes, where people are overlapping, the method tends to miss some targets. Occlusion can also lead to pose estimation errors. There will be

Fig. S5: Visualization of deformable attention sampling points. The results are from models trained for different HCP tasks. Left: detection and segmentation. Middle: detection and pose. Right: detection, pose, segmentation and attribute.

high keypoint localization errors on non-typical poses (*e.g.* upside-down cases). Due to age/gender bias in the dataset, the method may have some erroneous predictions on human attributes. Statues and toys also frequently lead to false positive errors. Most of these issues could be mitigated by adding related data for model training. For example, negative examples could help the network distinguish between humans and other humanoid figures. Adding occluded keypoint annotations could help predict body parts more accurately in occluded scenes.

S4 More Implementation Details

S4.1 Loss Functions

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

For human detection, we follow DINO [52] to apply focal loss [24] for classification L_{cls}^{focal} and detection loss (L1 regression loss L_{det}^{reg} and GIOU loss [37] L_{det}^{giou}). For human pose estimation, we follow PETR [38] 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

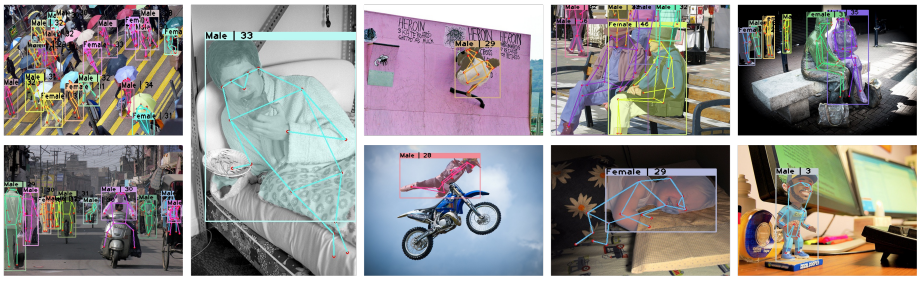

Fig. S6: Common failure cases: (a) missing detection in crowded scenes, (b) false pose detection in occluded scenes, (c) rare pose or appearance, (d) inaccurate or biased age estimation, (e) false positives on statues or toys.

loss [33] for age estimation L_{age}^{mean} and L_{age}^{var}. For mesh estimation, we use L1 regression loss for pose estimation L_{pose}^{reg}, shape estimation L_{shape}^{reg} and the 3D joints regressed from the body model L_{3d}^{reg}.

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

$$\begin{aligned} L = & \lambda_{cls}^{focal} L_{cls}^{focal} + \lambda_{det}^{reg} L_{det}^{reg} + \lambda_{det}^{giou} L_{det}^{giou} \\ & + \lambda_{2d}^{focal} L_{2d}^{focal} + \lambda_{2d}^{reg} L_{2d}^{reg} + \lambda_{2d}^{oks} L_{2d}^{oks} + \lambda_{2d}^{hm} L_{2d}^{hm} \\ & + \lambda_{pose}^{reg} L_{pose}^{reg} + \lambda_{shape}^{reg} L_{shape}^{reg} + \lambda_{3d}^{reg} L_{3d}^{reg} \\ & + \lambda_{seg}^{bce} L_{seg}^{bce} + \lambda_{seg}^{dice} L_{seg}^{dice} \\ & + \lambda_{gender}^{bce} L_{gender}^{bce} + \lambda_{age}^{mean} L_{age}^{mean} + \lambda_{age}^{var} L_{age}^{var}, , \end{aligned}$$

where λs are corresponding loss weights. Detailed settings for the loss weights can be found in Table S5.

S4.2 Details about Training

We follow the setting of DINO [52] 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 [17] 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.

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 [42]. Unlike DINO and Mask DINO which also pre-train models on

Table S5: Loss weights for training our models.

Detection	λ_{cls}^{focal}	1.0
	λ_{det}^{reg}	5.0
	λ_{det}^{giou}	2.0
Pose	λ_{kpt}^{focal}	1.0
	λ_{kpt}^{reg}	50.0
	λ_{kpt}^{oks}	1.5
	λ_{kpt}^{hm}	4.0
	λ_{pose}^{reg}	5.0
	λ_{shape}^{reg}	10.0
	λ_{3d}^{reg}	10.0
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

Objects365 [60], 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 [19,52], 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 [61], we use independent auxiliary heads to refine the multi-task predictions at each decoder layer.

S4.3 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).

S5 Details about Baseline Models

Details about human detection baselines. For human detection, we compare several baseline approaches, *i.e.* Faster-RCNN [36], IterDETR [58] and DINO [52]. Note that Faster-RCNN and DINO are originally trained to handle general 80 classes (marked with * in Table S3). For fair comparisons, we use

MMDetection [4] 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 representative top-down methods (SBL [48], HRNet [40], Swin [28], ViTPose [50] and PRTR [20]), bottom-up approaches (HrHRNet [6], DEKR [11], and SWAHR [29]) and single-stage approaches (FCPose [30], InsPose [39], PETR [38] and CID [45]). Note that the results of Swin (Swin-L), ViTPose (ViT-L) and CID (R-50-FPN) are from MMPose [8], 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, Swin and ViTPose use the same person detector provided by [48], which is a strong Faster-RCNN [36] 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 [40] 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 [13] 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 HQNet with state-of-the-art general and human-specific instance segmentation methods. Mask R-CNN [12] 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 [49], MEInst [54], YOLACT [3], and CondInst [44]. Results of PolarMask, MEInst, YOLACT, CondInst are from [56], 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 under-explored in the literature. We establish baselines using StrongBL [15] and Mask R-CNN [12]. 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.

Details about mesh estimation baseline. For human mesh estimation, we compare HQNet with state-of-the-art one-stage monocular method ROMP [41], and two-stage method HMR [16] and HMR+ [34]. Following official settings, we train these models on COCO-UniHuman with ResNet50 backbone. As two-stage models require human bbox as the input, in the experiment, we use GT bbox for comparisons.

S6 Discussion about Unifying HCP Tasks

S6.1 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 [46]) and Transformer-based backbone networks (*e.g.* TCFormer [51] are proposed for general human-centric visual tasks. Other works focus on designing network heads to unify different HCP tasks. For example, UniHead [22] designs a novel perception head with unified keypoint representations that can be used in different HCP tasks. Point-Set Anchors [47] 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.

S6.2 Pre-training on HCP tasks

There are also works [5, 14, 42] on pre-training on diverse human-centric tasks with large-scale data. HCMoCo [14] introduces a versatile multi-modal (RGB-D) pre-training framework for single-person pose estimation and segmentation. SOLIDER [5] presents a self-supervised learning framework to learn a general human representation with more semantic information. HumanBench [42] builds a large-scale human-centric pre-training dataset and introduces the projector-assisted pre-training method with hierarchical weight sharing. More recently, UniHCP [7] 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 multi-task manner.

S6.3 Co-learning on HCP tasks

Many works have investigated the correlations between pairs of HCP tasks [25, 31,32,43,53]. For example, [43] explore to integrate fine-grained person attribute learning into the pipeline of pedestrian detection. Mask-RCNN [12] extends Faster-RCNN by adding extra keypoint localization or segmentation branch to handle pose estimation and instance segmentation respectively. Pose2Seg [55] 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 [35] 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.

S7 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.

S7.1 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 [21], 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 [21],

PETA [9], Market1501-Attr [26, 59], RAP-2.0 [18] and PA-100K [27]) 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 [57]) 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.

S7.2 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, 15], local parts based [21], and visual attention based [27] 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.

S8 Limitations and Future Work

While our work focuses on RGB image based HCP tasks, tasks about video data or multi-modality data (*e.g.* IR and Depth) also hold significant potential. We encourage future research to explore more comprehensive multi-task human-centric perception.

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. Bolya, D., Zhou, C., Xiao, F., Lee, Y.J.: Yolact: Real-time instance segmentation. In: Int. Conf. Comput. Vis. pp. 9157–9166 (2019)
4. Chen, K., Wang, J., Pang, J., Cao, Y., Xiong, Y., Li, X., Sun, S., Feng, W., Liu, Z., Xu, J., et al.: Mmdetection: Open mmlab detection toolbox and benchmark. arXiv preprint arXiv:1906.07155 (2019)
5. 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)

6. 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)
7. 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)
8. Contributors, M.: Openmmlab pose estimation toolbox and benchmark. `https://github.com/open-mmlab/mmpose` (2020)
9. Deng, Y., Luo, P., Loy, C.C., Tang, X.: Pedestrian attribute recognition at far distance. In: ACM Int. Conf. Multimedia. pp. 789–792 (2014)
10. 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)
11. 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)
12. He, K., Gkioxari, G., Dollár, P., Girshick, R.: Mask r-cnn. In: Int. Conf. Comput. Vis. pp. 2961–2969 (2017)
13. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: IEEE Conf. Comput. Vis. Pattern Recog. (2016)
14. 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)
15. 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)
16. 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)
17. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
18. Li, D., Zhang, Z., Chen, X., Huang, K.: A richly annotated pedestrian dataset for person retrieval in real surveillance scenarios. IEEE Trans. Image Process. **28**(4), 1575–1590 (2019)
19. 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)
20. 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)
21. Li, Y., Huang, C., Loy, C.C., Tang, X.: Human attribute recognition by deep hierarchical contexts. In: Eur. Conf. Comput. Vis. (2016)
22. Liang, J., Song, G., Leng, B., Liu, Y.: Unifying visual perception by dispersible points learning. In: Eur. Conf. Comput. Vis. pp. 439–456 (2022)
23. Lin, T.Y., Dollár, P., Girshick, R., He, K., Hariharan, B., Belongie, S.: Feature pyramid networks for object detection (2017)
24. Lin, T.Y., Goyal, P., Girshick, R., He, K., Dollár, P.: Focal loss for dense object detection. In: Int. Conf. Comput. Vis. pp. 2980–2988 (2017)
25. 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)
26. Lin, Y., Zheng, L., Zheng, Z., Wu, Y., Hu, Z., Yan, C., Yang, Y.: Improving person re-identification by attribute and identity learning. Pattern Recognition (2019)

27. 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)
28. 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)
29. 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)
30. 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)
31. 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)
32. 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)
33. Pan, H., Han, H., Shan, S., Chen, X.: Mean-variance loss for deep age estimation from a face. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 5285–5294 (2018)
34. 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)
35. 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)
36. 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)
37. Rezatofighi, H., Tsoi, N., Gwak, J., Sadeghian, A., Reid, I., Savarese, S.: Generalized intersection over union: A metric and a loss for bounding box regression. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 658–666 (2019)
38. 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)
39. 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)
40. 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)
41. 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)
42. 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)
43. 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)
44. Tian, Z., Shen, C., Chen, H.: Conditional convolutions for instance segmentation. In: Eur. Conf. Comput. Vis. pp. 282–298 (2020)

45. Wang, D., Zhang, S.: Contextual instance decoupling for robust multi-person pose estimation. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 11060–11068 (2022)
46. 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)
47. 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)
48. Xiao, B., Wu, H., Wei, Y.: Simple baselines for human pose estimation and tracking. In: Eur. Conf. Comput. Vis. (2018)
49. 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)
50. 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)
51. 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)
52. 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)
53. 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)
54. 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)
55. 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)
56. 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)
57. Zhang, Y., Liu, L., Li, C., Loy, C.C.: Quantifying facial age by posterior of age comparisons. In: Brit. Mach. Vis. Conf. (2017)
58. 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)
59. Zheng, L., Shen, L., Tian, L., Wang, S., Wang, J., Tian, Q.: Scalable person re-identification: A benchmark. In: Int. Conf. Comput. Vis. (2015)
60. Zhou, X., Wang, D., Krähenbühl, P.: Objects as points. arXiv preprint arXiv:1904.07850 (2019)
61. 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)