UniHCP: A Unified Model for Human-Centric Perceptions

Yuanzheng Ci[1]*, Yizhou Wang[2,3]*, Meilin Chen[2], Shixiang Tang[1], Lei Bai[3]†, Feng Zhu[4], Rui Zhao[4,5], Fengwei Yu[4], Donglian Qi[2], Wanli Ouyang[3]
[1]The University of Sydney, [2]Zhejiang University, [3]Shanghai AI Laboratory, [4]SenseTime Research, [5]Qing Yuan Research Institute, Shanghai Jiao Tong University, Shanghai, China
yuanzheng.ci@sydney.edu.au, yizhouwang@zju.edu.cn, bailei@pjlab.org.cn

Abstract

*Human-centric perceptions (e.g., pose estimation, human parsing, pedestrian detection, person re-identification, etc.) play a key role in industrial applications of visual models. While specific human-centric tasks have their own relevant semantic aspect to focus on, they also share the same underlying semantic structure of the human body. However, few works have attempted to exploit such homogeneity and design a general-propose model for human-centric tasks. In this work, we revisit a broad range of human-centric tasks and unify them in a minimalist manner. We propose UniHCP, a **Uni**fied Model for **H**uman-**C**entric **P**erceptions, which unifies a wide range of human-centric tasks in a simplified end-to-end manner with the plain vision transformer architecture. With large-scale joint training on 33 human-centric datasets, UniHCP can outperform strong baselines on several in-domain and downstream tasks by direct evaluation. When adapted to a specific task, UniHCP achieves new SOTAs on a wide range of human-centric tasks, e.g., 69.8 mIoU on CIHP for human parsing, 86.18 mA on PA-100K for attribute prediction, 90.3 mAP on Market1501 for ReID, and 85.8 JI on CrowdHuman for pedestrian detection, performing better than specialized models tailored for each task. The code and pretrained model are available at https://github.com/OpenGVLab/UniHCP.*

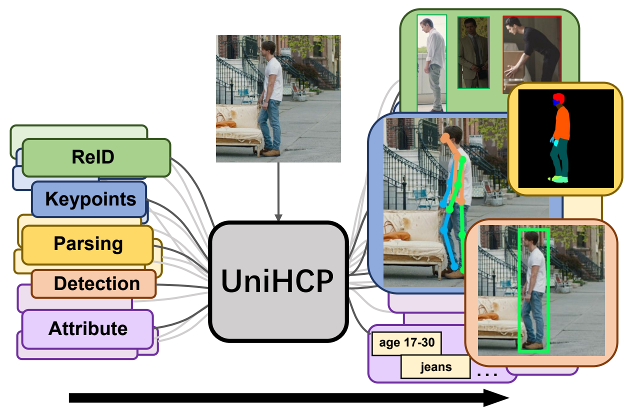

Figure 1. UniHCP unifies 5 human-centric tasks under one model and is trained on a massive collection of human-centric datasets.

1. Introduction

Research on human-centric perceptions has come a long way with tremendous advancements in recent years. Many methods have been developed to enhance the performance of pose estimation [9, 25, 60, 91], pedestrian detection [4, 62, 63, 76], person re-identification [42, 86, 101] (ReID), and many other human-centered tasks. These significant progress play a key role in advancing the applications of visual models in numerous fields, such as sports analysis [11], autonomous driving [97], and electronic retailing [27].

Although different human-centric perception tasks have their own relevant semantic information to focus on, those semantics all rely on the same basic structure of the human body and the attributes of each body part [64, 81]. In light of this, there have been some attempts trying to exploit such homogeneity and train a shared neural network jointly with distinct human-centric tasks [28,29,46,48,61,71,77,87,98]. For instance, human parsing has been trained in conjunction with human keypoint detection [46, 61, 98], pedestrian attribute recognition [87], pedestrian detection [48] or person re-identification [28]. The experimental results of these works empirically validate that some human-centric tasks may benefit each other when trained together. Motivated by these works, a natural expectation is that a more versatile all-in-one model could be a feasible solution for general human-centric perceptions, which can utilize the homogeneity of human-centric tasks for improving performance, enable fast adaption to new tasks, and decrease the burden of memory cost in large-scale multitask system deployment compared with specific models to specific tasks.

However, unifying distinct human-centric tasks into a

*Equal contribution.
†Corresponding author.

general model is challenging considering the data diversity and output structures. From the data's perspective, images in different human-centric tasks and different datasets have different resolutions and characteristics (e.g., day and night, indoor and outdoor), which calls for a robust representative network with the capability to accommodate them. From the perspective of output, the annotations and expected outputs of different human-centric tasks have distinct structures and granularities. Although this challenge can be bypassed via deploying separate output heads for each task/dataset, it is not scalable when the number of tasks and datasets is large.

In this work, we aim to explore a simple, scalable formulation for unified human-centric system and, for the first time, propose a Unified model for Human-Centric Perceptions (UniHCP). As shown in Figure.1, UniHCP unifies and simultaneously handles five distinct human-centric tasks, namely, pose estimation, semantic part segmentation, pedestrian detection, ReID, and person attribute recognition. Motivated by the extraordinary capacity and flexibility of the vision transformers [43, 94], a simple yet unified encoder-decoder architecture with the plain vision transformer is employed to handle the input diversity, which works in a simple feedforward and end-to-end manner, and can be shared across all human-centric tasks and datasets to extract general human-centric knowledge. To generate the output for different tasks with the unified model, UniHCP defines Task-specific Queries, which are shared among all datasets with the same task definition and interpreted into different output units through a Task-guided Interpreter shared across different datasets and tasks. With task-specific queries and the versatile interpreter, UniHCP avoids the widely used task-specific output heads, which minimizes task-specific parameters for knowledge sharing and make backbone-encoded features reusable across tasks.

Own to these designs, UniHCP is suitable and easy to perform multitask pretraining at scale. To this end, we pretrained an UniHCP model on a massive collection of 33 labeled human-centric datasets. By harnessing the abundant supervision signals of each task, we show such a model can simultaneously handle these in-pretrain tasks well with competitive performance compared to strong baselines relying on specialized architectures. When adapted to a specific task, both in-domain and downstream, our model achieves new SOTAs on several human-centric task benchmarks. In summary, the proposed model has the following properties:

- Unifying five distinct human-centric tasks and handling them simultaneously.
- Shared encoder-decoder network based on plain transformer.
- Simple task-specific queries identifying the outputs.
- Maximum weight sharing (99.97% shared parameters) with a task-guided interpreter.
- Trainable at scale and demonstrates competitive performance compared to task-specialized models.

2. Related Works

2.1. Human-Centric Perceptions

Human-centric perceptions are essential for substantial real-world applications. Depending on the targeted visual concept, the way of decoding output from image features varies across tasks. Specifically, pose estimation and pedestrian detection are both localization tasks that can be solved by either regression-based methods [37, 96] or heatmap-based methods [33, 34, 84]. Human parsing, as a fine-grained segmentation problem, is usually solved by per-pixel classification. While contour-based methods [65, 85] can also obtain segmentation masks, it requires instance-level mask annotations, which are not always available. PAR is treated as a multi-label classification task [104], and ReID is treated as a feature learning task [74].

Recently, several transformer-based solutions have been proposed for these human-centric tasks, with attention block designs on both backbone [19, 88, 93] and decoding network [40, 44, 54, 59, 87, 102]. However, these methods involve *different* task-specific designs and thus cannot be integrated into one model seamlessly. Built upon the general success of these works, we take a further step and unify human-centric tasks under the *same* architecture based on plain vision transformer.

2.2. Unified Models

A general-purpose model that can handle different tasks in a unified manner has long been a coveted alternative to models specifically tailored for different tasks. Pioneering works regarding Natural Language Processing (NLP) [66], vision-language [58], and basic vision tasks [30, 68] have shown the effectiveness of such kind of unified cross-task models. ExT5 [2] and OFA [80] further provide a degree of promise for the performance benefits of large-scale multitask co-training. Among models supporting visual tasks, UniHead [45] and UViM [31] propose a unified architecture for several vision tasks. However, they are only trained and evaluated in a single-task manner.

For methods supporting multitask co-training, Uni-Perceiver [106] focuses on tasks in which the desired output is inherently language or labels, which does not fit human-centric tasks. While UniT [21], OFA [80], Unified-IO [57], and Pix2Seq v2 [6] further extend the support for detection, keypoint detection, segmentation, and many other visual tasks, they rely on *independent decoder heads* [21, 80] or *autoregressive* modeling [6, 57]. These works do not focus on human-centric vision tasks. Differently, our work introduces a *shared decoder head* (task-guided interpreter) in a *parallelly feedforward* manner for human-centric vision

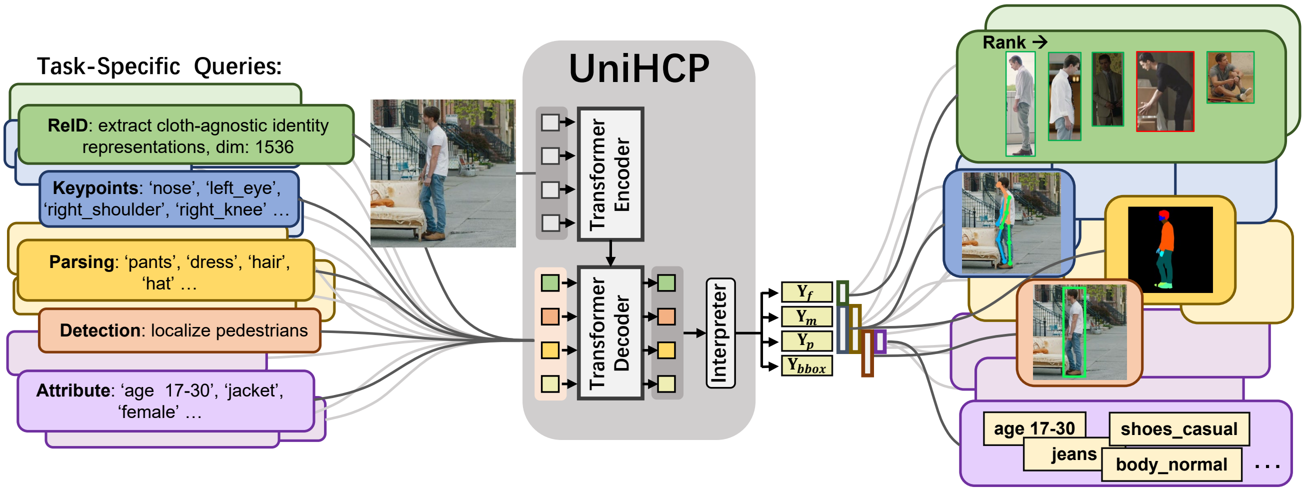

Figure 2. UniHCP handles a massive collection of human-centric tasks uniformly by task-specific queries and a task-guided interpreter, all predictions are yielded in parallel through a simple encoder-decoder transformer architecture.

Table 1. Network details of UniHCP

	Layers	Dimension	Params
Encoder	12	768	91.1M
Decoder	9	256	14.5M
Task-guided Interpreter			3.5M
Task-specific queries		256	<0.03M
Total			109.1M
Task-agnostic params / total params			99.97%

tasks, which is simple yet maximizes the parameter sharing among different tasks.

In the case of human-centric tasks, many works have shown great success by co-training a pair of human-centric tasks [28, 29, 46, 48, 61, 71, 77, 87, 98]. However, there is no work exploring a general unified model that can handle all representative human-centric tasks. Our work is the first attempt at designing, training, and evaluating a unified human-centric model with a large-scale multitask setting.

3. UniHCP

To share the most knowledge among various human-centric tasks, we attempt to maximize weight sharing among all tasks in UniHCP. Specifically, our UniHCP, as shown in Figure 2, consists of three components: (1) A task-agnostic transformer encoder E to extract image features. (2) A transformer decoder D that attends to task-specific information according to task-specific queries $\{\mathbf{Q}^t\}$, where t denotes a specific task. (3) A task-guided interpreter $\mathcal{I}$ produces output units, in which we decompose the output of multiple human-centric perception tasks into sharable units of diverse granularities, *i.e.,* feature representation, local probability map, global probability, bounding box coordinates. Since only the queries to the decoders are not shared among tasks, we can learn human-centric knowledge across different granularities by the designed interpreters and achieve maximum parameter sharing among all tasks, i.e., **99.97%** shared parameters, as shown in Table 1. The pipeline for our UniHCP is described as follows.

Step 1: Given an image $\mathbf{X}$ sampled from the dataset in task t, extract encoded features $\mathbf{F}$ by the task-agnostic transformer encoder E (Sec. 3.1).

Step 2 : A transformer decoder D with task-specific queries $\mathbf{Q}^t$ extracts task-specific features from encoded features $\mathbf{F}$ (Sec. 3.2).

Step 3: Generate output units according to the queried task, i.e., attended features $\mathbf{Y}_f$, local probability map $\mathbf{Y}_m$, global probability $\mathbf{Y}_p$ and bounding box coordinates $\mathbf{Y}_{bbox}$ by a task-guided interpreter $\mathcal{I}$ (Sec. 3.3). For example, for human parsing, two units: local probability map $\mathbf{Y}_m$ (for semantic part segmentation) and global probability $\mathbf{Y}_p$ (for existence of body part in the image), are generated.

Step 4: Calculate the loss of the corresponding task for optimizing the encoder E, the decoder D, the task-specific queries $\mathbf{Q}^t$ and task-guided interpreter $\mathcal{I}$ by backward propagation (Sec. 3.4).

3.1. Task-agnostic Transformer Encoder

UniHCP uses a plain Vision Trasnformer [14] (ViT) as the encoder. To handle input images of different resolutions, we use a shared learnable positional embedding with the size of 84×84 and interpolate it based on the spatial size of the input image after patch projection. The encoded feature $\mathbf{F}$ can be mathematically calculated as

$$\mathbf{F} = E(\mathbf{X}, \mathbf{P}_E), \tag{1}$$

where $\mathbf{P}_E$ is the positional embedding after interpolation and E denotes the task-agnostic transformer encoder.

3.2. Decoder with Task-specific Queries

To obtain the most discriminative feature for each task while maximizing knowledge sharing, we design task-specific queries to guide the transformer decoder only attending to task-relevant information.

Task-specific Queries. Task queries for task t are denoted as

$$\mathbf{Q}^t = [\mathbf{q}_1^t, \mathbf{q}_2^t, ..., \mathbf{q}_{N^t}^t], \tag{2}$$

where N^t denotes the number of queries representing N^t different semantic meanings in task t. For pedestrian attribute recognition, pose estimation, human parsing, and ReID, the number of queries respectively equals to the number of attributes, the number of pose joints, the number of semantic parsing classes, and the length of desired ReID features. For pedestrian detection, we follow the implementation in [82], with details provided in the supplementary material. We randomly initialize the task-specific query $\mathbf{Q}^t$ as learnable embeddings $\mathbf{Q}_0^t$ and refine it with the following decoder blocks.

Following the common practice as in [8, 78, 82], all $\mathbf{Q}^t$ are also associated with a positional embedding $\mathbf{Q}_p^t$, which has the same dimension as $\mathbf{Q}^t$ and is not shared across tasks. Different from $\mathbf{Q}^t$ that will be progressively refined in the decoder blocks, $\mathbf{Q}_p^t$ is shared across decoder blocks. For tasks other than pedestrian detection, $\mathbf{Q}_p^t$ is simply a learnable positional embedding that is randomly initialized. For pedestrian detection, we have

$$\mathbf{Q}_p^t = proj(\mathcal{A}_{\mathbf{Q}}), \tag{3}$$

where $\mathcal{A}_{\mathbf{Q}} \in \mathbb{R}^{N^t \times 2}$ refers to N^t learnable anchor points that are initialized with a uniform distribution following [82], and $proj$ is a projection from coordinates to positional embeddings (more details about the projector are elaborated in the supplementary materials).

Decoder. The transformer decoder aims to attend to task-specific features according to the task queries. We follow the standard design of transformer decoders [78]. In the decoder, each transformer block D_l for $l = 1, 2, ..., L$ consists of a cross-attention module, a self-attention module, and a feed-forward module (FFN), where L denotes the number of transformer blocks. We place cross-attention before self-attention as adopted by [8, 36]. For each block D_l, we attend to task-specific information from the encoded feature by task queries, which can be formulated as

$$\mathbf{Q}_l^t = D_l(\mathbf{Q}_{l-1}^t, \mathbf{Q}_p^t, \mathbf{F}, \mathbf{F}_p), \tag{4}$$

$$\text{where } \mathbf{F}_p = proj(\mathcal{A}_{\mathbf{F}}), \tag{5}$$

$\mathcal{A}_{\mathbf{F}} \in \mathbb{R}^{H_{\mathbf{F}} W_{\mathbf{F}} \times 2}$ is the coordinates with respect to the original image for all feature tokens in $\mathbf{F} \in R^{H_{\mathbf{F}} \times W_{\mathbf{F}}}$. For the cross-attention in the decoder D_l, the query is $\hat{\mathbf{Q}}_l^t = \mathbf{Q}_{l-1}^t + \mathbf{Q}_p^t$, the key is $\hat{\mathbf{K}} = \mathbf{F}' + \mathbf{F}_p$, and the value is $\hat{\mathbf{V}} = \mathbf{F}'$, where $\mathbf{F}'$ is linearly projected from the features of the encoder $\mathbf{F}$ to align channel dimensions. The result of cross-attention is then passed for self-attention in D_l.

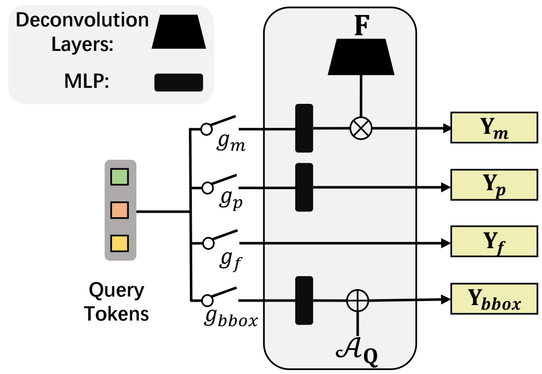

Figure 3. Task-guided interpreter. $\otimes$ denotes a dynamic convolution module [7] that takes the projected query feature as the kernel and takes the tokens $\mathbf{F}$ from the encoder as the feature map, where $\mathbf{F}$ is upscaled to the desired resolution $H' \times W'$, $\oplus$ denotes addition, for which the inputs are the projected query feature in the format of $[\nabla cx, \nabla cx, h, w]$ and $\mathcal{A}_{\mathbf{Q}}$, which contains the anchor point $[cx, cy]$ (see supplementary materials for details).

3.3. Task-guided Interpreter

Task-guided interpreter $\mathcal{I}$ interprets query tokens $\mathbf{Q}^t$ into four output units subject to the request of a specific task. As shown in Figure 3, these four output units are as follows:

$$\begin{aligned} \text{feature vector unit} &: \mathbf{Y}_f \in \mathbb{R}^{N^t \times C} \\ \text{global probability unit} &: \mathbf{Y}_p \in \mathbb{R}^{N^t \times 1} \\ \text{local probability map unit} &: \mathbf{Y}_m \in \mathbb{R}^{N^t \times H' \times W'} \\ \text{bounding box unit} &: \mathbf{Y}_{bbox} \in \mathbb{R}^{N^t \times 4}, \end{aligned} \tag{6}$$

where C is the output dimension of the decoder, $H' \times W'$ denotes the desired resolution for the local probability map. Given task t and output interpreter $\mathcal{I}$, the output of the UniHCP is defined as follows:

$$\{\mathbf{Y}_u | g_u^{\mathbf{t}_t} = 1, u \in \{f, p, m, bbox\}\} = \mathcal{I}(\mathbf{Q}^t, \mathbf{g}^{\mathbf{t}_t}), \tag{7}$$

where $\mathbf{t}_t \in \{reid, \ldots, pose\}$ denotes the task type of task t, $\mathbf{g}^{\mathbf{t}} = \{g_u^{\mathbf{t}}\}$ is a set of task-specific binary gates ($g \in \{0, 1\}$) that represents the desired output units for task type $\mathbf{t}$.

Guidance from tasks to output units. For human parsing, local probability map (for semantic part segmentation) and global probability (for existence of body part in the image) are activated, corresponding to $g_m^{seg} = 1$ and $g_p^{seg} = 1$ respectively. For person ReID, feature vectors are used, corresponding to $g_f^{reid} = 1$. For pose estimation, $g_m^{pose} = 1$

(for localizing key points) and $g_p^{pose} = 1$ (for existence of keypoints in the image). For detection, $g_{bbox}^{det} = 1$ (for bounding box prediction) and $g_p^{det} = 1$ (for existence of object). For pedestrian attribute prediction, $g_p^{par} = 1$ (for existence of attributes in the image). Therefore, the output unit of global probabilities is shared among pose estimation, human parsing, pedestrian detection, and attribute recognition. The output unit of local probability maps is shared among pose estimation and human parsing.

Discussion. The task-guided interpreter interprets each query token independently. Previous works focused on autoregressive decoding with tokenization [6, 57] or task-specific heads [21, 92] to handle different output units required by specific tasks. In contrast, the task-guided interpreter can handle tasks involving a varying number of classes, yield all results in parallel, and do not require task-specific heads. This is achieved by two designs in our UniHCP framework: 1) Class/instance information is self-contained in queries. As mentioned in Section 3.2, a query represents a particular semantic class in pose estimation, attribute prediction, human parsing, and pedestrian detection. We only need to retrieve a scalar probability value from a query to obtain the confidence information for a particular class/human instance. 2) Outputs of the same modality share the same output unit. For example, the heatmap for a particular joint in pose estimation and the heatmap for a particular body part in human parsing have the same dimension. Although these outputs have different meanings, experimental results in Section 4.3 show that it is suitable to obtain them through the same output unit and fully let the task-specific queries handle the differences in preferred information to be represented.

3.4. Objective Functions

In this section, we will introduce the objective functions for training diverse human-centric tasks together and illustrate how these objectives are related to the output units defined in Eq. 6. Unless otherwise specified, we omit the GT inputs in loss functions for brevity.

Overall Objective Function. Given a collection of datasets $\mathbb{D} = \{\mathcal{D}|\mathbf{t}_\mathcal{D} \in \{reid, \ldots, pose\}\}$, where $\mathbf{t}_\mathcal{D}$ denotes the task type of dataset $\mathcal{D}$, we also note $t_\mathcal{D}$ as the task of dataset $\mathcal{D}$, we have the overall loss defined as:

$$\mathcal{L} = \sum_{\mathcal{D}\in\mathbb{D}} w_D \mathcal{L}_{\mathbf{t}_\mathcal{D}}(\mathcal{I}(\mathbf{Q}^{t_\mathcal{D}}, \mathbf{g}^{\mathbf{t}_\mathcal{D}})), \tag{8}$$

where $w_\mathcal{D}$ is the loss weight for dataset $\mathcal{D}$, which is calculated based on the task type and batch size (calculations are elaborated in supplementary materials).

ReID. Person ReID is a feature learning task for extracting identification information. Therefore, we directly supervised the features after the decoder by identity annotations. Specifically, for ReID task, the extracted feature is a simple concatenation of all feature vectors $\mathbf{Y}_f = [y_f^1; \ldots; y_f^{N^t}]$, where $N^t = 6$ by default. The loss function is a combination of ID loss [103] and triplet loss [52] written as follows:

$$\mathcal{L}_{reid} = \mathcal{L}_{ID}(\mathbf{Y}_f) + \mathcal{L}_{triplet}(\mathbf{Y}_f). \tag{9}$$

PAR. Pedestrian attribute recognition only predicts whether an attribute exists in the global image. Therefore, we only supervise the output unit of global probabilities $\mathbf{Y}_p$ from the task-guided interpreter. Specifically, following the common practice [40, 75], we adopt the weighted binary cross-entropy loss. Given the probability predictions $\mathbf{Y}_p$ associated with N^t attributes, we have:

$$\begin{aligned} \mathcal{L}_{par} &= \sum_{n=1}^{N_t} w_n (y_n \log(y_p^n) + (1-y_n)\log(1-y_p^n)), \\ w_n &= y_n e^{1-\gamma_n} + (1-y_n)e^{\gamma_n}, \end{aligned} \tag{10}$$

where y_n denotes the annotation of n-th attribute and γ_n denotes the positive example ratio of n-th attribute.

Human Parsing. Human parsing can be considered as semantic segmentation of human part. We view the presence of semantic classes as predictable attributes since the semantic classes are not always present in an image. Therefore, the global probability $\mathbf{Y}_p$ and local probability map $\mathbf{Y}_m$ are selected from the output units to describe whether a semantic part exists on the image level (global) and pixel level (local), respectively. Given a query $\mathbf{q}_l$ defined in Eq. 2 which corresponds to a semantic class in human parsing, we adopt the binary cross entropy loss as $\mathcal{L}_{par}$ in pedestrian attribute recognition to constrain the global probability $\mathbf{Y}_p$, and a combination of binary cross-entropy loss and dice loss [8] to supervised local probability map $\mathbf{Y}_m$ as follows:

$$\mathcal{L}_{seg} = \lambda_{par}\mathcal{L}_{par}(\mathbf{Y}_p) + \mathcal{L}_{bce}(\mathbf{Y}_m) + \mathcal{L}_{dice}(\mathbf{Y}_m),$$

where λ_{par} denotes the loss weight for $\mathcal{L}_{par}(\mathbf{Y}_p)$.

Pose Estimation. We follow the common top-down setting for pose estimation, i.e., predicting keypoints based on the cropped human instances. We predict the heatmap w.r.t. the keypoints via mean-squared error. Similar to human parsing formulation, we also select the global probability $\mathbf{Y}_p$ and local probability map $\mathbf{Y}_m$ to predict whether a keypoint exists in the image level and pixel level, respectively. Mathematically, we have:

$$\mathcal{L}_{pose} = \lambda_{par}\mathcal{L}_{par}(\mathbf{Y}_p) + \mathcal{L}_{mse}(\mathbf{Y}_m). \tag{11}$$

Pedestrian Detection. Pedestrian Detection is a local prediction task but in a sparse manner. Following the widely adopted designs in end-to-end transformer-based detection [5, 102], ground-truth for N^t query features in $\mathbf{Q}_l$ are determined by optimal bipartite matching between all N^t predictions and GT boxes. Given output units $\mathbf{Y}_p$ and $\mathbf{Y}_{bbox}$, we adopt the identical cost formulation and loss as in [102],

$$\mathcal{L}_{peddet} = \lambda_{cls}\mathcal{L}_{cls}(\mathbf{Y}_p) + \lambda_{iou}\mathcal{L}_{iou}(\mathbf{Y}_{bbox}) + \lambda_{L1}\mathcal{L}_{L1}(\mathbf{Y}_{bbox}). \tag{12}$$

where $\mathcal{L}_{cls}, \mathcal{L}_{iou}$ and $\mathcal{L}_{L1}$ are focal loss [50], GIoU loss [67], and $L1$ loss, respectively. Their corresponding loss weights λ are also identically set as in [102].

4. Experiments

4.1. Implementation details

Datasets. To enable general human-centric perceptions, we pretrain the proposed UniHCP at scale on a massive and diverse collection of human-centric datasets. Specifically, the training splits of 33 publically available datasets are gathered to form the training set for UniHCP, including nine datasets for pose estimation and six datasets for ReID, Human Parsing, Attribute Prediction, Pedestrain Detection, seraprately. For ReID, there are two different subtasks: general ReID and cloth-changing ReID, where the difference is how cloth-changing is treated. We empirically found it is best to view them as different tasks with different task queries, hence, we opted for this setup.

We carefully follow the de-duplication practices as introduced in [95] to remove the samples that could appear in the evaluation datasets. We also remove images whose groundtruth labels are not given, leading to 2.3M distinct training samples in total. For evaluation, apart from the available validation or test splits of the 33 training sets, we also included several out-of-pretrain downstream datasets for each type of human-centric task. More details about dataset setups can be found in supplementary materials.

Training. We use the standard ViT-B [14] as the encoder network and initialize it with the MAE pretrained [18] weights following [43, 88]. For the main results, we use a batch size of 4324 in total, with the dataset-specific batch size being proportional to the size of each dataset. Unless otherwise specified, the image resolution used in pretraining is 256×192 for pose estimation and attribute prediction, 256×128 for ReID, 480×480 for human parsing, and a maximum height/width of 1333 for pedestrian detection.

For computational efficiency, each GPU only runs one specific task, and each task can be evenly distributed to multiple GPUs whereas a single GPU is not capable of handling its workloads. To further save the GPU memory during the training time, we adopt the gradient checkpointing [3] in the encoder forward pass among all tasks and additionally use accumulative gradients for detection tasks. Due to the high GPU-memory demand of detection datasets, the batch size for the detection task is timed by 0.6.

Table 2. Representative datasets used in multitask co-training.

Task Type	Datasets	Number of samples
ReID (6 datasets)	CUHK03 [41] PRCC [89] ...	268,002
Pose Estimation (9 datasets)	COCO-Pose [51] AI Challenger [83] ...	1,261,749
Human Parsing (6 datasets)	LIP [16] DeepFashion2 [15] ...	384,085
Attribute Prediction (6 datasets)	PA-100K [56] RAPv2 [35] ...	242,880
Pedestrian Detection (6 datasets)	COCO-Person [51] CrowdHuman [69] ...	170,687

We use Adafactor [70] optimizer and follow the recommended modifications [94] for adopting it to ViT, we set $\beta_1 = 0.9$, β_2 clipped at 0.999, disables the parameter scaling and decoupled weight decay to 0.05. We linearly warm up the learning rate for the first 1.5k iterations to 1e-3, after which the learning rate is decayed to 0 following a cosine decay scheduler. We also use a drop-path rate of 0.2 and layer-wise learning rate decay [43, 88] of 0.75 in the ViT-B encoder. The whole training process takes 105k iterations which are approximately 130 epochs for detection datasets and 200 epochs for other datasets. The whole training takes 120 hours in total on 88 NVIDIA V100 GPUs.

4.2. Main Results

To demonstrate the capability of UniHCP as a unified model for human-centric perceptions, we first evaluate our UniHCP on thirteen datasets that appear in the pretraining stage (in Section 4.2.1), *e.g.*, CIHP. Furthermore, we employ five datasets whose training splits are not included in the pretraining stage to evaluate the cross-datasets transferability of UniHCP (in Section 4.2.2). We also demonstrate that UniHCP has the potential to efficiently transfer to new datasets that do not appear in pretraining with only a few images (in Section 4.2.3). For detailed evaluation configuration, please refer to the supplementary.

4.2.1 In-pretrain Dataset Results

We conduct extensive evaluations on thirteen in-pretrain datasets to demonstrate the effectiveness of our UniHCP. Table 3-7 summarize the evaluation results of UniHCP on five representative human-centric tasks, *i.e.*, person ReID,

Table 3. Person ReID evaluation on Market1501, MSMT, CUHK03 with mAP. †indicates using additional camera IDs.

Method	Market1501	MSMT17	CUHK03
HOReID [79]	84.9	-	-
MNE [39]	-	-	77.7
SAN [26]	88.0	55.7	76.4
TransReID [19]	86.8	61.0	-
TransReID† [19]	88.9	**67.4**	-
UniHCP (direct eval)	80.7	55.2	68.6
UniHCP (finetune)	**90.3**	67.3	**83.1**

Table 4. Pedestrian attribute recognition evaluation on PA-100K and RAPv2 test sets with mA.

Method	PA-100K	RAPv2
SSC [23]	81.87	-
C-Tran [32]	81.53	-
Q2L [54]	80.72	-
L2L [40]	82.37	-
DAFL [24]	83.54	81.04
UniHCP (direct eval)	79.32	77.20
UniHCP (finetune)	**86.18**	**82.34**

Table 5. Human parsing evaluation on Human3.6M, LIP and CIHP val sets with mIoU.

Method	H3.6M	LIP	CIHP
HCMOCO [20]	62.50	-	-
SNT [22]	-	54.73	60.87
PCNet [99]	-	57.03	61.05
SCHP [38]	-	59.36	-
CDGNet [53]	-	60.30	65.56
UniHCP (direct eval)	65.90	63.80	68.60
UniHCP (finetune)	**65.95**	**63.86**	**69.80**

Table 6. Pedestrian detection evaluation on CrowdHuman val set. Compared with the SOTA, UniHCP achieves comparable mAP and better JI.

Method	mAP	$MR^{-2}(\downarrow)$	JI
DETR [5]	75.9	73.2	74.4
PEDR [49]	91.6	43.7	83.3
Deformable-DETR [105]	91.5	57.0	83.1
Sparse-RCNN [73]	91.3	44.8	81.3
Iter-Deformable-DETR [102]	92.1	**41.5**	84.0
Iter-Sparse-RCNN [102]	**92.5**	41.6	83.3
UniHCP (direct eval)	90.0	46.6	82.2
UniHCP (finetune)	**92.5**	41.6	**85.8**

Table 7. Pose estimation evaluation on COCO, Human3.6M, AI Challenge and OCHuman. Following [88], we report the results on COCO val set, Human3.6M, AI Challenge val set, and OCHuman test set. †denotes the results reported by MMPose [10]. ‡denotes the results achieved using multi-dataset training.

Method	COCO/mAP	H3.6M/EPE(↓)	AIC/mAP	OCHuman/mAP
HRNet-w32† [72]	74.4	9.4	-	-
HRNet-w48† [72]	75.1	7.4	-	-
TokenPose-L/D24 [44]	75.9	-	-	-
HRFormer-B [93]	75.6	-	-	-
ViTPose-B [88]	75.8	-	-	-
ViTPose-B‡ [88]	**77.1**	-	32.0	87.3
UniHCP (direct eval)	76.1	6.9	32.5	**87.4**
UniHCP (finetune)	76.5	**6.6**	**33.6**	N/A

pedestrian attribute recognition, human parsing, pedestrian detection, and pose estimation. We report two kinds of evaluation results of our UniHCP: (1) **direct evaluation**, where the pre-trained model with cross-task shared encoder-decoder weights and task-specific queries are directly used for evaluation on the target dataset, and (2) **finetuning**, where the pretrained UniHCP are first finetuned with the train split of the target dataset and then evaluated.

As observed, the direct evaluation results of UniHCP show promising performance on most human-centric tasks, especially on human parsing and pose estimation tasks, which show better or on-par performance with the State-Of-The-Art (SOTA). The exception is the person ReID task, which observes noticeable performance gaps with the SOTA. We suggest this is due to its huge disparity from other tasks and can be remedied with quick finetuning.

With finetuning, our UniHCP achieves new SOTAs on nine out of the total twelve datasets and on par performance on the rest three datasets, even without task-specific design in architecture or task-specific priors, showing that UniHCP extracts complementary knowledge among human-centric tasks. Concretely, Table 4 shows that in the human attribute recognition task, UniHCP significantly surpasses previous SOTA DAFL [24] by **+3.79%** mA on PA-100K and **+1.20%** mA on RAPv2 datasets, respectively, which indicates that UniHCP well extracts the shared attribute information among using the output unit of global probabilities in the interpreter. Second, UniHCP also pushes

Table 8. Transfer performance on ATR [47], SenseReID [100], Caltech [13], MPII [1] and PETA [12]. Results with †are achieved by using additional data. DE - direct evaluation. FT - finetuning.

Methods	Parsing	ReID	Detection	Pose	Attribute
	ATR	SenseReID	Caltech(↓)	MPII	PETA
SOTA	97.39 [53]	34.6 [100]	46.6 [17]	92.3 [90]	87.07 [24]
SOTA†	-	-	28.8 [17]	**93.3** [88]	-
UniHCP (DE)	-	**46.0**	37.8	-	-
UniHCP (FT)	**97.74**	N/A	**27.2**	93.2	**88.78**

the performance of another important human task, *i.e.*, human parsing, to a new level. Specifically, **+3.45%** mIoU, **+3.56%** mIoU, and **+4.24%** mIoU performance gains are observed on Human3.6M, LIP, and CIHP datasets, respectively. We suggest the newly-added global supervision $\mathcal{L}_{par}$ will help UniHCP to suppress the false prediction on not appeared semantic parts. UniHCP also shows superior performance to previous methods on pose estimation. On person ReID, UniHCP outperforms TransReid [19] on Market1501 and MNE [39] on CUHK03 without the help of any additional camera information and training images during evaluation. For pedestrian detection, our UniHCP achieves **+1.8%** JI performance gain compared with Iter-Deformable-DETR [102] and on-par performance with the Iter-Sparse-RCNN [102] on mAP. These strong performances on diverse datasets across five tasks demonstrate the feasibility and powerfulness of the unified human-centric model and large-scale pretraining.

Table 9. One-shot human parsing and human pose estimation transfer results under different tuning settings. Every method uses only 1 image per class to transfer. We repeat each experiment for 10 times and report the mean and standard deviation.

Methods	Learnable params ratio	Parsing	Pose
		ATR/pACC	MPII/mAP
One-shot finetuning	100%	90.49±1.22	70.6±7.53
One-shot prompt tuning	<1%	93.65±0.77	83.8±5.08
Full-data finetuning	100%	97.74	93.2

Table 10. Comparison of different parameter-sharing schemes. We report the average scores of direct evaluation results on in-pretrain human-centric datasets. "by $\mathbf{t}_t$" denotes sharing decoder and interpreter across task types $\mathbf{t}_t$. For more detailed results on each dataset, please refer to the supplementary.

Methods	Total params.	Shared params.	Shared module			Avg.
			Encoder	Decoder	Task-guided Interpreter	
Baseline	109.32M	109.08M	✓	✓	✓	67.4
(a)	156.17M	105.60M	✓	✓		67.4
(b)	489.67M	91.07M	✓			60.6
(c)	170.83M	109.08M	✓	by $\mathbf{t}_t$	by $\mathbf{t}_t$	65.0

4.2.2 Cross-datasets Transfer Results

As the task-guided interpreter formulates all task requests into four output units, knowledge learned behind these units can be easily transferred to unseen datasets. We conduct evaluations on five datasets that do not appear during pretraining to evaluate the transferability of UniHCP. UniHCP is finetuned to adapt to new datasets except for SenseReID, on which the performance is tested by direct evaluation. As shown in Table 8, UniHCP outperforms existing SOTAs in 4 out of 5 datasets. Specifically, UniHCP achieves **+0.35%** pACC, **+11.4%** top-1, **-1.6%** heavy occluded $\mathrm{MR}^{-2}(\downarrow)$, **+0.1%** mAP, and **+1.71%** mA on ATR, SenseReID, Caltech, MPII, and PETA, respectively. On MPII, UniHCP achieves on-par performance with multi-datasets trained SOTA while improving single-dataset trained SOTA by **+0.9%** mAP. Notably, even without finetuning, UniHCP achieves a **-8.8%** heavy occluded $\mathrm{MR}^{-2}(\downarrow)$ performance gain on single-dataset trained SOTA. Consistent improvements on transfer tasks provide strong support to the decent transferability of UniHCP.

4.2.3 Data-Efficient Transferring

As UniHCP achieves SOTAs on full-data finetuning setting, we further evaluate its potential for transferring to new datasets with extremely scarce training images, *e.g.*, only one image per class for training. As summarized in Table 9, by conducting prompt tuning with one image per class, UniHCL achieves **93.65%** pACC on ATR for parsing and **83.8%** mAP on MPII for pose estimation, respectively. For prompt tuning on ATR, we follow [55]. For prompt tuning on MPII, we only update queries and their associate position embeddings. The prompt tuning results are close to that of the full-data finetuning setting and suppress the results of finetuning the whole model with one image per class for a large margin. Moreover, UniHCP with prompt tuning shows much lower standard deviations than one-shot finetuning on human parsing and pose estimation tasks, verifying that UniHCP learns generic human-centric representation which is beneficial for data-efficient transferring with low computation cost.

4.3. Ablation Study on Weight Sharing

As UniHCP achieves desirable performance on various human-centric tasks while sharing most parameters among different tasks, one problem remains whether more task-specific parameters benefit learning. To answer the question, we ablate three weight sharing variants of UniHCP during pretraining using a 60k-iteration training schedule with 1k batch size. Results in Table 10(b) show that compared to the original UniHCP *i.e.*, the *Baseline*), unifying task-guided interpreters among all tasks resulted in an average performance on par with using specific heads while reducing about **30%** of the parameters. We also note that using task-specific or task-type-specific decoders and interpreters leads to an obvious (**-6.8%** and **-2.4%**, respectively) performance drop on average when compared to the original UniHCP (see results in Table 10(b) and (c)). We speculate that in these ablation settings, complementary human-centric knowledge can not be properly shared among tasks, which leads to performance drops on most tasks.

5. Conclusions

In this work, we present a Unified Model for Human-Centric Perceptions (UniHCP). Based on a simple query-based task formulation, UniHCP can easily handle multiple distinctly defined human-centric tasks simultaneously. Extensive experiments on diverse datasets demonstrate that UniHCP pretrained on a massive collection of human-centric datasets delivers a competitive performance compared with task-specific models. When adapted to specific tasks, UniHCP obtains a series of SOTA performances over a wide spectrum of human-centric benchmarks. Further analysis also demonstrate the capability of UniHCP on parameter and data-efficient transfer and the benefit of weight sharing designs. We hope our work can motivate more future works on developing general human-centric models.

Acknowledgement. This paper was supported by the Australian Research Council Grant DP200103223, Australian Medical Research Future Fund MRFAI000085, CRC-P Smart Material Recovery Facility (SMRF) – Curby Soft Plastics, and CRC-P ARIA - Bionic Visual-Spatial Prosthesis for the Blind.

References

[1] Mykhaylo Andriluka, Leonid Pishchulin, Peter Gehler, and Bernt Schiele. 2d human pose estimation: New benchmark and state of the art analysis. In *Proceedings of the IEEE Conference on computer Vision and Pattern Recognition*, pages 3686–3693, 2014. 7

[2] Vamsi Aribandi, Yi Tay, Tal Schuster, Jinfeng Rao, Huaixiu Steven Zheng, Sanket Vaibhav Mehta, Honglei Zhuang, Vinh Q Tran, Dara Bahri, Jianmo Ni, et al. Ext5: Towards extreme multi-task scaling for transfer learning. *arXiv preprint arXiv:2111.10952*, 2021. 2

[3] FairScale authors. Fairscale: A general purpose modular pytorch library for high performance and large scale training. https://github.com/facebookresearch/fairscale, 2021. 6

[4] Jiale Cao, Yanwei Pang, Jin Xie, Fahad Shahbaz Khan, and Ling Shao. From handcrafted to deep features for pedestrian detection: a survey. *IEEE transactions on pattern analysis and machine intelligence*, 2021. 1

[5] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *European conference on computer vision*, pages 213–229. Springer, 2020. 6, 7

[6] Ting Chen, Saurabh Saxena, Lala Li, Tsung-Yi Lin, David J Fleet, and Geoffrey Hinton. A unified sequence interface for vision tasks. *arXiv preprint arXiv:2206.07669*, 2022. 2, 5

[7] Yinpeng Chen, Xiyang Dai, Mengchen Liu, Dongdong Chen, Lu Yuan, and Zicheng Liu. Dynamic convolution: Attention over convolution kernels. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11030–11039, 2020. 4

[8] Bowen Cheng, Ishan Misra, Alexander G Schwing, Alexander Kirillov, and Rohit Girdhar. Masked-attention mask transformer for universal image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1290–1299, 2022. 4, 5

[9] Xiao Chu, Wei Yang, Wanli Ouyang, Cheng Ma, Alan L Yuille, and Xiaogang Wang. Multi-context attention for human pose estimation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1831–1840, 2017. 1

[10] MMPose Contributors. Openmmlab pose estimation toolbox and benchmark. https://github.com/open-mmlab/mmpose, 2020. 7

[11] Emily E Cust, Alice J Sweeting, Kevin Ball, and Sam Robertson. Machine and deep learning for sport-specific movement recognition: A systematic review of model development and performance. *Journal of sports sciences*, 37(5):568–600, 2019. 1

[12] Yubin Deng, Ping Luo, Chen Change Loy, and Xiaoou Tang. Pedestrian attribute recognition at far distance. In *Proceedings of the 22nd ACM international conference on Multimedia*, pages 789–792, 2014. 7

[13] Piotr Dollar, Christian Wojek, Bernt Schiele, and Pietro Perona. Pedestrian detection: An evaluation of the state of the art. *IEEE transactions on pattern analysis and machine intelligence*, 34(4):743–761, 2011. 7

[14] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 3, 6

[15] Yuying Ge, Ruimao Zhang, Xiaogang Wang, Xiaoou Tang, and Ping Luo. Deepfashion2: A versatile benchmark for detection, pose estimation, segmentation and re-identification of clothing images. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5337–5345, 2019. 6

[16] Ke Gong, Xiaodan Liang, Dongyu Zhang, Xiaohui Shen, and Liang Lin. Look into person: Self-supervised structure-sensitive learning and a new benchmark for human parsing. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 932–940, 2017. 6

[17] Irtiza Hasan, Shengcai Liao, Jinpeng Li, Saad Ullah Akram, and Ling Shao. Generalizable pedestrian detection: The elephant in the room. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11328–11337, 2021. 7

[18] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16000–16009, 2022. 6

[19] Shuting He, Hao Luo, Pichao Wang, Fan Wang, Hao Li, and Wei Jiang. Transreid: Transformer-based object re-identification. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 15013–15022, 2021. 2, 7

[20] Fangzhou Hong, Liang Pan, Zhongang Cai, and Ziwei Liu. Versatile multi-modal pre-training for human-centric perception. *arXiv preprint arXiv:2203.13815*, 2022. 7

[21] Ronghang Hu and Amanpreet Singh. Unit: Multimodal multitask learning with a unified transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1439–1449, 2021. 2, 5

[22] Ruyi Ji, Dawei Du, Libo Zhang, Longyin Wen, Yanjun Wu, Chen Zhao, Feiyue Huang, and Siwei Lyu. Learning semantic neural tree for human parsing. In *European Conference on Computer Vision*, pages 205–221. Springer, 2020. 7

[23] Jian Jia, Xiaotang Chen, and Kaiqi Huang. Spatial and semantic consistency regularizations for pedestrian attribute recognition. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 962–971, 2021. 7

[24] Jian Jia, Naiyu Gao, Fei He, Xiaotang Chen, and Kaiqi Huang. Learning disentangled attribute representations for robust pedestrian attribute recognition. 2022. 7

[25] Sheng Jin, Lumin Xu, Jin Xu, Can Wang, Wentao Liu, Chen Qian, Wanli Ouyang, and Ping Luo. Whole-body human pose estimation in the wild. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, Au-*

gust 23–28, 2020, Proceedings, Part IX 16, pages 196–214. Springer, 2020. 1

[26] Xin Jin, Cuiling Lan, Wenjun Zeng, Guoqiang Wei, and Zhibo Chen. Semantics-aligned representation learning for person re-identification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 11173–11180, 2020. 7

[27] Yannis Kalantidis, Lyndon Kennedy, and Li-Jia Li. Getting the look: clothing recognition and segmentation for automatic product suggestions in everyday photos. In *Proceedings of the 3rd ACM conference on International conference on multimedia retrieval*, pages 105–112, 2013. 1

[28] Mahdi M Kalayeh, Emrah Basaran, Muhittin Gökmen, Mustafa E Kamasak, and Mubarak Shah. Human semantic parsing for person re-identification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1062–1071, 2018. 1, 3

[29] Sameh Khamis, Cheng-Hao Kuo, Vivek K Singh, Vinay D Shet, and Larry S Davis. Joint learning for attribute-consistent person re-identification. In *European conference on computer vision*, pages 134–146. Springer, 2014. 1, 3

[30] Iasonas Kokkinos. Ubernet: Training a universal convolutional neural network for low-, mid-, and high-level vision using diverse datasets and limited memory. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6129–6138, 2017. 2

[31] Alexander Kolesnikov, André Susano Pinto, Lucas Beyer, Xiaohua Zhai, Jeremiah Harmsen, and Neil Houlsby. Uvim: A unified modeling approach for vision with learned guiding codes. *arXiv preprint arXiv:2205.10337*, 2022. 2

[32] Jack Lanchantin, Tianlu Wang, Vicente Ordonez, and Yanjun Qi. General multi-label image classification with transformers. *arXiv preprint arXiv:2011.14027*, 2020. 7

[33] Hei Law and Jia Deng. Cornernet: Detecting objects as paired keypoints. In *Proceedings of the European conference on computer vision (ECCV)*, pages 734–750, 2018. 2

[34] Hei Law, Yun Teng, Olga Russakovsky, and Jia Deng. Cornernet-lite: Efficient keypoint based object detection. *arXiv preprint arXiv:1904.08900*, 2019. 2

[35] Dangwei Li, Zhang Zhang, Xiaotang Chen, and Kaiqi Huang. A richly annotated pedestrian dataset for person retrieval in real surveillance scenarios. *IEEE transactions on image processing*, 28(4):1575–1590, 2019. 6

[36] Feng Li, Hao Zhang, Shilong Liu, Lei Zhang, Lionel M Ni, Heung-Yeung Shum, et al. Mask dino: Towards a unified transformer-based framework for object detection and segmentation. *arXiv preprint arXiv:2206.02777*, 2022. 4

[37] Jiefeng Li, Siyuan Bian, Ailing Zeng, Can Wang, Bo Pang, Wentao Liu, and Cewu Lu. Human pose regression with residual log-likelihood estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11025–11034, 2021. 2

[38] Peike Li, Yunqiu Xu, Yunchao Wei, and Yi Yang. Self-correction for human parsing. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020. 7

[39] Suichan Li, Dapeng Chen, Bin Liu, Nenghai Yu, and Rui Zhao. Memory-based neighbourhood embedding for visual recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6102–6111, 2019. 7

[40] Wanhua Li, Zhexuan Cao, Jianjiang Feng, Jie Zhou, and Jiwen Lu. Label2label: A language modeling framework for multi-attribute learning. In *European Conference on Computer Vision*, pages 562–579. Springer, 2022. 2, 5, 7

[41] Wei Li, Rui Zhao, Tong Xiao, and Xiaogang Wang. Deepreid: Deep filter pairing neural network for person re-identification. In *CVPR*, 2014. 6

[42] Xulin Li, Yan Lu, Bin Liu, Yating Liu, Guojun Yin, Qi Chu, Jinyang Huang, Feng Zhu, Rui Zhao, and Nenghai Yu. Counterfactual intervention feature transfer for visible-infrared person re-identification. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXVI*, pages 381–398. Springer, 2022. 1

[43] Yanghao Li, Hanzi Mao, Ross Girshick, and Kaiming He. Exploring plain vision transformer backbones for object detection. *arXiv preprint arXiv:2203.16527*, 2022. 2, 6

[44] Yanjie Li, Shoukui Zhang, Zhicheng Wang, Sen Yang, Wankou Yang, Shu-Tao Xia, and Erjin Zhou. Tokenpose: Learning keypoint tokens for human pose estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11313–11322, 2021. 2, 7

[45] Jianming Liang, Guanglu Song, Biao Leng, and Yu Liu. Unifying visual perception by dispersible points learning. *arXiv preprint arXiv:2208.08630*, 2022. 2

[46] Xiaodan Liang, Ke Gong, Xiaohui Shen, and Liang Lin. Look into person: Joint body parsing & pose estimation network and a new benchmark. *IEEE transactions on pattern analysis and machine intelligence*, 41(4):871–885, 2018. 1, 3

[47] Xiaodan Liang, Chunyan Xu, Xiaohui Shen, Jianchao Yang, Si Liu, Jinhui Tang, Liang Lin, and Shuicheng Yan. Human parsing with contextualized convolutional neural network. In *Proceedings of the IEEE international conference on computer vision*, pages 1386–1394, 2015. 7

[48] Chunze Lin, Jiwen Lu, and Jie Zhou. Multi-grained deep feature learning for pedestrian detection. In *2018 IEEE International Conference on Multimedia and Expo (ICME)*, pages 1–6. IEEE, 2018. 1, 3

[49] Matthieu Lin, Chuming Li, Xingyuan Bu, Ming Sun, Chen Lin, Junjie Yan, Wanli Ouyang, and Zhidong Deng. Detr for crowd pedestrian detection. *arXiv preprint arXiv:2012.06785*, 2020. 7

[50] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988, 2017. 6

[51] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014. 6

[52] Hao Liu, Jiashi Feng, Meibin Qi, Jianguo Jiang, and Shuicheng Yan. End-to-end comparative attention networks

for person re-identification. *IEEE Transactions on Image Processing*, 26(7):3492–3506, 2017. 5

[53] Kunliang Liu, Ouk Choi, Jianming Wang, and Wonjun Hwang. Cdgnet: Class distribution guided network for human parsing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4473–4482, 2022. 7

[54] Shilong Liu, Lei Zhang, Xiao Yang, Hang Su, and Jun Zhu. Query2label: A simple transformer way to multi-label classification. *arXiv preprint arXiv:2107.10834*, 2021. 2, 7

[55] Xiao Liu, Kaixuan Ji, Yicheng Fu, Zhengxiao Du, Zhilin Yang, and Jie Tang. P-tuning v2: Prompt tuning can be comparable to fine-tuning universally across scales and tasks. *arXiv preprint arXiv:2110.07602*, 2021. 8

[56] Xihui Liu, Haiyu Zhao, Maoqing Tian, Lu Sheng, Jing Shao, Shuai Yi, Junjie Yan, and Xiaogang Wang. Hydraplus-net: Attentive deep features for pedestrian analysis. In *Proceedings of the IEEE international conference on computer vision*, pages 350–359, 2017. 6

[57] Jiasen Lu, Christopher Clark, Rowan Zellers, Roozbeh Mottaghi, and Aniruddha Kembhavi. Unified-io: A unified model for vision, language, and multi-modal tasks. *arXiv preprint arXiv:2206.08916*, 2022. 2, 5

[58] Jiasen Lu, Vedanuj Goswami, Marcus Rohrbach, Devi Parikh, and Stefan Lee. 12-in-1: Multi-task vision and language representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10437–10446, 2020. 2

[59] Weian Mao, Yongtao Ge, Chunhua Shen, Zhi Tian, Xinlong Wang, Zhibin Wang, and Anton van den Hengel. Poseur: Direct human pose regression with transformers. *arXiv preprint arXiv:2201.07412*, 2022. 2

[60] Tewodros Legesse Munea, Yalew Zelalem Jembre, Halefom Tekle Weldegebriel, Longbiao Chen, Chenxi Huang, and Chenhui Yang. The progress of human pose estimation: a survey and taxonomy of models applied in 2d human pose estimation. *IEEE Access*, 8:133330–133348, 2020. 1

[61] Xuecheng Nie, Jiashi Feng, and Shuicheng Yan. Mutual learning to adapt for joint human parsing and pose estimation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 502–517, 2018. 1, 3

[62] Wanli Ouyang, Xingyu Zeng, and Xiaogang Wang. Learning mutual visibility relationship for pedestrian detection with a deep model. *International Journal of Computer Vision*, 120:14–27, 2016. 1

[63] Wanli Ouyang, Hui Zhou, Hongsheng Li, Quanquan Li, Junjie Yan, and Xiaogang Wang. Jointly learning deep features, deformable parts, occlusion and classification for pedestrian detection. *IEEE transactions on pattern analysis and machine intelligence*, 40(8):1874–1887, 2017. 1

[64] Seyoung Park, Bruce Xiaohan Nie, and Song-Chun Zhu. Attribute and-or grammar for joint parsing of human pose, parts and attributes. *IEEE transactions on pattern analysis and machine intelligence*, 40(7):1555–1569, 2017. 1

[65] Sida Peng, Wen Jiang, Huaijin Pi, Xiuli Li, Hujun Bao, and Xiaowei Zhou. Deep snake for real-time instance segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8533–8542, 2020. 2

[66] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, Peter J Liu, et al. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, 21(140):1–67, 2020. 2

[67] Hamid Rezatofighi, Nathan Tsoi, JunYoung Gwak, Amir Sadeghian, Ian Reid, and Silvio Savarese. Generalized intersection over union: A metric and a loss for bounding box regression. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 658–666, 2019. 6

[68] Pierre Sermanet, David Eigen, Xiang Zhang, Michaël Mathieu, Rob Fergus, and Yann LeCun. Overfeat: Integrated recognition, localization and detection using convolutional networks. *arXiv preprint arXiv:1312.6229*, 2013. 2

[69] Shuai Shao, Zijian Zhao, Boxun Li, Tete Xiao, Gang Yu, Xiangyu Zhang, and Jian Sun. Crowdhuman: A benchmark for detecting human in a crowd. *arXiv preprint arXiv:1805.00123*, 2018. 6

[70] Noam Shazeer and Mitchell Stern. Adafactor: Adaptive learning rates with sublinear memory cost. In *International Conference on Machine Learning*, pages 4596–4604. PMLR, 2018. 6

[71] Chi Su, Fan Yang, Shiliang Zhang, Qi Tian, Larry Steven Davis, and Wen Gao. Multi-task learning with low rank attribute embedding for multi-camera person re-identification. *IEEE transactions on pattern analysis and machine intelligence*, 40(5):1167–1181, 2017. 1, 3

[72] Ke Sun, Bin Xiao, Dong Liu, and Jingdong Wang. Deep high-resolution representation learning for human pose estimation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5693–5703, 2019. 7

[73] Peize Sun, Rufeng Zhang, Yi Jiang, Tao Kong, Chenfeng Xu, Wei Zhan, Masayoshi Tomizuka, Lei Li, Zehuan Yuan, Changhu Wang, et al. Sparse r-cnn: End-to-end object detection with learnable proposals. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14454–14463, 2021. 7

[74] Yifan Sun, Changmao Cheng, Yuhan Zhang, Chi Zhang, Liang Zheng, Zhongdao Wang, and Yichen Wei. Circle loss: A unified perspective of pair similarity optimization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6398–6407, 2020. 2

[75] Chufeng Tang, Lu Sheng, Zhaoxiang Zhang, and Xiaolin Hu. Improving pedestrian attribute recognition with weakly-supervised multi-scale attribute-specific localization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4997–5006, 2019. 5

[76] Yi Tang, Baopu Li, Min Liu, Boyu Chen, Yaonan Wang, and Wanli Ouyang. Autopedestrian: An automatic data augmentation and loss function search scheme for pedestrian detection. *IEEE transactions on image processing*, 30:8483–8496, 2021. 1

[77] Yonglong Tian, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Pedestrian detection aided by deep learning semantic tasks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5079–5087, 2015. 1, 3

[78] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 4

[79] Guan'an Wang, Shuo Yang, Huanyu Liu, Zhicheng Wang, Yang Yang, Shuliang Wang, Gang Yu, Erjin Zhou, and Jian Sun. High-order information matters: Learning relation and topology for occluded person re-identification. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6449–6458, 2020. 7

[80] Peng Wang, An Yang, Rui Men, Junyang Lin, Shuai Bai, Zhikang Li, Jianxin Ma, Chang Zhou, Jingren Zhou, and Hongxia Yang. Ofa: Unifying architectures, tasks, and modalities through a simple sequence-to-sequence learning framework. In *International Conference on Machine Learning*, pages 23318–23340. PMLR, 2022. 2

[81] Wenguan Wang, Hailong Zhu, Jifeng Dai, Yanwei Pang, Jianbing Shen, and Ling Shao. Hierarchical human parsing with typed part-relation reasoning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8929–8939, 2020. 1

[82] Yingming Wang, Xiangyu Zhang, Tong Yang, and Jian Sun. Anchor detr: Query design for transformer-based detector. In *Proceedings of the AAAI conference on artificial intelligence*, volume 36, pages 2567–2575, 2022. 4

[83] Jiahong Wu, He Zheng, Bo Zhao, Yixin Li, Baoming Yan, Rui Liang, Wenjia Wang, Shipei Zhou, Guosen Lin, Yanwei Fu, et al. Large-scale datasets for going deeper in image understanding. In *2019 IEEE International Conference on Multimedia and Expo (ICME)*, pages 1480–1485. IEEE, 2019. 6

[84] Bin Xiao, Haiping Wu, and Yichen Wei. Simple baselines for human pose estimation and tracking. In *Proceedings of the European conference on computer vision (ECCV)*, pages 466–481, 2018. 2

[85] Enze Xie, Peize Sun, Xiaoge Song, Wenhai Wang, Xuebo Liu, Ding Liang, Chunhua Shen, and Ping Luo. Polarmask: Single shot instance segmentation with polar representation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12193–12202, 2020. 2

[86] Jing Xu, Rui Zhao, Feng Zhu, Huaming Wang, and Wanli Ouyang. Attention-aware compositional network for person re-identification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2119–2128, 2018. 1

[87] Shilin Xu, Xiangtai Li, Jingbo Wang, Guangliang Cheng, Yunhai Tong, and Dacheng Tao. Fashionformer: A simple, effective and unified baseline for human fashion segmentation and recognition. *arXiv preprint arXiv:2204.04654*, 2022. 1, 2, 3

[88] Yufei Xu, Jing Zhang, Qiming Zhang, and Dacheng Tao. Vitpose: Simple vision transformer baselines for human pose estimation. *arXiv preprint arXiv:2204.12484*, 2022. 2, 6, 7

[89] Qize Yang, Ancong Wu, and Wei-Shi Zheng. Person re-identification by contour sketch under moderate clothing change. *IEEE transactions on pattern analysis and machine intelligence*, 43(6):2029–2046, 2019. 6

[90] Sen Yang, Zhibin Quan, Mu Nie, and Wankou Yang. Transpose: Keypoint localization via transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11802–11812, 2021. 7

[91] Wei Yang, Shuang Li, Wanli Ouyang, Hongsheng Li, and Xiaogang Wang. Learning feature pyramids for human pose estimation. In *proceedings of the IEEE international conference on computer vision*, pages 1281–1290, 2017. 1

[92] Lu Yuan, Dongdong Chen, Yi-Ling Chen, Noel Codella, Xiyang Dai, Jianfeng Gao, Houdong Hu, Xuedong Huang, Boxin Li, Chunyuan Li, et al. Florence: A new foundation model for computer vision. *arXiv preprint arXiv:2111.11432*, 2021. 5

[93] Yuhui Yuan, Rao Fu, Lang Huang, Weihong Lin, Chao Zhang, Xilin Chen, and Jingdong Wang. Hrformer: High-resolution vision transformer for dense predict. *Advances in Neural Information Processing Systems*, 34:7281–7293, 2021. 2, 7

[94] Xiaohua Zhai, Alexander Kolesnikov, Neil Houlsby, and Lucas Beyer. Scaling vision transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12104–12113, 2022. 2, 6

[95] Xiaohua Zhai, Xiao Wang, Basil Mustafa, Andreas Steiner, Daniel Keysers, Alexander Kolesnikov, and Lucas Beyer. Lit: Zero-shot transfer with locked-image text tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18123–18133, 2022. 6

[96] Shanshan Zhang, Rodrigo Benenson, Mohamed Omran, Jan Hosang, and Bernt Schiele. How far are we from solving pedestrian detection? In *Proceedings of the iEEE conference on computer vision and pattern recognition*, pages 1259–1267, 2016. 2

[97] Shifeng Zhang, Yiliang Xie, Jun Wan, Hansheng Xia, Stan Z Li, and Guodong Guo. Widerperson: A diverse dataset for dense pedestrian detection in the wild. *IEEE Transactions on Multimedia*, 22(2):380–393, 2019. 1

[98] Song-Hai Zhang, Ruilong Li, Xin Dong, Paul Rosin, Zixi Cai, Xi Han, Dingcheng Yang, Haozhi Huang, and Shi-Min Hu. Pose2seg: Detection free human instance segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 889–898, 2019. 1, 3

[99] Xiaomei Zhang, Yingying Chen, Bingke Zhu, Jinqiao Wang, and Ming Tang. Part-aware context network for human parsing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8971–8980, 2020. 7

[100] Haiyu Zhao, Maoqing Tian, Shuyang Sun, Jing Shao, Junjie Yan, Shuai Yi, Xiaogang Wang, and Xiaoou Tang. Spindle net: Person re-identification with human body region

guided feature decomposition and fusion. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1077–1085, 2017. 7

[101] Rui Zhao, Wanli Ouyang, and Xiaogang Wang. Unsupervised salience learning for person re-identification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3586–3593, 2013. 1

[102] Anlin Zheng, Yuang Zhang, Xiangyu Zhang, Xiaojuan Qi, and Jian Sun. Progressive end-to-end object detection in crowded scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 857–866, 2022. 2, 6, 7

[103] Zhedong Zheng, Liang Zheng, and Yi Yang. A discriminatively learned cnn embedding for person reidentification. *ACM transactions on multimedia computing, communications, and applications (TOMM)*, 14(1):1–20, 2017. 5

[104] Jianqing Zhu, Shengcai Liao, Zhen Lei, and Stan Z Li. Multi-label convolutional neural network based pedestrian attribute classification. *Image and Vision Computing*, 58:224–229, 2017. 2

[105] Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable detr: Deformable transformers for end-to-end object detection. *arXiv preprint arXiv:2010.04159*, 2020. 7

[106] Xizhou Zhu, Jinguo Zhu, Hao Li, Xiaoshi Wu, Hongsheng Li, Xiaohua Wang, and Jifeng Dai. Uni-perceiver: Pre-training unified architecture for generic perception for zero-shot and few-shot tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16804–16815, 2022. 2