UniHCP: A Unified Model for Human-Centric Perceptions
Supplementary Materials

A. One-shot Transfer Results

In this section, we provide details and full results for one-shot fine-tuning and prompt tuning on human parsing and pose estimation. For each experiment, we sample ten sets of images with different random seeds; we also grid search on both iterations and learning rates until performance converges. The reported results are based on the best config found for each setting.

Data sampling. In one-shot transfer experiments, only one image per class is used for a task [9]. Table 1 shows the number of sampled images on one-shot transfer tasks. Note that in UniHCP, classification tasks are multi-label classification for human parsing, pose estimation, and attribute recognition, where each query performs binary classification via the global probability unit. Therefore, we also make sure the presence of cases where a class is absent is covered in our samples. Such handling avoids the query simply learning to output 1 when the corresponding class always presents within the sampled images. On the other hand, when a class does appear in most of the images, *e.g.*, all keypoint joints in pose estimation or the background class in human parsing, we are able to achieve reasonably good results without such handling, thus we do not intentionally sample "not present" case for keypoint joints and background class in our experiments.

Table 1. Number of sampled images on one-shot transfer tasks. As we can easily find pose samples with all keypoint joints present in the image and do not have to consider the case where a joint is absent as explained above, we only need one sample to perform one-shot transfer on pose estimation.

	Parsing/ATR [19]	Pose/MPII [2]
Sampled images	$3 \sim 4$	1

Number of tunable parameters. For fine-tuning settings, all parameters are tuned. For prompt tuning on human parsing, we follow [21, 44] and add learnable prompt tokens in decoder layers. We update queries, additional prompt tokes, and layer normalization weights. For prompt tuning on pose estimation, we only update queries and their associate position embeddings. Table. 2 shows the number of parameters of each learnable component in prompt tuning.

Table 2. Number of tunable parameters for prompt tuning on human parsing, pose estimation, and pedestrian attribute recognition.

	Parsing/ATR	Pose/MPII	Attribute/PETA
Query	9216	8704	35840
Deep prompt [21, 44]	32256	-	-
LN [44]	16128	-	-
Learnable parameter ratio	0.053%	0.008%	0.033%

A.1. Human Parsing

Table 3 shows the full one-shot results for fine-tuning and prompt tuning on human parsing.

Table 3. One-shot human parsing results on ATR, evaluated by pACC. FT - finetuning, PT - prompt tuning.

	1	2	3	4	5	6	7	8	9	10	avg.	std.
FT	91.28	91.21	90.75	87.90	91.48	92.14	89.67	89.36	90.67	90.48	90.49	1.22
PT	93.31	92.99	93.41	92.31	93.89	95.16	93.41	93.81	94.01	94.23	93.65	0.77

A.2. Pose Estimation

Table 4 shows the full one-shot results for fine-tuning and prompt tuning on pose estimation.

Table 4. One-shot pose estimation results on MPII, evaluated by mAP. FT - finetuning, PT - prompt tuning.

	1	2	3	4	5	6	7	8	9	10	avg.	std.
FT	64.18	78.68	78.18	60.52	73.71	67.80	70.44	57.20	79.26	76.07	70.60	7.53
PT	87.32	86.13	87.33	77.44	85.91	81.16	88.29	71.97	87.45	85.29	83.83	5.08

B. Few-shot Transfer Results for Pedestrian Attribute Recognition

In this section, we provide the few-shot transfer results for finetuning and prompt tuning on pedestrian attribute recognition. Different from human parsing and pose estimation datasets, the targeted downstream pedestrian attribute recognition dataset PETA [6] contains images from ten different domains. Randomly sampling only one image per class may mislead the queries to extract domain-biased representation, and we found the one-shot result is poor for both finetuning and prompt tuning under this setting. Therefore, we loosen the data constraint to few-shot setting to evaluate the data-efficient transfer performance on pedestrian attribute recognition. Similar to one-shot experiments, we conduct the experiment on ten different sets of images, grid search on hyperparameters, and report results based on the best config for each setting.

Data sampling. PETA has ten different domains and 35 different attributes. For each domain, we sample images until both "present" and "not present" cases appeared at least once for each attribute; we sample multiple times and take the one with the least samples as a few-shot dataset. It takes **68** $\sim$ **75** samples to satisfy this constraint in our experiments.

Number of tunable parameters. All parameters are tuned for finetuning. For prompt tuning, we only update queries and their associate position embeddings. The number of tunable parameters in prompt tuning is shown in Table 2.

Results. Table 5 shows the full few-shot results for pedestrian attribute recognition; prompt tuning achieves better performance with a smaller standard deviation.

Table 5. Few-shot pedestrian attribute recognition results on PETA, evaluated by mA. FT - finetuning, PT - prompt tuning.

	1	2	3	4	5	6	7	8	9	10	avg.	std.
FT	59.41	61.03	59.17	61.73	59.11	61.30	59.31	60.46	60.30	61.38	60.32	0.96
PT	61.71	61.53	62.41	62.52	61.19	63.29	61.58	61.66	62.94	63.12	62.20	0.72

C. Full Ablation Results on Weight Sharing

In Table 6, we provide full results for the ablation study in Section 4.3. UniHCP achieves comparable performance with using task-specific interpreters while sharing most of the parameters among different human-centric tasks.

D. Additional Architecture Details

D.1. Task-guided Interpreter

Since the task-guided interpreter decodes each query token independently, we formulate the interpreter design by describing the generation of each output unit element $y \in \mathbf{Y}$ from query token $q \in Q^t$.

Feature vector unit $\mathbf{Y}_f$: as the query token is already in a feature space, we do not add any additional postprocessing. we have $y_f = q, y_f \in \mathbb{R}^C$, where C is the output dimension of the decoder.

Global probability unit $\mathbf{Y}_p$: we apply a 1-lyr MLP (i.e. linear projector) followed by a `sigmoid` function σ, on top of query token q to yield global probability $y_p \in \mathbb{R}^1$.

Local probability map unit $\mathbf{Y}_m$: We denoted visual tokens from the encoder as $\mathbf{F} \in \mathbb{R}^{C_e \times H/16 \times W/16}$, where C_e denotes the output dimension of the encoder, $H \times W$ denoted the original image size and 16 is the patch size of ViT-B. $\mathbf{F}$ is forwarded through two consecutive deconvolution layers with hidden dimension C_e to upscale the feature map to $\tilde{\mathbf{F}} \in \mathbb{R}^{C \times H/4 \times W/4}$. The query token q is applied with a 3-lyr MLP to get the embedding $\tilde{q} \in \mathbb{R}^C$. We obtain the final probability logit map $y_m \in \mathbb{R}^{H/4 \times W/4}$ by calculating the dot product between $\tilde{q}$ and $\tilde{\mathbf{F}}$, broadcasted in the spatial dimensions.

Bounding box unit $\mathbf{Y}_{bbox}$: Similar with [43], the query token q is applied with a 3-lyr MLP to get the box offset prediction logits $\tilde{q} = [\alpha_{\nabla cx}, \alpha_{\nabla cx}, \alpha_h, \alpha_w], \tilde{q} \in \mathbb{R}^4$. With its associated anchor point $\mathcal{A}_q = [cx, cy]$, we yield the final box prediction $y_{bbox} = [\sigma(\alpha_{\nabla cx} + \sigma^{-1}(cx)), \sigma(\alpha_{\nabla cy} + \sigma^{-1}(cy)), \sigma(\alpha_h), \sigma(\alpha_w)]$, where σ^{-1} denotes the inversed `sigmoid` function.

D.2. Positional Embedding for Encoder

The positional embedding for the encoder is shared across tasks and is interpolated according to the spatial size of the patch projected input image. The maximum image resolution during training is 1333×800 (or 800×1333), which will then be padded to 1344×800 before patch projection (rounded up to be divisible by patch size 16). Thus, the maximum H/W dimension for images after patch projection is 84. Accordingly, we set the number of tokens for learnable positional embedding to $84 \times 84 = 7056$.

D.3. Decoder Positional Embedding Projector

The positional embedding projector $proj$ follows the design in [30]. The coordinate is first encoded by sine-cosine position encoding function [27] and then projected by a simple 2-Layer MLP.

D.4. Auxiliary Loss:

Apart from the loss for Q_L^t after L-th decoder block, we also add auxiliary losses to intermediate queries for pose estimation, human parsing, and pedestrian detection following the best practices in [4, 5, 38]. For pose estimation and human parsing, the auxiliary loss is calculated on Q_l^t for $l \in \{0, ..., L-1\}$ following [5]. For pedestrian detection, the auxiliary loss is calculated on Q_l^t for $l \in \{1, ..., L-1\}$ following [4, 38].

D.5. Pose Estimation

For pose estimation, we set $\lambda_{par} = 0.001$. During the inference time, when the metric requires a confidence score for keypoint filtering and NMS (e.g. mAP), we additionally multiply the global probability prediction y_p to the confidence score and lower the visibility threshold to 0.05 accordingly.

Table 6. Detailed results for different parameter-sharing methods.

Methods	Shared module			Parsing/mIoU			ReID/mAP			Detection//mAP	Pose/mAP			Attribute/mA		Average
	Encoder	Decoder	Task heads	H3.6	LIP	CIHP	Market1501	MSMT17	CUHK03	CrowdHuman	COCO	AIC	OCHuman	PA-100K	RAPv2	
Baseline	✓	✓	✓	64.6	61.9	64.4	82.1	59.0	59.9	80.5	73.5	29.0	77.0	81.0	75.3	67.4
(a)	✓	✓		65.4	61.6	64.1	82.7	59.9	62.1	82.2	73.5	27.9	74.9	81.3	73.0	67.4
(b)	✓			64.2	59.8	61.1	76.9	51.3	51.0	36.2	71.3	25.6	69.0	81.9	78.7	60.6
(c)	✓	by $\mathbf{t}_t$	by $\mathbf{t}_t$	64.1	61.6	63.0	79.4	54.4	56.3	68.4	72.7	26.8	71.3	82.1	79.8	65.0

E. Additional Training Details

Loss Weight $w_{\mathcal{D}}$: for dataset $\mathcal{D}'$, its loss weight $w_{\mathcal{D}'}$ is calculated as follows:

$$w_{\mathcal{D}'} = \frac{b_{\mathcal{D}'} w_{\mathbf{t}_{\mathcal{D}'}}}{\sum_{\mathcal{D}\in\mathbb{D}} b_{\mathcal{D}} w_{\mathbf{t}_{\mathcal{D}}}}, \tag{1}$$

where $b_{\mathcal{D}}$ denotes the batch size allocated to dataset $\mathcal{D}$ and $w_{\mathbf{t}_{\mathcal{D}}}$ denotes the sample weight for task type $\mathbf{t}_{\mathcal{D}}$. The loss weight is normalized so that it only controls the relative weight for each dataset. Samples belonging to the same task type are treated with equal importance. Since different task types have different loss functions, image input resolution, number of samples, and convergence pattern, their loss weight should be set differently. For a reasonable loss weight trade-off between tasks, we gradually add task types one at a time in a small 10k iteration joint training setup and sweep sample weights for the newly added task type. After the hyperparameter search, we set $w_{reid} = 10, w_{par} = 1 \times 10^{-2}, w_{seg} = 5, w_{pose} = 2 \times 10^3, w_{peddet} = 2$.

Dataset-wise Configurations: we provide detailed dataset-wise training configurations in Table 7. In addition to these training datasets, downstream datasets are ATR [19], SenseReID [37], Caltech [7], MPII [2] and PETA [6].

F. Additional Finetuning Details

We provide major finetuning configurations in Table 8; other settings are identical to the training config.

G. Ethics

In this work, we proposed a model to unify multiple human-centric tasks and trained the model on a huge collection of public and widely used human-centric datasets. We acknowledge that the resulting model demonstrates good performance on public ReID benchmarks and thus may be associated with potential identity information leaking without consent if misused. Therefore, the pretrained model will be released only on a case-by-case basis, and the requester must sign an agreement limiting the usage to research purposes only. In addition, the pretrained query tokens for ReID tasks will be excluded from the model release.

Table 7. UniHCP joint training setup. †the batch size for pedestrian detection is reduced due to high GPU consumption.

Task Type	Dataset $\mathcal{D}$	Batch Size $b_{\mathcal{D}}$	Batch Size per GPU	Dataset Epoch	$b_{\mathcal{D}} w_{\mathbf{t}_{\mathcal{D}}}$	GPUs	Number of Samples	Sample Weight $w_{\mathbf{t}_{\mathcal{D}}}$
Pedestrian Detection	CrowdHuman [24] EuroCity Persons [3] CityPersons [34] WiderPerson [35] WiderPedestrian [23] COCO-Person [20]	212†	4	130.19†	424	53	170,687	2
Person ReID	Market-1501 [39] CUHK03 [17] MSMT17 [31]	96	96	199.06	960	1	50,549	10
	DGMarket [41] PRCC [33] LaST [25]	415	415	200.04	4150	1	217,453	10
Pose Estimation	COCO-Pose [20]	286	286	200.1	572000	1	149,813	2000
	AI Challenger [32]	720	240	199.46	1440000	3	378,352	2000
	PoseTrack [1]	185	185	199.55	3710000	1	97,174	2000
	MHP [15]	77	77	199.59	154000	1	40,437	2000
	3DPW [29]	131	131	199.98	262000	1	68,663	2000
	UpennAction [36]	66	66	200.66	132000	1	34,475	2000
	JRDB-Pose [28]	266	266	200.03	532000	1	139,385	2000
	Halpe [8]	79	79	200.69	158000	1	41,263	2000
	Human3.6M (pose) [13]	596	298	200.11	1192000	2	312,187	2000
Human Parsing	LIP [12]	58	58	199.57	290	1	30,462	5
	CIHP [11]	54	54	200.14	270	1	28,280	5
	Deep fashion [10]	364	52	198.75	1820	7	191,961	5
	VIP [42]	35	35	198.63	175	1	18,469	5
	ModaNet [40]	100	50	200.62	500	2	52,245	5
	Human3.6M (parse) [13]	120	40	200.71	600	3	62,668	5
Pedestrian Attribute Recognition	PA-100K [22]	172	172	200.32	1.72	1	90,000	0.01
	RAPv2 [14]	130	130	200.55	1.3	1	67,943	0.01
	HARDHC [18]	54	54	199.75	0.54	1	28,336	0.01
	UAV-Human [16]	31	31	200.78	0.31	1	16,183	0.01
	Parse27k [26]	52	52	198.33	0.52	1	27,482	0.01
	Market-1501 (attribute) [39]	25	25	202.57	0.25	1	12,936	0.01
Summary	/	Total: 4324	/	Avg.: 200.00 (excluding det.)	/	Total: 88	Total: 2,327,403	/

Table 8. Detailed finetuning configs for human-centric tasks.

Task Type	Dataset	Learning Rate	Batch Size	Iterations	Backbone lr Multiplier	Drop Path Rate	Layer Decay Rate	Weight Decay
Pedestrian Detection	CrowdHuman [24]	2.00E-04	32	160k	1.0	0.2	0.75	0.05
	Caltech [7]	1.00E-05	32	30k	0.1	0.2	0.75	0.05
Person ReID	Market-1501 [39]	1.00E-04	64	40k	0.4	0.05	0.75	0.5
	CUHK03 [17]	5.00E-05	64	20k	0.9	0.1	0.95	0.5
	MSMT17 [31]	1.00E-04	64	40k	0.9	0.05	0.75	0.5
Pose Estimation	COCO-Pose [20]	1.00E-04	512	20k	0.9	0.25	0.75	0.05
	AI Challenger [32]	1.00E-03	512	10k	0.9	0.2	0.75	0.05
	Human3.6M (Pose) [13]	5.00E-06	512	10k	0.9	0.3	0.75	0.05
	MPII [2]	7.00E-05	512	7.5k	0.9	0.3	0.75	0.05
Human Parsing	LIP [12]	5.00E-05	64	30k	1.0	0.3	0.75	0.05
	CIHP [11]	1.00E-04	64	35k	1.4	0.3	0.65	0.05
	Human3.6M (parse) [13]	1.00E-05	64	25k	1.3	0.3	0.85	0.05
	ATR [19]	1.00E-04	64	15k	0.7	0.3	0.85	0.05
Pedestrian Attribute Recognition	PA-100K [22]	3.00E-03	128	10k	0.05	0.2	0.85	0.05
	RAPv2 [14]	5.00E-04	128	4k	0.5	0.3	0.75	0.05
	PETA [6]	1.00E-03	128	20k	0.2	0.3	0.75	0.05

References

[1] Mykhaylo Andriluka, Umar Iqbal, Eldar Insafutdinov, Leonid Pishchulin, Anton Milan, Juergen Gall, and Bernt Schiele. Posetrack: A benchmark for human pose estimation and tracking. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5167–5176, 2018. 4

[2] 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. 1, 3, 4

[3] Markus Braun, Sebastian Krebs, Fabian Flohr, and Dariu M Gavrila. Eurocity persons: A novel benchmark for person detection in traffic scenes. *IEEE transactions on pattern analysis and machine intelligence*, 41(8):1844–1861, 2019. 4

[4] 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. 2

[5] 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. 2

[6] 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. 1, 3, 4

[7] 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. 3, 4

[8] Hao-Shu Fang, Jiefeng Li, Hongyang Tang, Chao Xu, Haoyi Zhu, Yuliang Xiu, Yong-Lu Li, and Cewu Lu. Alphapose: Whole-body regional multi-person pose estimation and tracking in real-time. *arXiv preprint arXiv:2211.03375*, 2022. 4

[9] Li Fei-Fei, Robert Fergus, and Pietro Perona. One-shot learning of object categories. *IEEE transactions on pattern analysis and machine intelligence*, 28(4):594–611, 2006. 1

[10] 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. 4

[11] Ke Gong, Xiaodan Liang, Yicheng Li, Yimin Chen, Ming Yang, and Liang Lin. Instance-level human parsing via part grouping network. In *Proceedings of the European conference on computer vision (ECCV)*, pages 770–785, 2018. 4

[12] 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. 4

[13] Catalin Ionescu, Dragos Papava, Vlad Olaru, and Cristian Sminchisescu. Human3.6m: Large scale datasets and predictive methods for 3d human sensing in natural environments. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(7):1325–1339, jul 2014. 4

[14] 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. 4

[15] Jianshu Li, Jian Zhao, Yunchao Wei, Congyan Lang, Yidong Li, Terence Sim, Shuicheng Yan, and Jiashi Feng. Multiple-human parsing in the wild. *arXiv preprint arXiv:1705.07206*, 2017. 4

[16] Tianjiao Li, Jun Liu, Wei Zhang, Yun Ni, Wenqian Wang, and Zhiheng Li. Uav-human: A large benchmark for human behavior understanding with unmanned aerial vehicles. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16266–16275, 2021. 4

[17] Wei Li, Rui Zhao, Tong Xiao, and Xiaogang Wang. Deepreid: Deep filter pairing neural network for person re-identification. In *CVPR*, 2014. 4

[18] Yining Li, Chen Huang, Chen Change Loy, and Xiaoou Tang. Human attribute recognition by deep hierarchical contexts. In *European Conference on Computer Vision*, 2016. 4

[19] 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. 1, 3, 4

[20] 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. 4

[21] 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. 1

[22] 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. 4

[23] Chen Change Loy, Dahua Lin, Wanli Ouyang, Yuanjun Xiong, Shuo Yang, Qingqiu Huang, Dongzhan Zhou, Wei Xia, Quanquan Li, Ping Luo, et al. Wider face and pedestrian challenge 2018: Methods and results. *arXiv preprint arXiv:1902.06854*, 2019. 4

[24] 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. 4

[25] Xiujun Shu, Xiao Wang, Xianghao Zang, Shiliang Zhang, Yuanqi Chen, Ge Li, and Qi Tian. Large-scale spatio-temporal person re-identification: Algorithms and benchmark. *IEEE Transactions on Circuits and Systems for Video Technology*, 2021. 4

[26] Patrick Sudowe, Hannah Spitzer, and Bastian Leibe. Person attribute recognition with a jointly-trained holistic cnn model. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 87–95, 2015. 4

[27] 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. 2

[28] Edward Vendrow, Duy Tho Le, and Hamid Rezatofighi. Jrdb-pose: A large-scale dataset for multi-person pose estimation and tracking. *arXiv preprint arXiv:2210.11940*, 2022. 4

[29] Timo von Marcard, Roberto Henschel, Michael Black, Bodo Rosenhahn, and Gerard Pons-Moll. Recovering accurate 3d human pose in the wild using imus and a moving camera. In *European Conference on Computer Vision (ECCV)*, sep 2018. 4

[30] 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. 2

[31] Longhui Wei, Shiliang Zhang, Wen Gao, and Qi Tian. Person transfer gan to bridge domain gap for person reidentification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 79–88, 2018. 4

[32] 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. 4

[33] Qize Yang, Ancong Wu, and Wei-Shi Zheng. Person reidentification by contour sketch under moderate clothing change. *IEEE transactions on pattern analysis and machine intelligence*, 43(6):2029–2046, 2019. 4

[34] Shanshan Zhang, Rodrigo Benenson, and Bernt Schiele. Citypersons: A diverse dataset for pedestrian detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3213–3221, 2017. 4

[35] 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. 4

[36] Weiyu Zhang, Menglong Zhu, and Konstantinos G Derpanis. From actemes to action: A strongly-supervised representation for detailed action understanding. In *Proceedings of the IEEE international conference on computer vision*, pages 2248–2255, 2013. 4

[37] 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. 3

[38] 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

[39] Liang Zheng, Liyue Shen, Lu Tian, Shengjin Wang, Jingdong Wang, and Qi Tian. Scalable person re-identification: A benchmark. In *Proceedings of the IEEE international conference on computer vision*, pages 1116–1124, 2015. 4

[40] Shuai Zheng, Fan Yang, M Hadi Kiapour, and Robinson Piramuthu. Modanet: A large-scale street fashion dataset with polygon annotations. In *Proceedings of the 26th ACM international conference on Multimedia*, pages 1670–1678, 2018. 4

[41] Zhedong Zheng, Xiaodong Yang, Zhiding Yu, Liang Zheng, Yi Yang, and Jan Kautz. Joint discriminative and generative learning for person re-identification. In *proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2138–2147, 2019. 4

[42] Qixian Zhou, Xiaodan Liang, Ke Gong, and Liang Lin. Adaptive temporal encoding network for video instance-level human parsing. In *Proceedings of the 26th ACM international conference on Multimedia*, pages 1527–1535, 2018. 4

[43] 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. 2

[44] Xizhou Zhu, Jinguo Zhu, Hao Li, Xiaoshi Wu, Hongsheng Li, Xiaohua Wang, and Jifeng Dai. Uni-perceiver: Pretraining 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. 1