HumanBench: Towards General Human-centric Perception with Projector Assisted Pretraining

Shixiang Tang[1,4*], Cheng Chen[4*], Qingsong Xie[4], Meilin Chen[2,4], Yizhou Wang[2,4], Yuanzheng Ci[1], Lei Bai[3†], Feng Zhu[4], Haiyang Yang[4], Li Yi[4], Rui Zhao[4,5], 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, China
stan3906@uni.sydney.edu.au, chencheng1@sensetime.com, bailei@pjlab.org.cn

1. Details of HumanBench

In the main text, we briefly introduce the number of images and number of tasks in the pretraining dataset of HumanBench. For the evaluation of HumanBench, we introduce the evaluation scenario and evaluation protocols. In this section, we present detailed information on the pretraining dataset and evaluation dataset and discuss the ethical issues of these datasets.

1.1. Dataset Statistics of HumanBench

HumanBench collects 37 publicly available datasets of 5 human-centric tasks, including person ReID, human parsing, pose estimation, pedestrian detection, and pedestrian attribute. More details can be seen in Table 1. The existing distribution of datasets includes large numbers of human-centric cropped images in ReID, video frames in person pose estimation, and human parsing. In particular, to avoid information reduction, we select a single frame from every 8 video frames. Particularly, except for using training images in all datasets, we also use all/partial test images in some datasets. Specifically, for the person Reid task, we use all test images in LaST and partial test images in the PRCC dataset; for the human parsing task, we only use train images and publicly released images in DeepFashion($\sim$half of the dataset reported in [21]). For the pedestrian detection dataset, we remove the images in which there is no person. For the pose estimation datasets, we only use train images. For the pedestrian attribute recognition dataset, we only use partial test images in the UAV-Human dataset and do not contain test images in other pedestrian attribute recognition datasets. All the images in the pretraining dataset have been de-duplicated with the testing datasets to be a meaningful benchmark of our HumanBench.

1.2. Discussion of Ethical Issues

The usage of HumanBench might bring several risks, such as privacy, and problematic content. We discuss these risks and their mitigation strategies as follows.

Copyright. All images in this paper and dataset are collected by publicly available. We claim the dataset:

- Copy and redistribute the material in any medium or format.
- Remix, transform and build on the material for any purpose, even commercially.

Referring to OmniBenchmark [41], MS-COCO [19], Kinetics-700 [3], we only present the lists of URLs and their corresponding meta information to our HumanBench.

1.3. Details of HumanBench-Subset

Due to the significant computational cost when we pretrain the model on the full dataset, we select 17 subsets from 37 full datasets for ablation study, which contains 1,270,186 images as a similar number with ImageNet-1K($\sim$1.28M). Table 1 summarizes the statistics of HumanBench-Subset. For the person ReID task, we select widely-used Market1501 and CUHK03 datasets, and the clothes-changing ReID dataset PRCC, forming a total of 38,197 images. For the human parsing task, we select widely used Human3.6M, LIP, CIHP, VIP datasets and one clothes parsing dataset, *i.e.,* ModaNet, with a total of 192,124 images. For the pose estimation task, we select widely-used COCO, AIC, and PoseTrack datasets with a total of 748,812 images. For the attribute task, we select PA-100K, RAPv2, and Market1501-Attribute datasets with a total of 170,879. Due to the significant resource cost, for the pedestrian detection task, we only select one widely used dataset CrowdHuman.

*Equal contribution. This work was done in SenseTime.
†Corresponding author.

Table 1. Dataset statistics of pretraining datasets

Partition	Task	Name	Number of images / samples	Task	Name	Number of images / samples
Full	ReID	Market1501 [44]	12,936	Detection	WIDER Pedestrian [22]	57,999
		CUHK03 [16]	7,365	Pose	COCO [19]	262,465
		MSMT [32]	30,248		AIC [33]	378,374
		LaST [26]	71,248		PoseTrack [1]	107,973
		PRCC [36]	17,896		JRDB [29]	310,035
		DGMarket [47]	128,306		MHP [12]	41,128
		LUPerson-NL [5]	5,178,420		UppenAction [40]	163,839
	Parsing	Human3.6M [10]	62,668		Halpe [4]	41,712
		LIP [7]	30,462		3dpw [30]	74,620
		CIHP [6]	28,280		MPI-INF-3DHP [24]	1,031,701
		VIP [48]	18,469		Human3.6M [10]	312,187
		Paper Doll [35]	1,035,825		AIST++ [13]	1,015,257
		DeepFashion [21]	191,961	Attribute	PA100K [20]	90,000
		ModaNet [46]	52,245		RAPv2 [11]	67,943
	Detection	CrowdHuman [25]	15,000		HARDHC [17]	28,336
		WiderPerson [39]	9,000		UVA-Human [14]	16,183
		COCO-person [19]	64,115		Parse27k [27]	27,482
		EuroCity Persons [2]	21,795		Market1501-Attribute [44]	12,936
		CityPersons [38]	2,778	**Total**	**11,019,187**	
Subset	ReID	Market1501 [44]	12,936	Pose	COCO [19]	262,465
		CUHK03 [16]	7,365		AIC [33]	378,374
		PRCC [36]	17,896		PoseTrack [1]	107,973
	Parsing	Human3.6M	62,668	Detection	CrowdHuman [25]	15,000
		LIP [7]	30,462	Attribute	PA100K [20]	90,000
		CIHP [6]	28,280		RAPv2 [11]	67,943
		VIP [48]	18,469		Market1501-Attribute [44]	12,936
		ModaNet [46]	52,245	**Total**	**1,165,012**	

Table 2. Results of the publicly released MAE and CLIP on HumanBench.

		Human Parsing				Person ReID				Pedestrian Detection	
		H3.6M	LIP	CIHP	ATR	Market1501	MSMT	CUHK03	SenseReID	CrowdHuman	Caltech (↓)
ViT-B	MAE	62.0	57.2	61.7	97.4	79.2	51.5	65.8	44.6	89.6	48.1
	MAE (Head FT)	40.4	31.7	37.1	94.4					75.7	66.2
	MAE (Partial FT)	50.5	42.0	48.0	96.4	43.8	22.5	33.2	21.2	82.6	70.2
	CLIP	58.2	53.4	61.7	97.0	78.6	53.6	66.9	43.6	82.1	78.6
	CLIP (Head FT)	28.4	11.7	14.2	85.8					33.2	98.5
	CLIP (Partial FT)	32.1	24.8	30.0	90.8	34.5	10.9	15.2	25.3	28.4	97.1
	Ours (FT)	65.0	61.4	66.8	97.5	89.5	69.1	82.6	56.8	90.6	30.1
	Ours (Head FT)	64.1	59.9	63.3	97.1					90.0	31.1
	Ours (Partial FT)	63.7	60.0	63.1	97.2	88.7	66.1	79.5	57.2	90.9	28.3
ViT-L	MAE (Partial FT)	21.94	10.74	11.98	85.8	44.9	23.6	64.5	23.5	10.1	99.8
	CLIP (Partial FT)	14.68	10.05	4.37	81.4	33.6	12.9	16.8	27.6	6.5	99.4
	Ours (Partial FT)	66.2	62.6	67.5	97.4	91.8	74.7	86.0	66.8	90.8	28.7

		Pose Eestimation				Pedestrian Attribute Recognition			Counting (unseen task)	
		COCO	H3.6M (↓)	AIC	MPII	PA100K	Rapv2	PETA	ShTech PartA (↓)	ShTech PartB (↓)
ViT-B	MAE	75.8	8.2	31.8	90.1	82.3	80.8	84.6	102.1	15.5
	MAE (Head FT)	60.9	7.9	19.2	84.6	55.9	55.1	61.4	156.2	32.5
	MAE (Partial FT)	69.2	8.0	26.9	88.5	78.5	82.3	85.0	135.6	26.8
	CLIP	74.4	9.9	31.1	88.1	76.1	77.0	81.2	117.9	16.3
	CLIP (Head FT)	28.4	11.7	14.2	85.8	51.3	51.5	54.9	198.5	36.8
	CLIP (Partial FT)	32.1	24.8	30.0	90.8	72.8	76.3	81.5	168.6	32.3
	Ours (FT)	76.3	6.2	35.0	93.3	85.0	81.2	88.0	91.7	10.8
	Ours (Head FT)	75.2	6.1	31.6	92.7	77.4	72.4	79.0	97.6	13.8
	Ours (Partial FT)	76.0	6.1	33.3	93.0	86.9	83.1	89.8	94.3	14.0
ViT-L	MAE (Partial FT)	76.5	6.6	34.2	92.2	67.2	51.3	51.7	141.3	27.9
	CLIP (Partial FT)	62.3	10.8	14.8	73.0	55.0	58.5	50.1	167.6	30.6
	Ours (Partial FT)	77.1	5.8	36.3	93.7	90.8	87.4	90.7	91.3	11.5

2. More Results of MAE and CLIP on HumanBench

In this section, we provide more results about the publicly released pretraining models, *i.e.,* MAE and CLIP, in Table 2. We can see two conclusions from Table 2. First, we can see models pretrained on natural images can not naturally increase the performance of human-centric tasks. Second, although CLIP leverages the vision-language pair and more images, it achieves worse performance than MAE which only uses 1.28M images, which illustrates that the languages in the existing datasets may not describe fine-grained information about human bodies and therefore can not be helpful to human-centric tasks.

3. Visualization of Task-Specific Features

To visualize the features attended by the task-specific projectors, we plot the heatmap of L2-normalization of the channels of the attended features. The red color in Figure 1, 2, 3 show the important region, which leads to three conclusions. First, the highlighted regions in the pose estimation and the human parsing locates at the joints of human bodies, which shows that these two tasks are very similar. Second, the heatmap for pedestrian detection includes the whole person, which is consistent with the goal of pedestrian detection to detect all people. Third, for the pedestrian attribute recognition, we can see that the heatmap highlights the attributes, *e.g.,* gloves, bags. These highlighted regions instead of the whole body are also consistent with the goal of pedestrian attribute recognition to recognize attributes.

4. Detailed Design of PATH

4.1. Structure of Attention Module in Projector

In this section, we describe the structure of attention modules in the projector, which includes a squeeze-and-excitation module and a self-attention module. Specifically, given the feature maps $\mathbf{F}$ and $\mathbf{f}_l \in \mathcal{R}^{C\times H\times W}$ extracted from l-th layer, *i.e.,* $\mathbf{F} = (\mathbf{f}_1, \mathbf{f}_2, ..., \mathbf{f}_8)$, the squeeze-and-excitation layer $\mathcal{E}$ transforms the feature $\mathbf{f}_l$ as

$$\mathbf{e}_l = \mathcal{E}(\mathbf{f}_l) = \mathcal{F}_{sq}(\frac{1}{H\times W}\sum_{u=1}^{H}\sum_{v=1}^{W} f_l(:,u,v)) \odot \mathbf{f}_l, \quad (1)$$

where $\mathcal{F}_{sq}$ is the 1-D convolution operation and $\odot$ is the element-wise multiplication of two tensors. Here $\mathcal{F}_{sq}(\frac{1}{H\times W}\sum_{u=1}^{H}\sum_{v=1}^{W} f_l(:,u,v))$ is the channel-wise attention calculated by the squeeze-and-excitation layer.

Next, we feed $\mathbf{s}_l$ into the self-attention module $\mathcal{A}$, which exactly follows [28], which mathmatically can be defined as

$$\mathbf{p} = \mathcal{A}(\mathbf{e}). \quad (2)$$

4.2. Task Head and Objective Functions

In this section, we present the task head and the loss designs in Sec. 4.3 in the main text. Given the features $\mathbf{P}_j^t$ after the projector of all image $\mathbf{X}$ for the j-th dataset in t-th task $\mathcal{D}_j^t$, we compute the losses according to different tasks.

4.2.1 Person ReID

Task Head. Following [23], the task head of person ReID is a Synchronized BatchNorm [9]. Mathematically, the activation $\mathbf{Z}_j^t$ is defined as

$$\mathbf{Z}_j^t = \text{BatchNorm}(\mathbf{P}_j^t). \quad (3)$$

Objective Function. We use the triplet loss [8] and cross-entropy [42] to supervise the ReID task. Mathematically,

$$\mathcal{L}_{\text{reid}} = \sum_{t=1}^{T}\sum_{j=1}^{N_t}\mathcal{L}_{\text{ce}}(\mathbf{Z}_j^t, \mathbf{Y}_j^t) + \sum_{t=1}^{T}\sum_{j=1}^{N_t}\mathcal{L}_{\text{triplet}}(\mathbf{Z}_j^t), \quad (4)$$

where $\mathcal{L}_{\text{ce}}$ is the cross-entropy loss, $\mathbf{Y}_j^t$ is the labels and N_j^t is the number of images in $\mathcal{D}_j^t$. The triplet loss enlarges the distance between negative pairs and minimizes the distance between positive pairs, which can be mathematically defined as

$$\mathcal{L}_{\text{triplet}} = [d_p - d_n + \alpha]_+, \quad (5)$$

where d_p and d_n are feature distances of positive and negative pairs. α is the margin of triple loss, and $[\cdot]$ equals $max(\cdot, 0)$.

4.2.2 Pose Estimation

Task Head. Following [34], the task head is lightweight, processes the features after the task-specific features, and localizes the keypoints. We use the structure of classic decoders in [34], which consists of two deconvolution blocks, each of which contains one deconvolution layer followed by layer normalization and ReLU. Following the common setting of previous methods in pose estimation, each block upsamples the feature maps by 2 times. Mathematically, the activation (the localization heatmaps) can be defined as

$$\mathbf{Z}_j^t = \text{Conv}_{1\times1}(\text{Deconv}(\text{Deconv}(\mathbf{P}_j^t))), \quad (6)$$

where $\mathbf{Z}_j^t \in \mathcal{R}^{\frac{H}{4}\times\frac{W}{4}\times N_k}$, H is the height of the image, W is the width of the image, and N_k is the number of keypoints.

Objective Function. We leverage the mean square error (MSE) for pose estimation, *i.e.,*

$$\mathcal{L}_{\text{pose}} = \sum_{t=1}^{T}\sum_{j=1}^{N_t}\text{MSE}(\mathbf{Z}_j^t, \mathbf{Y}_j^t), \quad (7)$$

where $\mathbf{Y}_j^t$ is the ground-truth heatmap of keypoints.

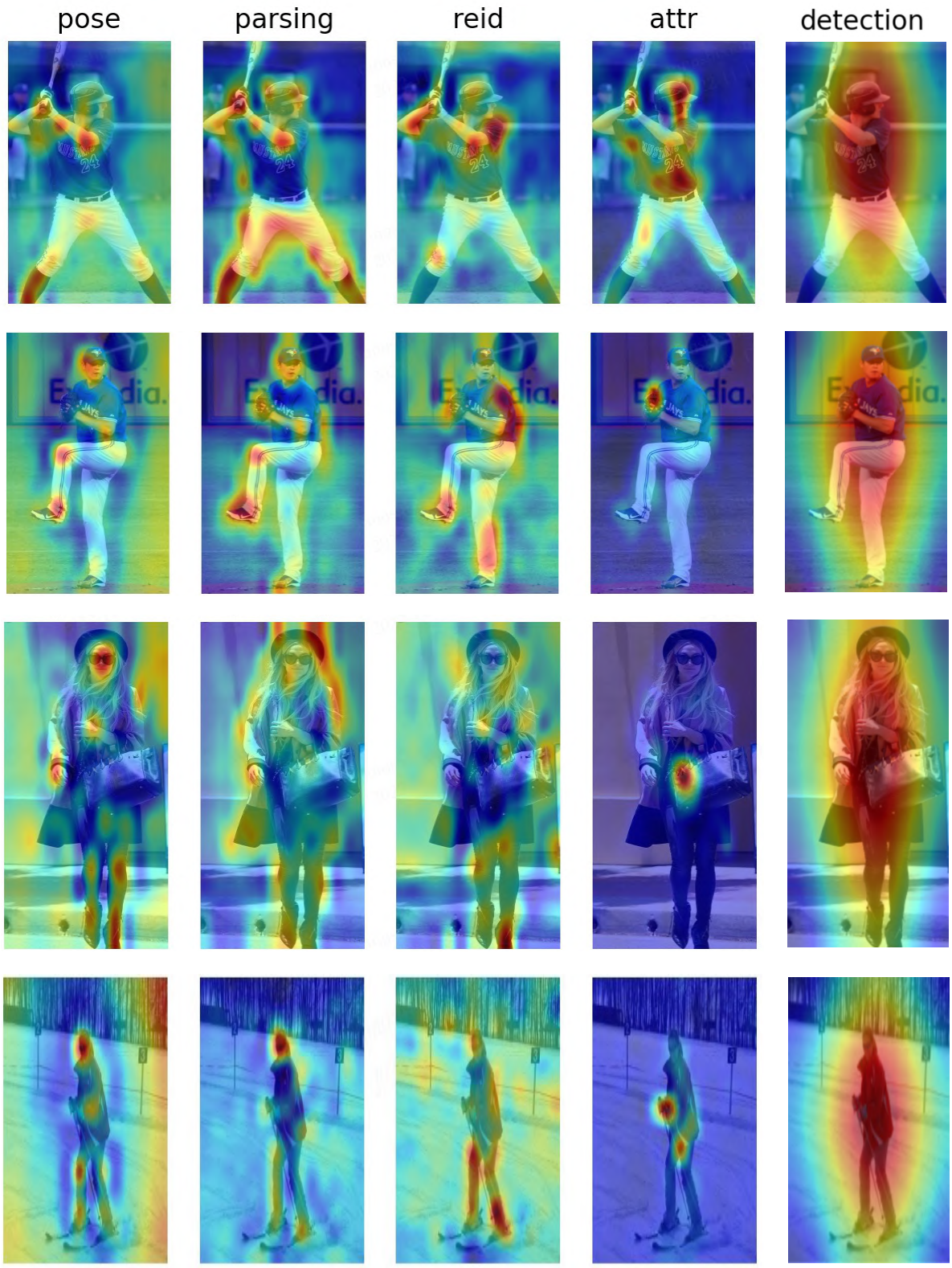

Figure 1. Visualization of features after the task-specific projectors

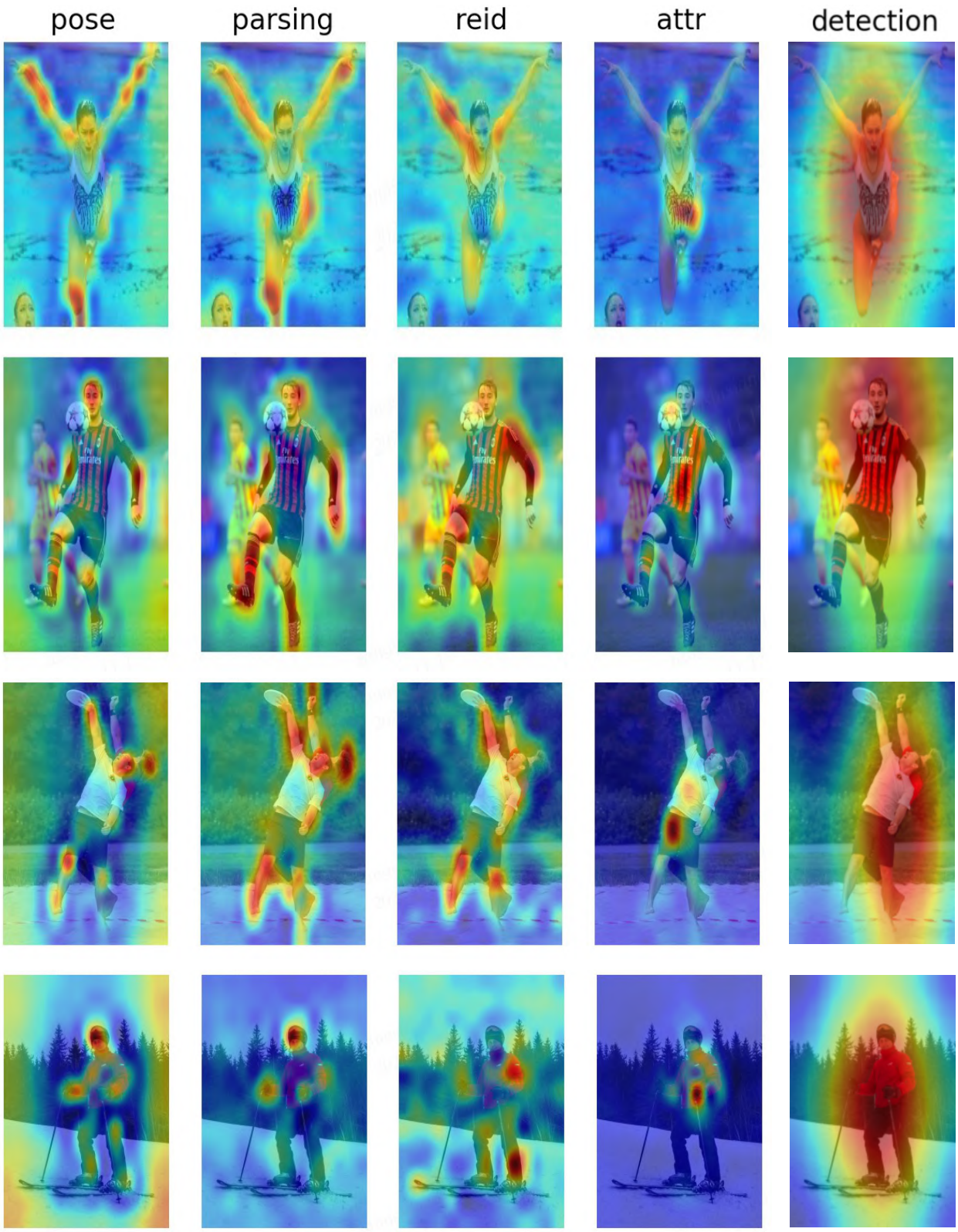

Figure 2. Visualization of features after the task-specific projectors

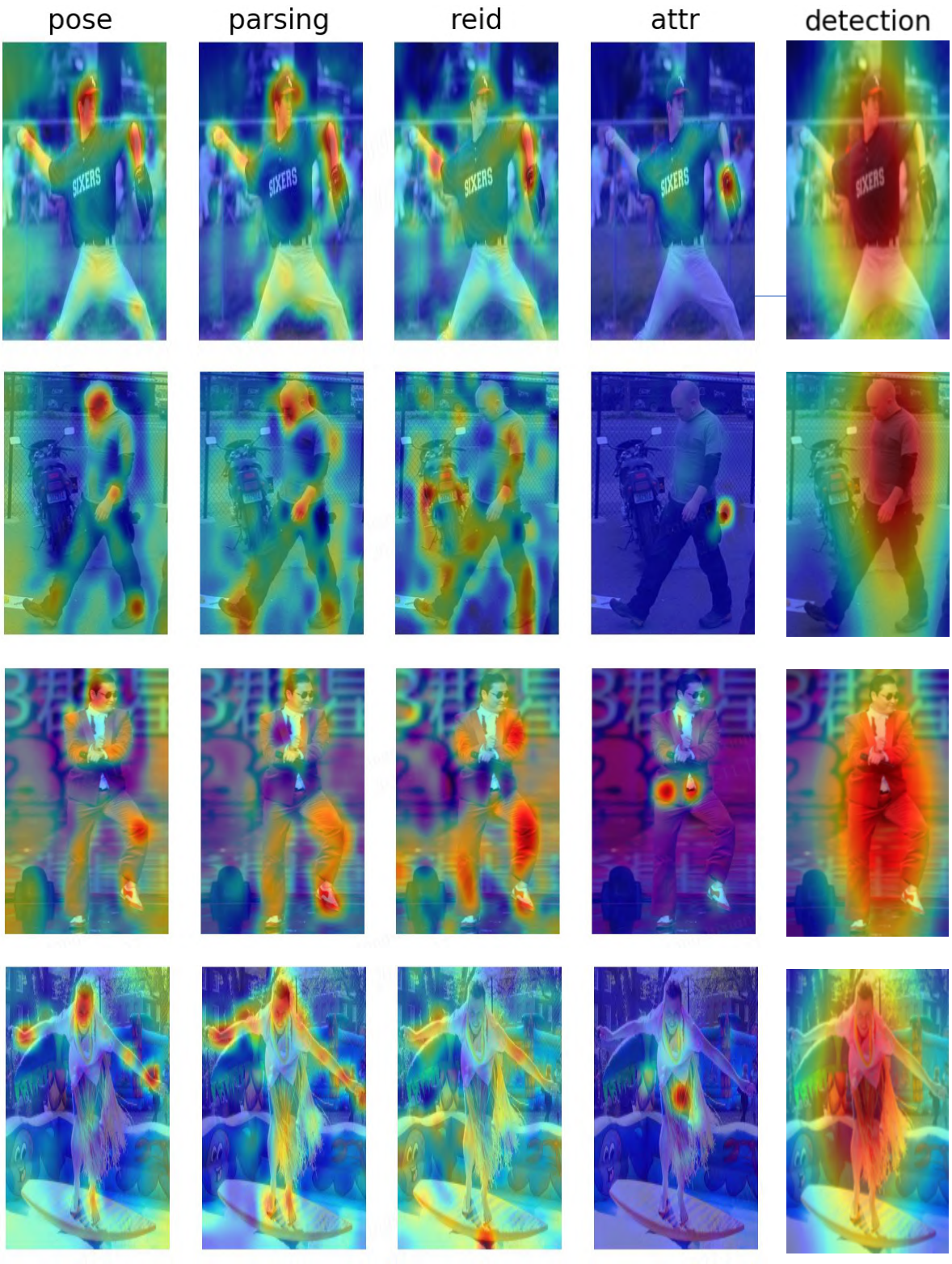

Figure 3. Visualization of features after the task-specific projectors

4.2.3 Human Parsing

Task Head. We follow the naive head design of [45] for human parsing. Specifically, the naive head first projects the features after the task-specific projectors to the dimension of category number (*e.g.,* 20 in LIP [18]). For this, we adopt a simple 2-layer network with architecture: 1×1 Conv+LayerNorm+ReLU+1×1Conv. After that, we simply bilinearly upsample the output to the full image resolution, followed by a classification layer with pixel-wise cross-entropy loss. Mathematically, the task head can be defined as

$$\mathbf{Z}_j'^t = \text{Conv}_{1\times1}(\text{LayerNorm}(\text{ReLU}(\text{Conv}_{1\times1}(\mathbf{P}_j^t)))), \quad (8)$$

$$\mathbf{Z}_j^t = \text{Upsample}(\mathbf{Z}_j'^t), \quad (9)$$

where $\mathbf{Z}_j^t$ is upsampled to the size of input images.

Objective Function. Following common implementations in [37], we use the cross-entropy loss to supervise the human parsing. Specifically, the objective function can be defined as

$$\mathcal{L}_{\text{parsing}} = \sum_{t=1}^{T}\sum_{j=1}^{N_t} \text{CE}(\mathbf{Z}_j^t, \mathbf{Y}_j^t), \quad (10)$$

where $\mathbf{Y}_j^t \in \mathcal{R}^{H\times W\times N_c}$ is the annotation map whose elements represent the label of the pixel.

4.2.4 Pedestrian Attribute Recognition

Task Head. Following the common implementations in [15], we only use a fully-connected layer followed by a sigmoid function to project the feature to the activation, which can be mathematically defined as

$$\mathbf{Z}_j^t = \text{Sigmoid}(\text{FC}(\mathbf{Y}_j^t)), \quad (11)$$

where $\mathbf{Z}_j^t \in \mathcal{R}^{N\times N_c}$ Fc is a fully-connected layer, and N_c is the number of attributes in the dataset.

Objective Function. Our objective function is the binary cross-entropy loss between the activation and the ground-truth label, which can be mathematically defined as

$$\mathcal{L}_{\text{attribute}} = \sum_{t=1}^{T}\sum_{j=1}^{N_t} \text{BCE}(\mathbf{Z}_j^t, \mathbf{Y}_j^t). \quad (12)$$

4.2.5 Pedestrian Detection

Task Head. Following Anchor Detr [31], the task head consists of 9 transformer decoder layers, *i.e.,* $\mathcal{D} = \{\mathcal{D}_1, \mathcal{D}_2, ..., \mathcal{D}_9\}$. The every transformer decoder layer $\mathcal{D}_i$ includes a cross-attention layer, a self-attention layer, and a feedforward network. Therefore, features processed by the decoder $\mathcal{D}_l$ are defined as

$$\mathbf{P}_l = \mathcal{D}_l(\mathbf{Q}_{l-1}^t, \mathbf{Q}_p^t, \mathbf{P}_j^t, \mathbf{P}_p), \quad (13)$$

where $\mathbf{P}_p = proj(\mathcal{A}_\mathbf{P})$, $proj$ is a linear projection, and $\mathcal{A}_\mathbf{P}$ is the coordinates of all tokens in the task-specific feature $\mathbf{P}_j^t$. Similarly, $\mathbf{Q}_p^t = proj(\mathcal{A}_\mathbf{Q})$ refers to a linear projection of the coordinates of learnable anchor points initialized with a uniform distribution following [31].

Objective Function. Given the features $\mathbf{P}_L$ after the decoder, we use the classification loss, GIoU loss and bounding box loss to supervise the pedestrian detection, *i.e.,*

$$\begin{aligned}\mathcal{L}_{peddet} =& \lambda_{cls}\mathcal{L}_{cls}(\mathbf{Z}_{cls}, \mathbf{Y}_{cls}) + \lambda_{iou}\mathcal{L}_{iou}(\mathbf{Z}_{bbox}, \mathbf{Y}_{bbox}) \\ &+ \lambda_{L1}\mathcal{L}_{L1}(\mathbf{Z}_{bbox}, \mathbf{Y}_{bbox}),\end{aligned} \quad (14)$$

where $\mathcal{L}_{cls}$ is the classificatin loss, λ_{iou} is the GIoU loss, λ_{L1} is L1 loss of the bounding boxes, and $\mathbf{Y}_{cls}$, $\mathbf{Y}_{bbox}$ are annotations of classes and bounding boxes. Here, $\mathbf{Z}_{cls} = f_{cls}(\mathbf{P}_L)$, $\mathbf{Z}_{bbox} = f_{bbox}(\mathbf{P}_L)$ are linearly projections of $\mathbf{P}_L$, f_{cls} and f_{bbox} are two fully connected layers.

4.2.6 Crowd Counting

Task Head. Table 3 details the configurations of counting head for regressing the density map. In this table, "Conv{k(3,3),c64,s1}-BN-R" represents the convolutional operation with kernel size of 3×3, output channels of 64, and stride size of 1. The "BN" and "ReLU" mean that the Batch Normalization and ReLU layer are added to this convolutional layer. Specifically, we denote the task head of counting using layers in Table 3 as $\mathcal{H}_{\text{count}}$, *i.e.,*

$$\mathbf{Z}_j^t = \mathcal{H}_{\text{count}}(\mathbf{P}_j^t). \quad (15)$$

Objective Function. We leverage the MSE between the activation and the ground-truth heatmap to supervise the learning of crowd counting, *i.e.,*

$$\mathcal{L}_{\text{counting}} = \text{MSE}(\mathbf{Z}_j^t, \mathbf{Y}_j^t), \quad (16)$$

where $\mathbf{Y}_j^t$ is the ground-truth heatmap of crowd counting.

Counting Head
Upsample(scale_factor=2)
Conv{k=(3,3),c=64,s=1}-BN-ReLU
Conv{k=(3,3),c=32,s=1}-BN-ReLU
Upsample(scale_factor=2)
Conv{k=(3,3),c=16,s=1}-BN-ReLU
Conv{k=(3,3),c=1,s=1}-ReLU

Table 3. Detailed architecture of counting head.

5. Details of Implementations in Pretraining

During pretraining, we collect in total 39 datasets from person ReID, human parsing, pose estimation, pedestrian attribute recognition, and pedestrian detection. To pretrain the model in a distributed manner, we only train a dataset in each GPU. We pretrain our model using 64 V100-32G GPUs. In the following, we present the task-agnostic parameters and task-specific parameters.

5.1. Task-agnostic Hyperparameters

Table 4 illustrates the learning hyper-parameters utilized in our pretraining stage. Specifically, we train our model for 80000 iterations in total. During pretraining, we use **STEP** learning rate decay strategy with a warm-up from $1e^{-7}$ to $5e^{-4}$ during 1500 iterations. we multiply the learning rate $5e^{-4}$ by 0.5, 0.2 and 0.1 at the 40000-th, 60000-th and 76000-th iteration, respectively. The backbone multiplier and the positional multiplier are the ratios of the actual learning rate of the backbone and the positional embedding, respectively, which are all set as 1.0.

5.2. Task-specific Hyperparameters

Table 5 presents the task-specific hyper-parameters of each dataset, including batch size per GPU, the number of GPUs, sample weights, and loss weights. Specifically, the dataset weights are related to sample weights and the number of GPUs:

$$\text{loss weight} = \text{sample weight} \times \text{images per GPU} \times \text{number of GPUs}. \tag{17}$$

The loss weights of the pose estimation are larger than other tasks because the loss functions used in pose estimation are MSE loss between the predicted heatmaps of keypoints and the heatmaps of the ground truth whose value is very small. For tasks other than pose estimation, the difference between different datasets among different tasks are relatively small.

5.3. Data Augmentation

We apply augmentation techniques to human-centric images, ranging from scene images in pedestrian detection to cropped images in person ReID. Here, we list the augmentations below for different tasks.

Person ReID. For person ReID, we use the same augmentation as in [23]. Specifically, we use the random horizontal flip and random erasing for pretraining. Finally, we resize the input image to size 256×128.

Pose Estimation. For pose estimation, we use the same sugmentation as ViTPose [34]. Specifically, we use random horizontal flip, half body transform and random scale rotation for pretraining. Finally, we resize the input image to size 256×192.

Table 4. Detailed description of task-agnostic hyper-parameters in the pretraining stage.

lr_schedule	type	Step
	base_lr	1.00E-07
	warmup_steps	1500
	warmup_lr	5.00E-04
	lr_mults	[0.5, 0.2, 0.1]
	lr_steps	[40000, 60000, 76000]
	max_iter	80000
	backbone_multiplier	1.0
	pos_embed_multiplier	1.0
optimizer	type	Adafactor_dev
	beta1	0.9
	clip_beta2	0.999
	clip_threshold	0.5
	decay_rate	-0.8
	scale_parameter	FALSE
	relative_step	FALSE
	weight_decay	0.05
layer_decay	num_layers	12
	layer_decay_rate	0.75

Human Parsing. For human parsing, we use the same augmentation as in [6]. Specifically, we use random crop, random image rotation, and photometric distortion augmentation for pretraining. Particularly, for the human parsing dataset, we also use horizontal random flip augmentation, *e.g.,* Human3.6M, LIP, CIHP, LIP, VIP. Finally, we resize the input image to size 480×480.

Pedestrian Attribute Recognition. For pedestrian attribute recognition, we use the same augmentation as in [15]. Specifically, we use random crop and random horizontal flip augmentation for pretraining. Finally, we resize the input image to size 256×192.

Pedestrian Detection. For pedestrian detection, we use the same augmentation as in [43]. Specifically, we use random horizontal flips and random crop augmentation for pretraining. Finally, we random resize the input image with the longest side bound of 1333 and the shortest side bound of 800 while keeping the height and width ratio.

Crowd Counting. For the crowd counting dataset, we use random horizontal flip, random scaling ($0.5\times \sim 2\times$), and random cropping augmentation for pretraining.

6. Details of Implementations in Evaluation

For full finetuning, we carefully tune the learning rate $\{1e^{-3}, 5e^{-4}, 1e^{-4}\}$, the weight decay $\{0.05, 0.1, 0.3\}$, drop path rate $\{0.1, 0.3, 0.5\}$, the backbone multiplier

Table 5. Detailed Implementation about Task-specific Hyper-parameters

Task	Dataset	Batch Size Per GPU	GPU	Sample Weight	Loss Weight
ReID	Market1501+MSMT+CUHK03	112	1	5	560
	DGMarket+LaST+PRCC	96	1	0.1	9.6
	LUPerson-NL	192	2	1	384
Pose	COCO	224	2	8000	3584000
	AIC	224	2	6000	2688000
	PoseTrack	224	1	6000	1344000
	JRDB	224	1	4000	896000
	MHP	96	1	4000	384000
	UppenAction	128	1	4000	512000
	MPI-INF-3DHP	128	1	4000	512000
	Halpe	64	1	2000	128000
	3dhp	128	1	2000	256000
	Human3.6M	128	1	2000	256000
	AIST++	128	1	2000	256000
Parsing	Human3.6M	26	3	20	1560
	LIP	18	2	20	720
	CIHP	24	2	20	960
	VIP	16	1	20	320
	Paper Doll	24	2	15	720
	DeepFashion	32	2	15	960
	ModaNet	32	1	15	480
Attribute	rap2+pa100k	128	1	0.1	12.8
	HARDHC+UAV-Human+Parse27k+Market1501-Attribute	116	1	0.1	11.6
Detection	CrowdHuman	2	16	10	320
	WidePerson+COCO-person+EuroCity Persons+CityPersons	2	16	10	320

$\{0.1, 0.3, 0.5\}$, and report the best performance. We will provide the exact hyperparameters in our released repository after acceptance. For head finetuning and partial finetuning, we specifically set the weight decay as 0, which empirically proved very important in our experiments.

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. 2

[2] Markus Braun, Sebastian Krebs, Fabian Flohr, and Dariu M Gavrila. The eurocity persons dataset: A novel benchmark for object detection. *arXiv preprint arXiv:1805.07193*, 2018. 2

[3] Joao Carreira, Eric Noland, Chloe Hillier, and Andrew Zisserman. A short note on the kinetics-700 human action dataset. *arXiv preprint arXiv:1907.06987*, 2019. 1

[4] 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. 2

[5] Dengpan Fu, Dongdong Chen, Jianmin Bao, Hao Yang, Lu Yuan, Lei Zhang, Houqiang Li, and Dong Chen. Unsupervised pre-training for person re-identification. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14750–14759, 2021. 2

[6] 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. 2, 8

[7] 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. 2

[8] Alexander Hermans, Lucas Beyer, and Bastian Leibe. In defense of the triplet loss for person re-identification. *arXiv preprint arXiv:1703.07737*, 2017. 3

[9] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International conference on machine learning*, pages 448–456. PMLR, 2015. 3

[10] 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. 2

[11] 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, 2018. 2

[12] 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. 2

[13] Ruilong Li, Shan Yang, David A Ross, and Angjoo Kanazawa. Ai choreographer: Music conditioned 3d dance generation with aist++. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13401–13412, 2021. 2

[14] 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. 2

[15] 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. 7, 8

[16] Wei Li, Rui Zhao, Tong Xiao, and Xiaogang Wang. Deepreid: Deep filter pairing neural network for person re-identification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 152–159, 2014. 2

[17] Yining Li, Chen Huang, Chen Change Loy, and Xiaoou Tang. Human attribute recognition by deep hierarchical contexts. In *European conference on computer vision*, pages 684–700. Springer, 2016. 2

[18] 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. 7

[19] 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. 1, 2

[20] 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. 2

[21] Ziwei Liu, Ping Luo, Shi Qiu, Xiaogang Wang, and Xiaoou Tang. Deepfashion: Powering robust clothes recognition and retrieval with rich annotations. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1096–1104, 2016. 1, 2

[22] 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. 2

[23] Hao Luo, Youzhi Gu, Xingyu Liao, Shenqi Lai, and Wei Jiang. Bag of tricks and a strong baseline for deep person re-identification. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, pages 0–0, 2019. 3, 8

[24] Dushyant Mehta, Helge Rhodin, Dan Casas, Pascal Fua, Oleksandr Sotnychenko, Weipeng Xu, and Christian Theobalt. Monocular 3d human pose estimation in the wild using improved cnn supervision. In *2017 international conference on 3D vision (3DV)*, pages 506–516. IEEE, 2017. 2

[25] 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. 2

[26] 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. 2

[27] Patrick Sudowe, Hannah Spitzer, and Bastian Leibe. Person Attribute Recognition with a Jointly-trained Holistic CNN Model. In *ICCV'15 ChaLearn Looking at People Workshop*, 2015. 2

[28] 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. 3

[29] 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. 2

[30] Timo Von Marcard, Roberto Henschel, Michael J Black, Bodo Rosenhahn, and Gerard Pons-Moll. Recovering accurate 3d human pose in the wild using imus and a moving camera. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 601–617, 2018. 2

[31] 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. 7

[32] Longhui Wei, Shiliang Zhang, Wen Gao, and Qi Tian. Person transfer gan to bridge domain gap for person re-identification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 79–88, 2018. 2

[33] 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. 2

[34] Yufei Xu, Jing Zhang, Qiming Zhang, and Dacheng Tao. Vitpose: Simple vision transformer baselines for human pose estimation. *arXiv preprint arXiv:2204.12484*, 2022. 3, 8

[35] Kota Yamaguchi, M Hadi Kiapour, and Tamara L Berg. Paper doll parsing: Retrieving similar styles to parse clothing items. In *Proceedings of the IEEE international conference on computer vision*, pages 3519–3526, 2013. 2

[36] 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. 2

[37] Y Yuan, F Rao, H Lang, W Lin, C Zhang, X Chen, and J Wang. Hrformer: High-resolution transformer for dense prediction. arxiv 2021. *arXiv preprint arXiv:2110.09408*. 7

[38] 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. 2

[39] 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. 2

[40] 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. 2

[41] Yuanhan Zhang, Zhenfei Yin, Jing Shao, and Ziwei Liu. Benchmarking omni-vision representation through the lens of visual realms. In *European Conference on Computer Vision*, pages 594–611. Springer, 2022. 1

[42] Zhilu Zhang and Mert Sabuncu. Generalized cross entropy loss for training deep neural networks with noisy labels. *Advances in neural information processing systems*, 31, 2018. 3

[43] 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. 8

[44] 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. 2

[45] Sixiao Zheng, Jiachen Lu, Hengshuang Zhao, Xiatian Zhu, Zekun Luo, Yabiao Wang, Yanwei Fu, Jianfeng Feng, Tao Xiang, Philip HS Torr, et al. Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6881–6890, 2021. 7

[46] 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. 2

[47] 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. 2

[48] 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. 2