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

Abstract

*Human-centric perceptions include a variety of vision tasks, which have widespread industrial applications, including surveillance, autonomous driving, and the metaverse. It is desirable to have a general pretrain model for versatile human-centric downstream tasks. This paper forges ahead along this path from the aspects of both benchmark and pretraining methods. Specifically, we propose a **HumanBench** based on existing datasets to comprehensively evaluate on the common ground the generalization abilities of different pretraining methods on 19 datasets from 6 diverse downstream tasks, including person ReID, pose estimation, human parsing, pedestrian attribute recognition, pedestrian detection, and crowd counting. To learn both coarse-grained and fine-grained knowledge in human bodies, we further propose a **P**rojector **A**ssis**T**ed **H**ierarchical pretraining method (**PATH**) to learn diverse knowledge at different granularity levels. Comprehensive evaluations on HumanBench show that our PATH achieves new state-of-the-art results on 17 downstream datasets and on-par results on the other 2 datasets. The code will be publicly at https://github.com/OpenGVLab/HumanBench.*

1. Introduction

Human-centric perception has been a long-standing pursuit for computer vision and machine learning communities. It encompasses massive research tasks and applications including person ReID in surveillance [16, 17, 60, 96, 112], human parsing and pose estimation in the metaverse [47, 48, 61, 79, 90, 92], and pedestrian detection in autonomous driving [8, 51, 87]. Although significant progress has been made, most existing human-centric studies and pipelines are task-specific for better performances, leading to huge costs in representation/network design, pretraining, parameter-tuning, and annotations. To promote real-world deployment, we ask: *whether a general human-centric pretraining model can be developed that can benefit diverse human-centric tasks and be efficiently adapted to downstream tasks?*

Intuitively, we argue that pretraining such general human-centric models is possible for two reasons. First, there are obvious correlations among different human-centric tasks. For example, both human parsing and pose estimation predict the fine-grained parts of human bodies [29, 49] with differences in annotation granularities. Thus, the annotations in one human-centric task may benefit other human-centric tasks when trained together. Second, recent achievements in foundation models [5, 11, 38, 65, 66, 80] have shown that large-scale deep neural networks (*e.g.*, transformers [13]) have the flexibility to handle diverse input modalities and the capacity to deal with different tasks. For example, Uni-Percevier [115] and BEITv3 [85] are applicable to multiple vision and language tasks.

Despite the opportunities of processing multiple human-centric tasks with one pretraining model, there are two obstacles for developing general human-centric pretraining models. First, although there are many benchmarks for every single human-centric task, there is still no benchmark to fairly and comprehensively compare various pretraining methods on a common ground for a broad range of human-centric tasks, data distributions, and application scenarios. Second, different from most existing general foundation models trained by unified global vision-language consistencies, pretraining human-centric models are required to learn both global (*e.g.,* person ReID and pedestrian detection) and fine-grained semantic features (*e.g.,* pose estimation and human parsing) of human bodies from diverse annotation granularity simultaneously.

In this paper, we first build a benchmark, called **HumanBench**, based on existing datasets to enable pretrain-

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

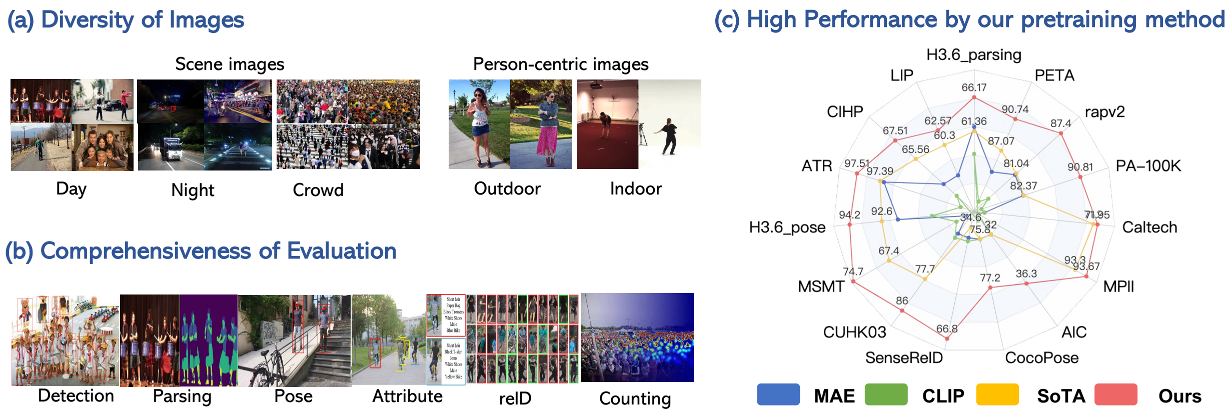

Figure 1. (a-b) Overview of our proposed HumanBench. HumanBench includes diverse images, including scene images and person-centric images. Our HumanBench also has comprehensive evaluation. Specifically, it evaluates pretraining models on 6 tasks, including pedestrian detection, human parsing, pose estimation, pedestrian attribute recognition, person ReID, and crowd counting. (c) High performances are achieved by our pretraining method on HumanBench. We report 1-heavy occluded MR^{-2} and 1-EPE for Caltech and H3.6pose.

ing and evaluating human-centric representations that can be generalized to various downstream tasks. HumanBench has two appealing properties. **(1) Diversity.** The images in our HumanBench include diverse image properties, ranging from person-centric cropped images to scene images with crowd pedestrians, ranging from indoor scenes to outdoor scenes (Fig. 1(a)), and from surveillance to metaverse. **(2) Comprehensiveness.** Humanbench covers comprehensive image-based human-centric tasks in both pretraining datasets and downstream tasks (Fig. 1(b)). For pretraining, we include 11 million images from 37 datasets across five representative human-centric tasks, *i.e.,* person ReID, pose estimation, human parsing, pedestrian attribute recognition, and pedestrian detection. For evaluation, HumanBench evaluates the generalization abilities on 12 pretraining datasets, 6 unseen datasets of pretraining tasks, and 2 datasets out of pretraining tasks, ranging from global prediction, *i.e.,* ReID, to local prediction, *i.e.,* human parsing and pose estimation. Results on our HumanBench (Fig. 1(c)) lead to two interesting findings. First, compared with datasets with natural images for general pretrained models, HumanBench is more effective for human-centric perception tasks. Second, as human-centric pretraining requires to learn features of diverse granularity, supervised pretraining methods with proper designs can learn from diverse annotations in HumanBench and perform better than the existing unsupervised pretraining methods, for which details will be shown in Sec. 5.3.

Based on HumanBench, we further investigate how to learn a better human-centric supervised pretraining model from diverse datasets with various annotations. However, naive multitask pretraining may easily suffer from the task conflicts [53, 97] or overfitting to pretrained annotations [67, 107], losing the desirable generalization ability of pretraining. Inspired by [86], which suggests adding an MLP projector before the task head can significantly enhance the generalization ability of supervised pretraining, we propose **P**rojector **A**ssis**T**ed **H**ierarchical Pre-training (**PATH**), a projector assisted pretraining method with hierarchical weight sharing to tackle the task conflicts of supervised pretraining from diverse annotations. Specifically, the weights of backbones are shared among all datasets, and the weights of projectors are shared only for datasets of the same tasks, while the weights of the heads are shared only for a single dataset – forming a hierarchical weight-sharing structure. During the pretraining stage, we insert the task-specific projectors before dataset heads but discard them when evaluating models on downstream tasks. With the hierarchical weight-sharing strategy, our pretraining method enforces the backbone to learn the shared knowledge pool, the projector to attend to the task-specific knowledge, and the head to focus on the dataset with specific annotation and data distribution.

In summary, our contributions are two folds: (1) we build HumanBench, a large-scale dataset for human-centric pretraining including diverse images and comprehensive evaluations. (2) To tackle the diversity of input images and annotations of various human-centric datasets, we propose PATH, a projector-assisted hierarchical weight-sharing method for pretraining the general human-centric representations. We achieve state-of-the-art results by PATH on 15 datasets throughout 6 downstream tasks (Fig. 1(c)), on-par results on 2 datasets, and slightly lower results on 2 datasets on HumanBench when using ViT-Base. Experiments with ViT-Large backbone show that our method can further achieve considerable gains over ViT-Base, achieving another 2 new state-of-the-art results and showing the promising scalability of our method. We hope our work can shed light on future research on pretraining human-centric representations, such as unified structures.

Table 1. Statistics of Pretraining Datasets

Task	Number of datasets	Number of Images
Person ReID	7	5,446,419
Pose estimation	11	3,739,291
Human parsing	7	1,419,910
Pedestrian Attribute	6	242,880
Pedestrian Detection	6	170,687
In total	37	11,019,187

2. Related Work

General Vision Models. The vision community has witnessed the emergence of general vision models [69, 85, 98, 115], which could learn complex patterns from large-scale data and are powerful when adapted for downstream tasks. Although they achieve promising results, they pretrain on natural images, which are sub-optional for human-centric perception tasks. Furthermore, most of them leverage contrastive learning or masked reconstruction on global images, but do not learn multi-scale representations of human bodies with a diverse granularity that is important for diverse human-centric perceptions. In this paper, we tackle this problem by supervised pretraining from large-scale publicly available datasets with released annotations. The massive human body annotations with diverse granularity enable the model to learn multi-scale features, which is preferred in various human-centric tasks. SL-MLP [86] is an improved supervised pretraining method that proposes to insert an MLP projector to increase the generalization ability of pretraining models. However, it does not tackle the task conflict problem in a multitask pretraining as our method.

Pretraining on Human-centric Tasks. Human-centric perception has been studied for decades. However, there are only a few works on pretraining diverse human-centric tasks. HCMoCo [29] pretrains human representations for only pose estimation and human parsing with human-centric images. Our work pretrains a model for more diverse human-centric tasks, ranging from global identification to local prediction, from tasks on cropped human-centric images to scene images. Furthermore, different from HCMoCo which mainly learns human-centric representation from the multi-modal (RGB-D) representation of the same image, we learn human representations from multi-datasets where RGB images with one or two kinds of annotation, which is a more practical and data-scalable setting for pretraining. For example, publicly available datasets for ReID, attribute, and counting do not have RGB-D data.

Multi-task Multi-dataset Pretraining. Multi-task and Multi-dataset Pretraing is a popular framework for pretraining models [27, 84, 115]. However, they may easily suffer from task conflicts. There are generally two routes to reduce task conflicts, including network designs [25, 37, 55, 59, 63, 78] and optimization strategies [7, 20, 36, 55, 68]. Different from existing methods to tackle task conflicts from the balance of losses, Our PATH explores a new network optimization strategy from framework design at the training stage. Specifically, we insert a task-specific projector into the backbone and the dataset head, and design the parameters in the backbone, the parameters in the task-specific projector and the head are shared among all datasets, datasets of the same task and not shared. With this strategy, the dataset-specific heads and the task-specific projectors are the regularizations for the backbone of learning shared knowledge across various human-centric tasks. Our method does not design new networks because we only evaluate the generalization of the backbone.

3. HumanBench

3.1. Pretraining Datasets

According to biologists [9], nonverbal communication in daily life includes identity, visual appearance, and posture information. Following this domain knowledge, we select person ReID as the identification task, pedestrian attribute recognition, pedestrian detection, human parsing as the visual appearance task, and pose estimation as the posture task in HumanBench. 37 datasets containing 11,019,187 images [1] are collected for pretraining. Tab. 1 presents the number of datasets and images in each task. For the selected datasets, we leverage their original annotations except for the noisy labeled person ReID dataset, *i.e.,* LUPerson-NL. In LUPerson-NL, we observe that identities with relatively few images are accurate. Therefore, we only select the identities that contain 15 to 200 images in LUPerson-NL, corresponding to 151,595 identities and 5,178,420 images.

To ensure no data leakage and small information redundancy, we further de-duplicate the pretraining dataset from two aspects. First, we remove all potential duplicates from pretraining datasets that may appear in the evaluation datasets (detailed in Sec. 3.2) to enable a meaningful evaluation of generalization. Specifically, we first utilize the Difference Hash [77] to calculate the hash code of images in the evaluation datasets and pretraining datasets. Then, we delete the images in the pretraining datasets that have the same hash code as any image in the evaluation datasets. Second, some images come from some video-based datasets, *e.g.,* AIST++ [41] and UppenAction [102], which contain large information redundancy between consecutive frames. In this case, we select only one image from every 8 consecutive frames to reduce redundancy.

3.2. Evaluation Scenarios and Protocols

Evaluation Scenarios. Our benchmark comprehensively quantifies the generalization ability of human-centric representation on 6 human-centric tasks from 19 datasets.

[1] Full list of these 37 datasets are given in the supplementary.

Table 2. Summary of datasets for in-dataset evaluations, out-of-dataset evaluations, and unseen-task evaluations.

Task	Datasets	in-dataset evaluations	out-of-dataset evaluations	unseen-task evaluations
ReID	Market1501 [109]	✓		
	MSMT [88]	✓		
	CUHK03 [45]	✓		
	SenseReID [106]		✓	
Pose	COCO [52]	✓		
	Human3.6M [33]	✓		
	AIC [89]	✓		
	MPII [2]		✓	
Parsing	Human3.6M [33]	✓		
	LIP [19]	✓		
	CIHP [18]	✓		
	ATR [50]		✓	
Attribute	PA-100K [56]	✓		
	RAPv2 [39]	✓		
	PETA [10]		✓	
Detecton	CrowdHuman [70]	✓		
	Caltech [12]		✓	
Counting	ShTech PartA [104]			✓
	ShTech PartB [104]			✓

Specifically, we establish three evaluation scenarios for HumanBench: (1) *in-dataset evaluation:* we select 12 representative datasets whose training subsets are allocated to the pretraining dataset and evaluation subsets assigned to the evaluation dataset to evaluate the performance of a general pretrained model on diverse seen datasets (meaning similar data distribution for training and evaluation). (2) *out-of-dataset evaluation:* we select 5 datasets that do not appear in pretraining but belong to the seen task for evaluating the ability of the pretrained model on unseen datasets (meaning potentially different data distribution for training and evaluation). (3) *unseen-task evaluation:* we further add 2 datasets for crowd counting to evaluate the generalization ability to unseen tasks. More detailed distributions of these evaluation datasets are presented in Tab. 2.

Evaluation Protocols. For each evaluation scenario, we expect a good representation can generalize to specific human-centric tasks without updating the feature extractor or being a good starting point when adapted to any specific human-centric tasks by finetuning. Therefore, we present three evaluation protocols for the experiments presented in Sec. 5.

Full Finetuning. Full Finetuning evaluates the generalization ability when pretraining models serve as training starting points. In this case, we load the pretrained backbone and finetune all layers by supervision from downstream tasks.

Head Finetuning. Head Finetuning is very similar to linear probing [24] in self-supervised learning on natural image classification. It evaluates the generalization ability of pretrained models without updating. Therefore, we keep the pretrained backbone frozen and learn simple task heads for downstream datasets.

Partial Finetuning. Partial Fintuning is a setting between head finetuning and full finetuning [23], which finetunes the last several layers while freezing the others. This evaluation protocol can take advantage of both full finetuning and head finetuning, *i.e.,* it can efficiently evaluate the opportunity of pursuing strong but non-linear features.

4. Methodology

We now introduce our proposed projector-assisted pretraining method (PATH) with hierarchical weight sharing. Our method is motivated by [86], which reveals that inserting an MLP projector before the objective function can significantly increase the generalization ability of supervised pretraining. To avoid task conflicts among various tasks, we improve this method by inserting task-specific projectors between the backbone and the head of every dataset and designing a new hierarchical weight-sharing strategy. Concretely, the projectors are very lightweight modules composed of an attention module and a gating module (Sec. 4.2). The hierarchical weight-sharing strategy enforces that parameters of backbone, projectors, and heads are shared among all datasets across different tasks, shared among datasets in the same task, and not shared, respectively (Sec. 4.1). As such, we expect the backbone to learn general representations of all human-centric tasks, the projector to attend to the task-specific features from the general representations, and the head to supervise the network optimization by the annotations of every dataset. To evaluate the generalization ability of the pretrained backbone, we discard the projectors and heads, using the backbone only.

4.1. Hierarchical Weight Sharing

We design a hierarchical weight-sharing strategy to reduce task conflicts among various annotations. Specifically, our model consists of three components: a single backbone shared by all datasets, T task-specific projectors shared by all datasets in the same task, and N dataset-specific heads that are not shared, where $N = N_1 + N_2 + ... + N_T$ is the number of datasets in the pretraining dataset and N_t is the number of datasets in the t-th task.

Backbone. The backbone $\mathcal{F}$ is implemented by a plain vision transformer [13] in the experiments. The parameters of the backbone are shared by all datasets regardless of tasks.

Task-specific projector. Each task-specific projector $\mathcal{P}^t$ consists of sets of attention modules and gating modules, which link with the backbone $\mathcal{F}$, where $t \leq T$ and T is the number of tasks. Since the parameters of the task-specific projector are shared among the datasets with the same task, the attention modules in the task-specific network can be considered as selecting features from the shared backbone network, whilst the shared backbone network learns a compact global feature pool across all datasets.

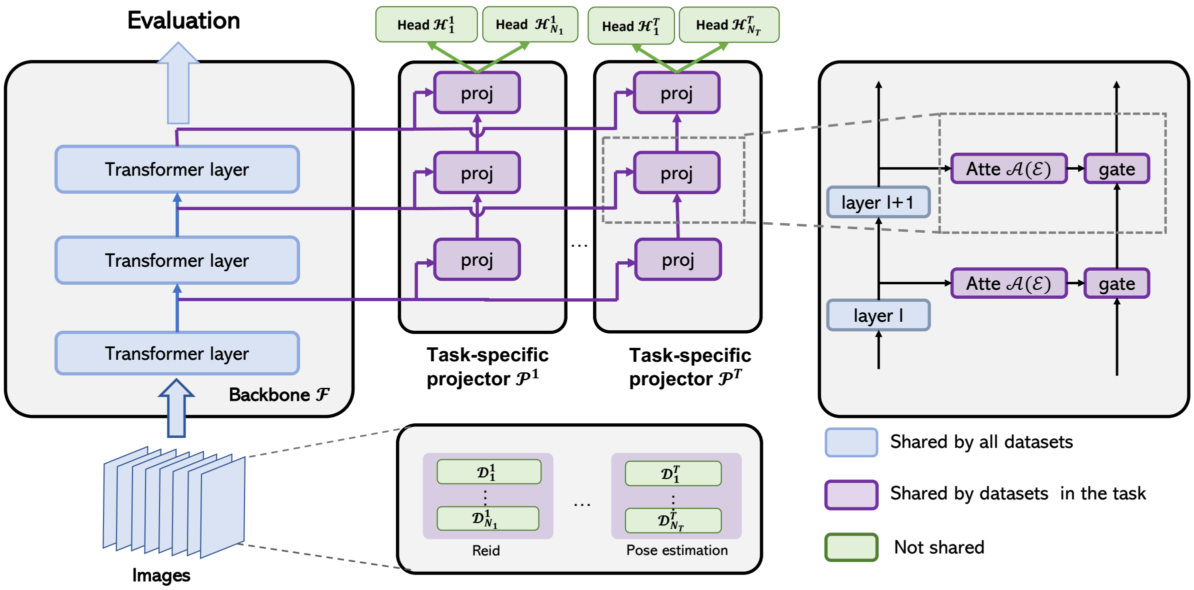

Figure 2. Overview of our proposed pretraining method, PATH. Images from various datasets are fed into the same backbone to extract the general features, and then the task-specific projector attends to the task-specific features from the general features. The dataset-specific head is imposed to predict dataset-specific results, which are fed into the loss function for training.

Dataset-specific head. To tackle the possible data distribution shift in different datasets, we still preserve the dataset-specific head $\mathcal{H}_j^t$, whose parameters are not shared. Here, $\mathcal{H}_j^t$ is the j-th dataset in the t-th task.

Figure 2 shows a detailed visualization of our PATH. The detailed pipeline is described as follows.

Step1: Extract the general features of images in the pretraining dataset. Given an image $\mathbf{x}$ sampled from $\mathcal{D}_j^t$ which is the j-th dataset in the t-th task in the pretraining dataset, extract the intermediate and final feature maps $\mathbf{F}$ by the backbone, which will be fed into the projectors.

Step2: Attend the task-specific features by the task-specific projector (Sec. 4.2). Given the feature maps $\mathbf{F}$ from the backbone, we attend the task-specific features $\mathbf{p} = \mathcal{P}^t(\mathbf{F})$ by the t-th task-specific projector.

Step3: Calculate the activations by dataset-specific heads, losses by the activations, and optimize the parameters of the backbone, the projector, and the head simultaneously by backward propagation (Sec. 4.3).

During the evaluation stage, we discard the projectors and evaluate the generalization ability of the backbone $\mathcal{F}$ using the protocols in Sec. 3.2.

4.2. Design of Task-specific Projector

The task-specific projector is designed to attend to task-specific features from backbone outputs, by applying an alternating chain of the attention module and gating module to the features in the shared backbone. Given an image $\mathbf{x}$ sampled from the j-th dataset in the t-th task, *i.e.*, $\mathcal{D}_j^t$ and its intermediate feature maps $\mathbf{f}_l$ in the l-th transformer block, we leverage a squeeze-and-excitation layer [31] to implement the channel attention and a self-attention module [80] to implement spatial attention to generate the attended feature maps $\mathbf{z}_l$. Mathematically, $\mathbf{z}_l = \mathcal{A}^t(\mathcal{E}^t(\mathbf{f}_l))$, where $\mathcal{A}^t$ and $\mathcal{E}^t$ respectively denote standard self-attention blocks [80] and squeeze-and-excitation blocks [31] for the t-th task. We will detail the structure of the Squeeze-and-excitation module and self-attention module in the supplementary materials.

In order to effectively aggregate features from different layers of the backbone, we design a gating module to dynamically aggregate features from different layers. Specifically, given the feature map $\mathbf{z}_l$ after the attention module and the gated feature maps $\mathbf{p}_{l-1}$ in the $(l-1)$-th layer, the gating function aggregates features as follows:

$$\mathbf{p}_l = \mu_l \mathbf{z}_l + (1 - \mu_l)\mathbf{p}_{l-1}, \tag{1}$$

where $\mathbf{p}_1 = \mathbf{z}_1$, $\mu_l = \sigma(\alpha_l / T)$ is a gate parameterized with a learnable zero-initilalized scalar α_i and temperature T(=0.1), and σ is the sigmoid function. By iteratively computing Eq. 1 from $l = 1$ to L, we generate the final feature maps *i.e.*, $\mathbf{p} = \mathbf{p}_L$ for the dataset head.

4.3. Dataset-specific Head and Objective Functions

Dataset-specific heads aim at transforming task-specific features into activations for computing losses of every dataset. In general multi-dataset learning with N datasets, the features $\mathbf{P}_i$ after the projector of all images $\mathbf{X}_i$ and labels $\mathbf{Y}_i$, $i = 1, 2, ..., N$ in i-th dataset, the objective function is defined as $\mathcal{L} = \sum_{i=1}^{N} \lambda_i \mathcal{L}_i(\mathbf{Z}_i, \mathbf{Y}_i)$, where $\mathbf{Z}_i$ is the activation generated by the dataset-specific head. This is the linear combination of dataset-specific losses $\mathcal{L}_i$ with task weightings λ_i. In this paper, we follow some basic head and loss function designs of all pretraining tasks we include. Specifically, we follow the head and loss function designs in VitPose [92] for pose estimation, in TransReID [26] for person ReID, in Segformer [91] for human parsing, in Anchor Detr [87] for pedestrian detection, in Label2Label [44] for pedestrian attribute recognition, and in DR.VIC [21] for crowd counting. More details of these head and loss designs will be elaborated in the supplementary materials.

4.4. Technical Details

Replacing all Batchnorm with Layernorm in pose and parsing decoders. Generally, the original feature normalization method in pose estimation and human parsing tasks is batch normalization with CNN backbone, which renders the model to learn powerful feature distribution from the statistics of batch inputs when trained on a single domain. However, in HumanBench, each task has different datasets, which may have domain gaps, resulting in inaccurate normalization statistics when the dataset-specific head is fed with features from the task-share projector. To reduce the inaccurate statistics, we replace the Normalization method from BatchNorm [32] to LayerNorm [3] and experimentally find that it can improve feature representation.

Sharing Positional Embedding among All Datasets. In HumanBench, the input image size of different tasks varies largely, resulting in different numbers of patch embeddings and positional embeddings after projecting an image to patch embedding. As a result, different tasks can not share positional embeddings when the model is trained in a distributed manner. To tackle this problem, we parameterize positional embeddings as 224×224 in all tasks and interpolate its size according to the actual input image size of each dataset during the pretraining stage.

5. Experiment

5.1. Experimental Setup

The backbone used for experiments is the plain ViT-base. It has 12 transformer blocks with the dimension of patch embedding 768 and 12 attention heads. In the pre-train stage, each GPU is responsible for one dataset independently for training in a distributed manner. We use Adafactor [71] optimizer with base learning rate of 5e-4 and weight decay of 0.05. We linearly warmup the learning rate from 1e-7 to 5e-4 for the first 1500 iteration steps. Step learning rate decay of 0.5 is used in 50%, 75%, 95% iterations. For the ViT-Base encoder, we set a layer-wise learning rate decay of 0.75 for 12 transformer blocks and the model is trained for 80k iterations.

5.2. Experimental Results

As detailed in Sec. 3.2, we implement in-dataset evaluation, out-of-dataset evaluation, and unseen-task evaluation on **HumanBench**. Both in-dataset evaluation, out-of-dataset evaluation include 5 human-centric tasks, *i.e.,* person ReID, pose estimation, human parsing, pedestrian attribute recognition, and pedestrian detection. The unseen downstream task which is not in the pretraining tasks, *i.e,* crowd counting, evaluates the generalization ability to unseen tasks. The compared methods are the state-of-the-art methods of each task and two popular pretraining models, *i.e.,* MAE [23] and CLIP [64]. MAE is a newly proposed vision self-supervised pretraining method. Pretrained on ImageNet-1K, MAE achieves excellent results for many visual tasks. CLIP learns generic and transferable representations from a dataset of 400 million (image, text) pairs. We summarize our experimental results with 3 evaluation scenarios and 3 evaluation protocols in Tab. 3.

In-dataset Evaluation. In-dataset evaluation quantifies the ability of the pretraining method when it is evaluated on the data with similar data distribution and pretrained tasks. As shown in Tab. 3, compared with SoTA methods used in different papers for their specific tasks, our HumanBench with full finetuning achieves better performance on 8 datasets. Specifically, for human parsing, we improve the current state-of-the-art results by **+2.5%** mIOU, **+1.1%** mIOU and **+1.2%** mIOU on Human3.6M, LIP and CIHP, respectively. We also improve the person ReID by **+4.9%** mAP on CUHK03 datasets. We notice our results are lower than PASS [116] on Market1501 and MSMT, probably because PASS uses techniques, *i.e.,* part models [76, 83], that are time-consuming (120 hours using 8 A100 GPUs) but specifically effective for ReID. Besides, we improve pose estimation by **+3.0%** AP, **-1.2%** $\mathrm{MR}^{-2}(\downarrow)$ on AIC and Human3.6m, respectively. Furthermore, we improve pedestrian attribute recognition **+1.5%** mA and **+0.2%** mA on PA-100K and RAPv2 datasets, respectively.

To evaluate the generalization of different methods when all backbone parameters or most of the backbone parameters are frozen, we further evaluate our HumanBench with head finetuning and partial finetuning with 100% of the downstream data. We observe that our method with only head finetuning can be on par with and even surpasses the SoTAs in 12 seen datasets, such as **-1.2%** heavy occluded $\mathrm{MR}^{-2}(\downarrow)$ **+1.6%** mIOU on Human3.6m pose estimation and human parsing tasks. Our HumanBench with partial finetuning performs better than full finetuning in 2 Pedestrian Attribute Recognition datasets (PA-100K and RapV2) of 12 seen datasets, possibly because these two datasets have fewer data.

We also use ViT-Large to verify the model scalability of our method PATH on HumanBench in Tab. 3. Results show that the results with a large backbone under partial

Table 3. Experimental results of our PATH and recent state-of-the-art methods (SoTA in the table) on 6 human-centric tasks. The results include 12 *in-dataset evaluations*, 5 *out-of-dataset evaluations* (columns w. gray) and 2 *unseen task evaluations* on the unseen counting task. Following the most commonly-used metrics, for human parsing tasks, we report pACC for ATR, mIoU for others. † indicates that the results are obtained with additional information, multitask learning, or stronger models. We highlight the **best** using ViT-Base and ViT-Large backbone, respectively. We also highlight these best results in red if they outperform SoTAs.

		Human Parsing				Person ReID				Pedestrian Detection	
		Human3.6M	LIP	CIHP	ATR	Market1501	MSMT	CUHK03	SenseReID	CrowdHuman	Caltech (↓)
	SoTA	62.5 [30]	60.3 [54]	65.6 [54]	97.4 [54]	86.8 [26]	61.0 [26]	76.4 [35]	34.6 [106]	92.1 [108]	46.6 [22]
	SoTA †	-	-	-	-	93.0 [116]	71.8 [116]	77.7 [42]	-	92.5 [108]	28.8 [22]
ViT-B	MAE	62.0	57.2	62.9	97.4	79.2	51.5	65.8	43.8	89.6	48.1
	CLIP	58.2	53.4	61.7	97.0	78.6	53.6	66.9	42.5	82.1	-
	PATH (w/o FT)	63.9	56.3	63.9	-	88.6	66.3	77.2	-	89.1	-
	PATH (FT)	**65.0**	**61.4**	**66.8**	**97.5**	**89.5**	**69.1**	**82.6**	47.7	90.6	30.1
	PATH (Head FT)	64.1	59.9	63.3	97.1	-	-	-	-	90.0	31.1
	PATH (Partial FT)	63.7	60.0	63.1	97.2	88.7	66.1	79.5	**48.2**	**90.9**	**28.3**
ViT-L	PATH (w/o FT)	65.0	**62.9**	67.1	-	91.6	72.7	83.7	-	89.4	-
	PATH (Partial FT)	**66.2**	62.6	**67.5**	**97.4**	**91.8**	**74.7**	**86.0**	**60.0**	**90.8**	**28.7**

		Pose Estimation				Pedestrian Attribute Recognition			Counting (unseen task)	
		COCO	Human3.6M (↓)	AIC	MPII	PA-100K	RAPv2	PETA	ShTech PartA (↓)	ShTech PartB (↓)
	SoTA	75.8 [92]	7.4 [75]	-	92.3 [95]	83.5 [34]	81.0 [34]	87.1 [34]	94.3 [72]	11.0 [72]
	SoTA †	77.1 [92]	-	32.0 [92]	93.3 [92]	-	-	-	-	-
ViT-B	MAE	75.8	8.2	31.8	90.1	82.3	80.8	84.6	102.1	15.5
	CLIP	74.4	9.9	31.1	88.1	76.1	77.0	81.2	117.9	16.3
	PATH (w/o FT)	75.0	6.9	31.1	-	-	-	-	-	-
	PATH (FT)	**76.3**	6.2	**35.0**	**93.3**	85.0	81.2	88.0	**91.7**	**10.8**
	PATH (Head FT)	75.2	**6.1**	31.6	92.7	77.4	72.4	79.0	-	-
	PATH (Partial FT)	76.0	**6.1**	33.3	93.0	**86.9**	**83.1**	**89.8**	-	14.0
ViT-L	PATH (w/o FT)	74.7	7.1	25.6	-	-	-	-	-	-
	PATH (Partial FT)	**77.1**	**5.8**	**36.3**	**93.7**	**90.8**	**87.4**	**90.7**	-	-

finetuning can further achieve considerable gains over the best ViT-Base results, showing the promising scalability of our proposed pretraining method PATH on HumanBench.

Out-of-dataset Evaluation. To quantify the generalization ability of pretrained models on tasks with potentially different data distribution but the same task in the pretraining dataset, we implement out-of-dataset evaluations on 5 datasets, *i.e.,* ATR, SenseReID, Caltech, MPII, PETA, one dataset for each pretraining task. As shown in Tab. 3, our pretraining method PATH performs better than previous methods in 4 of 5 unseen datasets and comparable in the remaining one. To be concrete, our method improves by **+0.1%**pACC, **+4.4%**Top1 accuracy, **-0.5%** heavy occluded $\text{MR}^{-2}(\downarrow)$ and **+2.7%** mA on ATR (human parsing), SenseReID (person ReID), Caltech (pedestrian detection) and PETA datasets (pedestrian attribute recognition), respectively. These significant and consistent performance gains across different datasets verify the generalization ability of our pretrained model to tasks with potentially different data distributions. We also observe the results when we only finetune the last two layers are already on par or even better than the results by full finetuning. Especially, the results of SenseReID, Caltech and PETA by partial finetuning are better than that of full finetuning by **+0.5%**Top1 accuracy, **-1.8%**heavy occluded $\text{MR}^{-2}(\downarrow)$ and **+1.8%**mA, showing the good generalization of our pretrained models and its easy deployment in the real world. Similar to the results in the *out-of-dataset evaluation*, partial finetuing performs better than full finetuning when the dataset is small in Caltech (4250 images) and PETA (9500 images). Therefore, partial finetuning can be a choice when the downstream dataset has few samples.

Unseen-task Evaluations. To evaluate the generalization ability to unseen tasks, we construct an *unseen task evaluation*, in which the task is not involved in the pretraining tasks, *i.e,* crowd counting. As presented in Tab. 3, our pretrained model achieves significant performance gains than the MAE pretrained model by **-10.4%** MSE(↓) and **-4.7%** MSE(↓). Our HumanBench improves previous SoTAs that are specially designed for crowd counting by **-2.6%** MSE(↓) and **-0.2%** MSE(↓) on ShTech PartA and PartB datasets, respectively. These consistent improvements validate the generalization ability of our learned representations.

Comparsion with MAE and CLIP Models. We also compare our pretrained method with other popular pretrained models, *i.e.,* MAE and CLIP, on our proposed HumanBench. In Tab. 3, we find our pretraining method performs considerably better than CLIP and MAE under the full finetuning evaluation protocol on all tasks. Interestingly, the

Table 4. Ablation results. "A", "S", and "T" respectively denote all shared, specific, and task-shared projector. † indicates the results are reported as 1-heavy occluded MR^{-2} for averaging.

		(a)	(b)	(c)	(d)
	Shared Pos. Embedding				✓
	Projector Share Type	A	S	T	T
Detection	Caltech †	60.6	57.5	60.2	60.9
Attribute	PA100K	84.4	84.6	84.6	84.4
	PETA	87.6	87.2	87.9	87.5
Pose	MPII	92.0	92.4	92.5	92.4
Parsing	LIP	59.7	59.7	60.5	61.0
ReID	Market1501	86.2	86.5	87.1	87.6
	MSMT	65.8	66.0	66.1	66.8
	On average	**76.6**	**76.3**	**76.9**	**77.2**

Table 5. Comparison with self-supervised pretraining methods on ImageNet and the subset of our HumanBench. † indicates the results are repoted as 1-heavy occluded MR^{-2} for averaging.

Pretraining data		ImageNet-1K	Our subset		
Method		MAE	MAE	MOCOv3	Ours
Detection	Caltech †	51.9	58.2	57.5	60.9
Attribute	PA100K	82.3	83.5	82.9	84.4
	PETA	84.6	85.3	84.3	87.5
Pose	MPII	90.1	91.3	90.4	92.4
Parsing	LIP	57.2	60.1	58.6	61.0
ReID	Market1501	79.2	84.6	86.8	87.6
	MSMT	51.5	64.5	67.2	66.8
On average		**71.0**	**75.4**	**75.4**	**77.2**

performance of CLIP[2] is lower than MAE, which shows that more data on natural images and languages may not naturally benefit a variety of human-centric tasks, which empirically validates the importance of our HumanBench for further research on human-centric pretraining.

5.3. Ablation Study

Due to the significant computation cost with the large-scale full datasets, as summarized in Table 1 in **Supplementary Material**, we sample a subset containing a similar number of images as ImageNet-1K ($\sim$ 1.28 M) from the full training set. We pretrain our models on this subset to verify the effectiveness of our designs by default in this section, and implement 4 in-dataset evaluations (PA-100K, LIP, Market1501, MSMT) and 3 out-of-dataset evaluations (Caltech, PETA, MPII) full dataset finetuning.

Effectiveness of hierarchical weight sharing. To verify the effectiveness of our hierarchical weight sharing, we adapt the three projector share strategies: (1) all shared projector (**A**): sharing the projector parameters across all the tasks and datasets (Table 4 (a)); (2) task-shared projector (**T**): sharing the projector parameters across all datasets in a single task, while maintaining an independent projector for each task (Table 4 (c)); and (3) specific projector (**S**): maintaining an independent projector for each dataset (Table 4 (b)). The results show that the task-shared projector is better than the other two. We speculate that the projector is the core component to map the general human-centric features to task-specific features. Therefore, all datasets in the same task are supposed to share the same mapping functionality while different tasks should operate differently due to the existing task gaps.

Effectiveness of shared positional embedding. Experiments (c) and (d) in Tab. 4 ablate whether positional embeddings are shared or not across the different tasks. The results show that shared positional embedding helps to learn general human-centric representations and leads to **+0.3%** improvement on average when the five tasks are considered. Since the backbone is shared, we conjecture that independent positional embedding would cause inconsistency between different tasks, create barriers in learning shared backbone across tasks, and result in difficulties in learning the model.

Comparison with self-supervised pretraining methods. As shown in Table 5, we first ablate the effectiveness of our dataset on downstream tasks. With almost the same number of images, the MAE pretrained on our subset for 800 epochs surpasses ImageNet pretrained MAE by +4.4%, which shows that by combining the diverse human-centric data across various human-centric tasks, our dataset is more suitable to learn human-centric features. Second, pretrained on our subset, our supervised pretraining method, i.e. PATH, performs better than both MAE (800 epochs) and MOCOv3 (800 epochs) by +1.8%. Different from MAE and MOCOv3 which ignore the general properties of the human body and the potential association between the data in different tasks, our PATH is designed to capture the potential complementary knowledge between different tasks, leading to learning more general human-centric representations to improve the performance on various human-centric tasks.

6. Conclusion

In this paper, we investigate the opportunities and challenges in pretraining on various human-centric tasks, and propose a new HumanBench with the existing publicly available datasets. Based on HumanBench, we design a projector-assisted pretraining with hierarchical weight sharing (PATH) to learn human-centric information from annotations with different granularities. We hope our HumanBench can facilitate future works such as unified network structure design and multi-task/supervised/self-supervised learning methods on a broad variety of human-centric tasks.

[2]We carefully tune the learning rate, drop path rate, and weight decay of CLIP pretrained ViT-B and report the best results we have ever achieved.

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. 15

[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. 4

[3] Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. Layer normalization. *arXiv preprint arXiv:1607.06450*, 2016. 6

[4] 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. 15

[5] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020. 1

[6] 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. 14

[7] Zhao Chen, Vijay Badrinarayanan, Chen-Yu Lee, and Andrew Rabinovich. Gradnorm: Gradient normalization for adaptive loss balancing in deep multitask networks. In *International conference on machine learning*, pages 794–803. PMLR, 2018. 3

[8] Xuangeng Chu, Anlin Zheng, Xiangyu Zhang, and Jian Sun. Detection in crowded scenes: One proposal, multiple predictions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12214–12223, 2020. 1

[9] Charles Darwin and Phillip Prodger. *The expression of the emotions in man and animals*. Oxford University Press, USA, 1998. 3

[10] 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. 4

[11] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018. 1

[12] 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. 4

[13] 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. 1, 4

[14] 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. 15

[15] 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. 15

[16] Yixiao Ge, Dapeng Chen, and Hongsheng Li. Mutual mean-teaching: Pseudo label refinery for unsupervised domain adaptation on person re-identification. *arXiv preprint arXiv:2001.01526*, 2020. 1

[17] Yixiao Ge, Feng Zhu, Dapeng Chen, Rui Zhao, et al. Self-paced contrastive learning with hybrid memory for domain adaptive object re-id. *Advances in Neural Information Processing Systems*, 33:11309–11321, 2020. 1

[18] 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, 15, 21

[19] 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, 15

[20] Michelle Guo, Albert Haque, De-An Huang, Serena Yeung, and Li Fei-Fei. Dynamic task prioritization for multitask learning. In *Proceedings of the European conference on computer vision (ECCV)*, pages 270–287, 2018. 3

[21] Tao Han, Lei Bai, Junyu Gao, Qi Wang, and Wanli Ouyang. Dr. vic: Decomposition and reasoning for video individual counting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3083–3092, 2022. 6

[22] 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

[23] 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. 4, 6

[24] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9729–9738, 2020. 4

[25] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969, 2017. 3

[26] 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. 6, 7

[27] Yinan He, Gengshi Huang, Siyu Chen, Jianing Teng, Kun Wang, Zhenfei Yin, Lu Sheng, Ziwei Liu, Yu Qiao, and Jing Shao. X-learner: Learning cross sources and tasks for universal visual representation. In *European Conference on Computer Vision*, pages 509–528. Springer, 2022. 3

[28] Alexander Hermans, Lucas Beyer, and Bastian Leibe. In defense of the triplet loss for person re-identification. *arXiv preprint arXiv:1703.07737*, 2017. 19

[29] Fangzhou Hong, Liang Pan, Zhongang Cai, and Ziwei Liu. Versatile multi-modal pre-training for human-centric perception. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16156–16166, 2022. 1, 3

[30] 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

[31] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7132–7141, 2018. 5

[32] 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. 6, 19

[33] 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, 15

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

[35] 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

[36] Alex Kendall, Yarin Gal, and Roberto Cipolla. Multi-task learning using uncertainty to weigh losses for scene geometry and semantics. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7482–7491, 2018. 3

[37] 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. 3

[38] Dmitry Lepikhin, HyoukJoong Lee, Yuanzhong Xu, Dehao Chen, Orhan Firat, Yanping Huang, Maxim Krikun, Noam Shazeer, and Zhifeng Chen. Gshard: Scaling giant models with conditional computation and automatic sharding. *arXiv preprint arXiv:2006.16668*, 2020. 1

[39] 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. 4, 15

[40] 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. 15

[41] 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. 3, 15

[42] 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

[43] 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. 15

[44] 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. 6, 20, 21

[45] 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. 4, 15

[46] 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. 15

[47] Yanghao Li, Chao-Yuan Wu, Haoqi Fan, Karttikeya Mangalam, Bo Xiong, Jitendra Malik, and Christoph Feichtenhofer. Mvitv2: Improved multiscale vision transformers for classification and detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4804–4814, 2022. 1

[48] 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. 1

[49] 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, 19

[50] 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. 4

[51] 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. 1

[52] 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, 14, 15

[53] Bo Liu, Xingchao Liu, Xiaojie Jin, Peter Stone, and Qiang Liu. Conflict-averse gradient descent for multi-task learning. *Advances in Neural Information Processing Systems*, 34:18878–18890, 2021. 2

[54] 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

[55] Shikun Liu, Edward Johns, and Andrew J Davison. End-to-end multi-task learning with attention. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1871–1880, 2019. 3

[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. 4, 15

[57] 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. 14, 15

[58] 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. 15

[59] Yongxi Lu, Abhishek Kumar, Shuangfei Zhai, Yu Cheng, Tara Javidi, and Rogerio Feris. Fully-adaptive feature sharing in multi-task networks with applications in person attribute classification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5334–5343, 2017. 3

[60] 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. 1, 19, 21

[61] Liqian Ma, Xu Jia, Qianru Sun, Bernt Schiele, Tinne Tuytelaars, and Luc Van Gool. Pose guided person image generation. *Advances in neural information processing systems*, 30, 2017. 1

[62] 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. 15

[63] Ishan Misra, Abhinav Shrivastava, Abhinav Gupta, and Martial Hebert. Cross-stitch networks for multi-task learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3994–4003, 2016. 3

[64] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763. PMLR, 2021. 6

[65] Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al. Improving language understanding by generative pre-training. 2018. 1

[66] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019. 1

[67] Mert Bulent Sariyildiz, Yannis Kalantidis, Diane Larlus, and Karteek Alahari. Concept generalization in visual representation learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9629–9639, 2021. 2

[68] Ozan Sener and Vladlen Koltun. Multi-task learning as multi-objective optimization. *Advances in neural information processing systems*, 31, 2018. 3

[69] Jing Shao, Siyu Chen, Yangguang Li, Kun Wang, Zhenfei Yin, Yinan He, Jianing Teng, Qinghong Sun, Mengya Gao, Jihao Liu, et al. Intern: A new learning paradigm towards general vision. *arXiv preprint arXiv:2111.08687*, 2021. 3

[70] 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, 15

[71] 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

[72] Weibo Shu, Jia Wan, Kay Chen Tan, Sam Kwong, and Antoni B Chan. Crowd counting in the frequency domain. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19618–19627, 2022. 7

[73] 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. 15

[74] 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. 15

[75] 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

[76] Yifan Sun, Liang Zheng, Yi Yang, Qi Tian, and Shengjin Wang. Beyond part models: Person retrieval with refined

part pooling (and a strong convolutional baseline). In *Proceedings of the European conference on computer vision (ECCV)*, pages 480–496, 2018. 6

[77] Deepa Suryawanshi et al. *Image Recognition: Detection of nearly duplicate images*. PhD thesis, California State University Channel Islands, 2018. 3

[78] Marvin Teichmann, Michael Weber, Marius Zoellner, Roberto Cipolla, and Raquel Urtasun. Multinet: Real-time joint semantic reasoning for autonomous driving. In *2018 IEEE intelligent vehicles symposium (IV)*, pages 1013–1020. IEEE, 2018. 3

[79] Hanyue Tu, Chunyu Wang, and Wenjun Zeng. Voxelpose: Towards multi-camera 3d human pose estimation in wild environment. In *European Conference on Computer Vision*, pages 197–212. Springer, 2020. 1

[80] 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. 1, 5, 19

[81] 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. 15

[82] 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. 15

[83] Guanshuo Wang, Yufeng Yuan, Xiong Chen, Jiwei Li, and Xi Zhou. Learning discriminative features with multiple granularities for person re-identification. In *Proceedings of the 26th ACM international conference on Multimedia*, pages 274–282, 2018. 6

[84] 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. 3

[85] Wenhui Wang, Hangbo Bao, Li Dong, Johan Bjorck, Zhiliang Peng, Qiang Liu, Kriti Aggarwal, Owais Khan Mohammed, Saksham Singhal, Subhojit Som, et al. Image as a foreign language: Beit pretraining for all vision and vision-language tasks. *arXiv preprint arXiv:2208.10442*, 2022. 1, 3

[86] Yizhou Wang, Shixiang Tang, Feng Zhu, Lei Bai, Rui Zhao, Donglian Qi, and Wanli Ouyang. Revisiting the transferability of supervised pretraining: an mlp perspective. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9183–9193, 2022. 2, 3, 4

[87] 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. 1, 6, 20

[88] 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. 4, 15

[89] 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, 15

[90] 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. 1

[91] Enze Xie, Wenhai Wang, Zhiding Yu, Anima Anandkumar, Jose M Alvarez, and Ping Luo. Segformer: Simple and efficient design for semantic segmentation with transformers. *Advances in Neural Information Processing Systems*, 34:12077–12090, 2021. 6

[92] Yufei Xu, Jing Zhang, Qiming Zhang, and Dacheng Tao. Vitpose: Simple vision transformer baselines for human pose estimation. *arXiv preprint arXiv:2204.12484*, 2022. 1, 6, 7, 19, 21

[93] 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. 15

[94] 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. 15

[95] 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

[96] Shijie Yu, Feng Zhu, Dapeng Chen, Rui Zhao, Haobin Chen, Shixiang Tang, Jinguo Zhu, and Yu Qiao. Multiple domain experts collaborative learning: Multi-source domain generalization for person re-identification. *arXiv preprint arXiv:2105.12355*, 2021. 1

[97] Tianhe Yu, Saurabh Kumar, Abhishek Gupta, Sergey Levine, Karol Hausman, and Chelsea Finn. Gradient surgery for multi-task learning. *Advances in Neural Information Processing Systems*, 33:5824–5836, 2020. 2

[98] 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. 3

[99] 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*. 19

[100] 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. 15

[101] 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. 15

[102] 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. 3, 15

[103] 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. 14

[104] Yingying Zhang, Desen Zhou, Siqin Chen, Shenghua Gao, and Yi Ma. Single-image crowd counting via multi-column convolutional neural network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 589–597, 2016. 4

[105] 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. 19

[106] 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. 4, 7

[107] Nanxuan Zhao, Zhirong Wu, Rynson WH Lau, and Stephen Lin. What makes instance discrimination good for transfer learning? *arXiv preprint arXiv:2006.06606*, 2020. 2

[108] 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. 7, 21

[109] 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, 15

[110] 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. 19

[111] 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. 15

[112] Yi Zheng, Shixiang Tang, Guolong Teng, Yixiao Ge, Kaijian Liu, Jing Qin, Donglian Qi, and Dapeng Chen. Online pseudo label generation by hierarchical cluster dynamics for adaptive person re-identification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8371–8381, 2021. 1

[113] 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. 15

[114] 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. 15

[115] Jinguo Zhu, Xizhou Zhu, Wenhai Wang, Xiaohua Wang, Hongsheng Li, Xiaogang Wang, and Jifeng Dai. Uni-perceiver-moe: Learning sparse generalist models with conditional moes. *arXiv preprint arXiv:2206.04674*, 2022. 1, 3

[116] Kuan Zhu, Haiyun Guo, Tianyi Yan, Yousong Zhu, Jinqiao Wang, and Ming Tang. Pass: Part-aware self-supervised pre-training for person re-identification. In *European Conference on Computer Vision*, pages 198–214. Springer, 2022. 6, 7