Structure-Aware Masked Image Modeling for Human-Centric Perception

Junkun Yuan[1, 2], Xinyu Zhang[2], Hao Zhou[2], Jian Wang[2], Zhongwei Qiu[3], Zhiyin Shao[4], Shaofeng Zhang[5], Sifan Long[6], Kun Kuang[1], Kun Yao[2], Junyu Han[2], Errui Ding[2], Lanfen Lin[1], Fei Wu[1], Jingdong Wang[2]

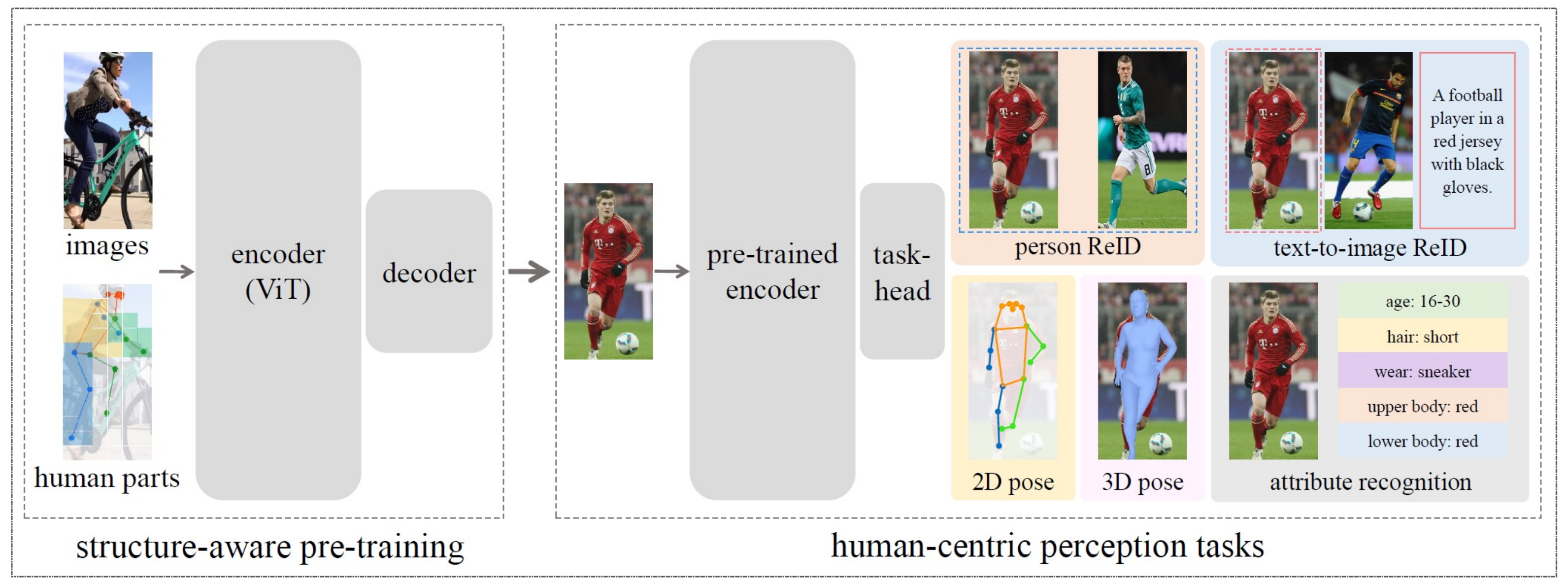
[1]Zhejiang University, [2]Baidu VIS, [3]University of Science and Technology Beijing, [4]South China University of Technology, [5]Shanghai Jiao Tong University, [6]Jilin University

NEURAL INFORMATION PROCESSING SYSTEMS

Motivation

- Human-centric perception includes a broad range of human-related tasks, including person ReID, human pose estimation, attribute recognition, etc.
- Due to the independent nature of these tasks, the efficiency of data utilization and training is limited, and the performance is suboptimal.

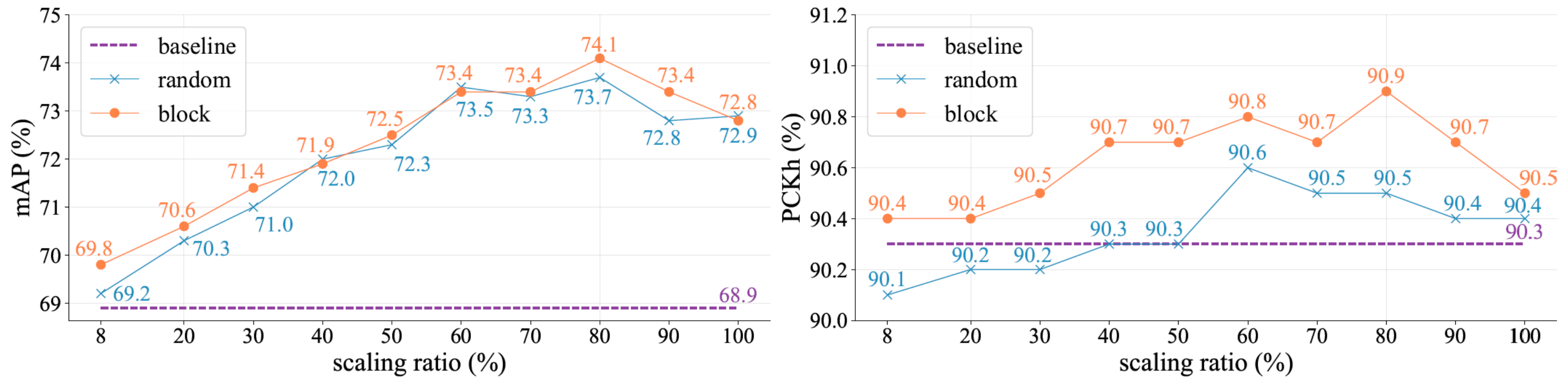
Human-Centric Pre-Training



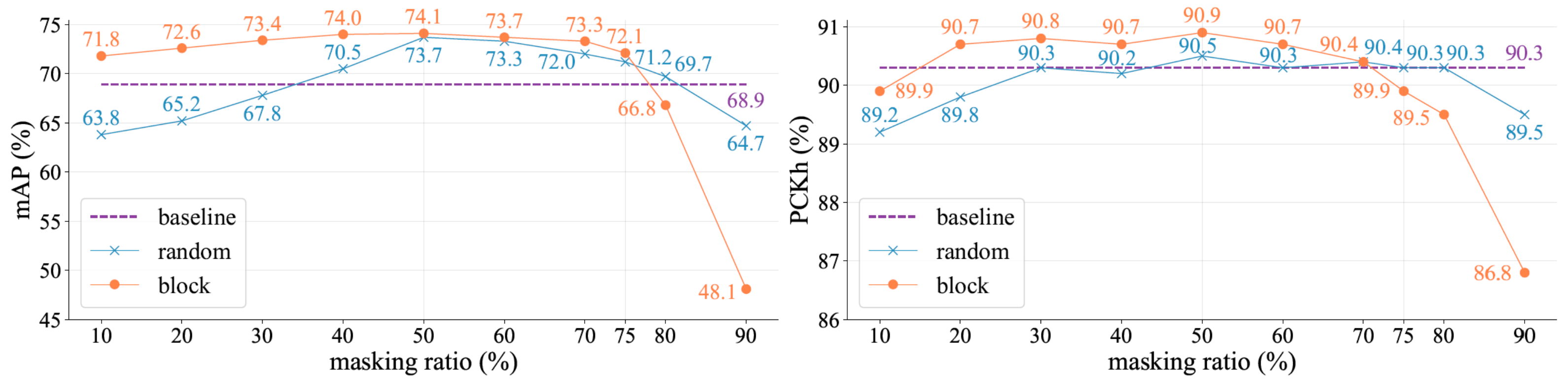
We propose a novel human-centric pre-training framework named HAP:

structure-aware pre-training + downstream fine-tuning

Great Potential of Human Structure Priors



Analysis of scaling ratio for (left) person ReID and (right) 2D pose estimation.



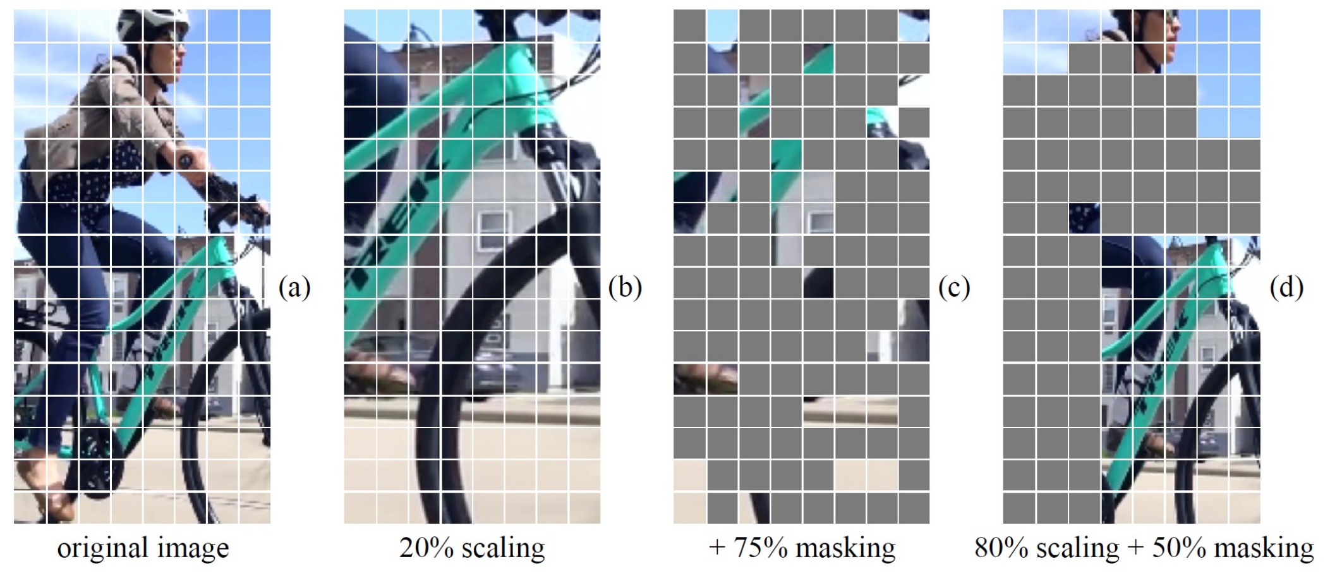
Analysis of masking ratio for (left) person ReID and (right) 2D pose estimation.

Empirical study shows great potential of human structure-related training factors:

(i) High scaling ratio (ranging from 60% to 90%)

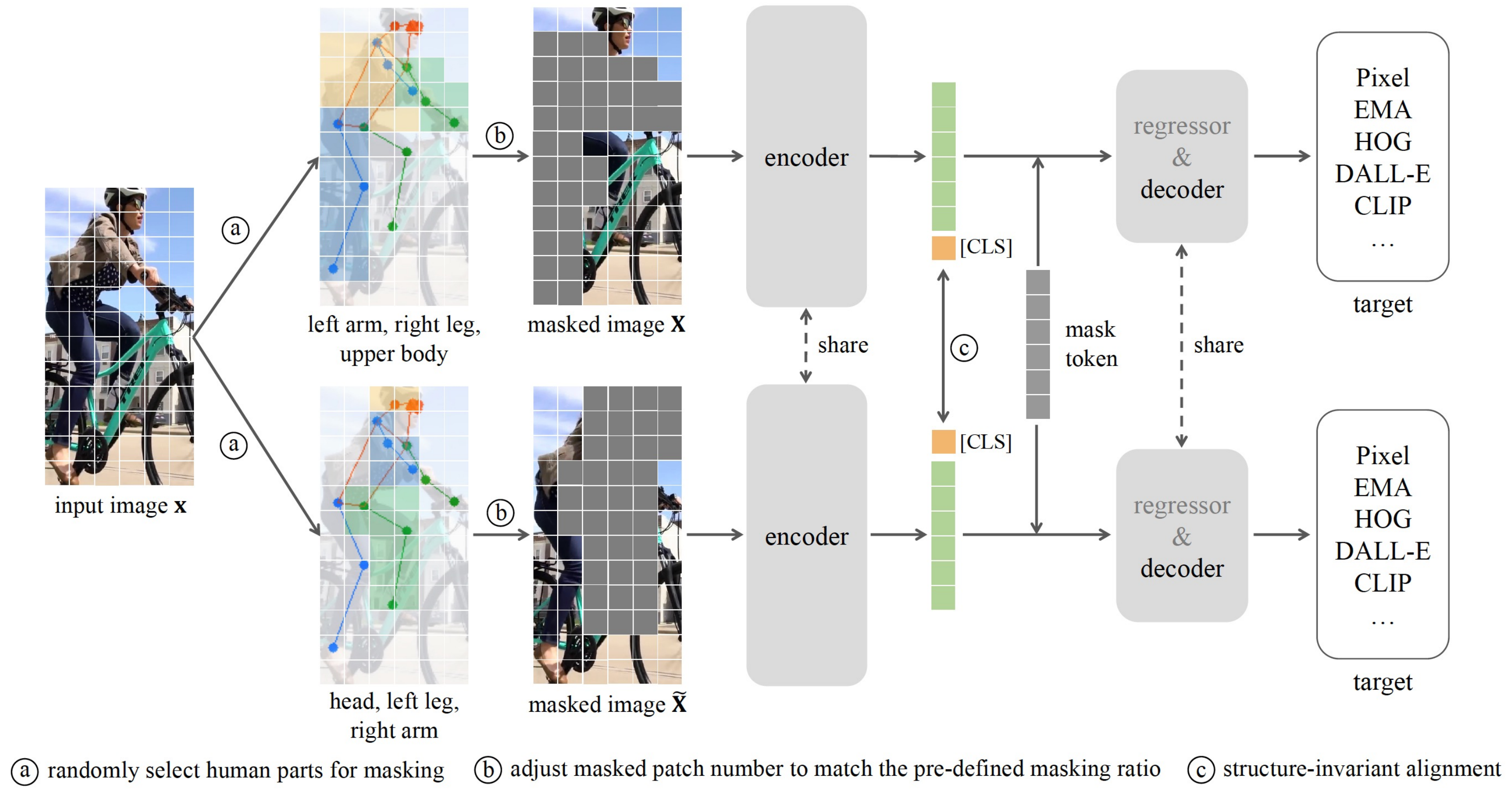
(ii) Mediate masking ratio (ranging from 40% to 60%)

(iii) Block-wise masking



- For a given image (a), the baseline of MAE uses 20% scaling ratio (b) and 75% masking ratio (c) with random mask sampling strategy, yielding a meaningless image with little human structure information.
- We adopt 80% scaling ratio and 50% masking ratio with block-wise mask sampling (d), maintaining the overall body structure.

HAP: Structure-Aware Pre-Training

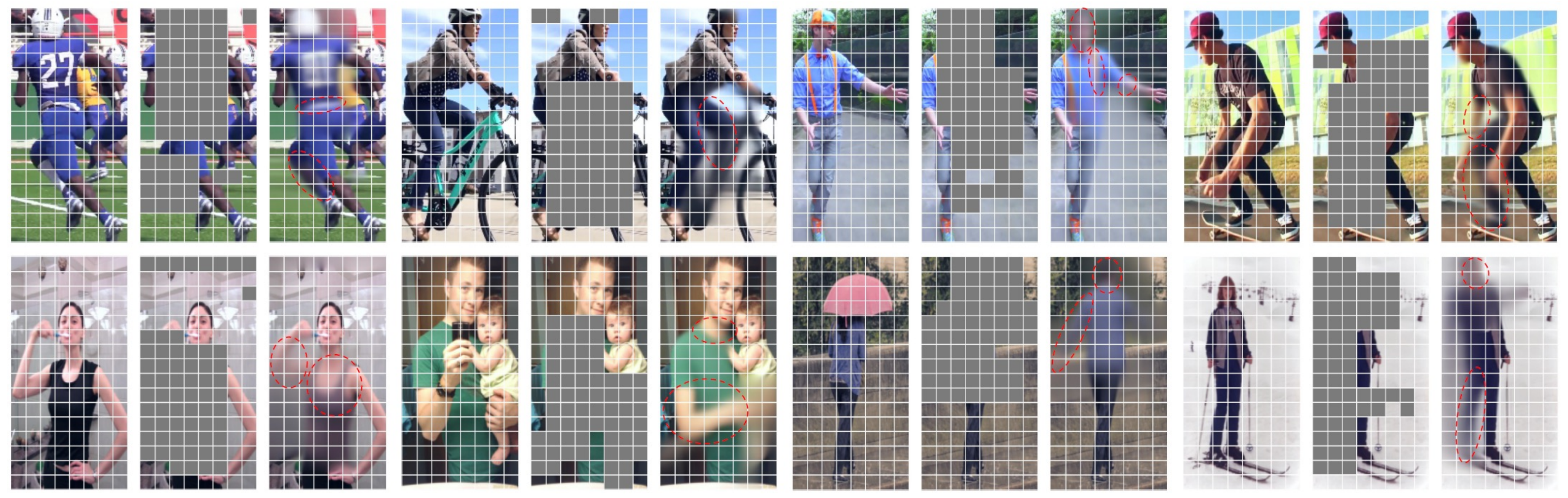


$\mathcal{L}_{\text{recon}}$: human body parts are randomly masked out to reconstruct

$\mathcal{L}_{\text{align}}$: [CLS] tokens of different view through random part masking are aligned

$$\mathcal{L} = \mathcal{L}_{\text{recon}} + \gamma\mathcal{L}_{\text{align}}$$

Reconstruction of Corrupted Images



HAP generates semantically reasonable body parts.

Results on Human-Centric Perception Tasks

statistics of human-centric pre-training methods

method	publication	datasets	samples
LiftedCL	ICLR 2023	1	~150K
SOLIDER	CVPR 2023	1	~4.2M
HCMoCo	CVPR 2022	2	~82K
UniHCP	CVPR 2023	33	~2.3M
PATH	CVPR 2023	37	~11.0M
HAP	NeurIPS 2023	1	~2.1M

HAP is simple: two modalities, one dataset, fewer training samples.

person ReID

method	MSMT17	Market-1501
PASS	71.8	93.0
MALE	73.0	92.2
PATH	69.1	89.5
UniHCP	67.3	90.3
MAE	62.0	82.9
HAP	**78.0**	**93.8**

text-to-image person ReID

method	CUHK-PEDES	ICFG-PEDES
LBUL	61.95	-
CAIBC	64.43	-
SSAN	61.37	54.23
SRCF	64.04	57.18
MAE	60.19	53.68
HAP	**68.05**	**61.80**

2D human pose estimation

method	MPII	COCO	AIC
HRNet-w48	90.1	75.1	33.5
ViTPose	93.3	77.1	32.0
HRFormer	-	77.2	-
LiftedCL	89.3	71.1	-
PATH	93.3	76.3	35.0
UniHCP	-	76.5	33.6
SOLIDER	-	76.6	-
MAE	89.6	75.7	31.3
HAP	**93.6**	**78.2**	**38.1**

3D human pose and shape estimation

method	MPJPE	PA-MPJPE	MPVPE
Pose2Mesh	89.5	56.3	105.3
3DCrowdNet	81.7	51.5	98.3
MAE	95.6	58.0	112.7
HAP	**90.1**	**56.0**	**106.3**

pedestrian attribute recognition

method	PA-100K	RAP	PETA
PATH	85.0	81.2	88.0
UniHCP	86.18	82.34	-
SOLIDER	86.37	-	-
MAE	79.56	75.73	80.82
HAP	**86.54**	**82.91**	**88.36**

HAP achieves SOTA on 11 human-centric benchmarks.

code

project

References

1. He, Kaiming, et al. "Masked autoencoders are scalable vision learners." Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2022.
2. Ci, Yuanzheng, et al. "UniHCP: A Unified Model for Human-Centric Perceptions." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2023.
3. Chen, Weihua, et al. "Beyond Appearance: a Semantic Controllable Self-Supervised Learning Framework for Human-Centric Visual Tasks." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2023.