Beyond Appearance:
a Semantic Controllable Self-supervised Learning Framework for Human-Centric Visual Tasks

Weihua Chen, Xianzhe Xu, Jian Jia, Hao Luo, Yaohua Wang, Fan Wang, Rong Jin, Xiuyu Sun

Paper ID: 8746

Session Tag: WED-PM-257

JUNE 18-22, 2023
CVPR
VANCOUVER, CANADA

Alibaba Group 阿里巴巴集团

SOLIDER

Beyond Appearance: a Semantic Controllable Self-supervised Learning Framework for Human-Centric Visual Tasks

Weihua Chen, Xianzhe Xu, Jian Jia, Hao Luo, Yaohua Wang, Fan Wang, Rong Jin, Xiuyu Sun

Paper ID: 8746

Session Tag: WED-PM-257

JUNE 18-22, 2023 CVPR VANCOUVER, CANADA

1. MOTIVATION

Background

1. Human-centric tasks are **curial in application**;
2. **Massive unlabeled** human images available;
3. **Semantic information** is important.
4. Different downstream tasks require **different needs** of semantic.

1. MOTIVATION

Background

1. Human-centric tasks are **curial in application**;
2. **Massive unlabeled** human images available;
3. **Semantic information** is important.
4. Different downstream tasks require **different needs** of semantic.

Goal

1. Learn **a general human representation** from **massive unlabeled** data.
2. The pre-trained model can benefit to downstream tasks and meet their **different semantic requirements**. (**One pretrained model adapt to All**.)

1. MOTIVATION

Background

1. Human-centric tasks are **curial in application**;
2. **Massive unlabeled** human images available;
3. **Semantic information** is important.
4. Different downstream tasks require **different needs** of semantic.

Goal

1. Learn **a general human representation** from **massive unlabeled** data.
2. The pre-trained model can benefit to downstream tasks and meet their **different semantic requirements**. (**One pretrained model adapt to All**.)

Main Contributions

1. Build a **semantic supervision from human prior** to train SOLIDER.
2. Design a **semantic controller** to adjust the semantic rate in the pre-trained model.

2. METHOD

Contribution 1. Semantic Supervision from Human Prior

The framework of SOLIDER

2. METHOD

Contribution 1. Semantic Supervision from Human Prior

The framework of SOLIDER

Contribution 2. Semantic Controller

3. EXPERIMENTS

3.1. Can SOLIDER learn semantic ?

DINO space:
Different parts from same person are closer.

SOLIDER space:
Same semantic parts from different people are closer.

3. EXPERIMENTS

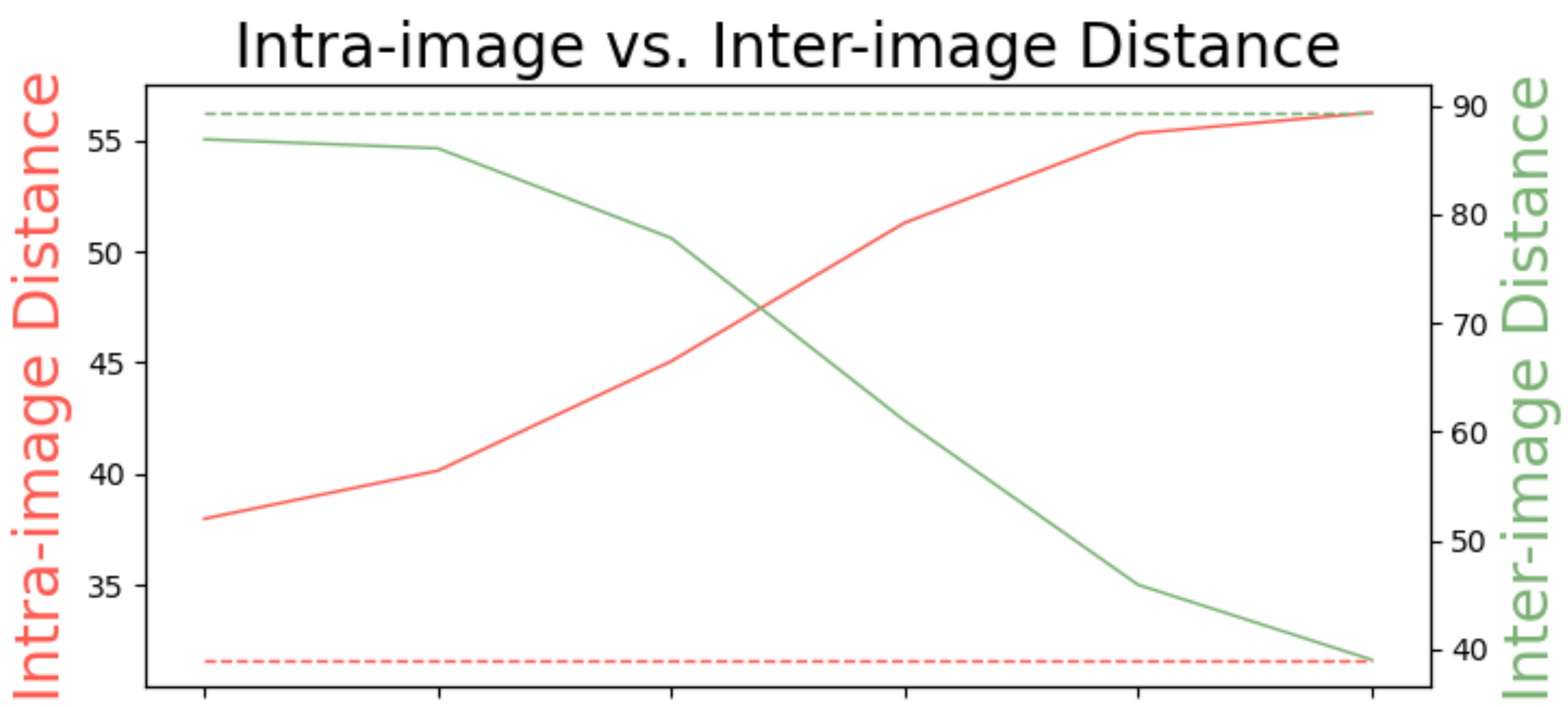
3.2. Can λ control semantic rate ?

[Exp 1] *With λ increased:*

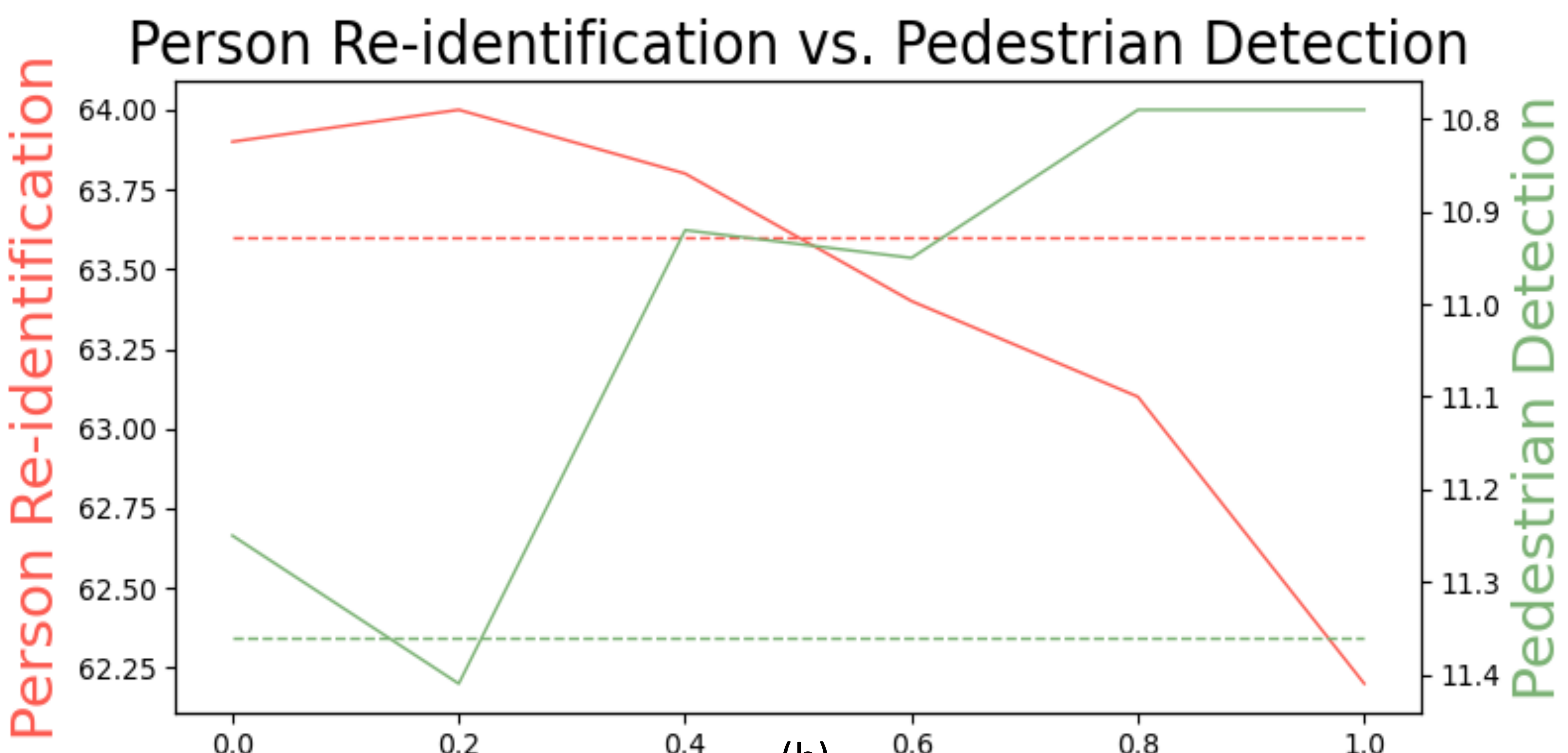
Observation 1. Different parts from same person (intra-distance) gets further.
Observation 2. Same semantic parts from different people (inter-distance) gets closer.

[Exp 2] *With λ increased:*

Observation 1. The person re-identification performance is lower.
Observation 2. The pedestrian detection performance is higher.



(a)



(b)

3. EXPERIMENTS

3.3. How is Ablation Study and SOTA Comparison?

Ablation Study

Pretrain Methods		Sup	DINO [6]	+ Clustering	+ Clustering&Controller
Pretrain Data		ImageNet	LUP1M	LUP1M	LUP1M
Person Re-identification mAP/Rank1 ↑ TransReID [36]	Market1501	78.1/90.2	89.6/95.9	89.5/95.5	**89.9/96.1**
	MSMT17	49.7/73.6	63.3/83.2	61.6/82.2	**63.9/83.8**
Attribute Recognition mA ↑ RethinkPAR [38]	$PETA_{zs}$	72.86	73.64	73.90	**74.20**
	RAP_{zs}	72.10	73.04	73.16	**73.21**
	PA100k	80.67	82.98	82.98	**84.15**
Person Search mAP/Rank1 ↑ SeqNet [45]	CUHK-SYSU	93.0/94.1	93.6/94.3	93.6/94.1	**94.0/94.7**
	PRW	50.0/84.4	52.9/84.7	53.0/84.0	**54.1/85.0**
Pedestrian Detection MR^{-2}(R/HO) ↓ CSP [73]	CityPerson	11.6/43.8	11.4/43.1	11.1/41.7	**10.8/40.7**
Human Parsing mIOU ↑ SCHP [43]	LIP	51.10	54.45	55.25	**55.45**
Pose Estimation AP/AR ↑ HRFormer [83]	COCO	72.4/78.2	73.1/78.5	73.4/78.7	**74.4/79.7**

SOTA Comparison

							SOLIDER		
Person Re-identification mAP/Rank1 ↑		SCSN [16]	ABDNet [9]	TransReID [36]	UP-ReID [81]	PASS [97]	Swin-T	Swin-S	Swin-B
	Market1501	88.5/95.7	88.3/95.6	89.5/95.2	91.1/97.1	93.3/**96.9**	91.6/96.1	93.3/96.6	**93.9/96.9**
	MSMT17	58.5/83.8	60.8/82.3	69.4/86.2	63.3/84.3	74.3/89.7	67.4/85.9	76.9/90.8	**77.1/90.7**
Attribute Recognition mA ↑		MsVAA [61]	VAC [30]	ALM [66]	JLAC [65]	RethinkPAR [38]	SOLIDER Swin-T	SOLIDER Swin-S	SOLIDER Swin-B
	$PETA_{zs}$	71.53	71.91	73.01	73.60	71.62	74.37	76.21	**76.43**
	RAP_{zs}	72.04	73.70	74.28	76.38	72.32	74.23	76.84	**77.06**
	PA100k	80.41	79.16	80.68	82.31	81.61	84.14	86.25	**86.37**
Person Search mAP/Rank1 ↑		NAE+ [7]	AlignPS+ [79]	TCTS [68]	SeqNet [45]	GLCNet [93]	SOLIDER Swin-T	SOLIDER Swin-S	SOLIDER Swin-B
	CUHK-SYSU	92.1/92.9	94.0/94.5	93.9/95.1	94.8/95.7	**95.8/96.2**	94.9/95.7	95.5/95.8	94.9/95.5
	PRW	44.0/81.1	46.1/82.1	46.8/87.5	47.6/87.6	47.8/**87.8**	56.8/86.8	**59.8**/86.7	59.7/86.8
Pedestrian Detection MR^{-2}(R/HO) ↓		RepLoss [71]	CSP [73]	NMS-Loss [56]	ACSP [70]	PedesFormer [33]	SOLIDER Swin-T	SOLIDER Swin-S	SOLIDER Swin-B
	CityPerson	13.2/56.9	11.0/49.3	10.8/-	9.3/46.3	**9.2/36.9**	10.3/40.8	10.0/39.2	9.7/39.4
Human Parsing mIOU ↑		JPPNet [46]	BraidNet [50]	CE2P [60]	PCNet [87]	SCHP [43]	SOLIDER Swin-T	SOLIDER Swin-S	SOLIDER Swin-B
	LIP	51.37	54.40	53.10	57.03	59.36	57.52	60.21	**60.50**
Pose Estimation AP/AR ↑		CPN [17]	SimpleBase [75]	TokenPose [44]	HRNet [63]	HRFormer [83]	SOLIDER Swin-T	SOLIDER Swin-S	SOLIDER Swin-B
	COCO	68.6/-	74.3/79.7	75.8/80.9	76.3/81.2	**77.2/82.0**	74.4/79.6	76.3/81.3	76.6/81.5

4. OPEN CODE

Project Website: *https://github.com/tinyvision/SOLIDER*

Alibaba Group 阿里巴巴集团

SOLIDER

a Semantic cOntrollable seLf-supervIseD lEaRning framework

from Alibaba Group

State of the Art | Pedestrian Attribute Recognition on PA-100K

State of the Art | Person Re-Identification on MSMT17 (using additional training data)

Ranked #3 | Person Re-Identification on Market-1501 (using additional training data)

Ranked #4 | Person Search on CUHK-SYSU

State of the Art | Person Search on PRW

Ranked #6 | Pedestrian Detection on CityPersons (using additional training data)

State of the Art | Semantic Segmentation on LIP val

Ranked #2 | Pose Estimation on COCO

Updates

- **[2023/04/24: Codes of attribute recognition task is released!]** NEW
  - Training details of our pretrained model on downstream person attribute recognition task is released.
- **[2023/03/28: Codes of 3 downstream tasks are released!]**
  - Training details of our pretrained model on 3 downstream human visual tasks, including person re-identification, person search and pedestrian detection, are released.
- **[2023/03/13: SOLIDER is accepted by CVPR2023!]**
  - The paper of SOLIDER is accepted by CVPR2023, and its offical pytorch implementation is released in this repo.