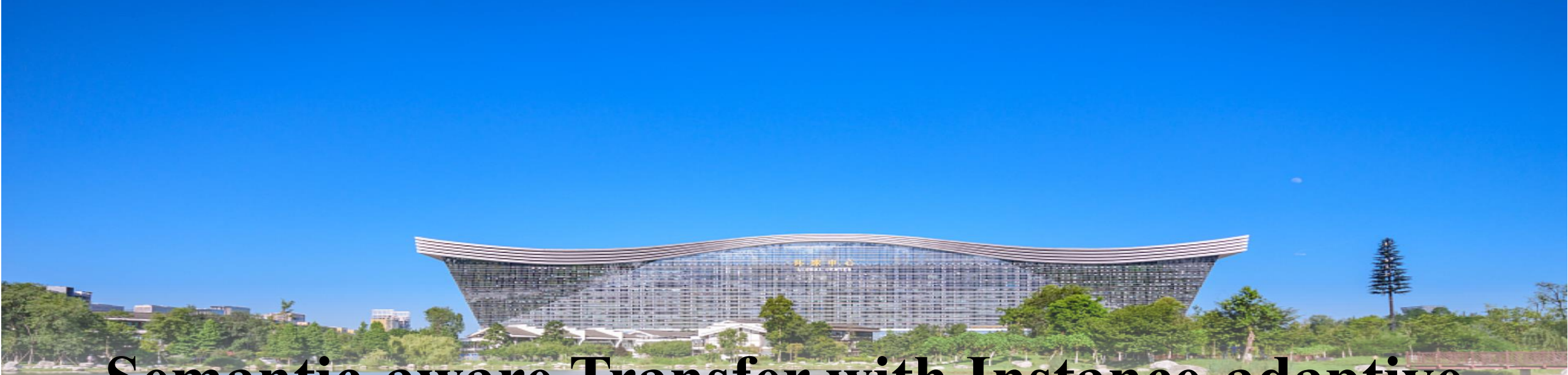


Semantic-aware Transfer with Instance-adaptive Parsing for Crowded Scenes Pose Estimation

Xuanhan Wang, Lianli Gao, Yan Dai,

Yixuan Zhou and Jingkuan Song

ACM multimedia

Chengdu, China OCT 20-24 2021

Wang, X., Gao, L., Dai,Y., Zhou, Y., & Song, J. (2021). Semantic-aware Transfer with Instance-adaptive Parsing for Crowded Scenes Pose Estimation. In *29th ACM International Conference on Multimedia.*

Outline

- Introduction
- Motivation
- Method
- Experiments and Results
- Summary

Introduction

Multi-Person Pose Estimation

Simultaneously detecting people and localizing their anatomical keypoints under complex scenes.

Bottom-up pipeline

- Instance-agnostic keypoints detection.
- Keypoints grouping.

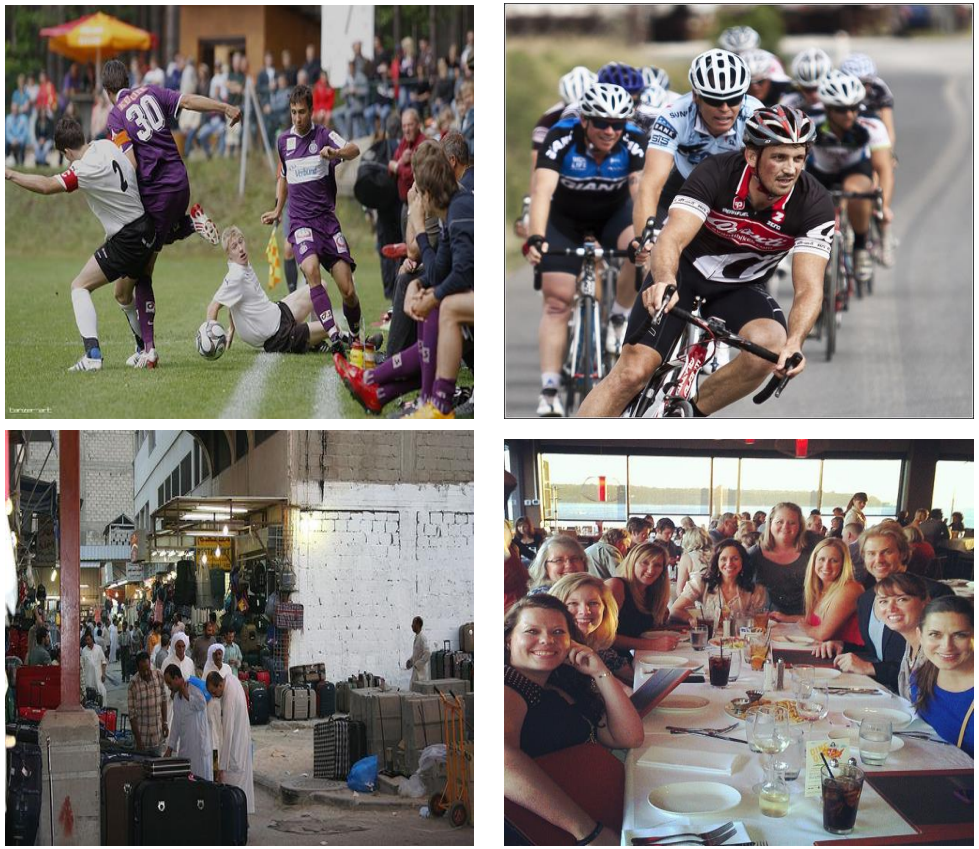
Top-down pipeline

- Human detection.
- Single-person pose estimation

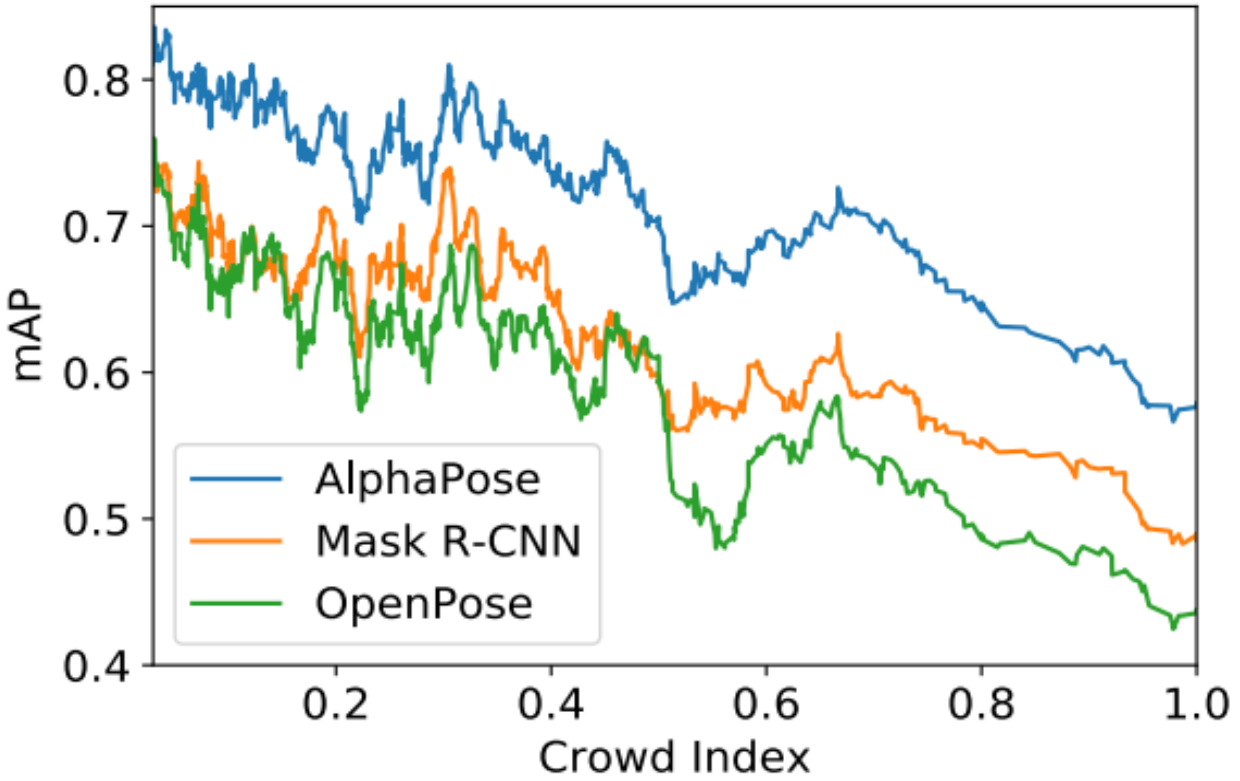
[1] https://cocodataset.org/#keypoints-2020

Motivation

Crowded Scenes Pose Estimation



Crowded scenes



Performance decline

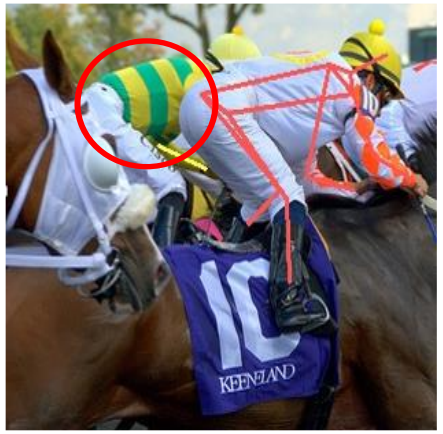
[1] Jiefeng Li, Can Wang, Hao Zhu, Yihuan Mao, Hao-Shu Fang and Cewu Lu. CrowdPose: Efficient Crowded Scenes Pose Estimation and A New Benchmark. In CVPR 2019.

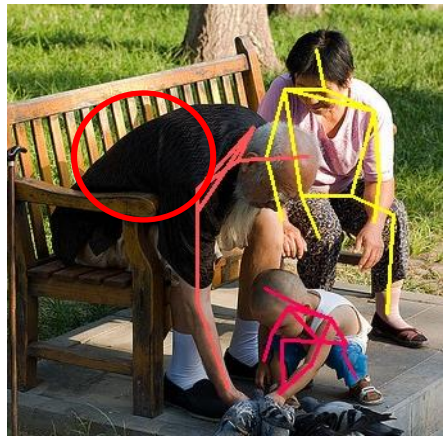
Motivation

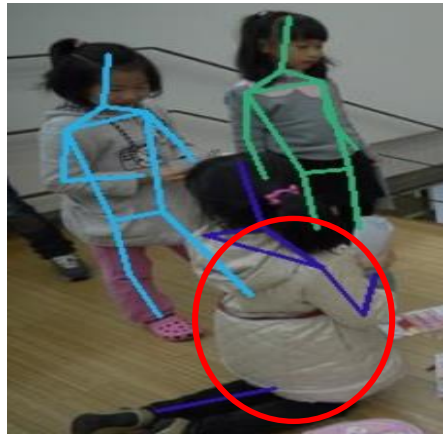
Crowded Scenes Pose Estimation → Missing keypoints




Crowded scenes







Missing keypoints

[1] Jiefeng Li, Can Wang, Hao Zhu, Yihuan Mao, Hao-Shu Fang and Cewu Lu. CrowdPose: Efficient Crowded Scenes Pose Estimation and A New Benchmark. In CVPR 2019.

Motivation

Crowded Scenes Pose Estimation → Ambiguously labeling

multiple instances in one bounding box

Instance A

Instance B

Pos label for Instance A

Neg label for Instance A

Pos label for Instance B

Pos label for Instance B

Neg label for Instance B

Pos label for Instance A

Motivation

Crowded Scenes Pose Estimation → Ambiguously labeling

Instance A

Instance B

Pos label for Instance A

Neg label for Instance A

Pos label for Instance B

Pos label for Instance B

Neg label for Instance B

Pos label for Instance A

labels for Instance A

labels for Instance B

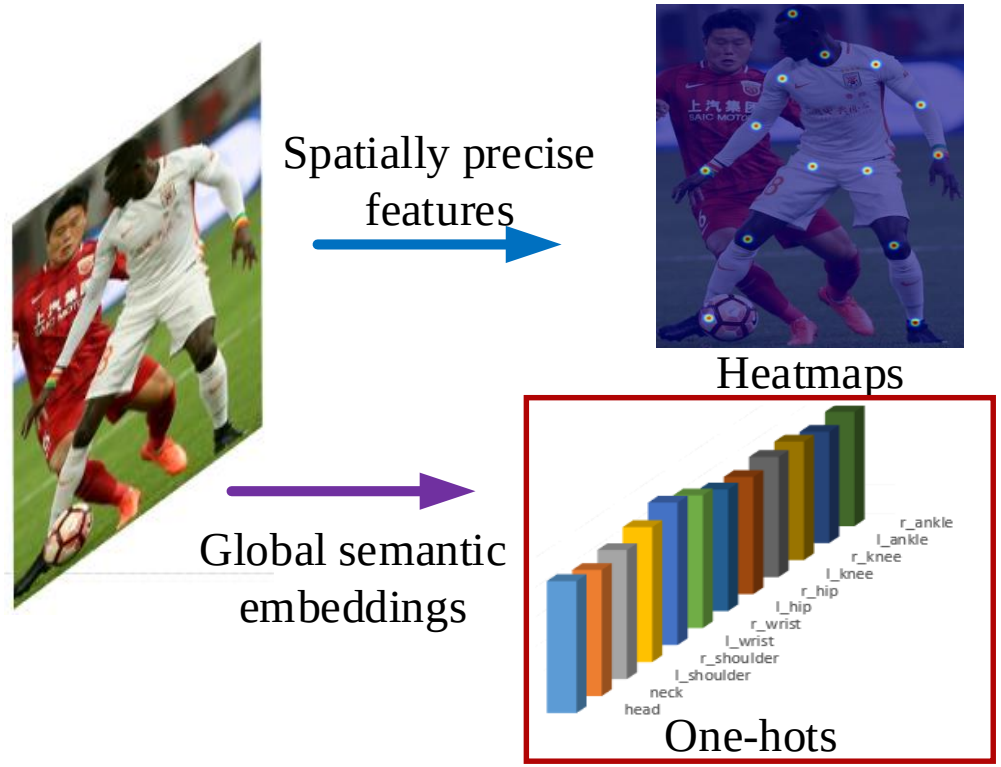
Pose Estimation Network

Conflicting supervision

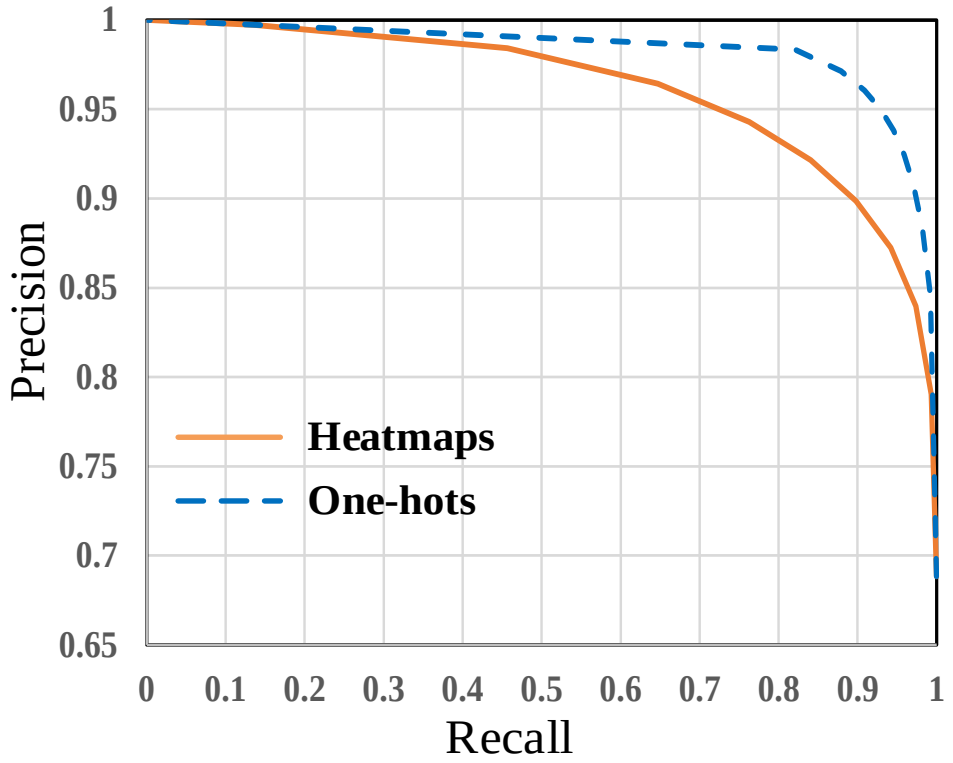
Motivation

Issues that we focus on

- *How to design an effective pipeline for pixel-level representation enhancement ?*
- *How to equip this pipeline with the ability to handle the ambiguously labeling ?*



Two Patterns of Keypoint Embeddings

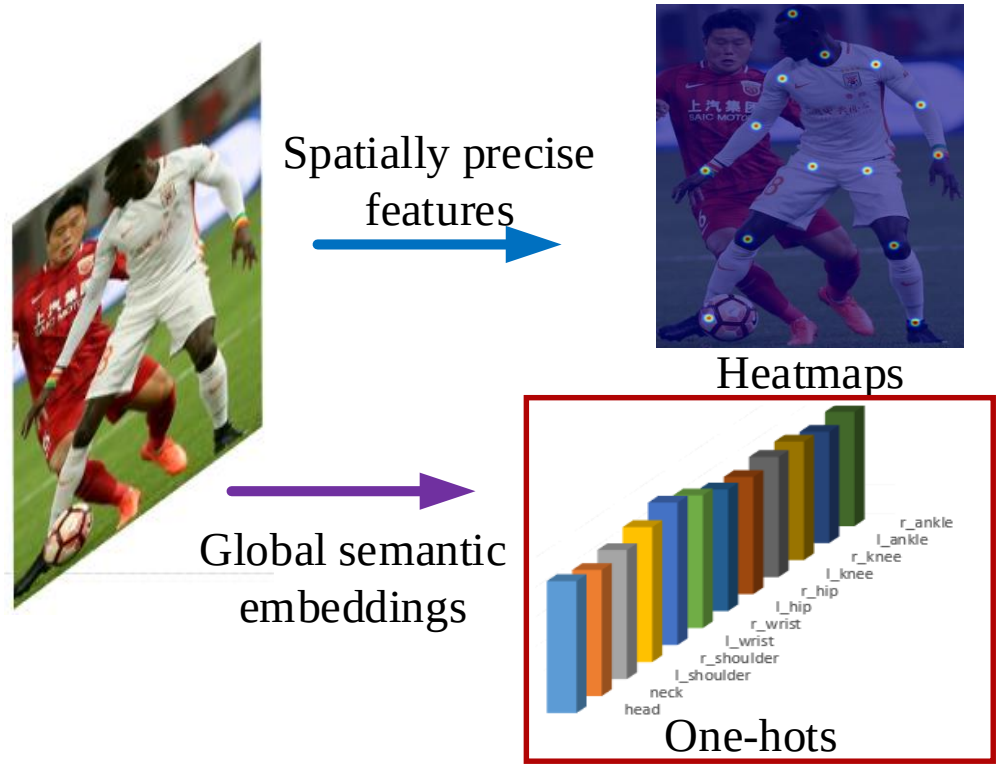


P-R Curve on CrowdPose Dataset

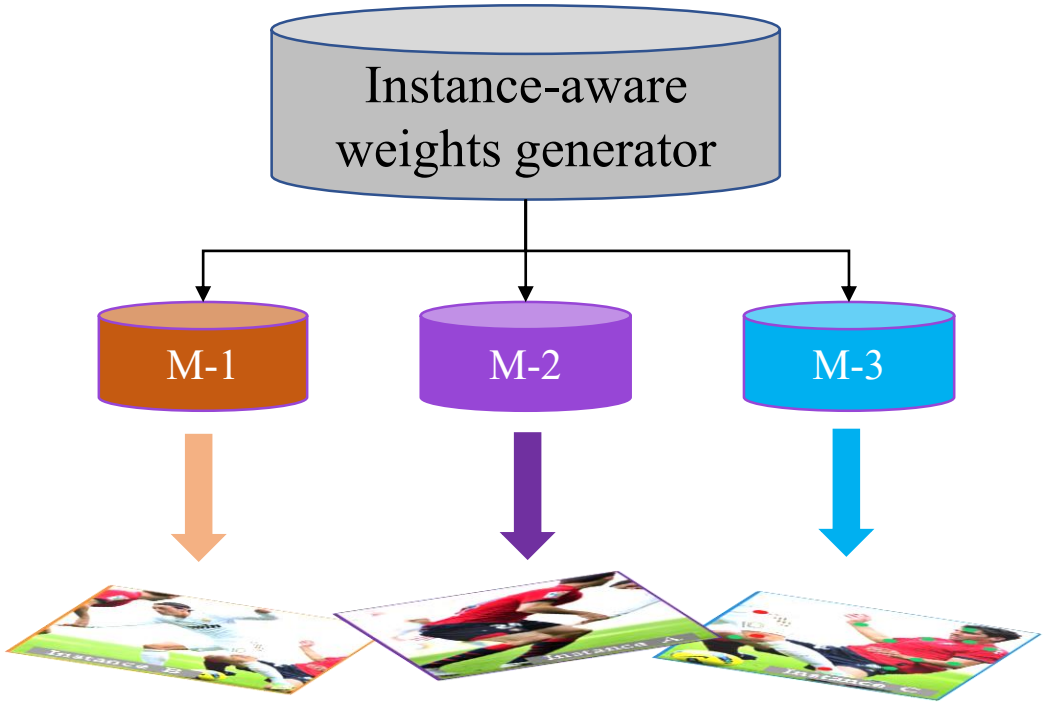
Motivation

Issues that we focus on

- *How to design an effective pipeline for pixel-level representation enhancement ?*
- *How to equip this pipeline with the ability to handle the ambiguously labeling ?*



Semantic-aware enhancement

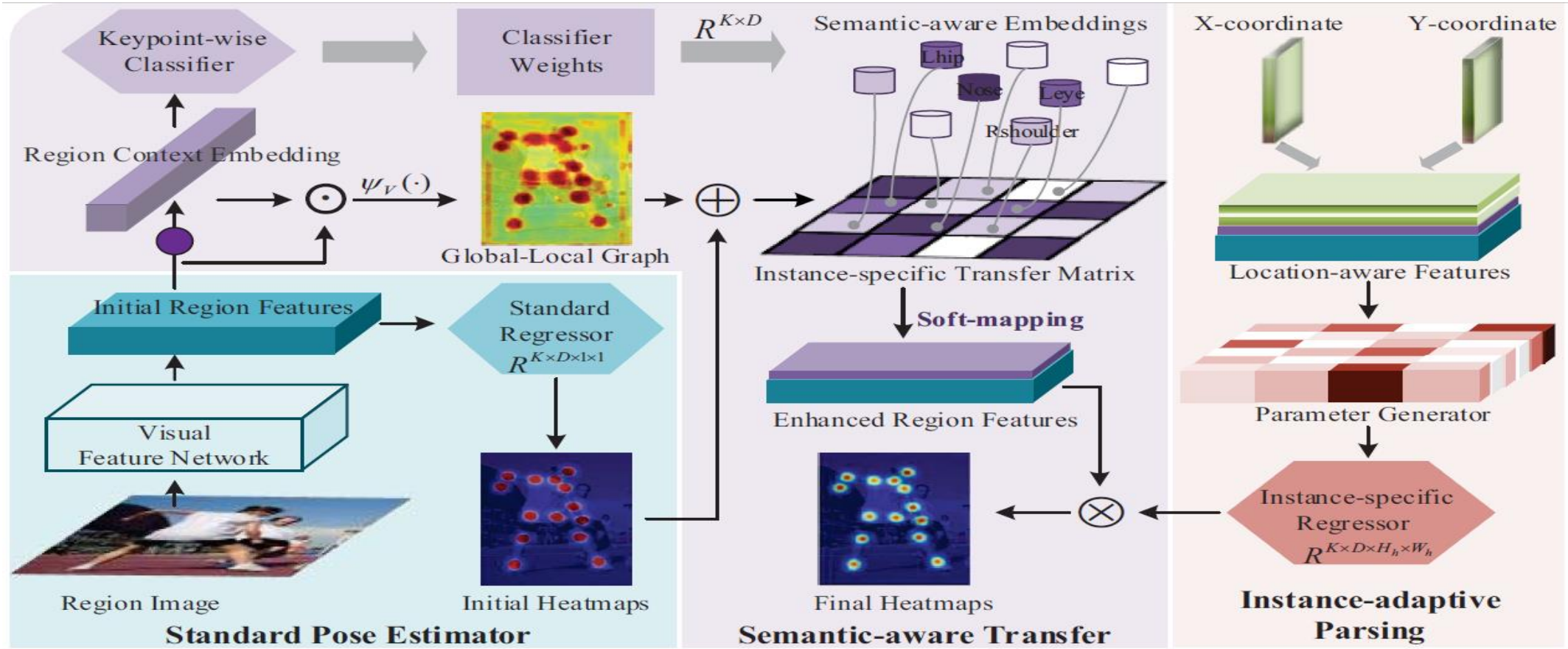


Instance-aware parsing

Method

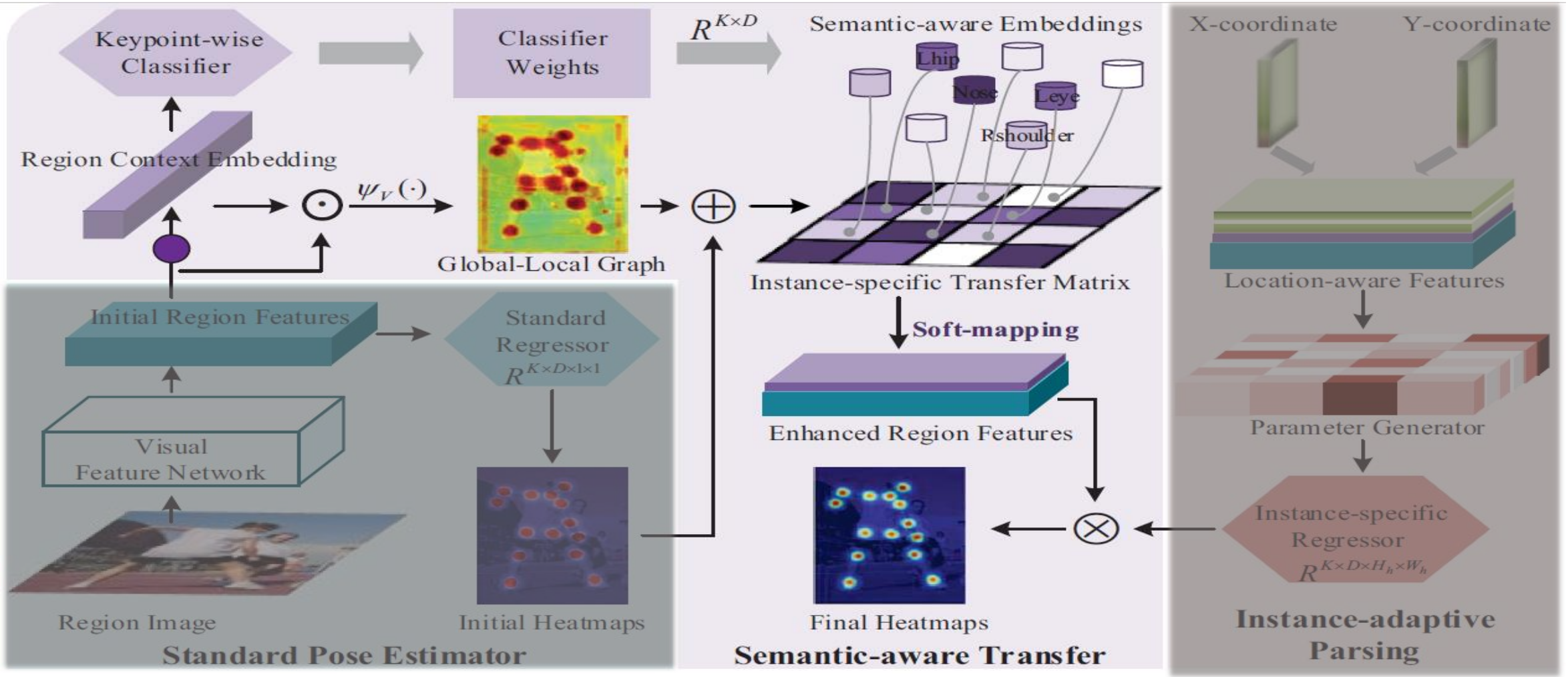
Semantic-aware Transfer with Instance-adaptive Parsing (STIP)

- Semantic-aware Transfer (SaT)
- Instance-adaptive Parsing (IaP)



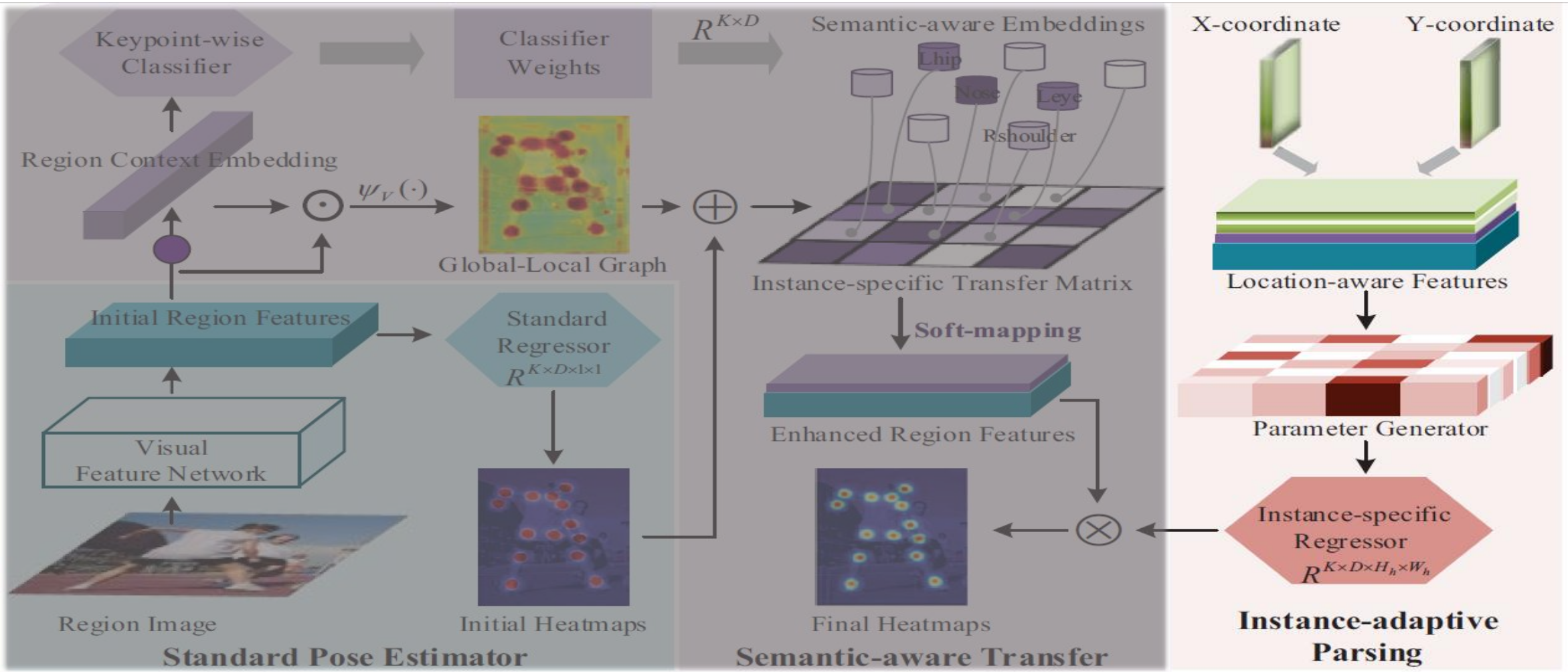
Method

Semantic-aware Transfer (SaT)



Method

Instance-adaptive Parsing (IaP)



Experiments and Results

Dataset: CrowdPose & MS-COCO

The **CrowdPose** dataset contains 20K images and 80K human annotations in total. It is split into two subsets: 12K images for training, 8K images for testing.
The **MS-COCO** dataset contains about 250K humans annotations. Moreover, it is split into two subsets: training set and validation set with 57K images and 5k images.

Evaluation Metric: Object Keypoints Similarity (OKS)

$$OKS_j = \frac{1}{|P_j|}\sum_{p\in P_j} exp\left(\frac{-g(i_p,\hat{i}_p)^2}{2k^2}\right)$$

P_j: a set of ground truth keypoints annotated on person instance j
i_p: the keypoint estimated by a model at p-th class
$\hat{i}_p$: the ground truth keypoint at p-th class
k : the variance of gaussian function
g : Euclidean metric

mAP: the mean of AP scores at a number of Object Keypoints Similarity (OKS) ranging from 0.5 to 0.95.

Experiments and Results

Ablation Study: Module effectiveness

Experiments on CrowdPose

Baseline	SaT	IaP	AP	AP_E	AP_M	AP_H
√			71.7%	79.6%	72.7%	61.5%
√	√		73.5%	81.1%	74.6%	63.6%
√		√	73.6%	81.1%	74.8%	63.6%
√	√	√	**74.1%**	**81.6%**	**75.1%**	**64.3%**

Experiments on MS-COCO

Baseline	SaT	IaP	AP	AP_E	AP_M	AP_H
√			74.4%	70.8%	81.0%	79.8%
√	√		75.6%	71.8%	82.4%	80.8%
√		√	75.6%	71.8%	82.5%	80.7%
√	√	√	**75.8%**	**72.1%**	**82.6%**	**81.0%**

CrowdPose:

STIP improves the baseline model by 3.4% AP score.

- Baseline + **SaT**: 71.7% -> 73.5% (+1.8%)
- Baseline + **IaP**: 71.7% -> 73.6% (+1.9%)

MS-COCO:

STIP improves the baseline model by 1.4% AP score.

- Baseline + **SaT**: 74.4% -> 75.6% (+1.2%)
- Baseline + **IaP**: 74.4% -> 75.6% (+1.2%)

Experiments and Results

Ablation Study: Missing Keypoints Reduction

Method	Threshold Values								
	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
HRNet-W32	99.3	97.4	94.2	89.8	84.0	76.3	64.7	45.7	13.5
+STIP	99.2	97.4	94.5	90.7	85.9	79.9	70.9	54.7	22.2
HRNet-W48	99.3	97.6	95.0	91.6	87.0	81.1	72.1	55.4	19.7
+STIP	99.3	97.7	95.4	92.2	88.0	82.9	75.3	60.2	25.4

The proposed method achieves higher recall score than baseline model at high threshold values.

Experiments and Results

The Generalizability of proposed method

Experiments on CrowdPose

Method	Backbone	Input size	mAP
HRNet	HRNet-W32	256x192	71.7%
+STIP	HRNet-W32	256x192	74.1% (+2.4)
HRNet	HRNet-W48	256x192	73.3%
+STIP	HRNet-W48	256x192	74.8% (+1.5)
HRNet	HRNet-W32	384x288	73.0%
+STIP	HRNet-W32	384x288	74.7% (+1.7)
HRNet	HRNet-W48	384x288	73.9%
+STIP	HRNet-W48	384x288	75.2% (+1.3)
SimpleBaseline	ResNet50	256x192	68.4%
+STIP	ResNet50	256x192	68.9% (+0.5)

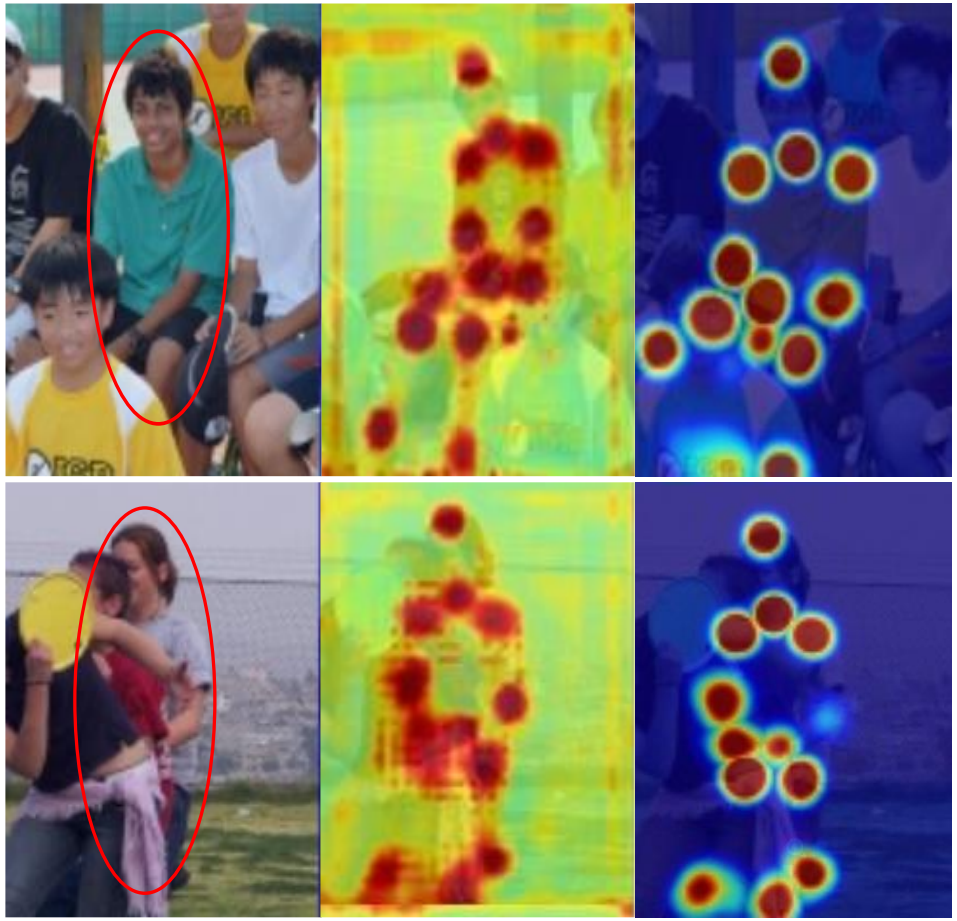
Experiments on MS-COCO

Method	Backbone	Input size	mAP
HRNet	HRNet-W32	256x192	74.4%
+STIP	HRNet-W32	256x192	75.8% (+1.4)
HRNet	HRNet-W48	256x192	75.1%
+STIP	HRNet-W48	256x192	76.1% (+1.0)
HRNet	HRNet-W32	384x288	75.8%
+STIP	HRNet-W32	384x288	76.5% (+0.7)
HRNet	HRNet-W48	384x288	76.3%
+STIP	HRNet-W48	384x288	76.8% (+0.5)
SimpleBaseline	ResNet50	256x192	70.4%
+STIP	ResNet50	256x192	71.4% (+1.0)
SimpleBaseline	ResNet101	256x192	71.4%
+STIP	ResNet101	256x192	72.1% (+0.7)

- HRNet series can be improved with the help of STIP, about ~1.3% AP gains.
- ResNet series can be improved with the help of STIP, about ~0.7% AP gains.

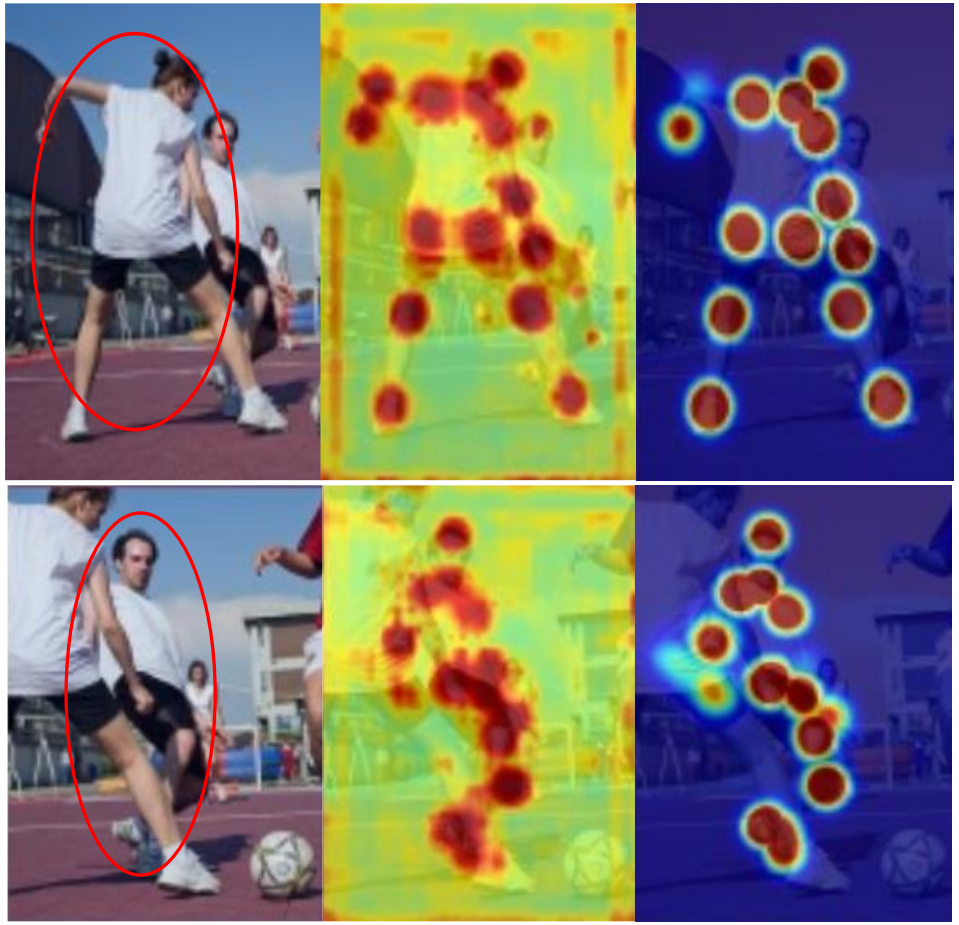
Experiments and Results

Visualization Analysis



global-local graph

parameter map



global-local graph

parameter map

Summary

Contributions

1. We propose an effective keypoints estimation method named **semantic-aware transfer with instance-adaptive parsing (STIP),** which tackles the problem of missing keypoints in crowded scenes and handles the ambiguously labeling during training.

2. **Semantic-aware Transfer (SaT)** that enhances the discriminative power of pixel-level features by transferring keypoint-wise semantic embeddings to pixels.

3. **Instance-adaptive parsing (IaP)** method is proposed to handle the ambiguously labeling by replacing a shared regressor with instance-adaptive regressors.

Thank you!

The code is released on GitHub:

https://github.com/stoa-xh91/STIP

If you have any questions, please e-mail us at:

wxuanhan@hotmail.com

ACM multimedia
Chengdu, China OCT 20-24 2021