ACM MM 21-mfp0403

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

Xuanhan Wang, Lianli Gao, Yan Dai, Yixuan Zhou, Jingkuan Song
Center for Future Media, University of Electronic Science and Technology of China
wxuanhan@hotmail.com

Introduction

Crowded Scenes Pose Estimation

Crowded scenes pose estimation refers to recognize and localize anatomical keypoints for each person instance from a highly complex scenario, which is a fundamental yet challenging task in multimedia applications. In crowded scenes pose estimation, countable instances with their keypoints are expected to be represented and resolved in a unified pipeline. The top-down mechanism has become the mainstream solution for general pose estimation and obtained impressive progress. However, simply applying this mechanism to crowded scenes pose estimation results in unsatisfactory performance due to several issues, in particular involving missing keypoints in crowds and ambiguously labeling during training. To tackle above two issues, we introduce a novel method named Semantic-aware Transfer with Instance-adaptive Parsing (STIP).

Motivation

Challenges in Crowded Scenes

✓ Multiple instances in one bounding box

Figure 1: Challenges existing in crowded scenes pose estimation:
1. **Missing keypoints:**
Partial salient keypoints are ignored.
2. **Ambiguously Labeling:**
Each keypoint appearing in an intersection of two proposals is assigned with two different labels.

Framework

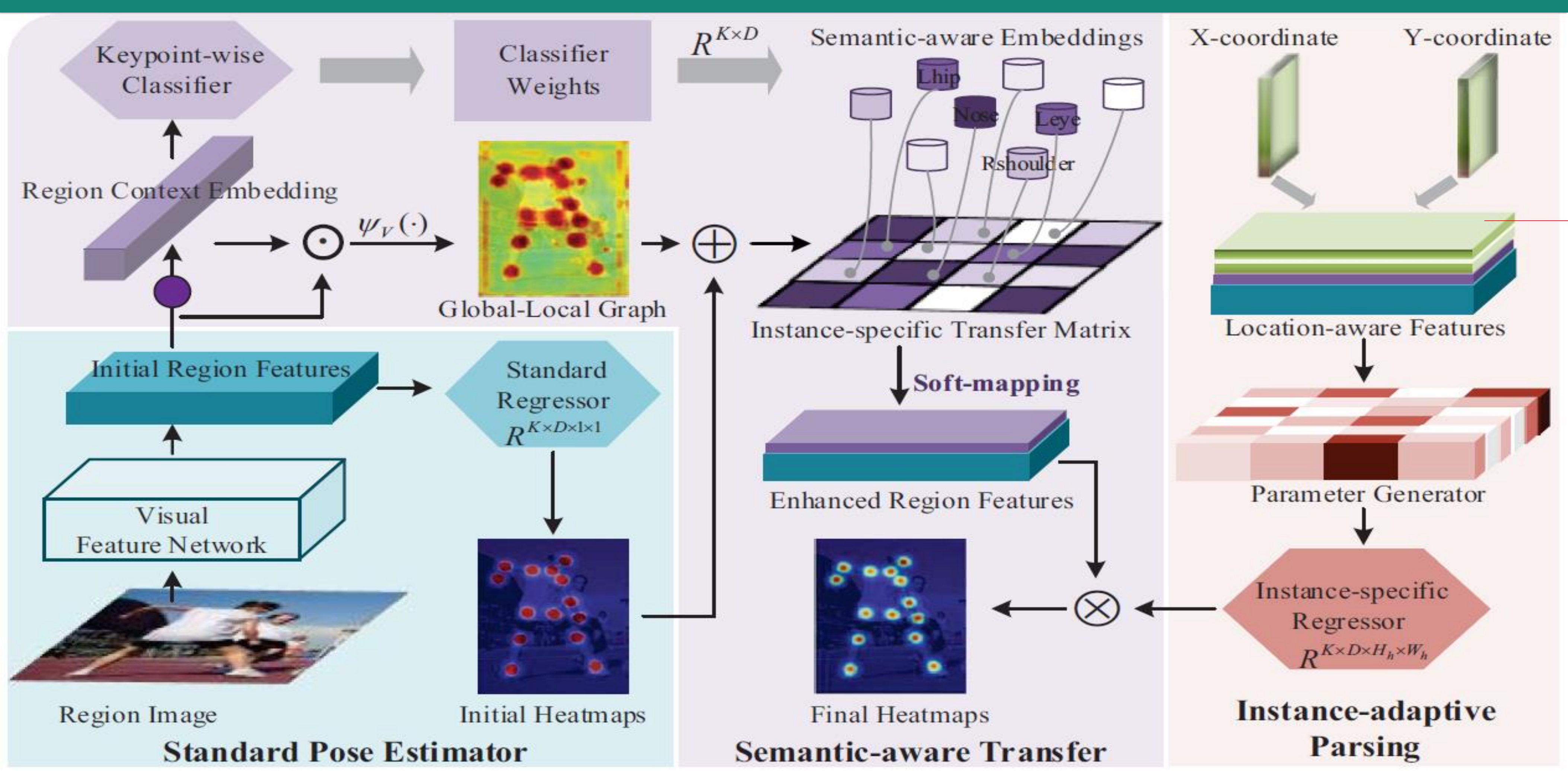

Figure 2: The overview of the proposed framework. Given visual features and initial heatmaps provided by a standard pose estimator, a semantic-aware transfer (SaT) module is utilized to transfer semantic-aware embeddings from a keypoint-wise classifier to pixels. With the location-aware features, instance-adaptive parsing (IaP) module is used to generate parameter maps for instance-specific regressors.

Contribution

Our work has following 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). To tackle missing keypoints, a semantic-aware transfer (SaT) is proposed to enhance the discriminative power of pixel-level features by transferring keypoint-wise semantic embeddings to pixels. Furthermore, we introduce an instance-adaptive parsing (IaP) method to handle the ambiguously labeling by replacing a shared regressor with instance-adaptive regressors. Notably, the STIP with above two techniques is flexible to be integrated into any top-down models.

(3). Extensive experiments conducted on challenging benchmarks (i.e., CrowdPose and MS-COCO) demonstrate the effectiveness and generalizability of proposed method.

Experiments

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%**

The proposed method improves baseline model by 2.7% AP score.
Baseline + SaT: 71.7% -> 73.5% (+1.8%)
Baseline + IaP: 71.7% -> 73.6% (+1.9%)

Experiments on MS-COCO

Baseline	SaT	IaP	AP	AP_M	AP_L	AR
√			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%**

The proposed method improves baseline model by 1.2% AP score.
Baseline + SaT: 74.4% -> 75.6% (+1.2%)
Baseline + IaP: 74.4% -> 75.6% (+1.2%)

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)

- ResNet series can be improved with STIP, about 0.7% AP gains.

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)

- HRNet **series** can be improved with STIP, about 1.3% AP gains.

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

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

Right: The proposed method improves baseline models at strict threshold values under all crowding levels.

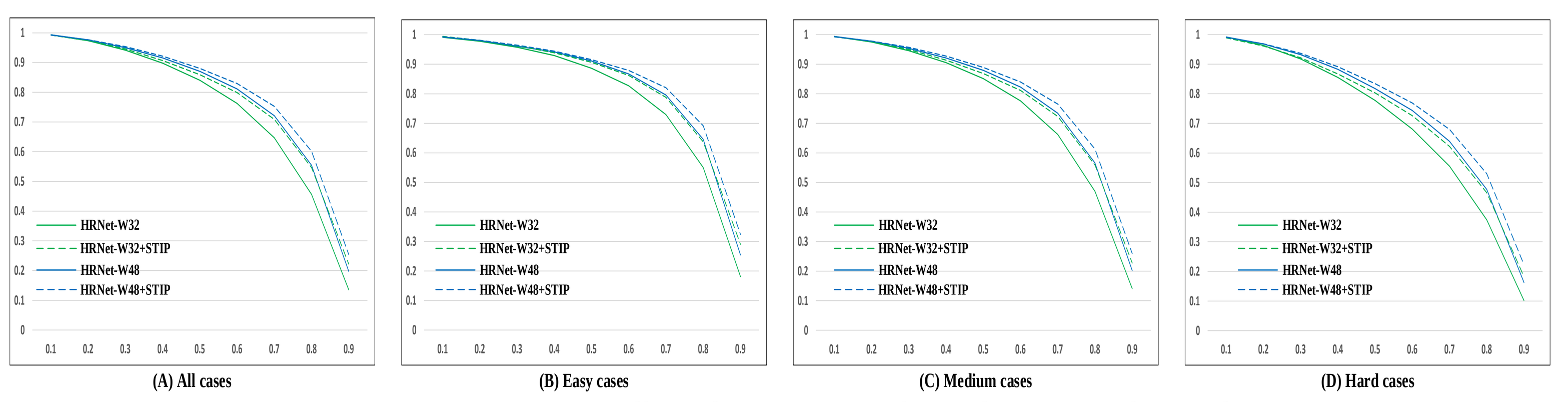

Figure 3: Keypoint-wise recall performance evaluated on CrowdPose dataset.

Figure 4: Qualitative comparison on CrowdPose test set.

Left: HRNet vs HRNet+STIP. The red circles spot the difference between two models. The yellow circles mark the positions where both models fail to estimate keypoints.

Right: Visualization of model predictions. For each example, it shows global-local graph and parameter map.

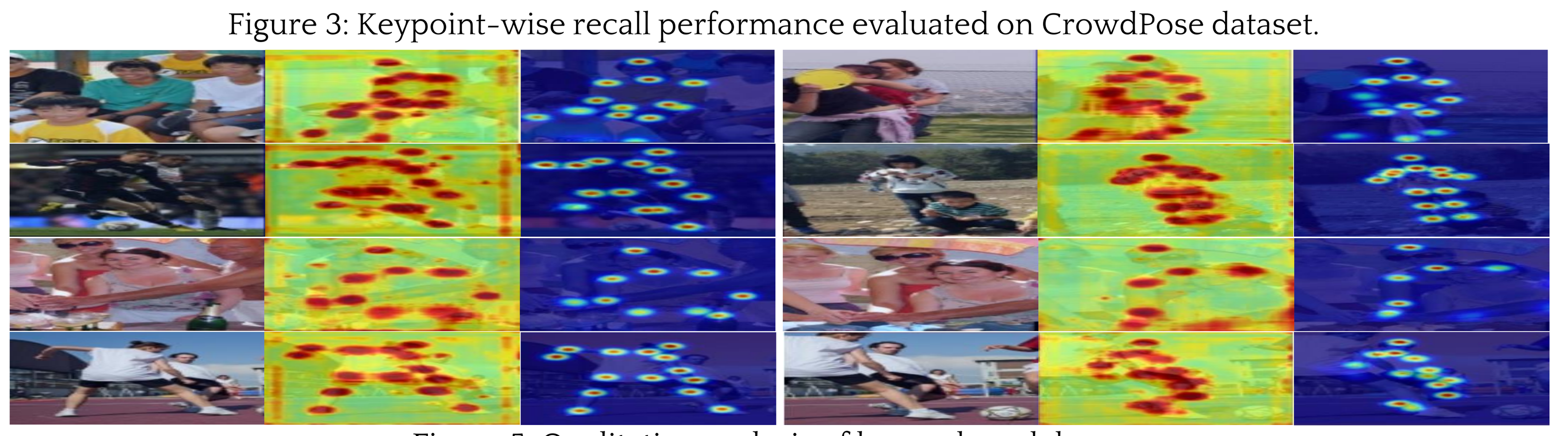

Figure 5: Qualitative analysis of learned model.