BoIR: Box-Supervised Instance Representation for Multi-Person Pose Estimation - Supplementary Material

Uyoung Jeong[1]
jeong.uyoung@unist.ac.kr
Sengryul Baek[1]
srbaek@unist.ac.kr
Hyung Jin Chang[2]
h.j.chang@bham.ac.uk
Kwang In Kim[3]
kimkin@postech.ac.kr

[1] Ulsan National Institute of Science and Technology
Ulsan, Republic of Korea
[2] University of Birmingham
Birmingham, United Kingdom
[3] Pohang University of Science and Technology
Pohang, Republic of Korea

Abstract

In this supplementary material, we provide further details about 1) Bbox Mask Loss 2) Architecture composition 3) Additional experiments 4) Visualization and comparative analysis with CID.

1 Bbox Mask Loss

From the instance embedding map e, we first apply L2 normalization on e. Then, we sample instance embedding p from e and compute respective loss terms as following:

$$\mathcal{L}_{pull}^{in} = \frac{1}{DN}\sum_{i=1}^{N}\sum_{d=1}^{D}\left(p_{(i,d)} - \bar{p}_{(i,d)}\right)^2 \tag{1}$$

$$\mathcal{L}_{push}^{out} = \frac{1}{N}\sum_{i=1}^{N}\exp\left\{-\frac{\beta}{D}\sum_{d=1}^{D}\left(p_{(i,d)} - \bar{p}_{(i,d)}^{c}\right)^2\right\} \tag{2}$$

$$\mathcal{L}_{push}^{inst} = \frac{1}{\frac{N(N-1)}{2}}\sum_{i=1}^{N}\sum_{j>i}^{N}\exp\left\{-\frac{\beta}{D}\sum_{d=1}^{D}\left(p_{(i,d)} - p_{(j,d)}\right)^2\right\} \tag{3}$$

D is the dimension of the instance embedding, N is the number of ground-truth instances in an image, and i, j represent the instance indices. $\beta = \frac{1}{2\sigma^2}$ is a scaling coefficient for the Gaussian kernel proposed in Associative Embedding [4].

Method	Backbone	Input size	AP	AP^{50}	AP^{75}	AP^{M}	AP^{L}	AR
Top-down methods								
SBL [8]	ResNet-152	384×288	74.3	89.6	81.1	70.5	81.6	79.7
HRNet [6]	HRNet-W32	384×288	75.8	90.6	82.5	72.0	82.7	80.9
Bottom-up methods								
HrHRNet [1]	HrHRNet-W32	512	67.1	86.2	73.0	61.5	76.1	-
HrHRNet [1]	HrHRNet-W48	640	69.9	87.2	76.1	65.4	76.4	-
DEKR [2]	HRNet-W32	512	68.0	86.7	74.5	62.1	77.7	73.0
DEKR [2]	HRNet-W48	512	71.0	88.3	77.4	66.7	78.5	76.0
SWAHR [3]	HrHRNet-W32	512	67.1	86.2	73.0	61.5	76.1	-
SWAHR [3]	HrHRNet-W48	640	69.9	87.2	76.1	65.4	76.4	-
Single stage methods								
PETR [5]	ResNet-101	800	70.0	88.5	77.5	63.6	79.4	-
ED-Pose [9]	ResNet-50	800	71.6	89.6	78.1	65.9	**79.8**	-
CID [7]	HRNet-W32	512	69.8	88.5	76.6	64.0	78.9	75.4
BoIR	HRNet-W32	512	**70.6**	**89.2**	**77.4**	**65.1**	**79.0**	**76.3**
BoIR	HRNet-W48	640	**72.5**	**89.9**	**79.1**	**68.2**	79.4	**78.3**

Table 1: Comparison with state-of-the-art methods on COCO `val` set. Best scores are marked as bold for small(e.g. HRNet-W32) and large(e.g. HRNet-W48) models respectively.

2 Architecture Composition

In the case of 512x512 input size, output heatmap size is set to 128x128. In the case of 640x640 input size, output heatmap size is 160x160. HRNet-W32 backbone outputs 480 channels, due to concatenation of all block outputs. Similarly, HRNet-W48 backbone outputs 720 channels.

In the case of auxiliary task heads, 1 residual block and 1 convolution layer are applied. Residual block receives 256 input channels and outputs 128 channels. The final convolution layer outputs task-specific output channels. In case of bottom-up keypoint head, it is the number of keypoints(17 in COCO, 14 in CrowdPose). In case of the bounding box head, it outputs 4 channels(left, top, right, bottom distance). In case of embedding head, it is D. All convolution layers in the auxiliary task head have a 3x3 kernel size.

In case of instance-wise keypoint regression head, 64 hidden channel size is applied for HRNet-W32 backbone. 96 hidden channel size is used for HRNet-W48 backbone.

3 Additional Experiments

We report full comparative evaluation results on COCO `val` set on Table 1. CID paper does not report full results, so we report the scores using the provided trained model weights. Since HRNet-W48 backbone model is not available, CID's HRNet-W48 results are not reported. BoIR outperforms all comparative stae-of-the-arts except for HRNet-W48 backbone on AP^{L}. Our method with HRNet-W32 backbone outperforms CID by 0.8 AP. Our method with HRNet-W48 even outperforms ED-Pose by 0.9 AP.

We report ablation experiments on AE's Gaussian kernel scaling coefficient β and embedding dimension D on COCO `val` set on Table 2. For fast training and simplicity, we

β	AP	AR
0.5	69.8	75.7
1	70.3	76.2
5	70.2	76.0
10	70.5	76.2
15	70.2	76.0

Emb. Dim	AP	AR
1	70.4	76.3
8	70.4	76.3
16	70.4	76.2
32	70.4	76.3
64	70.1	75.8
128	70.5	76.2

Table 2: Left: Ablation experiment of β on COCO `val` set. $D = 128$ by default. Right: Ablation experiment of embedding dimension D on COCO `val` set. $\beta = 10$ by default.

use Bbox Mask Loss and do not use bbox head during experiment. In case of β, 10 was the best among the options. In case of D, changing the embedding dimension shows little performance difference. We conjecture that AE loss for higher dimensions needs more refinement to benefit high dimensional representation. The primary cause is L2 normalization over embedding dimension before loss computation, which significantly drops the loss scale and floating point precision compared to the original AE loss formulation. Simply removing the normalization would cause unstable training with AMP, so further research is required to improve the current framework.

4 Visualization

We provide extensive outputs of our model in Fig. 1 and Fig. 2. We visualize keypoint prediction outputs along with instance center heatmap, t-SNE of backbone output feature, and feature similarity between top-1 confident instance parameter and the entire feature map. t-SNE is applied on the output backbone feature for 250 iterations with 3 output dimensions per pixel, which directly corresponds to normalized RGB values. Instance similarity is measured by computing the L2 distance between the top-1 confident instance's parameter and the feature map, and then applying a Gaussian Kernel.

We additionally report failure cases of BoIR in Fig. 3. In case of the left images, BoIR produces duplicated predictions on the same person, due to wide activation area of the center heatmap. In case of the right images, BoIR places occluded joints on implausible positions, while this does not affect evaluation performance. However, BoIR generally produces disentangled instance features and detects people better than CID.

References

[1] Bowen Cheng, Bin Xiao, Jingdong Wang, Honghui Shi, Thomas S. Huang, and Lei Zhang. Higherhrnet: Scale-aware representation learning for bottom-up human pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

[2] Zigang Geng, Ke Sun, Bin Xiao, Zhaoxiang Zhang, and Jingdong Wang. Bottom-up human pose estimation via disentangled keypoint regression. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14676–14686, June 2021.

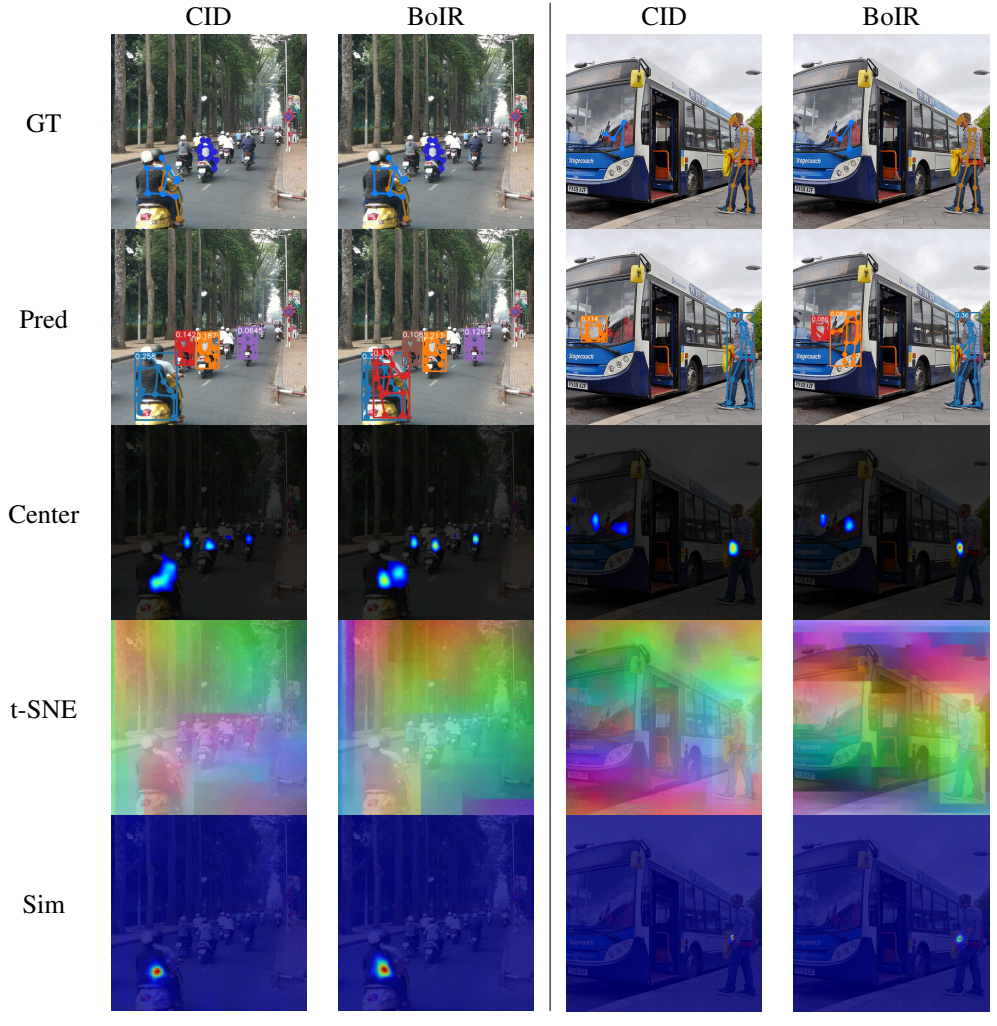

Figure 1: Comparative visualization on COCO `val` set.

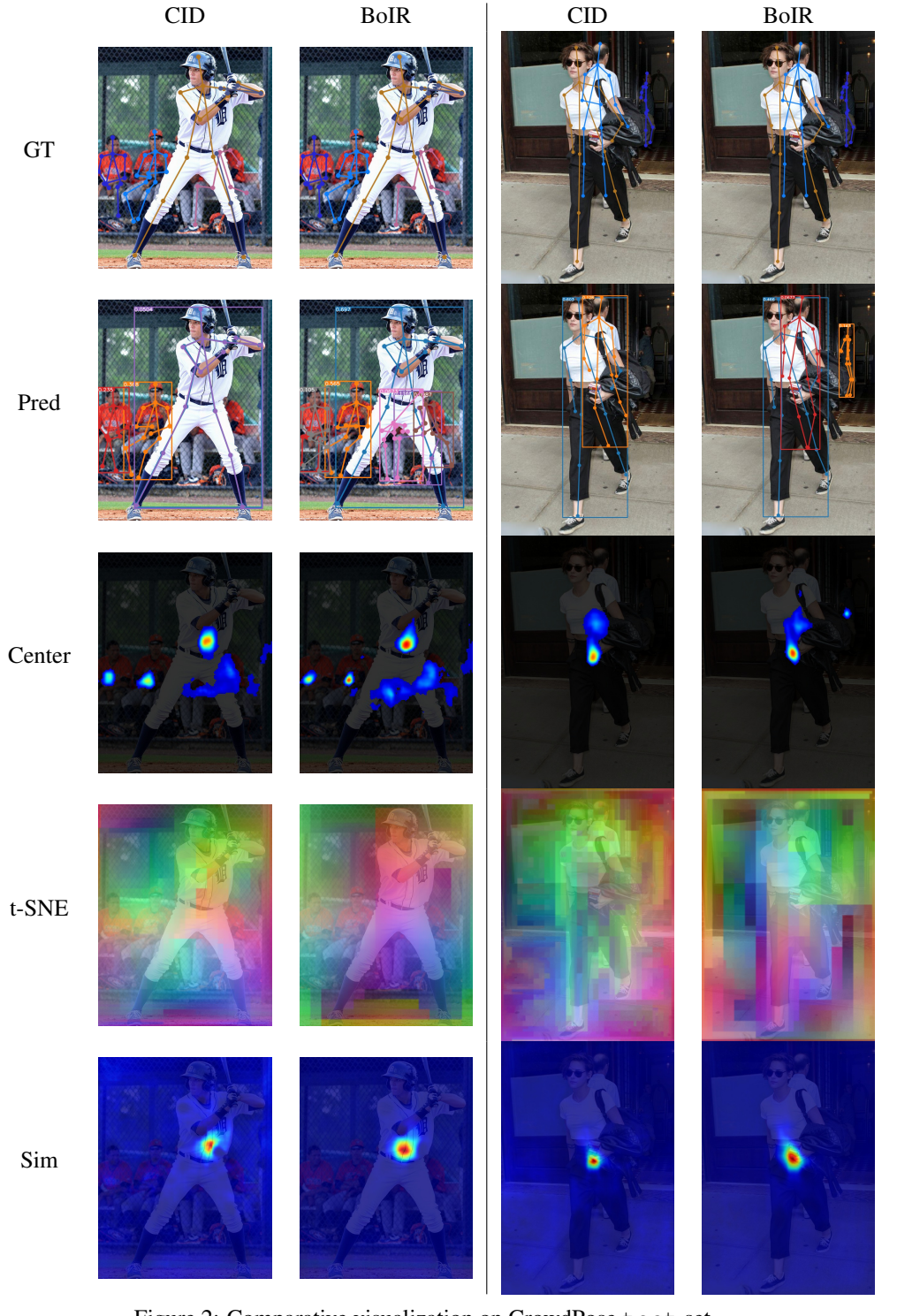

Figure 2: Comparative visualization on CrowdPose `test` set.

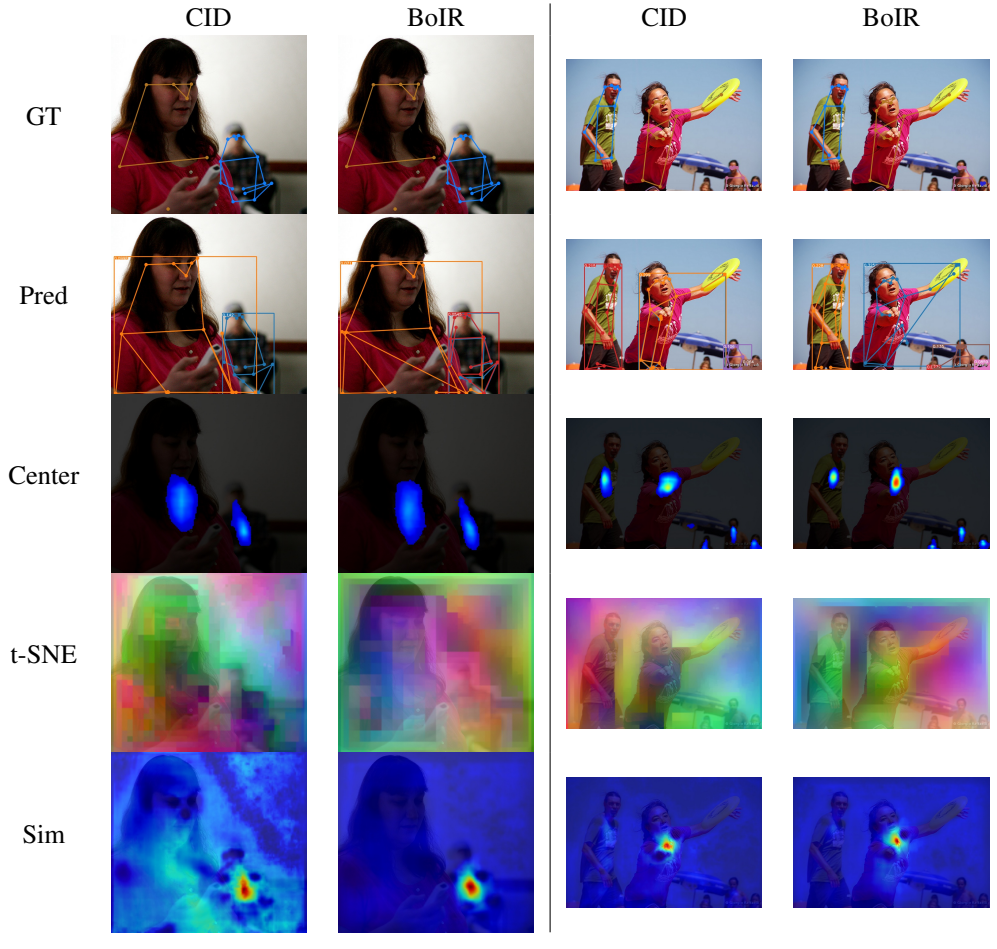

Figure 3: Failure cases on COCO `val` set.

[3] Zhengxiong Luo, Zhicheng Wang, Yan Huang, Liang Wang, Tieniu Tan, and Erjin Zhou. Rethinking the heatmap regression for bottom-up human pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13264–13273, June 2021.

[4] Alejandro Newell, Zhiao Huang, and Jia Deng. Associative embedding: End-to-end learning for joint detection and grouping. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper/2017/file/8edd72158ccd2a879f79cb2538568fdc-Paper.pdf.

[5] Dahu Shi, Xing Wei, Liangqi Li, Ye Ren, and Wenming Tan. End-to-end multi-person pose estimation with transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11069–11078, June 2022.

[6] Ke Sun, Bin Xiao, Dong Liu, and Jingdong Wang. Deep high-resolution representation learning for human pose estimation. In *CVPR*, 2019.

[7] Dongkai Wang and Shiliang Zhang. Contextual instance decoupling for robust multi-person pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11060–11068, June 2022.

[8] Bin Xiao, Haiping Wu, and Yichen Wei. Simple baselines for human pose estimation and tracking. In *European Conference on Computer Vision (ECCV)*, 2018.

[9] Jie Yang, Ailing Zeng, Shilong Liu, Feng Li, Ruimao Zhang, and Lei Zhang. Explicit box detection unifies end-to-end multi-person pose estimation. In *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=s4WVupnJjmX.