BoIR

Jeong, Uyoung; Baek, Seungryul; Chang, Hyung Jin; Kim, Kwang In

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

Jeong, Uyoung; Baek, Seungryul; Chang, Hyung Jin; Kim, Kwang In

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

BMVC 2023 Submission # 763

Abstract

Single-stage multi-person human pose estimation (MPPE) methods have shown great performance improvements, but existing methods fail to disentangle features by individual instances under crowded scenes. In this paper, we propose a bounding box-level instance representation learning called BoIR, which simultaneously solves instance detection, instance disentanglement and instance-keypoint association problems. Our new instance embedding loss provides learning signal on the entire area of the image with bounding box annotations, achieving globally consistent and disentangled instance representation. Our method exploits multi-task learning of bottom-up keypoint estimation, bounding box regression and contrastive instance embedding learning, without additional computational cost during inference. BoIR is effective for crowded scenes, outperforming state-of-the-arts on COCO (0.5 AP), CrowdPose (4.9 AP) and OCHuman (3.5 AP).

1 Introduction

Multi-person human pose estimation(MPPE) aims to localize 2D keypoint locations of multiple human instances from an image. It is useful not only for 3D pose estimation and activity recognition [40], but also for human-robot interaction [5], autonomous driving [42], augmented/virtual reality and surveillance applications. In wild scenarios, where severe inter-person occlusion and background clutter frequently occur, the capability of multi-person pose estimation becomes even more crucial.

Recent advances in single-stage MPPE methods [20, 34, 39] have shown significant performance improvements. Compared to top-down methods [11, 16, 37], they do not require off-the-shelf person detector and therefore robust to detection errors. Unlike bottom-up methods [4, 6, 19, 35, 38], they solve instance-keypoint association problem by explicitly detecting instances, usually using instance center locations.

While single stage methods showed promising results, they still suffer from instance-keypoint association under heavy inter-person occlusion, which often results in noisy predictions. We summarize the main reasons in two aspects. First, existing representation-based methods conceptually lack supervision to learn disentangled instance representation. Even if doing so would incur computational overhead during inference. Second, previous works have spatially sparse supervision. Many works apply learning signals only on ground-truth keypoint locations, which is too sparse for the model to holistically learn the entire image region, leading to noisy and globally inconsistent results. Although heatmap-based approaches apply Gaussian kernel to generate ground-truth keypoint heatmaps, it is still more sparse than conventional segmentation level supervision.

In this paper, we focus on effective instance representation learning method which can provide both conceptually and spatially rich supervision. First, we reformulate to apply embedding loss on separate embedding branch, which can effectively map nonlinear features of instances while primary task branch's performance is not degraded. Then, we design a new contrastive learning scheme, termed Bbox Mask Loss, using bounding box supervision. It contrasts instance embeddings on both inside and outside of the ground-truth bounding boxes, which provides learning signals on the entire image region. Combining with bounding box regression and bottom-up keypoint heatmap regression as auxiliary tasks, we apply multi-task learning scheme to learn effective instance representation for multiple keypoint estimation.

To summarize, our paper presents a new box level instance representation learning method, called BoIR, which simultaneously solves instance disentanglement and instance detection problem, without additional computational cost during inference.

- Bbox Mask Loss effectively disentangles features by instances in the embedding space using a new embedding loss with spatially rich bounding box level supervision.
- Auxiliary task heads enrich instance representation by sharing multiple aspects of the instance, while no additional computational cost is induced during inference.
- BoIR excels at challenging crowded scenes, surpassing comparative methods by 0.5 AP on COCO `test-dev`, 4.9 AP on CrowdPose `test`, and 3.5 AP on OCHuman `test`.

2 Related Works

2D Multi Person Human Pose Estimation. 2D MPPE methods can be roughly classified by instance handling approaches. Top-down methods use detectors [8, 26, 27] to get person bounding boxes and use cropped images as input. Bottom-up methods first detect keypoints and group them into instances. Single stage methods detect instances first, and then regress instance-wise keypoints. Single stage methods do not need to crop an image into multiple instance-wise images, and it does not need to group the keypoints into instances.

SimpleBaseline [37] and HRNet [31] are top-down methods, and generally used as backbone networks in various works. MIPNet [11] is one of the recent top-down approaches which consider multiple instances within a bounding box, by modulating channel dimension to regress individual keypoints.

OpenPose [1], PersonLab [24], and PifPaf [12] share similar idea of estimating a vector field which associates keypoints with instances. HigherHRNet [4] and its subsequent works [6, 19, 35, 38] are another class of bottom-up methods using Associative Embedding [22]. From the pixel-wise one dimensional embedding, they assign the detected keypoints to respective instance using off-the-shelf grouping algorithm [13]. These methods tend to lack capability of instance detection, since their training losses are mainly targeted for keypoint estimation.

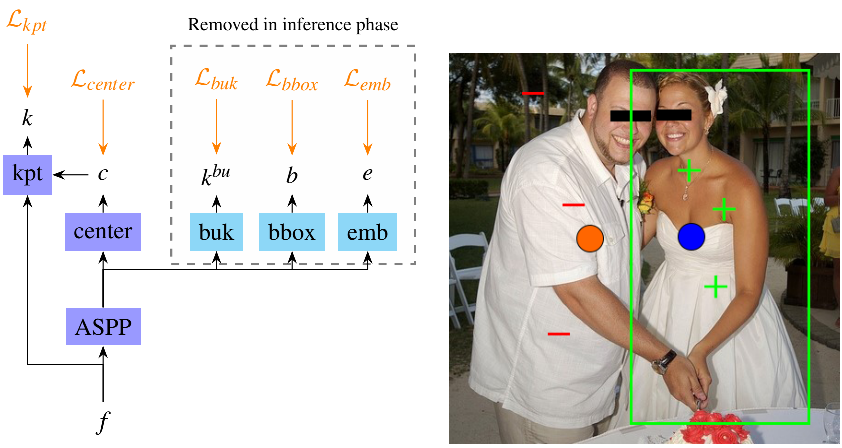

Figure 1: Overview of our framework. Left: keypoint(kpt) head and center head are primary regression heads for MPPE. bottom-up keypoint(buk) head, bounding box(bbox) head and embedding(emb) head are auxiliary multi-task regressors which are not used during inference. Right: Visual illustration of Bbox Mask Loss. Blue circle is a query instance center, where green plus signs represent positive samples within a bounding box. Red minus signs are negative samples. Orange circle is a negative instance's center.

There are several single stage methods based on Transformers [33]. PETR [29] is a bottom-up method based on Transformers architecture. Instead of using Hungarian algorithm for instance grouping, it randomly initializes query embeddings to regress keypoints. On the other hand, ED-Pose [39] extracts query embeddings via human detection decoder. It requires huge computational cost and inference time due to massive amount of learnable parameters, which is critical for real time pose estimation.

FCPose [20] and CID [34] are single stage methods using instance center map. FCPose first obtains instance proposals from one-stage person detector and applies instance-wise dynamic convolution on global feature. Similarly, CID estimates instance body center map to detect instances, and performs channel and spatial attention between sampled feature and global feature, but it does not explicitly perform bounding box regression. CID directly applies contrastive loss on the backbone network's output feature, which actually does not effectively disentangle features by instances, as discussed in SimCLR [3]. Also, CID's contrastive loss is spatially sparse since it is applied only on instance center locations. Instead, we introduce a separate embedding branch which does not hinder learning keypoint features, and also provide spatially and conceptually rich supervision. KAPAO [21] is another single stage method. It reformulates the task as object detection task, and jointly detects person and keypoint objects.

Representation Learning with Distance Metrics. Deep metric learning's objective is to learn a distance metric in embedding space for extracting better representation, generally composed with pull term for closing the distance among positive samples, and push term for disambiguating different classes. Push loss term is crucial for effective representation learning, so many works devoted to propose various negative sampling strategies. Contrastive loss [7], triplet loss [28], N-pair loss [30] and InfoNCE loss [23] are some of the approaches. SimCLR [3], MoCo [9] and CLIP [25] are representative works using variants of InfoNCE loss. All of these methods use cosine similarity as similarity metric.

3 Method

3.1 Framework Overview

Our framework can be decomposed into two main parts: auxiliary task branch and instance keypoint branch. Given an input image, backbone network outputs feature $f \in \mathbb{R}^{C,H,W}$, where H is height and W is width. Task-specific heads produce instance center heatmaps $c \in \mathbb{R}^{1,H,W}$, bounding box(bbox) predictions $b \in \mathbb{R}^{4,H,W}$, bottom-up keypoint heatmaps $k^{bu} \in \mathbb{R}^{K,H,W}$ and instance embedding map $e \in \mathbb{R}^{D,H,W}$. During inference, after detecting instances from the center map, instance embeddings p are sampled from backbone feature on respective center coordinates. p are used as conditions for regressing instance-wise keypoints k in the instance keypoint head, as proposed in [34]. We apply several modifications on the keypoint head including Layer Normalization and Instance Normalization for stable learning. b, k^{bu}, e are not estimated during inference.

3.2 Bbox Mask Loss for Spatial Richness

Existing instance representation learning methods such as Associative Embedding(AE) and CID's contrastive loss failed to handle multiple people in several aspects, leading to noisy results. First, existing methods only compare instance embeddings with ground-truth(GT) instance locations, so they cannot produce push loss term when only one GT instance is available for an image. Second, there are unlabeled instances in training datasets. Existing works simply ignore these unlabeled instances, inducing additional noise during inference. Third, the number of human instances per image in training datasets is too few to effectively learn instance representation. For example, COCO `train` set has 2.6 people per image, excluding labels with `iscrowd`=1. Similarly, CrowdPose `trainval` set has 4.2 people per image.

To alleviate aforementioned challenges, inspired from weakly supervised instance segmentation method [36], we introduce spatially rich supervision using box annotation, called Bbox Mask Loss. It disambiguates each instance embedding from outside of the bounding box region, which can handle arbitrary unlabeled instances and background clutter. It applies soft masking on the inside of the bounding box based on embedding similarity, which is effective for disentangling features under heavy cross-instance occlusion cases. Moreover, it can produce push loss term even when only a single GT instance is available in an image, serving as a simple but effective negative sampling method.

Bbox Mask Loss incorporates multitude of push and pull loss terms, including in-box pull $\mathcal{L}_{pull}^{in}$, out-box push $\mathcal{L}_{push}^{out}$, and cross-instance push $\mathcal{L}_{push}^{inst}$. First, given a GT instance and corresponding bounding box with height h and width w, we compute pixel-wise embedding similarity between embedding map and the instance embedding as defined in Equation 1:

$$s_i^{(x,y)} = \psi(d(e^{(x,y)}, p_i)), \quad (x,y) \in \mathcal{B}_i, \tag{1}$$

where d is a distance metric, and ψ is an inversion operator to convert the distance to similarity with [0,1] output range. From ablative experiment, as reported in Table 4, we find that L2 distance for d and Gaussian kernel for ψ outperforms cosine distance and cosine similarity. $\mathcal{B}_i$ is a set of coordinates inside the box b_i, where $i = 1, 2, ..., N$. As a pulling term inside the box, we want the model to produce similar embeddings on the foreground region of the same person. To realize the objective, we compare the embedding sampled from the box center with the mean instance embedding $\bar{p}_i$, as defined below:

$$\mathcal{L}_{pull}^{in} = \frac{d(p_i, \bar{p}_i)}{N}, \quad \bar{p}_i = \frac{\sum_{(x,y)\in\mathcal{B}_i} e^{(x,y)} s_i^{(x,y)}}{\sum_{(x,y)\in\mathcal{B}_i} s_i^{(x,y)}} \tag{2}$$

In order to decouple the instance embedding from the background, we define the out-box push loss using out-box mean embedding p_i^{out}, as defined in Equation 3:

$$\mathcal{L}_{push}^{out} = d(p_i, p_i^{Cout}), \quad p_i^{Cout} = \frac{\sum_{(x,y)\in\mathcal{B}_i^c} e_{(x,y)}}{|\mathcal{B}_i^c|} \tag{3}$$

Note that $\mathcal{B}_i^c$ is a set of coordinates outside the ith bounding box.

Lastly, cross-instance push term compares instance embeddings retrieved from ground-truths, which is the same as the existing losses.

$$\mathcal{L}_{push}^{inst} = d(p_i, p_{j\neq i}) \tag{4}$$

3.3 Auxiliary Tasks for Conceptual Richness

In order to encourage the features to have richer and more disentangled information for MPPE, we designed to incorporate multiple auxiliary tasks and instance representation learning in parallel. Our multi-task branch consists of shared layers and four separate regression heads, consisting of instance embedding, bottom-up keypoint, bounding box, and instance center.

We concurrently reduce dimensionality of the backbone feature and incorporate multi-resolution shared feature representation based on ASPPv2 [2]. It resolves the problem of regressing globally consistent instance features. Original ASPPv2 module suffers from heavy computational cost during fusion among multiple resolution features. We alleviate this by further squeezing the output channel size of each multi-resolution feature to 128, and then apply fusion layer to obtain final feature with 256 channel size. This design reduces the number of trainable parameters of ASPP by 50%. This shared bottleneck module design helps to prevent auxiliary tasks from dominating over the primary task, by restricting the amount of information flow to auxiliary tasks.

Each regression head comprises with one residual block and one output convolution layer for sufficient capability of learning nonlinear feature transformation. In case of bounding box regression, we adopt anchor free method [15] for efficient training. Note that we do not use the bounding box head outputs during inference, and our bbox head serves as an efficient and informative auxiliary task head.

3.4 Training Losses

In overall, we apply five loss functions: , instance-wise keypoint heatmap loss $\mathcal{L}_{kpt}$, center heatmap loss $\mathcal{L}_{center}$, bottom-up keypoint heatmap loss $\mathcal{L}_{buk}$, bounding box loss $\mathcal{L}_{bbox}$, and embedding loss $\mathcal{L}_{emb}$.

$$\mathcal{L} = \mathcal{L}_{kpt} + \mathcal{L}_{center} + \mathcal{L}_{buk} + \mathcal{L}_{bbox} + \mathcal{L}_{emb} \tag{5}$$

Specifically, Focal loss [14, 44] is used for $\mathcal{L}_{kpt}, \mathcal{L}_{center}$ and $\mathcal{L}_{buk}$, while CIoU loss [43] is used for $\mathcal{L}_{bbox}$. For embedding loss, we use four loss terms as defined in Equation 2,3,4. We use AE loss for calculating respective terms.

$$\mathcal{L}_{emb} = \mathcal{L}_{pull}^{in} + \mathcal{L}_{push}^{out} + \mathcal{L}_{push}^{inst} \tag{6}$$

Method	Backbone	Input size	AP	AP^{50}	AP^{75}	AP^{M}	AP^{L}	AR
Top-down methods								
SBL [37]	ResNet-152	384×288	73.7	91.9	81.1	70.3	80.0	-
HRNet [32]	HRNet-W32	384×288	74.9	92.5	82.8	71.3	80.9	-
Bottom-up methods								
HrHRNet [4]	HrHRNet-W32	512	66.4	87.5	72.8	61.2	74.2	-
DEKR [6]	HRNet-W32	512	67.3	87.9	74.1	61.5	76.1	72.4
SWAHR [19]	HrHRNet-W32	512	67.9	88.9	74.5	62.4	75.5	-
Single stage methods								
FCPose [20]	ResNet-101+FPN	800	65.6	87.9	72.6	62.1	72.3	-
PETR [29]	ResNet-101	800	68.5	90.3	76.5	62.5	77.0	-
ED-Pose [39]	ResNet-50	800	69.8	90.2	77.2	64.3	77.4	-
CID [34]	HRNet-W32	512	68.9	89.9	76.0	63.2	**77.7**	74.6
CID [34]	HRNet-W48	640	70.7	90.3	77.9	66.3	**77.8**	76.4
BoIR	HRNet-W32	512	**69.5**	**90.4**	**76.9**	**64.2**	77.3	**75.3**
BoIR	HRNet-W48	640	**71.2**	**90.8**	**78.6**	**67.0**	77.6	**77.1**

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

Method	Backbone	Input size	AP	AP^{50}	AP^{75}	AP^{E}	AP^{M}	AP^{H}
Top-down methods								
SBL [37]	ResNet-101	-	60.8	81.4	65.7	71.4	61.2	51.2
SPPE [16]	ResNet-101	320× 256	66.0	84.2	71.5	75.5	66.3	57.4
Bottom-up methods								
HrHRNet [4]	HrHRNet-W48	640	65.9	86.4	70.6	73.3	66.5	57.9
DEKR [6]	HrHRNet-W32	512	65.7	85.7	70.4	73.0	66.4	57.5
SWAHR [19]	HrHRNet-W48	640	71.6	88.5	77.6	78.9	72.4	63.0
Single stage methods								
PETR [29]	Swin-L	800	71.6	90.4	78.3	77.3	72.0	65.8
ED-Pose [39]	ResNet-50	800	69.9	88.6	75.8	77.7	70.6	60.9
CID [34]	HRNet-W32	512	71.3	90.6	76.6	77.4	72.1	63.9
CID [34]	HRNet-W48	640	72.3	90.8	77.9	78.7	73.0	64.8
BoIR	HRNet-W32	512	70.6	89.9	76.5	77.1	71.2	63.0
BoIR	HRNet-W48	640	71.2	90.3	76.7	77.8	71.8	63.5
BoIR*	HRNet-W32	512	**75.8**	**92.2**	**82.3**	**82.3**	**76.5**	**67.5**
BoIR*	HRNet-W48	640	**77.2**	**92.4**	**83.5**	**82.7**	**78.1**	**69.8**

Table 2: Comparison with state-of-the-art methods on CrowdPose `test` set. Best scores are marked as bold for small(e.g. HRNet-W32) and large(e.g. HRNet-W48) models respectively. Models with $*$ are trained on COCO and finetuned on CrowdPose.

Method	Backbone	COCO val		OCHuman val		OCHuman test	
		AP	AR	AP	AR	AP	AR
DEKR [6]	HRNet-W32	68.0	73.0	37.9	-	36.5	-
DEKR [6]	HRNet-W48	71.0	76.0	-	-	-	-
CID [34]	HRNet-W32	69.8	75.4	44.9	-	44.0	-
CID [34]	HRNet-W48	-	-	46.1	-	45.0	-
BoIR	HRNet-W32	**70.6**	**76.3**	**47.4**	**80.1**	**47.0**	**80.3**
BoIR	HRNet-W48	**72.5**	**78.3**	**49.4**	**80.8**	**48.5**	**80.7**

Table 3: Comparison with state-of-the-art methods on COCO `val` and OCHuman `val`, `test` set. OCHuman performance is evaluated with COCO pretrained model without fine-tuning.

4 Experiments

4.1 Datasets and Evaluation Metrics

COCO Keypoint 2017. [17] It comprises train(57K images), val(5K images), and test-dev(20K images) splits, annotated with 17 keypoints. We use train split for training, and val split for hyperparameter tuning.
CrowdPose. [16] It consists of 20K images and 80K instances, annotated with 14 keypoints. Following the evaluation protocol of [34], we use trainval split(12K images, 43.4K instances) for training and test split(8K images, 29K instances) for evaluation.
OCHuman. [41] OCHuman dataset is targeted for evaluation on crowded scenes with extreme conditions. 2,500 images are for `val` set, and 2,231 images are for `test` set. We evaluate our method following [10, 34].
Evaluation metrics. We follow COCO evaluation protocol, where AP(Average Precision) and AR(Average Recall) are computed based on OKS(Object Keypoint Similarity) with varying thresholds, including AP(averaged AP), AP^{50}(AP at OKS=0.5), and AP^{75}(AP at OKS=0.75). In case of CrowdPose, we additionally report metrics based on crowd index, including AP^{E}(easy), AP^{M}(medium), and AP^{H}(hard).

4.2 Implementation Details

Our implementation is based on [34]. We use HRNet-W32 and HRNet-W48 as backbone networks, and perform hyperparameter tuning with COCO `val` set results. We apply AdamW optimizer with initial learning rate 1.0e-3, weight decay 1.0e-2 and cosine learning rate scheduler with 10 warmup epochs, following [18]. For COCO, we train the model for 140 epochs on 4 GPUs(RTX 3090 for HRNet-W32 backbone, A6000 for HRNet-W48 backbone) with AMP, where 20 batch size is used for each device. For CrowdPose, similar to [34], we train the model for 310 epochs when training from scratch, while 100 epochs with 1 warmup epoch is applied for transfer learning. Following [4, 6, 34], we apply single scale test with flipping.

4.3 Comparison with State-of-the-arts

COCO Dataset Results. We report COCO `val` results in Table 3, and `test-dev` results in Table 1. Our method outperforms existing state-of-the-art under the same or similar back-

Bbox Mask Loss	Bbox Head	AP
		69.6
✓		70.2
	✓	70.4
✓	✓	70.6

Emb. Loss	Dist. Metric	AP
Contrastive	cosine	70.3
Contrastive	L2	70.2
AE	L2	70.6

Table 4: Left: Ablation study of Bbox Mask Loss and bounding box regression head on COCO `val` set. Right: Ablation study of embedding loss function and distance metric on COCO `val` set, where Bbox Mask Loss and bbox head are used.

Method	Backbone	Params(M)	GFLOPs	Time(ms)	AP
CID	HRNet-W32	29.3	42.8	86.7	69.8
CID	HRNet-W48	65.4	-	-	-
ED-Pose	ResNet-50	47.9	187.5	113.9	71.6
ED-Pose	Swin-L	218.8	2,615	272.1	74.3
BoIR	HRNet-W32	31.8	83.4	110.6	70.6
BoIR	HRNet-W48	68.9	227.7	167.3	72.5

Table 5: Computational cost comparison on COCO `val` set. Inference time is measured with single RTX 3090 and 1 batch size.

bone. Our method with HRNet-W32 backbone outperforms CID by 0.8 AP on `val` and 0.6 AP on `test-dev`. Similarly, we achieve 0.5 AP improvement on `test-dev` with HRNet-W48 backbone.

CrowdPose Dataset Results. We compare other methods on CrowdPose `test` in Table 2. BoIR is second best among state-of-the-art methods. Nontheless, our method suffers from performance drop by 0.7 AP on HRNet-W32 backbone and 1.1 AP on HRNet-W48 backbone. We speculate that as the model size increases, the model suffers from insufficient amount of training data on CrowdPose, as the performance difference between CID and ED-Pose on CrowdPose is also reversed on COCO. To validate the hypothesis, we introduce finetuning on CrowdPose using the model weights trained on COCO `train` set. Finetuning strategy is proven to be far more effective, surpassing existing state-of-the-art by 4.5 AP with HRNet-W32 backbone, and 4.9 AP with HRNet-W48 backbone.

OCHuman Results. Comparison on OCHuman is summarized in Table 3. Following the protocol in [10], we evaluate the model trained on COCO without finetuning on OCHuman. BoIR outperforms comparative methods on both `val` and `test` set by large margin. Therefore, our instance representation learning is effective especially for crowded scenes.

4.4 Ablation Study

We conduct ablative experiment as demonstrated in Table 4. Effectiveness of Bbox Mask Loss and Bbox Head is validated by enumerating four possible combinations, and the result shows that our proposed methods are useful. We additionally conduct ablative experiment on embedding losses and distance metrics. AE loss turns out to be superior than Contrastive loss. We hypothesize that L2 distance with Gaussian kernel used in AE loss is better suited for keypoint evaluation criteria, as claimed in [22]. We also extensively compare computational cost in Table 5. Our method manages to keep the computational cost within reasonable extent, compared to ED-Pose. For qualitative and visual analysis, we compare our method

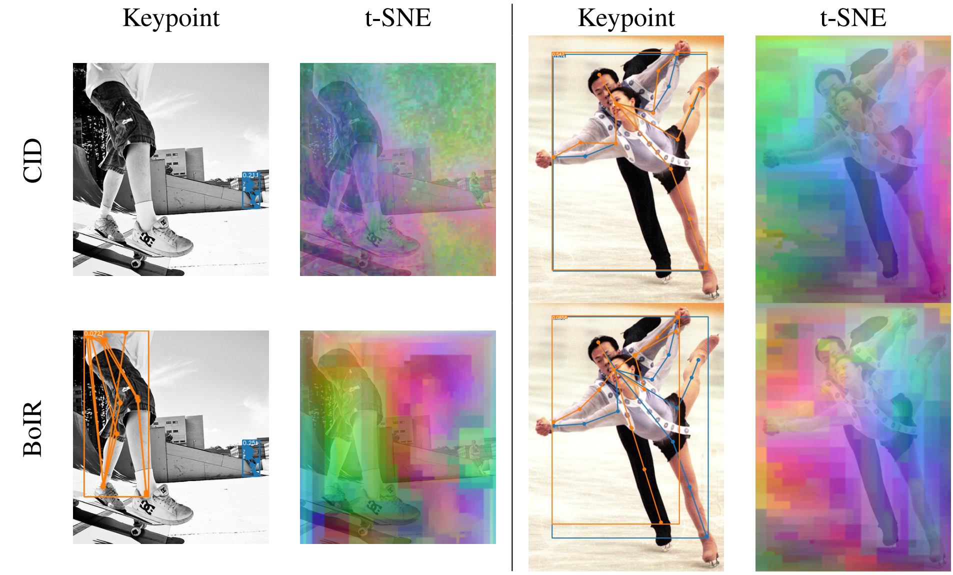

Figure 2: Qualitative results of our method. Left image is from COCO `val` set, and right image is from CrowdPose `test` set.
t-SNE is applied on the output backbone feature for 250 iterations with 3 output dimension per pixel, which directly corresponds to normalized RGB value.

with CID in Fig. 2. Our method better disentangles features by instances, effectively handling background noise and inter-person occlusion.

5 Conclusion

This paper proposes a new multi-person pose estimation method using bounding box-supervised instance representation learning, called BoIR. It provides rich spatial supervision, utilizing embedding similarity as a soft mask for positive sampling, and the background region as a negative sample. It also incorporates auxiliary tasks for conceptually richer representation learning, without additional computation cost during inference. Our instance embedding can effectively disentangle instances in crowded scenes, surpassing comparable state-of-the-art methods on multiple human pose estimation benchmarks. Despite notable performance improvement with transfer learning, effective representation learning on small training data is a remaining issue, and we plan to mitigate the limitation as a future work. We hope BoIR can motivate further instance representation learning methods for multi-person pose estimation.

References

[1] Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Realtime multi-person 2d pose estimation using part affinity fields. In *CVPR*, 2017.

[2] Liang-Chieh Chen, George Papandreou, Florian Schroff, and Hartwig Adam. Re-

thinking atrous convolution for semantic image segmentation, 2017. URL https://arxiv.org/abs/1706.05587.

[3] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR, 2020.

[4] 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.

[5] Qing Gao, Jinguo Liu, Zhaojie Ju, and Xin Zhang. Dual-hand detection for human–robot interaction by a parallel network based on hand detection and body pose estimation. *IEEE Transactions on Industrial Electronics*, 66(12):9663–9672, 2019. doi: 10.1109/TIE.2019.2898624.

[6] 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.

[7] Raia Hadsell, Sumit Chopra, and Yann LeCun. Dimensionality reduction by learning an invariant mapping. In *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, volume 2, pages 1735–1742. IEEE, 2006.

[8] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969, 2017.

[9] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9729–9738, 2020.

[10] Sheng Jin, Wentao Liu, Enze Xie, Wenhai Wang, Chen Qian, Wanli Ouyang, and Ping Luo. Differentiable hierarchical graph grouping for multi-person pose estimation. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part VII 16*, pages 718–734. Springer, 2020.

[11] Rawal Khirodkar, Visesh Chari, Amit Agrawal, and Ambrish Tyagi. Multi-instance pose networks: Rethinking top-down pose estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 3122–3131, October 2021.

[12] Sven Kreiss, Lorenzo Bertoni, and Alexandre Alahi. Pifpaf: Composite fields for human pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

[13] Harold W Kuhn. The hungarian method for the assignment problem. *Naval research logistics quarterly*, 2(1-2):83–97, 1955.

[14] Hei Law and Jia Deng. Cornernet: Detecting objects as paired keypoints. In *Proceedings of the European Conference on Computer Vision (ECCV)*, September 2018.

[15] Chuyi Li, Lulu Li, Hongliang Jiang, Kaiheng Weng, Yifei Geng, Liang Li, Zaidan Ke, Qingyuan Li, Meng Cheng, Weiqiang Nie, Yiduo Li, Bo Zhang, Yufei Liang, Linyuan Zhou, Xiaoming Xu, Xiangxiang Chu, Xiaoming Wei, and Xiaolin Wei. Yolov6: A single-stage object detection framework for industrial applications, 2022.

[16] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

[17] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft coco: Common objects in context. In David Fleet, Tomas Pajdla, Bernt Schiele, and Tinne Tuytelaars, editors, *Computer Vision – ECCV 2014*, pages 740–755, Cham, 2014. Springer International Publishing. ISBN 978-3-319-10602-1.

[18] Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A convnet for the 2020s. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11976–11986, 2022.

[19] 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.

[20] Weian Mao, Zhi Tian, Xinlong Wang, and Chunhua Shen. Fcpose: Fully convolutional multi-person pose estimation with dynamic instance-aware convolutions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9034–9043, June 2021.

[21] William McNally, Kanav Vats, Alexander Wong, and John McPhee. Rethinking keypoint representations: Modeling keypoints and poses as objects for multi-person human pose estimation. In Shai Avidan, Gabriel Brostow, Moustapha Cissé, Giovanni Maria Farinella, and Tal Hassner, editors, *Computer Vision – ECCV 2022*, pages 37–54, Cham, 2022. Springer Nature Switzerland. ISBN 978-3-031-20068-7.

[22] 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.

[23] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

[24] George Papandreou, Tyler Zhu, Liang-Chieh Chen, Spyros Gidaris, Jonathan Tompson, and Kevin Murphy. Personlab: Person pose estimation and instance segmentation with

a bottom-up, part-based, geometric embedding model. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 269–286, 2018.

[25] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021.

[26] Joseph Redmon and Ali Farhadi. Yolov3: An incremental improvement. *arXiv preprint arXiv:1804.02767*, 2018.

[27] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*, 28, 2015.

[28] Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 815–823, 2015.

[29] 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.

[30] Kihyuk Sohn. Improved deep metric learning with multi-class n-pair loss objective. In D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc., 2016. URL https://proceedings.neurips.cc/paper/2016/file/6b180037abbebea991d8b1232f8a8ca9-Paper.pdf.

[31] Ke Sun, Bin Xiao, Dong Liu, and Jingdong Wang. Deep high-resolution representation learning for human pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

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

[33] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[34] 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.

[35] Haixin Wang, Lu Zhou, Yingying Chen, Ming Tang, and Jinqiao Wang. Regularizing vector embedding in bottom-up human pose estimation. In Shai Avidan, Gabriel Brostow, Moustapha Cissé, Giovanni Maria Farinella, and Tal Hassner, editors, *Computer Vision – ECCV 2022*, pages 107–122, Cham, 2022. Springer Nature Switzerland. ISBN 978-3-031-20068-7.

[36] Adrian Wolny, Qin Yu, Constantin Pape, and Anna Kreshuk. Sparse object-level supervision for instance segmentation with pixel embeddings. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4402–4411, 2022.

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

[38] Nan Xue, Tianfu Wu, Gui-Song Xia, and Liangpei Zhang. Learning local-global contextual adaptation for multi-person pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13065–13074, June 2022.

[39] 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.

[40] Pengfei Zhang, Cuiling Lan, Wenjun Zeng, Junliang Xing, Jianru Xue, and Nanning Zheng. Semantics-guided neural networks for efficient skeleton-based human action recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

[41] Song-Hai Zhang, Ruilong Li, Xin Dong, Paul Rosin, Zixi Cai, Xi Han, Dingcheng Yang, Haozhi Huang, and Shi-Min Hu. Pose2seg: Detection free human instance segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 889–898, 2019.

[42] Jingxiao Zheng, Xinwei Shi, Alexander Gorban, Junhua Mao, Yang Song, Charles R. Qi, Ting Liu, Visesh Chari, Andre Cornman, Yin Zhou, Congcong Li, and Dragomir Anguelov. Multi-modal 3d human pose estimation with 2d weak supervision in autonomous driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 4478–4487, June 2022.

[43] Zhaohui Zheng, Ping Wang, Wei Liu, Jinze Li, Rongguang Ye, and Dongwei Ren. Distance-iou loss: Faster and better learning for bounding box regression. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 12993–13000, 2020.

[44] Xingyi Zhou, Dequan Wang, and Philipp Krähenbühl. Objects as points, 2019. URL https://arxiv.org/abs/1904.07850.