Learning Local-Global Contextual Adaptation for Multi-Person Pose Estimation

Nan Xue[1,2] Tianfu Wu[2*] Gui-Song Xia[1,3] Liangpei Zhang[3]
[1] School of Computer Science, Wuhan University [2] Department of ECE, NC State University
[3] State Key Lab. of Information Engieering in Surveying, Mapping and Remote Sensing, Wuhan University
Code& Models available at: https://github.com/cherubicXN/logocap

Abstract

This paper studies the problem of multi-person pose estimation in a bottom-up fashion. With a new and strong observation that the localization issue of the center-offset formulation can be remedied in a local-window search scheme in an ideal situation, we propose a multi-person pose estimation approach, dubbed as LOGO-CAP, by learning the LOcal-GlObal Contextual Adaptation for human Pose. Specifically, our approach learns the keypoint attraction maps (KAMs) from the local keypoints expansion maps (KEMs) in small local windows in the first step, which are subsequently treated as dynamic convolutional kernels on the keypoints-focused global heatmaps for contextual adaptation, achieving accurate multi-person pose estimation. Our method is end-to-end trainable with near real-time inference speed in a single forward pass, obtaining state-of-the-art performance on the COCO keypoint benchmark for bottom-up human pose estimation. With the COCO trained model, our method also outperforms prior arts by a large margin on the challenging OCHuman dataset.

1. Introduction

2D human pose estimation is a classical computer vision problem that aims to parsing articulated structures of human parts from natural images. With rich and longstanding studies, we have witnessed great successes on single-person pose estimation [21, 31] by convolutional neural networks. Therefore, it is of great interest in pushing the pose estimation from the single-person to multi-person configuration.

The problem of multi-person pose estimation from images has been extensively studied in top-down paradigms [21, 31, 36] that formulate the problem as single-person pose estimation with an off-the-shelf person detector with an impressively high AP (e.g, 56% on the COCO-2017 validation dataset [36]). Although the top-down paradigm has been dramatically pushed into a high-performing stages, it remains several problems in both aspects of efficiency and accuracy due to the dependency of detecting person bounding boxes. Motivated by this, we are interested in studying the problem of multi-person pose estimation without incurring extra priors of bounding boxes, which is conventionally termed as *bottom-up pose estimation*.

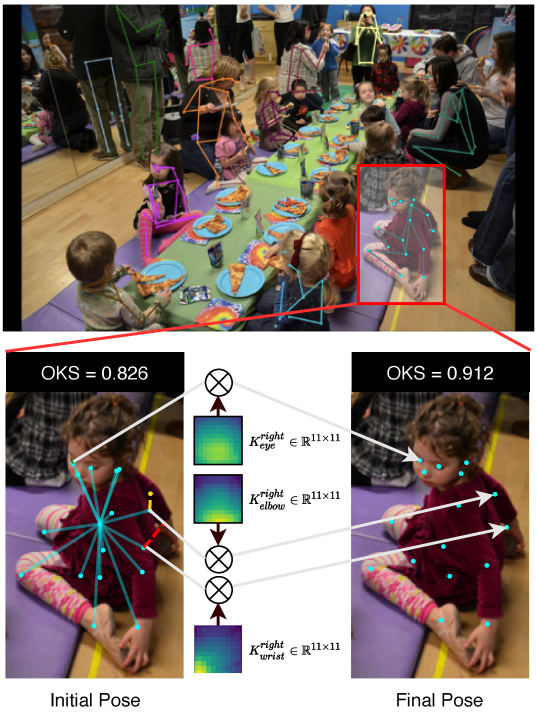

Figure 1. An illustrative example of multi-person pose estimation by the proposed LOGO-CAP with the HRNet-W32 backbone. For each initial pose obtained by the center-offset regression, LOGO-CAP learns 11×11 local filters for each joint, and then refines the initial keypoint by convolution with the learned kernels: *The local filters are learned to* ***refocusing*** *those initially-less-accurate pose keypoints towards better placement.* In more detailed, we show an example of the initial center-offset pose that only obtains the OKS of 0.826 due to the misplacement between the keypoints of `the right elbow` and `the right wrist`. We mark the residual vectors between the predictions and the groundtruth with yellow and red dash lines respectively. The LOGO-CAP improves the OKS by 10.4%. Please see text for details.

Bottom-up pose estimation approaches pay much atten-

*Corresponding author

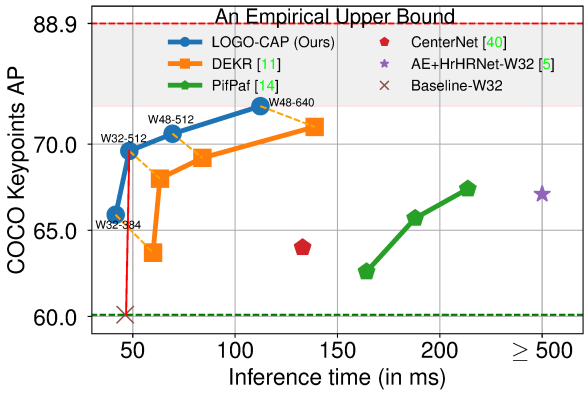

Figure 2. Our LOGO-CAP is motivated by a strong empirical observation that a vanilla center-offset baseline with an AP of 60.1 (marked by green dashed line) can be improved by leveraging a searching scheme in 11×11 local windows to an AP of 88.9 (marked by red dashed line). In the meanwhile, we illustrate the speed-accuracy comparisons between our LOGO-CAP and prior arts on the COCO val-2017 dataset. Wx-Y (e.g. W32-384) means that a model uses the backbone HRNet-Wx [31] and is tested with the image resolution Y in the short side.

tion on estimating pose parameters from any given multi-person image instead of using the cropped single-person ones, which poses several challenges on accurately identifying the learned bottom-level keypoints as person parts and grouping them into different individuals by learning part affinity fields [2, 3], part association fields [14], associative embedding [20]. Those grouping approaches are accurate but sophisticated due to the necessity of ad-hoc grouping/decoding schemes. Recently, many researchers attempted to learning center-offsets [6, 11, 30, 40] for pose estimation as its intrinsic simplicity and efficiency. However, most of the center-offset approaches are suffering from the main challenge of localization inaccuracy due to the large structural variations of human pose, thus leading to inferior performance than the grouping ones.

In this paper, we are interested in studying the bottom-up pose estimation using the center-offset formulation for its simplicity and efficiency. We directly address the aforementioned main challenges for center-offset based multi-person pose estimation. Our proposed approach is motivated by a surprisingly strong empirical observation via analyzing what the fundamental issues are with the vanilla center-offset pose estimation network.

An Empirical Observation. Our analyses are based on results in the fully-annotated subset of the COCO val-2017 dataset[1]. The vanilla center-offset regression method utilizes the HRNet-W32 [31] as the feature backbone to directly predict keypoints center heatmap and the offset vectors. As shown in Fig. 2, it obtains 60.1 average precision (AP), which is not great, but reasonably good. It clearly shows that the pose keypoints center and the offset vectors can be learned reasonably well. We ask: *How exactly bad are the center-offset estimation?* We want to know (i) whether some of keypoints are truly bad being far away from the underlying ground-truth (GT) locations, or (ii) whether most of them are already in the close proximity of the GT ones? We observe that the latter is true. To quantitatively characterize the close proximity, instead of directly utilizing the learned offset vectors for human pose estimation, we treat them as human pose keypoint initialization and do a local window search to compute the empirical upper-bound of performance. More specifically, based on the initially predicted human poses, by introducing a local window (e.g., 11×11) centered at each detected key point and by computing the single keypoint similarity with the ground-truth keypoint, **an empirical upper-bound of** 88.9 **AP is obtained**, which is significantly higher than the state of the art and shows the potential of improving the vanilla center-offset regression paradigm.

A Direct Solution Leveraging the Observation. The implication of the observation is significant: It reveals that the fundamental challenge of improving the center-offset approach for pose estimation is to resolve the local misplacement. To that end, a straightforward way is just to learn a local heatmap (e.g., 11×11) for each human pose keypoint based on the learned center and offset vectors, and then to compute the refined keypoints by taking $\arg\max$ within the local heatmap. Although appealing, this does not work as observed during our development of the LOGO-CAP. The underlying reason is also straightforward: if this can work, the original offset vector regression should work at the first place since no additional information is introduced through learning the local heatmap.

We hypothesize that on the one hand, in addition to the local heatmap, the structural relationship between different pose keypoints needs to be taken into account, and on the other hand, the intrinsic uncertainty of the local information in a local heatmap needs to be resolved. The former is the key challenge of structured output prediction problems. Many message passing algorithms have been developed in the literature. The latter can not be addressed by simply increasing the local window size. It entails learning stronger local-global information interaction and adaptation.

To verify the two hypotheses, the proposed LOGO-CAP lifts the initial keypoints via the center-offset prediction to keypoint expansion maps (KEMs) to counter their lack of localization accuracy in two modules (Section 3). The KEMs extend the star-structured representation of the center-offset formulation to the pictorial structure representation [9, 10]. As shown in Fig. 1, in our LOGO-CAP, one module computes local KEMs and learns to account for the

[1]Note that the COCO-val-2017 dataset contain many partially-annotated images (with only ground-truth bounding boxes), we use 2346 testing samples that are fully annotated with keypoint annotations.

structured output prediction nature of the human pose estimation problem, resulting in the keypoint attraction maps (KAMs), i.e., the local filters in Fig. 1. Another module computes global KEMs and learns to refine the global KEMs by integrating the KAMs.

Our LOGO-CAP is a fully end-to-end bottom-up human pose estimation method with near real-time inference speed. It obtains 70.0 AP in the fully-annotated subset of the COCO val-2017 dataset, which is an absolute increase of 9.9 AP compared to the vanilla center-offset method, making a significant step forward. Fig. 2 shows the advantage of the proposed LOGO-CAP in terms of overall speed-accuracy comparisons between our LOGO-CAP and prior arts. Meanwhile, we should notice that there is still a significant gap towards the empirical upper bound in Fig. 2, which encourages more work to be investigated.

2. Related Works and Our Contributions

There is a vast body of literature for human pose estimation. Many elegant representation schema have been developed for modeling articulated human pose in the traditional approaches such as the well-known pictorial structure model [9, 10] and its many variants [1, 25, 26, 28, 37]. With the resurgence of DNNs and the end-to-end learning, the performance of single-person pose estimation has been largely improved. We briefly review the recent deep learning based approaches for multi-person pose estimation.

Top-Down Pose Estimation. It essentially exploits single-person pose estimation approaches on cropped single human image patches [4, 7, 16, 21, 24, 31, 36], where human detection is often done by an off-the-shelf object detector (*e.g.*, Faster-RCNN [27]). Although excellent performance has been achieved in such a pipeline, they are suffering from efficiency issue due to the dependency of object detectors. Accordingly, there are work focusing on developing efficient backbone networks (*e.g.*, Lite-HRNet [19, 29, 38]) to reduce the inference latency on single-person pose estimation, often at the expense of degenerating the accuracy of pose estimation by large margins. Mask-RCNN [12] addresses this problem by learning heatmaps from features extracted by ROIAlign in object proposals, whose performance on pose estimation lags behind. To clarify the use of terminology, top-down approaches referred in this paper are those using the precomputed bounding boxes as done in the SimpleBaseline method [36]. Compared with the top-down approaches, our LOGO-CAP directly addressed the problem of multi-person pose estimation in a bottom-up way without incurring the regional image/feature context over the person boxes. Our proposed LOGO-CAP obtains competitive performance in terms of accuracy while achieving nearly real-time inference speed.

Limb-based Grouping Approaches. Motivated by the naturalness of modeling limbs based on keypoints, the problem of multi-person pose estimation has been extensively studied by grouping persons from the learned limb associations. Given a predefined limb configuration (*e.g.*, the COCO person skeleton template consisting of 19 limbs based on 17 keypoints), the grouping can be addressed by Part affinity field (PAF) [2, 3], Associative Embedding (AE) [20], mid-range offset fields in PersonLab [23] and the fields of Part Intensity and Association [14]. Typically, sophisticated designs are entailed to achieve good performance. For example, a bipartite graph matching is used in OpenPose [2]. In addition to be computationally expensive, another drawback of these methods is not fully end-to-end trainable. More recently, the differentiability issue was studied by the Hierarchical Graph Clustering (HGG) method [13], which utilizes graph convolution networks to repeatedly delineate pose parameters of multiple persons from a keypoint graph. HGG improves the performance compared to its baseline, the Associative Embedding method [20] at the expense of significantly increased computational cost. In contrast to those approaches, our proposed LOGO-CAP is fully end-to-end trainable and achieves near real-time inference speed.

Direct Regression based Approaches. This formulation has attracted much attention due to their conceptual simplicity [6, 11, 30, 32, 34], inspired by the recent remarkable success of direct bounding box regression in object detection such as the FCOS method [33] and CenterNets [6, 40]. As aforementioned, one main challenge is the difficulty of accurately regressing the offset vectors, especially for the long-range keypoints with respect to the center. Sophisticated post-processing schema are often entailed to improve the performance. For example, a method of matching the directly regressed poses to the nearest keypoints that are extracted from the global keypiont heatmaps is used in [40]. Although being simple, the performance of this line of work is usually inferior to the limb-based approaches. The mixture regression network [34] alleviated the issue of regression quality to some extent, but still remained an indispensable performance gap comparing with the grouping-based approaches. Most recently, Geng *et al.* presented the first competitive direct method, DEKR [11] with a novel pose-specific neural architecture for disentangled keypoint regression. To improve the performance, the DEKR method utilizes a lightweight rescoring network to recalibrate the pose scores that are computed based on the keypoint heatmaps , and thus is not fully end-to-end. The proposed LOGO-CAP retains the simplicity of the vanilla center-offset formulation and enjoys fully end-to-end training and fast inference speed.

Our Contributions. The proposed LOGO-CAP makes three main contributions to the field of bottom-up human pose estimation: (i) We address the drawback of the vanilla center-offset formulation while retaining its efficiency. It

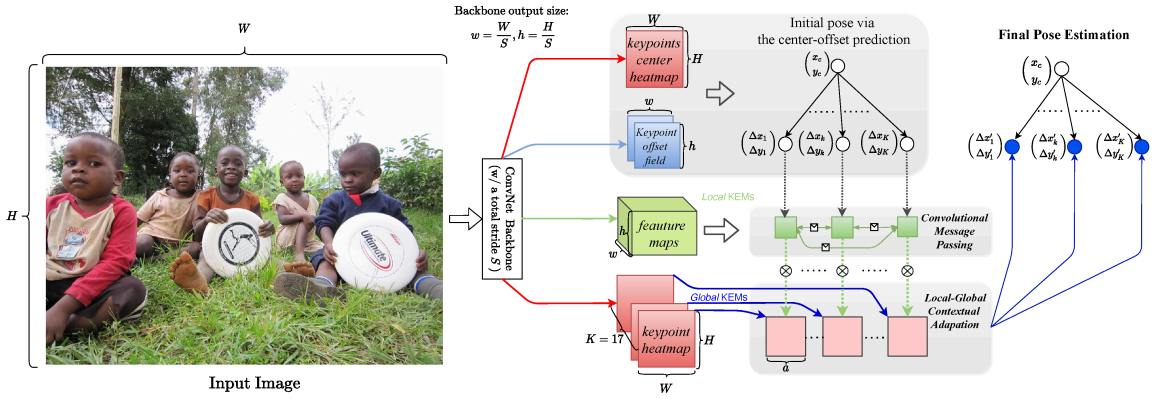

Figure 3. Schematic illustration of the proposed LOGO-CAP for bottom-up human pose estimation. See text for detail.

proposes the key idea of lifting a keypoint to a keypoint expansion map to counter the lack of localization accuracy. (ii) We present a novel local-global contextual adaptation formulation that accounts for the nature of structured output prediction in human pose estimation and harnesses local-global structural information integration. (iii) Our proposed LOGO-CAP obtains state-of-the-art performance on the COCO val-2017 and test-2017 datasets. It also shows the strong generalization ability with state-of-the-art performance on the OCHuman dataset.

3. The Proposed LOGO-CAP

Fig. 3 illustrates the proposed LOGO-CAP, which consists of three main components: a feature backbone, the initial center-offset human pose estimation, and the proposed local-global contextual adaptation component for the final human pose estimation.

3.1. Learning Local and Global Context

Backbone Network and Pose Initialization. Given an input image I, the output of the feature backbone is a C-dim feature map, denoted by $\mathbf{F} \in \mathbb{R}^{C\times h\times w}$, where C is the feature dimension of the last convolutional layer in the feature backbone, and the spatial size $h \times w$ depends on the total stride in the feature backbone. The feature map F is used as the shared features by two different head branches for the prediction of center heatmaps $\mathbf{C}$ and offsets fields $\mathbf{O}$. Then, the initial pose parameters are extracted using the top-N local maximal locations and the corresponding offset vectors, denoted by $\mathbf{p}_i \in \mathbb{R}^{17\times 2}$ for the i-th person.

Local Context. We first introduce the local keypoint expansion maps (KEMs) for the initial pose parameters. Specifically, for the j-type keypoint (*e.g.*, nose of a person), we follow the **empirical observation** in Sec. 1 to compute the local KEMs in 11×11 windows as $\mathcal{M} \in \mathbb{R}_{N\times 17\times 11\times 11\times 2}$, as shown in Algorithm 1.

Then, we encode the geometric mesh of KEMs $\mathcal{M}$ in a d-dim latent space (*e.g.*, $d = 64$ in our experiments), computed based on the feature backbone output. A pose

Algorithm 1: Computing KEMs in a PyTorch-like style

```
σ= coco_sigmas #17 × 1, keypoint sigmas
  provided in the COCO dataset.
def KptsExpansionCoco(P,ks=11)
    #Initial poses P:  Nx17x2
    r = ks // 2
    dy, dx = meshgrid(arange(-r,r),arange(-r,r))
    dy = dy.reshape(1, 1, ks, ks)
    dx = dx.reshape(1, 1, ks, ks)
    scale = σ.reshape(1, 17, 1, 1)/σ.min()
     #keypoint type specific expansion rate
    dy = dy ∗ scale #1x17x11x11
    dx = dx ∗ scale #1x17x11x11
    dxy = stack((dx,dy),dim=-1) #1x17x11x11x2
    M = P.reshape(N, 17, 1, 1, 2) + dxy
    return M
```

instance is represented by concatenating all the 17 keypoints. All initial keypoint-based poses are geometrically "expanded / lifted" and feature-activated (*i.e.*, sampling the features over $\mathcal{M}$), resulting the initial local context, $\mathcal{K} \in \mathbb{R}^{N\times(17\times d)\times 11\times 11}$.

Local Context Convolutional Message Passing. To facilitate the structural information flow between different latent codes of the keypoints of a pose instance, we propose a simple convolutional message passing (CMP) module with three layers of Conv+Norm+ReLU operations with the Attentive Norm [17] used in the second layer. The transformed latent code $\mathcal{K}'$ is decoded by a 1×1 Conv. layer as the local keypoints attraction maps (KAMs), $\mathbf{K} \in \mathbb{R}^{N\times 17\times 11\times 11}$ to measure the uncertainty of the initial poses.

Local-Global Contextual Adaptation. Through the CMP, we obtain the dynamic (a.k.a., data-driven) kernels for the 17 keypoints in a pose instance-sensitive way, which are used to refine the global heatmaps $\mathcal{H}$ for the 17 keypoints. In detail, we first compute another geometric mesh with window $a \times a$ (e.g., $a = 97$) for each keypoint of the N pose instances, and the entire mesh is denoted by

$\mathcal{M}_G \in \mathbb{R}^{N\times 17\times a\times a\times 2}$. The mesh can be interpreted as the global KEM. It is then instantiated with appearance features extracted from the global heatmaps, and we have,

$$\mathcal{H}^{\uparrow}_{(1:17)} \xrightarrow[\text{bi-linear}]{\mathcal{M}_G} \mathbb{H} \xrightarrow[\text{reweighing}]{\mathcal{G}_{a\times a}(0,\sigma)} \bar{\mathbb{H}} \in \mathbb{R}^{N\times 17\times a\times a}, \quad (1)$$

where to encode the Gaussian prior of keypoint heatmaps, the resulting pose-guided heatmaps $\mathbb{H}$ is reweighed by a Gaussian kernel $\mathcal{G}_{a\times a}(0, \sigma = \frac{a-1}{2\times 3})$ (e.g., $\sigma = 16$ when $a = 97$) in an element-wise way. By doing so, it means that the enlarged mesh follows the 3σ principle.

Then, we apply the learned keypoint 11×11 kernels $K_{n,i}$'s to convolve the reweighed $a \times a$ heatmap $\bar{\mathbb{H}}_{n,i}$ in a pose instance-sensitive and keypoint-specific way, leading to **LO**cal-**GlO**bal **C**ontextual **A**daptation,

$$\bar{\mathbb{H}} \xrightarrow[\text{LOGO-CA}]{K_{N\times 17\times 11\times 11}} \tilde{\mathbb{H}} \in \mathbb{R}^{N\times 17\times a\times a}, \quad (2)$$

which represents the refined heatmaps for the 17 human pose keypoints.

The Pose Estimation Output. With the local-global contextually adpated heatmaps $\tilde{\mathbb{H}}_{N\times 17\times a\times a}$, we maintain the top-2 locations for each keypoint within the $a \times a$ heatmap, and then utilize a convex average of the top-2 locations as the final predicted offset vectors (i.e. $(\Delta x'_i, \Delta y'_i)$'s in Fig. 3), and of their confidence scores as the prediction score, with a predefined weight λ for the top-1 location (0.75 in our experiments). Together with the predicted keypoints centers $\mathcal{C}_{N\times 3}$, the final prediction score for each keypoint is the product between the convex average confidence score and the center confidence score. We keep the keypoints whose final scores are greater than 0. We have,

$$\{\mathcal{C}_{N\times 3}, \tilde{\mathbb{H}}\} \xrightarrow[\text{Score thresholding}]{\text{Output}} \{\hat{L}^n_I; n = 1, \cdots N'\}, \quad (3)$$

where N' is the number of the final predicted pose instances in an image I.

3.2. Loss Functions in Training

In the fully end-to-end training, we need to define loss functions for the global heatmap $\mathcal{H}$, the refined local heatmap $\tilde{\mathbb{H}}$, the offset field $\mathcal{O}$, and the keypoint kernels.

The Heatmap Loss. The widely adopted mean squared error (MSE) loss is used. Denoted by $\mathcal{H}^{GT} \in \mathbb{R}^{18\times h\times w}$ the ground truth heatmaps in which each keypoint (including the center) is modeled by a 2-D Gaussian with dataset-provided mean and variance. Let $\mathbf{p} = (i, \mathbf{x})$ be the index of the domain D of dimensions $18 \times h \times w$. For the predicted heatmaps $\mathcal{H} \in \mathbb{R}^{18\times h\times w}$, the MSE loss is defined by,

$$\mathcal{L}_{\mathcal{H}} = 1/|D| \cdot \sum_{\mathbf{p}\in D} \|w(\mathbf{x})(\mathcal{H}(\mathbf{p}) - \hat{\mathcal{H}}(\mathbf{p}))\|^2_2, \quad (4)$$

where $w(\mathbf{x})$ represents the weight for the foreground and the background pixels. The foreground mask is provided by the dataset annotation. In our experiment, we set $w(\mathbf{x}) = 1/0.1$ for a foreground / background pixel respectively.

In defining the loss function $\mathcal{L}_{\tilde{\mathbb{H}}}$ for the refined local heatmap $\tilde{\mathbb{H}}$ (Eqn. 2), the ground-truth heatmap $\tilde{\mathbb{H}}^{GT}$ is generated on-the-fly based on the mesh $\mathcal{M}^G$ (Eqn. 1) and the ground-truth keypoints using a Gaussian model with mean being the displacement between the current predicted keypoints and the ground-truth ones, and variance σ (i.e., the standard deviation of the reweighing Gaussion prior model in Eqn. 1).

The Offset Field Loss. The widely adopted SmoothL1 loss [27] is used. Let $\mathcal{O}^{GT} \in \mathbb{R}^{34\times h\times w}$ be the ground-truth offset field, and $\mathcal{C}^{GT}$ be the non-empty set of ground-truth keypoints centers. For the predicted offset field $\mathcal{O}$, we have,

$$\mathcal{L}_{\mathcal{O}}(\mathbf{p}) = \mathcal{A}(\mathbf{p})\ell_1^{\text{smooth}}\left(\mathcal{O}(\cdot, \mathbf{p}), \mathcal{O}^{GT}(\cdot, \mathbf{p}); \beta\right), \quad (5)$$

for each foreground pixel $\mathbf{p} \in \mathcal{C}^{GT}$, and

$$\mathcal{L}_{\mathcal{O}} = \frac{1}{|\mathcal{C}^{GT}|} \cdot \sum_{\mathbf{p}\in\mathcal{C}^{GT}} \mathcal{L}_{\mathcal{O}}(\mathbf{p}) \quad (6)$$

where $\mathcal{A}(\mathbf{p})$ is the area of the person centered at the pixel $\mathbf{p}$, and β the cutting-off threshold (e.g., $\frac{1}{9}$ in our experiments).

The OKS Loss for the Keyoint Kernels. Consider a single predicted pose instance, learning the keypoint kernels, $K_{17\times 11\times 11}$ is the key to facilitate the local-global contextual adaptation. To that end, the figure of merits of the KEMs, $\mathcal{M}_{17\times 11\times 11\times 2}$ needs to directly reflect the task loss, i.e., the OKS loss. With respect to the N^{GT} ground-truth pose instances in an image, we can compute the similarity score per keypoint candidate in the KEMs, and obtain the score tensor $S_{17\times 11\times 11\times N^{GT}}$. The score tensor is further clamped with a threshold 0.5, i.e., $S_{17\times 11\times 11\times N^{GT}} = \max(S_{17\times 11\times 11\times N^{GT}}, 0.5)$. A mean reduction is applied to the first three dimensions of the clamped score tensor to compute the matching score for each of the N^{GT} pose instance. Then, the best ground-truth pose instance indexed by n^* is selected in terms of the matching score, and its matching score is denoted by s_{n^*}. Based on the selected ground-truth pose instance, we compute the per-keypoint similarity score for the predicted pose instance at hand, denoted by s_k ($k \in [1, 17]$). Then, the loss function fo the keypoint kernels are defined by,

$$\mathcal{L}_K = s_{n^*} \cdot \sum_{k,i,j} s_k \cdot |K_{k,i,j} - S_{k,i,j,n^*}|^2. \quad (7)$$

The Total Loss is then defined by $\mathcal{L} = \mathcal{L}_{\mathcal{H}} + \mathcal{L}_{\tilde{\mathbb{H}}} + \lambda\cdot(\mathcal{L}_{\mathcal{O}} + \mathcal{L}_K)$, where the trade-off parameter λ is used to balance the different loss items ($\lambda = 0.01$ in our experiments).

4. Experiments

In this section, we present detailed experimental results and analyses of the proposed LOGO-CAP.

Training and Testing Settings. We train two LOGO-CAP networks with the ImageNet pretrained HRNet-W32 and HRNet-W48 [31] as the feature backbone respectively on the COCO-train-2017 dataset [18]. Common training and testing specifications are used in experiments, which

Table 1. Evaluation results on the `COCO-val-2017` and `COCO-testdev-2017` dataset. For HGG [13] and SimplePose [15], the multi-scale inference† is applied on the testdev-2017 dataset. For DEKR [11] that uses an rescoring network to get the final predictions, we report both the performance with and without rescoring (which is the fair baseline for our LOGO-CAP). The numbers of SPM [22] and HGG [13] are extracted from their papers.

	Method	Backbone	COCO-val-2017					COCO-testdev-2017				
			AP [%]	AP^{50} [%]	AP^{75} [%]	AP^{M} [%]	AP^{L} [%]	AP [%]	AP^{50} [%]	AP^{75} [%]	AP^{M} [%]	AP^{L} [%]
Top-Down	HRNet [31]	HRNet-W48	76.3	90.8	82.9	72.3	83.4	75.5	92.5	83.3	71.9	81.5
	Lite-HRNet [38]	Lite-HRNet-30	70.4	88.7	77.7	76.2	92.8	69.7	90.7	77.5	66.9	75.0
	Mask-RCNN [12]	ResNet-50-FPN	64.2	86.6	69.7	58.7	73.0	63.1	87.3	68.7	57.8	71.4
Grouping	OpenPose [40]	VGG-19	61.0	84.9	67.5	56.3	69.3	61.8	84.9	67.5	57.1	68.2
	PifPaf [14]	ResNet-152	67.4	86.9	73.8	63.1	74.1	66.7	87.8	73.6	62.4	72.9
	PersonLab [23]	ResNet-152	66.5	86.2	71.9	62.3	73.2	66.5	88.0	72.6	62.4	72.3
	AE [5, 20]	HrHRNet-W32	67.1	86.2	73.0	61.5	76.1	66.4	87.5	72.8	61.2	74.2
		HrHRNet-W48	69.9	87.2	76.1	65.4	76.4	68.4	88.2	75.1	64.4	74.2
	HGG [13]	Hourglass	60.4	83.0	66.2	–	–	67.6†	85.1†	73.7†	62.7†	74.6†
	SimplePose [15]	IMHN	66.1	85.9	71.6	59.8	76.2	68.5†	86.7†	74.9†	66.4†	71.9†
Direct	SPM [22]	Hourglass	–	–	–	–	–	66.9	88.5	72.9	62.6	0.731
	CenterNet [40]	Hourglass	64.0	85.6	70.2	59.4	72.1	63.0	86.8	69.6	58.9	70.4
	DEKR [11]	HRNet-W32	68.0	86.7	74.5	62.1	77.7	67.3	87.9	74.1	61.5	76.1
	(w. Rescoring)	HRNet-W48	71.0	88.3	77.4	66.7	78.5	70.0	89.4	77.3	65.7	76.9
	DEKR [11]	HRNet-W32	67.2	86.3	73.8	61.7	77.1	66.6	87.6	73.5	61.2	75.6
	(w.o. Rescoring)	HRNet-W48	70.3	87.9	76.8	66.3	78.0	69.3	89.1	76.7	65.3	76.4
	LOGO-CAP (Ours)	HRNet-W32	69.6	87.5	75.9	64.1	78.0	68.2	88.7	74.9	62.8	76.0
		HRNet-W48	**72.2**	**88.9**	**78.9**	**68.1**	**78.9**	**70.8**	**89.7**	**77.8**	**66.7**	**77.0**

are provided in the supplementary material.

4.1. Datasets and Evaluation Metrics

Two datasets are used: **The COCO dataset [18]** is the most popular testbed for human pose estimation. It consists of 65k, 5k and 20k images with human pose well-annotated in the training, validation and testing datasets respectively. In all experiments, the proposed LOGO-CAP is trained using the 65k training images. **The OCHuman dataset** [39] is one popular *testing-only* dataset for evaluating human pose estimation under the occlusion scenarios. It consists of a total number of 4713 images with 8110 detailed annotated human pose instances using the COCO keypoint configuration. All the annotated 8110 human pose instances have occlusions with the maxIOU$\geq$ 0.5. Furthermore, 32% instances are more challenging with the maxIOU$\geq$ 0.75.

4.2. Results on the COCO Dataset

The proposed LOGO-CAP is compared with prior arts including the bottom-up approaches of OpenPose [2], PifPaf [14], PersonLab [23], AE [20] and DEKR [11], as well as the top-down approaches of Mask-RCNN [12], HRNet [31], and the light-weight LiteHRNet [38]. As reported in Table 1, the proposed LOGO-CAP outperforms all the bottom-up approaches and the efficiency-toward top-down approaches (*i.e.*, Lite-HRNet [38] and Mask-RCNN [12]) on both `validation and test-dev` datasets.

Fig. 4 shows some qualitative examples of human pose estimation by the proposed LOGO-CAP.

In comparisons to the best-performing grouping approach, AE [20] with a larger backbone HrHRNet-W48 [5], our LOGO-CAP obtains competitive performance with a smaller HRNet-32 backbone, and improves the AP score with HRNet-W48 backbone on the validation and test-dev datasets by `2.3` and `2.5` points, respectively. For the fully differentiable grouping approach HGG [13], our LOGO-CAP achieves better performance by a significantly large margin, more than `9.2` points on the validation set under the single-scale testing. Although the performance of HGG is improved by the multi-scale testing on the test-dev set, the performance of our LOGO-CAP is still significantly better without using the multi-scale testing scheme.

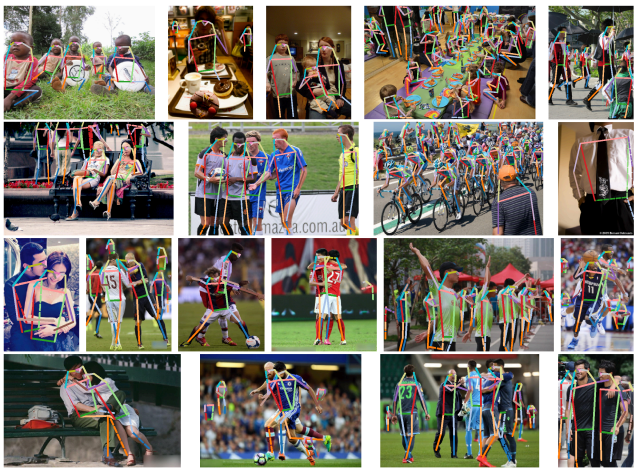

Figure 4. Qualitative results of our LOGO-CAP (HRNet-W32). All images were picked thematically without considering our algorithms performance. The first two rows display our approach on the COCO-val-2017 dataset and the last two ones show our results on the OCHuman test dataset.

In comparisons to the direct regression based approaches, our LOGO-CAP obtains the *best results* without incurring either the matching scheme used in CenterNet [40] or the additional rescoring network used in DEKR [11]. When we disable the rescoring network for DEKR [11] for fair comparisons, our LOGO-CAP significantly improves the AP on the validation and testdev datasets by `2.4` points and `1.6` points respectively when

Table 2. Results on the OCHuman datasets [39].

	Methods	Backbone	Val. AP [%]	Test AP [%]
Top-down	RMPE [8]	Hourglass	38.8	30.7
	SBL [36]	ResNet-50	37.8	30.4
	SBL [36]	ResNet-152	41.0	33.3
Bottom-up	AE [20]	Hourglass	32.1	29.5
	HGG [20]	Hourglass	35.6	34.8
	DEKR [11]	HRNet-W32	37.9	36.5
		HRNet-W48	38.8	38.2
	LOGO-CAP (Ours)	HRNet-W32	39.0	38.1
		HRNet-W48	**41.2**	**40.4**

HRNet-W32 is used as backbone. The larger backbone is beneficial for both DEKR and our method, which further improves the AP score of our LOGO-CAP to `72.2` and `70.8` on the validation and test-dev dataset respectively, outperforming DEKR by `1.9` and `1.5` respectively.

In comparisons to the top-down approaches, our LOGO-CAP outperforms the end-to-end Mask-RCNN [12] by 7.7 AP on the `test-dev-2017` dataset. For the efficiency-toward Lite-HRNet, we improves the AP on the test-dev split from 69.7 to 70.8. Although the performance of heavy top-down HRNet [31] is better than our LOGO-CAP, it should be noted that the headnet of our method is only an one-layer convolution network on the heatmap space, instead of leveraging a very deep model in the feature space of the cropped large-size images.

4.3. Results on the OCHuman Dataset

Table 2 shows that our LOGO-CAP achieves the best AP performance on both the validation and testing datasets by significant margins of `2.4` and `2.2` points in comparing with the bottom-up approaches. For the top-down approaches, although they obtain strong AP scores on the validation split, there exists a large performance gap between the validation and testing sets. In comparisons to DEKR [11] (with the rescoring network), our LOGO-CAP improves the performance from 37.9 to 39.0 and from 36.5 to 38.1 on the validation and testing splits with the same backbone HRNet-W32, respectively. The similar improvement is observed when the HRNet-W48 backbone is used, outperforming both bottom-up and top-down approaches.

4.4. Inference Speed

In comparing the inference speed, we test all the models on a single TITAN RTX GPU for its popularity in practice. The average inference speed, FPS (frames per second), over the 5000 images in COCO-val-2017 is used for the comparison. For DEKR [11], we re-implement their inference code with better speed obtained for fair comparisons at the algorithm level. For methods that have post-processing schema on CPU, only one thread is used. As shown in Table 3, our LOGO-CAP runs significantly faster than PifPaf [14] and AE [20]. The CenterNet [40] runs slower than DEKR and our LOCO-CAP as it requires a post-processing scheme to match the predicted offsets to the keypoints obtained from heatmaps. Comparing with DEKR, the speed improvement of our LOGO-CAP is from the lightweight design of head modules since the same backbones are used. For the comparisons in Table 2, we run the models with different resolutions of testing images.

Table 3. The single image inference speed comparison for bottom-up human pose estimation approaches.

Method	AP [%]	Backbone	Time ↓ [ms]	FPS ↑
PifPaf [14]	67.4	ResNet-152	213	4.68
AE [5, 20]	67.1	HrHRNet-W32	560	1.78
CenterNet [40]	64.0	Hourglass	147	6.80
DEKR [11]	68.0	HRNet-W32	63	15.8
DEKR [11]	71.0	HRNet-W48	139	7.21
LOGO-CAP	69.6	HRNet-W32	48	**20.7**
LOGO-CAP	72.2	HRNet-W48	112	8.95

Table 4. The breakdown of inference time of the proposed LOGO-CAP method. For each model, we seperately report the averaged inference time across 5000 images, the averaged inference time in the images that detect only one person, the avaraged inference time in the images that have 30 persons.

LOGO-CAP	#Persons	Backbone	Local KEMs	Local KAMs	Global KAMs
W32	-	38.6 ms	3.05 ms	2.49 ms	2.85 ms
	1		2.39 ms	1.14 ms	1.12 ms
	30		3.69 ms	3.49 ms	5.87 ms
W48	-	99.9 ms	4.18 ms	3.00 ms	3.34 ms
	1		3.19 ms	1.10 ms	1.07 ms
	30		2.97 ms	3.59 ms	5.97 ms

Table 5. Ablation studies on the three components of the LOGO-CAP: the OKS loss, the Gaussian reweighing method for heatmaps and the Attentive Normalization.

	OKS Loss	Reweigh	AttNorm	AP	AP^{50}	AP^{75}	AP^{M}	AP^{L}
baseline	-	-	-	60.0	84.4	66.4	54.0	71.1
(a)	✓	-	-	66.1	86.7	72.7	60.0	75.6
(b)	-	✓	-	67.6	87.0	74.3	62.1	76.7
(c)	✓	✓	-	69.0	87.0	75.2	63.4	77.5
(d)	✓	-	✓	65.8	86.8	72.3	59.3	75.4
(e)	-	✓	✓	67.5	86.6	74.1	62.2	76.7
(f)	✓	✓	✓	69.6	87.5	75.9	64.1	78.0

Furthermore, we analyze the inference time for different number of persons in the input image. As shown in Table 4, the main bottleneck of the inference time is the backbone. For the Local-Global Contextual Adaptation module, it only takes about 10 ms on average.

4.5. Ablation Studies on COCO Validation

In this section, we run a series of experiments to verify the effectiveness of the design for our proposed Local-Global Contextual Adaptation module. We use HRNet-W32 as our backbone for all ablation studies.

Designs for Contextual Adaptation. We train 6 models to study the effectiveness in Table 5 for the used components: (1) OKS Loss for learning local KAMs, (2) the Gaussian Reweighing scheme and (3) the Attentive Norm for convolutional message passing. Compared with the center-offset baseline, the model (a) trained with OKS loss for the local KAMs estimation obtains a large improvement of AP by 6.1 points, justifying that the potential of using local KAMs for better pose estimation. For (b), we set the factor of OKS Loss to 0 and train the model by using the end-to-end reweighing scheme. In this setting, the local KAMs are

Table 6. Ablation study of the different size of the local KEMs.

size of local KEM	Message Passing	AP	AP50	AP75	APM	APL	FPS
7×7	3+1	68.4	86.6	74.9	63.4	76.6	21.8
11×11	1	68.8	86.9	74.9	63.1	77.5	22.5
11×11	3+1	69.6	87.5	75.9	64.1	78.0	20.7
15×15	3+1	69.3	87.1	75.2	63.2	78.3	16.5
19×19	3+1	69.0	87.1	75.2	62.8	78.2	13.2

Table 7. Ablation study of using different priors in LOGO-CAP.

Type of the Prior	AP	AP50	AP75	APM	APL
KEMs only	59.4	80.8	62.8	50.9	71.6
KAMs only	65.7	86.0	72.3	60.6	74.0
LOGO-CAP (KEMs + KAMs)	69.6	87.5	75.9	64.1	78.0

learned only under the supervision of the contextual adaptation. Similar to model (a), a large improvement is also obtained compared against the baseline. Then, we train the model (c) that uses both OKS loss and the reweighing scheme while replacing the Attentive Norm [17] to BatchNorm in the convolutional message passing module. It is shown that the OKS loss and the reweighing scheme collaborate very well with further improvement obtained. For the effectiveness of Attentive Norm, it is shown in Table 5 (c-f) that its feature recalibration mechanism requires different information sources in human pose estimation. By enabling all the components, our LOGO-CAP-W32 finally obtains an AP of 69.6 on the COCO-val-2017 dataset.

Size of the Local KEMs and the Design of Convolutional Message Passing. We perform an ablation study with the results shown in Table 6, which confirms that the kernel size of 11×11 obtains the best performance. One possible explanation is that smaller kernel sizes fail to compensate the uncertainty of the initial keypoint estimation results, while larger kernel sizes may introduce more nuisances factors that affect the performance, such as the "collision" between different local KEMs of different keypoints from either the same person or adjacent different persons. Third, for different designs of the CMP module, we perform the ablation study for the different implementations. In detail, we replace the original architecture that is consist of 3 Conv+Norm+ReLU layers and 1 output Conv layer (denoted by `3+1` in Table 6) to a simpler architecture that only use 1 output Conv layer (denoted by `1`) for dimension reduction. It is shown that the `3+1` architecture performs better than the one only using the output Conv layer.

Different Type of the Priors for Contextual Adaptation. As our contextual adaptation takes both the global KEMs and KAMs as priors for final pose predictions, we quantitatively compare the possible designs for the contextual adaptation. In Table 7, we compare the performance on the COCO-val-2017 dataset by using either global KEMs or the learned global KAMs to the one that use two sources for prediction. On one hand, since the global KEMs are actually the standard Gaussian around each initial keypoint, it cannot provide more information for refinement. On another hand, when we enforce the use of local KAMs for adaptation with only global KEMs, the uncertainty from local KEMs will affect the results. That is the reason why only using global KEMs is worse than using global KAMs. When using both global KEMs and KAMs, our approach obtains the best performance.

4.6. Limitations and Potentials

One main limitation of the proposed method is that it has not close the gap between the empirical upper bound and the initial center-offset prediction. A traditional failure mode is observed when estimating the poses with heavy self-occlusion (e.g., lying down or squatting persons).

Consider the generic applicability of the center-offset formulation to many computer vision tasks as demonstrated in [40], we hypothesize that the proposed LOGO-CAP has a great potential to remedy the lack of sufficient accuracy using the vanilla center-offset method in those tasks. We also notice that the minimally-simple design in learning the contextual adaptation for refinement can be relaxed for different accuracy-speed trade-offs in practice. For example, for the convolutional message passing module, an alternative method could be the Transformer model [35], which potentially will further improve the performance at the expense of inference speed. We leave these for future work.

5. Conclusion

This paper presents a method of learning LOcal-GlObal Contextual Adaptation for Pose estimation in a bottom-up fashion, dubbed as LOGO-CAP. The proposed LOGO-CAP is built on the conceptually simple center-offset paradigm. The key idea of our LOG-CAP is to lift the center-offset predicted keypoints to keypoint expansion maps (KEMs) to address the inaccuracy of the initial keypoints. Two types of KEMs are introduced: Local KEMs are used to learn keypoint attraction maps (KAMs) via a convolutional message passing module that accounts for the structural information of human pose. Global KEMs are used to learn local-global contextual adaptation which convolves global KEMs using the KAMs as kernels. The refined global KEMs are used in computing the final human pose estimation. The proposed LOGO-CAP obtains state-of-the-art performance in COCO val-2017 and test-dev 2017 datasets for bottom-up human pose estimation. It also achieves state-of-the-art transferability performance in the OCHuman dataset with the COCO trained models.

Acknowledgments. This work was supported by National Nature Science Foundation of China under grant 62101390, 61922065, 41820104006 and 61871299, China National Postdoctoral Program for Innovative Talents under grant BX20200248. T. Wu was supported by NSF IIS-1909644, NSF CMMI-2024688, NSF IUSE-2013451 and DHHS-ACL Grant 90IFDV0017-01-00. We also thank Linxi Huan, Liang Dong, Fudong Wang and the anonymous reviewers for their helpful discussions and comments.

References

[1] Mykhaylo Andriluka, Stefan Roth, and Bernt Schiele. Pictorial structures revisited: People detection and articulated pose estimation. In *2009 IEEE conference on computer vision and pattern recognition*, pages 1014–1021. IEEE, 2009. 3

[2] Zhe Cao, Gines Hidalgo Martinez, Tomas Simon, Shih-En Wei, and Yaser A. Sheikh. Openpose: Realtime multi-person 2d pose estimation using part affinity fields. *IEEE Trans. on Pattern Analysis and Machine Intelligence (PAMI)*, 2019. 2, 3, 6

[3] Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Realtime multi-person 2d pose estimation using part affinity fields. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1302–1310, 2017. 2, 3

[4] Yilun Chen, Zhicheng Wang, Yuxiang Peng, Zhiqiang Zhang, Gang Yu, and Jian Sun. Cascaded pyramid network for multi-person pose estimation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7103–7112, 2018. 3

[5] 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 *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5385–5394. IEEE, 2020. 2, 6, 7

[6] Kaiwen Duan, Song Bai, Lingxi Xie, Honggang Qi, Qingming Huang, and Qi Tian. Centernet: Keypoint triplets for object detection. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 6568–6577, 2019. 2, 3

[7] Haoshu Fang, Shuqin Xie, Yu-Wing Tai, and Cewu Lu. RMPE: regional multi-person pose estimation. In *IEEE International Conference on Computer Vision (ICCV)*, pages 2353–2362, 2017. 3

[8] Haoshu Fang, Shuqin Xie, Yu-Wing Tai, and Cewu Lu. RMPE: regional multi-person pose estimation. In *IEEE International Conference on Computer Vision (ICCV)*, pages 2353–2362, 2017. 7

[9] Pedro F Felzenszwalb and Daniel P Huttenlocher. Pictorial structures for object recognition. *International journal of computer vision*, 61(1):55–79, 2005. 2, 3

[10] Martin A Fischler and Robert A Elschlager. The representation and matching of pictorial structures. *IEEE Transactions on computers*, 100(1):67–92, 1973. 2, 3

[11] Zigang Geng, Ke Sun, Bin Xiao, Zhaoxiang Zhang, and Jingdong Wang. Bottom-up human pose estimation via disentangled keypoint regression. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 2, 3, 6, 7

[12] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross B. Girshick. Mask R-CNN. In *IEEE International Conference on Computer Vision (ICCV)*, pages 2980–2988, 2017. 3, 6, 7

[13] 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 *European Conference on Computer Vision (ECCV)*, volume 12352, pages 718–734, 2020. 3, 6

[14] Sven Kreiss, Lorenzo Bertoni, and Alexandre Alahi. Pifpaf: Composite fields for human pose estimation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11977–11986, 2019. 2, 3, 6, 7

[15] Jia Li, Wen Su, and Zengfu Wang. Simple pose: Rethinking and improving a bottom-up approach for multi-person pose estimation. In *AAAI Conference on Artificial Intelligence (AAAI)*, pages 11354–11361, 2020. 6

[16] Ke Li, Shijie Wang, Xiang Zhang, Yifan Xu, Weijian Xu, and Zhuowen Tu. Pose recognition with cascade transformers. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1944–1953, 2021. 3

[17] Xilai Li, Wei Sun, and Tianfu Wu. Attentive normalization. In Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm, editors, *European Conference on Computer Vision (ECCV)*, volume 12362, pages 70–87, 2020. 4, 8

[18] Tsung-Yi Lin, Michael Maire, Serge J. Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO: common objects in context. In David J. Fleet, Tomás Pajdla, Bernt Schiele, and Tinne Tuytelaars, editors, *European Conference on Computer Vision (ECCV)*, volume 8693, pages 740–755, 2014. 5, 6

[19] Ningning Ma, Xiangyu Zhang, Hai-Tao Zheng, and Jian Sun. Shufflenet V2: practical guidelines for efficient CNN architecture design. In Vittorio Ferrari, Martial Hebert, Cristian Sminchisescu, and Yair Weiss, editors, *Computer Vision - ECCV 2018 - 15th European Conference, Munich, Germany, September 8-14, 2018, Proceedings, Part XIV*, volume 11218 of *Lecture Notes in Computer Science*, pages 122–138. Springer, 2018. 3

[20] Alejandro Newell, Zhiao Huang, and Jia Deng. Associative embedding: End-to-end learning for joint detection and grouping. In *Advances in Neural Information Processing Systems 30 (NeurIPS)*, pages 2277–2287, 2017. 2, 3, 6, 7

[21] Alejandro Newell, Kaiyu Yang, and Jia Deng. Stacked hourglass networks for human pose estimation. In *European Conference on Computer Vision (ECCV)*, pages 483–499, 2016. 1, 3

[22] Xuecheng Nie, Jiashi Feng, Jianfeng Zhang, and Shuicheng Yan. Single-stage multi-person pose machines. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 6950–6959, 2019. 6

[23] 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 *European Conference on Computer Vision (ECCV)*, pages 282–299, 2018. 3, 6

[24] George Papandreou, Tyler Zhu, Nori Kanazawa, Alexander Toshev, Jonathan Tompson, Chris Bregler, and Kevin Murphy. Towards accurate multi-person pose estimation in the wild. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3711–3719, 2017. 3

[25] Leonid Pishchulin, Mykhaylo Andriluka, Peter Gehler, and Bernt Schiele. Poselet conditioned pictorial structures. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 588–595, 2013. 3

[26] Deva Ramanan, David A Forsyth, and Andrew Zisserman. Strike a pose: Tracking people by finding stylized poses. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, volume 1, pages 271–278. IEEE, 2005. 3

[27] Shaoqing Ren, Kaiming He, Ross B. Girshick, and Jian Sun. Faster R-CNN: towards real-time object detection with region proposal networks. *IEEE Trans. Pattern Anal. Mach. Intell.*, 39(6):1137–1149, 2017. 3, 5

[28] Brandon Rothrock, Seyoung Park, and Song-Chun Zhu. Integrating grammar and segmentation for human pose estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3214–3221, 2013. 3

[29] Mark Sandler, Andrew G. Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. Mobilenetv2: Inverted residuals and linear bottlenecks. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4510–4520, 2018. 3

[30] Ke Sun, Zigang Geng, Depu Meng, Bin Xiao, Dong Liu, Zhaoxiang Zhang, and Jingdong Wang. Bottom-up human pose estimation by ranking heatmap-guided adaptive keypoint estimates. *CoRR*, abs/2006.15480, 2020. 2, 3

[31] Ke Sun, Bin Xiao, Dong Liu, and Jingdong Wang. Deep high-resolution representation learning for human pose estimation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5693–5703, 2019. 1, 2, 3, 5, 6, 7

[32] Zhi Tian, Hao Chen, and Chunhua Shen. Directpose: Direct end-to-end multi-person pose estimation. *CoRR*, abs/1911.07451, 2019. 3

[33] Zhi Tian, Chunhua Shen, Hao Chen, and Tong He. FCOS: fully convolutional one-stage object detection. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 9626–9635, 2019. 3

[34] Ali Varamesh and Tinne Tuytelaars. Mixture dense regression for object detection and human pose estimation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13083–13092, 2020. 3

[35] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. *arXiv preprint arXiv:1706.03762*, 2017. 8

[36] Bin Xiao, Haiping Wu, and Yichen Wei. Simple Baselines for Human Pose Estimation and Tracking. *Computer Vision and Pattern Recognition*, 2018. 1, 3, 7

[37] Yi Yang and Deva Ramanan. Articulated human detection with flexible mixtures of parts. *IEEE transactions on pattern analysis and machine intelligence*, 35(12):2878–2890, 2012. 3

[38] Changqian Yu, Bin Xiao, Changxin Gao, Lu Yuan, Lei Zhang, Nong Sang, and Jingdong Wang. Lite-hrnet: A lightweight high-resolution network. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10440–10450, 2021. 3, 6

[39] Song-Hai Zhang, Ruilong Li, Xin Dong, Paul L. Rosin, Zixi Cai, Xi Han, Dingcheng Yang, Haozhi Huang, and Shi-Min Hu. Pose2seg: Detection free human instance segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 889–898, 2019. 6, 7

[40] Xingyi Zhou, Dequan Wang, and Philipp Krähenbühl. Objects as points. *CoRR*, abs/1904.07850, 2019. 2, 3, 6, 7, 8