This CVPR paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

Contextual Instance Decoupling for Robust Multi-Person Pose Estimation

Dongkai Wang[1] Shiliang Zhang[1,2]
[1]School of Computer Science, Peking University
[2]Peng Cheng Laboratory
{dongkai.wang, slzhang.jdl}@pku.edu.cn

Abstract

Crowded scenes make it challenging to differentiate persons and locate their pose keypoints. This paper proposes the Contextual Instance Decoupling (CID), which presents a new pipeline for multi-person pose estimation. Instead of relying on person bounding boxes to spatially differentiate persons, CID decouples persons in an image into multiple instance-aware feature maps. Each of those feature maps is hence adopted to infer keypoints for a specific person. Compared with bounding box detection, CID is differentiable and robust to detection errors. Decoupling persons into different feature maps allows to isolate distractions from other persons, and explore context cues at scales larger than the bounding box size. Experiments show that CID outperforms previous multi-person pose estimation pipelines on crowded scenes pose estimation benchmarks in both accuracy and efficiency. For instance, it achieves 71.3% AP on CrowdPose, outperforming the recent single-stage DEKR by 5.6%, the bottom-up CenterAttention by 3.7%, and the top-down JC-SPPE by 5.3%. This advantage sustains on the commonly used COCO benchmark[†].

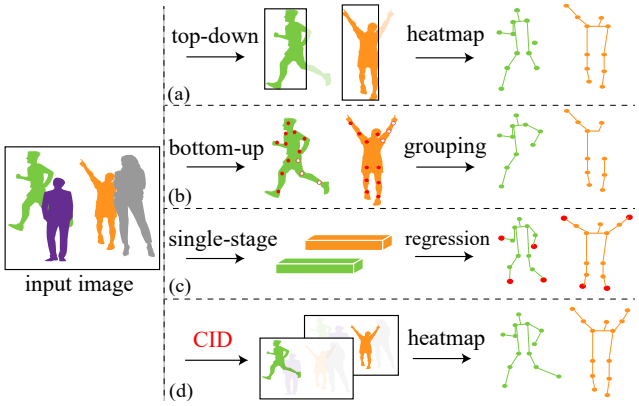

Figure 1. Illustration of different MPPE pipelines for differentiating persons. (a) Top-down methods use bounding box to crop person; (b) Bottom-up methods first detect all keypoints then group them into different persons; (c) Single-stage methods directly regress keypoint coordinates based on sampled feature vector. (d) The proposed Contextual Instance Decoupling (CID) first generates instance-aware feature maps, then infers heatmaps from each person. It has potential to enjoy better robustness to detection errors in (a), keypoint localization errors in (b), and alleviate the difficulty of long distance regression in (c).

1. Introduction

Multi-Person Pose Estimation (MPPE) detects all persons in an image, and locates keypoints for each of them. As an important step for human activity understanding, human-object interaction, human parsing, *etc.*, MPPE has attracted increasing attention. Current MPPE methods can be summarized into three categories according to their followed pipelines, *i.e.* i) top-down methods [7, 21, 26, 33], which detect person bounding boxes and perform pose estimation for each bounding box, ii) bottom-up methods [2, 9, 11, 17, 20, 23] that first detect body keypoints, then group them into corresponding persons, and iii) single-stage regression methods, which regress pose keypoints coordinates [6, 19, 28, 30, 35] based on person features. Fig. 1 (a), (b), and (c) illustrate those three pipelines, respectively.

[†]Code is available at https://github.com/kennethwdk/CID

The above pipelines show different properties and defects for MPPE in crowded scenes. Regression paradigm enjoys better efficiency, but suffers from the difficulty of long distance regression. Top-down and bottom-up pipelines rely on heatmaps to locate keypoints, *e.g.*, a likelihood heatmap can be generated for each keypoint, where the keypoint is located with argmax or soft-argmax [27] operations. Heatmap based methods need extra computations to differentiate persons. As shown in Fig. 1 (a) and (b), top-down methods adopt bounding box cropping to differentiate persons, while bottom-up methods utilize keypoint grouping. Bounding box cropping, *e.g.*, RoIAlign [7] is sensitive to detection errors, and cannot explore contexts outside the box. Keypoint grouping is complex and not capable of recover keypoint location errors because it discards contextual cues. Sec. 2 presents a more detailed review.

This work aims at studying a new MPPE pipeline to ef-

fectively separate instances while preserving rich context cues to estimate keypoint locations. As shown in Fig. 1 (d), given an image with multiple persons, we first decouple each person into a specific instance-aware feature map. Each instance-aware feature map contains all the necessary cues required for single-person keypoint localization. It is hence used to infer keypoints for a specific person. Compared with top-down and bottom-up piplelines, this new pipeline is robust to spatial detection errors, and is able to explore contextual cues at larger scale. Compared with the regression baseline, it encodes more spatial cues in feature maps, and alleviates the difficulty of long-range location regression.

This pipeline is implemented in our Contextual Instance Decoupling (CID) module. Given an image with multiple persons, CID first extracts the location and feature for each person via Instance Information Abstraction (IIA) module, which are hence used to decouple different persons in the Global Feature Decoupling (GFD) module. GFD computes spatial and channel attention with location and feature cues of each person to isolate the corresponding person. To improve the discriminative power of those instance features, we introduce a contrastive loss to ensure different persons present different features. The generated instance-aware feature maps are used to estimate the final heatmap for keypoint estimation.

We test CID on different multi-person pose estimation benchmarks, *i.e.*, COCO Keypoint [14], CrowdPose [13] and OCHuman [34]. Experiment results show that, our CID achieves superior performance compared with recent works following different pipelines. For instance, CID achieves 71.3% AP on CrowdPose with high efficiency, outperforming the recent single-stage DEKR [6] by 5.6%, the bottom-up CenterAttention [1] by 3.7%, and the top-down JC-SPPE [13] by 5.3%. It also shows competitive performance on the commonly used COCO benchmark, *e.g.*, achieves 68.9% AP, outperforming the DEKR [6] by 1.6%.

To the best of our knowledge, this is an original effort on contextual instance decoupling pipeline for MPPE. Compared with previous MPPE pipelines, CID shows better robustness to detection errors, and alleviates the complexity of keypoint grouping and long distance regression. Being able to encode more contextual and spatial cues at larger scales, it is also capable of isolating distractions from other persons. Those advantages make CID a more effective pipline for MPPE.

2. Related Work

Top-down methods first detect person bounding boxes by detectors like YOLO [25], then perform single-person pose estimation in the cropped region. Mask R-CNN [7] is a typical method that adds a keypoint detection branch on Faster R-CNN to use the RoIAlign feature. G-RMI [21] breaks top-down methods into two stages, and uses separate models for person detection and pose estimation, respectively. Most of top-down methods focus on designing better networks for locating keypoints. For instance, Hourglass [18], SimpleBaseline [33] and HRNet [26] have achieved superior performance with their proposed pose estimation networks.

Bottom-up methods first detect identity-free keypoints for all persons, then group keypoints into individual persons. Heatmap is widely adopted for keypoint detection and most bottom-up methods focus on keypoints grouping algorithms. The first line of grouping methods formulates grouping as integer linear programming, and DeepCut [23] is a representative work. The second line utilizes vector fields to encode keypoint relationships, and grouping is conducted by parsing those fields. OpenPose [2], PersonLab [20] and PifPaf [11] are representative methods. The third line clusters keypoints into poses. Associative Embedding [17] learns each keypoint with a tag embedding and conducts the grouping by clustering the tags. HGG [9] introduces a differentiable graph clustering to replace traditional offline clustering operation.

Single-stage methods tend to directly regress keypoint locations, thus differs with top-down and bottom-up methods, which require a two-stage computation. Single-stage regression also makes the whole pipeline end-to-end trainable. CenterNet [35] and DirectPose [28] are the early single-stage methods to directly estimate multi-person poses by regressing. Researchers have proposed several works to improve the performance of regression. SPM [19] proposes a structured pose representation to relieve the difficulty of long distance regression. DEKR [6] proposes the separating regression and adaptive convolution to improve the quality of regression, and have achieved comparable performance with two-stage methods.

Our CID differs with the above pipelines. Compared with top-down and bottom-up methods, it is end-to-end trainable, more robust to detection errors, and alleviates the challenge of keypoint grouping. It also avoids the difficulty of long distance regression faced by single-stage methods. Our experiments show that CID outperforms latest works of those pipelines in both efficiency and accuracy.

3. Method

3.1. Overview

Given an image $\mathcal{I}$ with multiple persons, the goal of multi-person pose estimation is to estimate locations of pose keypoints for each person and can be denoted as,

$$\{\mathcal{K}_j^{(i)}\}_{j=1,\dots,n}^{i=1,\dots,m} = \mathrm{MPPE}(\mathcal{I}),$$

where $\mathcal{K}_j^{(i)}$ denotes the j-th pose keypoint for the i-th person in image $\mathcal{I}$, m and n represent the number of persons in

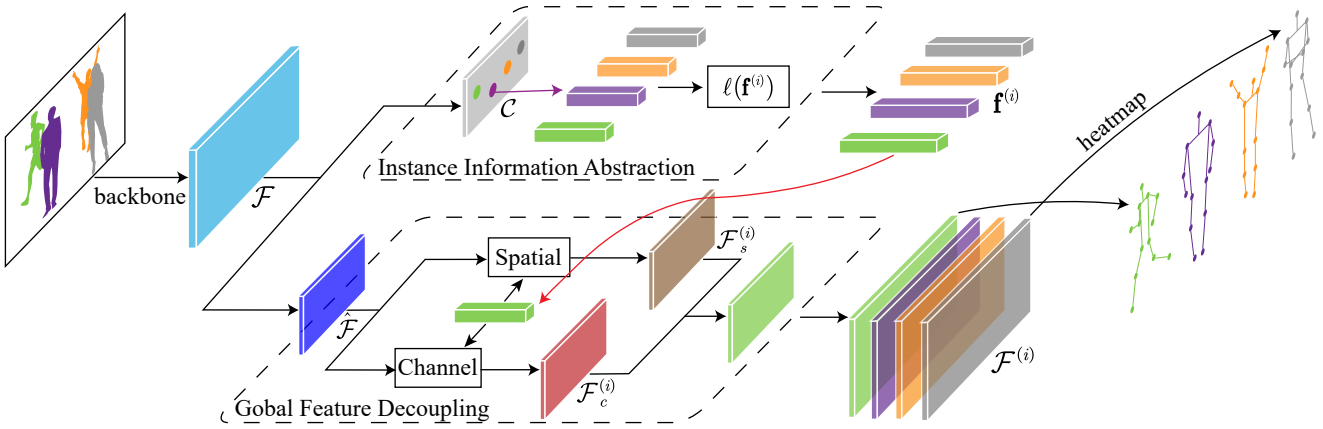

Figure 2. The pipeline of the proposed Contextual Instance Decoupling (CID). CID uses a CNN to extract the feature map. Instance Information Abstraction (IIA) extracts the location and feature to represent each person. Global Feature Decoupling (GFD) modulates the original feature map to produce instance-aware feature maps, each of which is used to estimate heatmap and keypoints for a person, respectively.

$\mathcal{I}$ and the number of keypoints for each person, *e.g.*, $n = 17$ for COCO Keypoint [34] and $n = 14$ for CrowdPose [13].

We adopt the heatmap to locate keypoints. The input to the heatmap module is a feature map extracted by a Convolutional Neural Network (CNN) and the output is a n-channel heatmap, indicating the probability distribution map for each keypoint, *i.e.*,

$$\{\mathcal{H}_j\}_{j=1}^{n} = \mathrm{HM}(\mathcal{F}), \mathcal{F} = \Phi(\mathcal{I}), \tag{1}$$

where $\mathrm{HM}(\cdot)$ denotes the heatmap module, $\{\mathcal{H}_j\}_{j=1}^{n}$ is a n-channel heatmap, where $\mathcal{H}_j$ is the heatmap for j-th keypoint. $\Phi(\cdot)$ denotes the CNN backbone and $\mathcal{F}$ is the extracted global feature map for input image $\mathcal{I}$. Keypoints can be decoded by finding local maxima in each channel of $\mathcal{H}$, *e.g.*,

$$\{\mathcal{K}_j^{(i)}\}_{i=1}^{m} = \mathrm{rank}(\mathcal{H}_j, m), \tag{2}$$

where m denotes the number of keypoints we want to decode from $\mathcal{H}_j$. $m = 1$ if $\mathcal{F}$ contains only one person and $m > 1$ for multiple persons.

In MPPE, cues of multiple persons are mixed in $\mathcal{F}$. Extra efforts are required to differentiate those persons, *e.g.*, through spatially detecting bounding boxes or grouping keypoints. Different from those strategies, Contextual Instance Decoupling (CID) decouples the multi-person feature map $\mathcal{F}$ into a set of instance-aware feature maps $\{\mathcal{F}^{(i)}\}_{i=1}^{m}$, where each map $\mathcal{F}^{(i)}$ represents cues of a specific person and preserves contextual cues to infer his/her keypoints.

Previous studies like SE [8] and CBAM [32] reveal that attention mechanism can modulate the feature map, making its certain spatial location or channel emphasize specific parts of the image. We also use attention mechanism to recalibrate $\mathcal{F}$ to generate the $\{\mathcal{F}^{(i)}\}_{i=1}^{m}$. CID first identifies all persons in $\mathcal{F}$ and describes each person with his/her appearance and spatial location. This procedure is finished in the Instance Information Abstraction (IIA) module, which can be denoted as,

$$\{\mathbf{f}^{(i)}, \mathbf{l}^{(i)}\}_{i=1}^{m} = \mathrm{IIA}(\mathcal{F}), \tag{3}$$

where $\mathbf{l}^{(i)}$ denotes the location of the i-th person, $\mathbf{f}^{(i)}$ is the representative feature encoding his/her appearance.

The $\{\mathbf{f}^{(i)}, \mathbf{l}^{(i)}\}_{i=1}^{m}$ supervises the attention mechanism to decouple the original feature map $\mathcal{F}$ into m instance-aware feature maps via Global Feature Decoupling (GFD),

$$\{\mathcal{F}^{(i)}\}_{i=1}^{m} = \mathrm{GFD}(\mathcal{F}, \{\mathbf{f}^{(i)}, \mathbf{l}^{(i)}\}_{i=1}^{m}), \tag{4}$$

where $\mathcal{F}^{(i)}$ is the decoupled feature map for i-th person.

$\mathcal{F}^{(i)}$ can be feed to the heatmap module of Eq. (1) to get the keypoint heatmap $\{\mathcal{H}_j^{(i)}\}_{j=1}^{n}$ for the i-th person. His/her keypoints can be obtained via Eq. (2) by simply setting $m = 1$. Note that, $\mathcal{F}^{(i)}$ contains more contextual cues than a bounding box. It also isolates distractions from other persons. Those prosperities benefit the heatmap module design, *e.g.*, a lightweight module can be implemented with good performance because it handles single person. The lightweight heatmap module ensures the efficiency of CID.

We learn CID by training the backbone $\Phi(\cdot)$, $\mathrm{IIA}(\cdot)$, $\mathrm{GFD}(\cdot)$ modules in an end-to-end manner. The overall training objective is denoted by

$$\mathcal{L} = \mathcal{L}_{IIA} + \lambda \mathcal{L}_{GFD}, \tag{5}$$

where $\mathcal{L}_{IIA}$ supervises to learn discriminative instance features for person decoupling. $\mathcal{L}_{GFD}$ supervises the instance heatmap estimation and λ is a weight to balance two losses. Fig. 2 illustrates the pipeline of the proposed CID. Following parts proceed to present the details of IIA, GFD and loss computation.

3.2. Instance Information Abstraction

As shown in Fig. 2, given the input feature map $\mathcal{F} \in \mathbb{R}^{C\times H\times W}$, IIA locates each person and generates corresponding features. Previous regression methods generate keypoint coordinates based on features at the person center point [28, 35]. We follow this intuition and use the feature at center point to represent each person. IIA regards each center point as a keypoint and use heatmap to locate center points. IIA estimates the center point with a heatmap module similar to the one in Eq. (1), *i.e.*,

$$\mathcal{C} = \mathrm{HM}_{center}(\mathcal{F}), \tag{6}$$

where $\mathcal{C}$ denotes the center heatmap indicating the confidence that each pixel is the center of a person. Fig. 3 illustrates the estimated center heatmap.

The center heatmap $\mathcal{C}$ is hence input into Eq. (2) to find locations of center points. We use $\{\mathbf{l}^{(i)}\}_{i=1}^{m}$ to denote the locations of center points of m persons, where $\mathbf{l}^{(i)} = (x_i, y_i)$ denotes the center coordinates of the i-th person. Features at the center points on feature map $\mathcal{F}$ are regraded as representative features for those persons. For the i-th person, his/her representative feature can be computed as,

$$\mathbf{f}^{(i)} = \mathcal{F}(\mathbf{l}^{(i)}). \tag{7}$$

The computed $\mathbf{f}^{(i)}$ is hence used to identify and decouple the i-th person from other persons. It is expected to have strong discriminative power to effective differentiate visual similar persons. In other words, if two neighboring or overlapped persons share similar appearance, their features may be similar, which leads to failure cases in person decoupling. To boost the discriminative power of person features, we train the IIA with a contrastive loss to ensure the discriminative power of each $\mathbf{f}^{(i)}$.

Given a set of person features $\{\mathbf{f}^{(i)}\}_{i=1}^{m}$, we constrain the i-th person feature by minimizing the similarity of it and other features, which can be calculated by,

$$\ell(\mathbf{f}^{(i)}) = -\log \frac{\exp(\bar{\mathbf{f}}^{(i)} \cdot \bar{\mathbf{f}}^{(i)}/\tau)}{\sum_{j=1}^{m} \exp(\bar{\mathbf{f}}^{(i)} \cdot \bar{\mathbf{f}}^{(j)}/\tau)}, \tag{8}$$

where $\bar{\mathbf{f}}^{(i)}$ denotes the l_2 normalized feature for i-th person and τ is a temperature coefficient, which is set to 0.05 in all experiments. The effectiveness of the contrastive loss will be validated in experiments.

3.3. Global Feature Decoupling

GFD is designed to decouple persons cues from the original global feature map $\mathcal{F}$ based on instance features and locations $\{\mathbf{f}^{(i)}, \mathbf{l}^{(i)}\}_{i=1}^{m}$. It jointly considers spatial-wise and channel-wise decoupling, *i.e.*, decouples persons into different spatial locations and channels of the feature map. This is achieved by first computing the spatial recalibration and channel recalibration, then applying the fused recalibration.

Figure 3. Visualization of center heatmap $\mathcal{C}$. IIA can identify each person and estimate their corresponding center location.

Spatial recalibration tends to decouple persons into different spatial locations. To spatially emphasize the i-th person on a feature map, a straightforward way is to increase the weights of features on his/her foreground and degrade others. GFD generates a spatial mask to represent the foreground for each person, and computes the spatial recalibration for the i-th person as,

$$\mathcal{F}_s^{(i)} = \mathcal{M}^{(i)} \cdot \hat{\mathcal{F}}, \tag{9}$$

where $\mathcal{F}_s^{(i)}$ denotes the generated feature map for the i-th person, and $\mathcal{M}^{(i)}$ indicates the foreground mask. $\hat{\mathcal{F}} = \mathrm{Conv}(\mathcal{F})$ is the transformed global feature map to adjust the channel size for saving computation and memory.

To compute the mask $\mathcal{M}^{(i)}$, we consider the location of this person $\mathbf{l}^{(i)} = (x_i, y_i)$ and generate a relative coordinates map $\mathcal{O}^{(i)}$ following [29]. We also compute the inner product of instance feature $\mathbf{f}^{(i)}$ and features at each spatial location on $\hat{\mathcal{F}}$. This leads to a map $\mathcal{M}_{sim}^{(i)}$ indicating the pixel-level feature similarity. $\mathcal{O}^{(i)}$ and $\mathcal{M}_{sim}^{(i)}$ are concatenated and convolved by a convolution layer to produce the spatial mask, which can be denoted as,

$$\mathcal{M}^{(i)} = \mathrm{Sigmoid}(\mathrm{Conv}([\mathcal{O}^{(i)}; \mathcal{M}_{sim}^{(i)}])), \tag{10}$$

where $\mathcal{M}^{(i)}$ is applied in Eq. (9) to indicate the spatial location of foreground region of the i-th person.

Channel recalibration is computed to separate persons into different channels of the feature map. Previous work [8] shows that channel plays an important role to encode contexts and each channel can be regraded as a feature detector. We hence re-weight the original feature map on channel dimension with person features and generate conditioned feature maps. Specifically, given the feature map $\hat{\mathcal{F}}$ and a person feature $\mathbf{f}^{(i)}$, GFD computes the channel recalibration for the i-th person with,

$$\mathcal{F}_c^{(i)} = \hat{\mathcal{F}} \otimes \mathbf{f}^{(i)}, \tag{11}$$

where $\otimes$ denotes the element-wise manipulation, and $\mathcal{F}_c^{(i)}$ is the channel-recalibrated feature map for the i-th person. Eq. (11) uses $\mathbf{f}^{(i)}$ to weight different channels, hence produces different $\mathcal{F}_c^{(i)}$ for different persons.

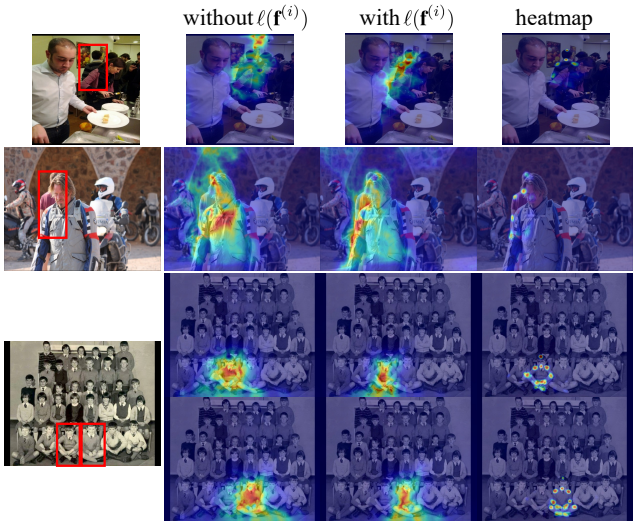

Figure 4. Visualization of instance-aware feature maps and keypoint heatmaps of persons highlighted by bounding boxes. $\ell(\mathbf{f}^{(i)})$ boosts the discriminative power of person feature, making CID more robust to occlusions and distractions from neighboring persons with similar appearance. With $\ell(\mathbf{f}^{(i)})$, the instance-aware feature map can better focus on each person foreground, and ensures the generation of reliable keypoint heatmaps.

Note that, applying element-wise manipulation to two dissimilar features leads a small vector. This enables Eq. (11) to keep cues of the i-th person based on $\mathbf{f}^{(i)}$, and discards cues of other persons. In other words, Eq. (11) highlights features showing similar channel-wise distribution with $\mathbf{f}^{(i)}$, and depresses others. Eq. (11) does not decouple a person into a specific channel of the feature map, but ensures different persons show different channel-wise distributions. Learning discriminative features in Eq. (8) further enforces the performance of channel recalibration.

Fused recalibration is computed based on $\mathcal{F}_s^{(i)}$ and $\mathcal{F}_c^{(i)}$ to produce the instance-aware feature map. The instance-aware feature map $\mathcal{F}^{(i)}$ for the i-th person can be computed as,

$$\mathcal{F}^{(i)} = \mathrm{ReLU}(\mathrm{Conv}([\mathcal{F}_c^{(i)}; \mathcal{F}_s^{(i)}])), \tag{12}$$

where $\mathcal{F}_s^{(i)}$ and $\mathcal{F}_c^{(i)}$ are fused to seek better discriminative power in decoupling persons. $\mathcal{F}^{(i)}$ is hence used to produce the heatmap and estimate keypoints for the i-th person.

Discussions: Compared with bottom-up methods relying on keypoint grouping, CID shows better performance and efficiency due to its end-to-end training property and robustness to detection errors. Recent single-stage methods [35] directly regress keypoint coordinates from a representative person feature $\mathbf{f}^{(i)}$. Such pipeline discards the spatial contexts in the original feature map. Differently, GFD generates keypoints from $\mathcal{F}^{(i)}$, which encodes more spatial contexts than the vector $\mathbf{f}^{(i)}$. Fig. 4 illustrates instance-aware feature maps. It is clear CID is robust to occlusions and distractions from neighboring persons.

GFD shares some similarity with several methods like SE [8] and CBAM [32] in attention computation. However, GFD differs with them in both motivation and implementation: 1) SE and CBAM aim to enhance the model capacity. GFD is more explainable and aims to decouple persons into instance-aware feature maps. 2) GFD leverages person feature and spatial location into attention computation, and boosts the efficiency in decoupling persons.

3.4. Loss Computation

To implement $\mathcal{L}_{IIA}$ and $\mathcal{L}_{GFD}$ for CID training, we follow previous works to generate ground truth heatmap. For a j-th keypoint with spatial coordinates (x_j, y_j), we compute its response on the ground truth heatmap $\mathcal{H}_j^*$ as,

$$\mathcal{H}_j^*(x,y) = \exp(-\frac{(x-x_j)^2 + (y-x_j)^2}{2\sigma_i^2}), \tag{13}$$

where σ_i indicates the person size-adaptive standard deviation in [12]. Eq. (13) is adopted to generate the multi-person groundtruth heatmap $\mathcal{H}^*$ and center map $\mathcal{C}^*$, and ground truth heatmap $\mathcal{H}^{(i)*}$ for the i-th person, respectively.

$\mathcal{L}_{IIA}$ is computed with multi-person groundtruth heatmap $\mathcal{H}^*$ and center map $\mathcal{C}^*$. It also combines the contrastive loss in Eq. (8), *i.e.*,

$$\mathcal{L}_{IIA} = \mathrm{FL}([\mathcal{H}; \mathcal{C}], [\mathcal{H}^*; \mathcal{C}^*]) + \frac{1}{m}\sum_{i=1}^{m} \ell(\mathbf{f}^{(i)}), \tag{14}$$

where $[;]$ denotes the feature concatenation, and $\mathrm{FL}(\cdot)$ computes the Focal Loss [12, 35], *i.e.*,

$$\mathrm{FL}(\mathcal{H}, \mathcal{H}^*) = \frac{-1}{N}\sum_{xy} \begin{cases} (1-\mathcal{H}_{x,y})^{\alpha}\log(\mathcal{H}_{x,y}) & \text{if } \mathcal{H}^*_{x,y}=1 \\ (1-\mathcal{H}^*_{x,y})^{\beta}(\mathcal{H}_{x,y})^{\alpha} & \\ \log(1-\mathcal{H}_{x,y}) & \text{otherwise} \end{cases} \tag{15}$$

where α and β are hyper-parameters and N is the number of points in $\mathcal{H}^*$ with value equal to 1. We adopt the default $\alpha = 2$ and $\beta = 4$ following [12, 35].

$\mathcal{L}_{GFD}$ measurs the differences between computed heatmap and ground truth of each person, *i.e.*,

$$\mathcal{L}_{GFD} = \frac{1}{m}\sum_{i=1}^{m} \mathrm{FL}(\mathcal{H}^{(i)}, \mathcal{H}^{(i)*}). \tag{16}$$

4. Experiments

4.1. Datasets and Evaluation Metric

We evaluate CID on three widely used multi-person pose estimation benchmarks, *i.e.*, COCO Keypoint [14], CrowdPose [13], and OCHuman [34].

COCO Keypoint [14] contains 64K images of 270K persons labeled with 17 keypoints. We use the `train` set containing 57K images, 150K persons for training. The `val`

set containing 5K images, 6.3K persons and `test-dev` set containing 20K images are used for evaluation.

CrowdPose [13] contains 20K images and 80K persons labeled with 14 keypoints. Following [3, 6], we use the `trainval` set (12K images, 43.4K persons) and for evaluation we use the `test` set (8K images, 29K persons).

OCHuman [34] is a benchmark to examine MPPE in more challenging scenarios. It consists of 4,731 images in total, including 2,500 images for `val` set and 2,231 images for `test` set. We report the results on OCHuman following the setting of previous work [24] and [9].

We follow the standard evaluation metric and use OKS-based metrics for MPPE. We report average precision with different thresholds: AP, AP^{50}, AP^{75}. In addition, for COCO we also report performance on different object sizes: AP^M and AP^L. For CrowdPose, results on different crowd-index are also reported: AP^E, AP^M and AP^H.

4.2. Implementation Details

All experiments are implemented on PyTorch [22]. We adopt HRNet-W32 [26] pretrained on ImageNet [4] as backbone for all experiments and follow the most configuration of [6]. Results of HRNet-W48 are also reported to verify the scale ability of CID. We set $\lambda = 4$ in Eq. (5).

Training procedure resizes each image to 512*512. It uses Adam [10] to optimize the model, and sets the learning rate to 0.001 for all layers. We train the model for 140 epochs on COCO and OCHuman, with learning rate dividing by 10 at 90th, 120th epoch. For ablation study, we train model 35 epochs on COCO. For CrowdPose, we train model with 300 epochs and divide learning rate by 10 at 200th, 260th epoch. The batch size is set to 20 OCHuman and 40 for CrowdPose and COCO. We adopt data augmentation strategies including random rotation (-30,30), scale ([0.75,1.5]), translation ([-40,40]) and flipping (0.5).

Testing procedure resizes the short side of each image to 512, and keeps the aspect ratio. We adopt single scale test with flipping following [3, 6] for all experiments.

4.3. Ablation Study

This section aims to investigate the contribution of each proposed components in CID, including the channel and spatial recalibration and the contrastive loss in IIA. We also present the runtime analysis and visualization of CID.

Component Analysis. We first analyze the effectiveness of each proposed component. The results are shown in Table 1. Simply applying spatial recalibration only obtains 17.9% AP on COCO `val` set, indicating its poor performance in separating different persons. Adding the contrastive loss in Eq. (8) significantly boosts the performance of spatial recalibrate to 64.6%. This demonstrates the importance of feature discriminative power in person decoupling, as well as the validity of our contrastive loss. We also notice that this loss is useful to channel recalibration, *e.g.*, improves the performance from 64.9% to 65.3%.

Table 1 also compares spatial and channel recalibration methods. CID with only channel recalibration obtains 64.9% AP on COCO `val` set. The spatial recalibration with contrastive loss achieves 64.6% AP. This indicates that both recalibration methods are effectively in separating persons. Fusing channel and spatial recalibration consistently gets the best performance for cases with and without contrastive loss. We conclude that learning discriminative person features, and jointly considering spatial-wise and channel-wise

$\ell(\mathbf{f}^{(i)})$	Spatial	Channel	AP	AP^{50}	AP^{75}	AP^M	AP^L
	✓		17.9	29.4	16.7	11.7	26.9
		✓	64.9	86.1	71.0	58.1	74.5
	✓	✓	65.3	86.4	71.4	59.3	74.8
✓	✓		64.6	86.0	70.8	58.5	74.1
✓		✓	65.3	85.9	71.9	59.1	75.3
✓	✓	✓	66.0	86.7	72.3	59.8	76.0

Table 1. Validity of contrastive loss, spatial and channel recalibration in CID on COCO `val` set.

# channels	AP	AP^{50}	AP^{75}	AP^M	AP^L
8	62.8	85.1	68.2	56.2	72.8
16	64.4	86.1	70.4	58.5	74.1
32	66.0	86.7	72.3	59.8	76.0
64	66.1	86.8	72.6	60.0	76.0

Table 2. Performance with different embedding dimensions on COCO `val` set.

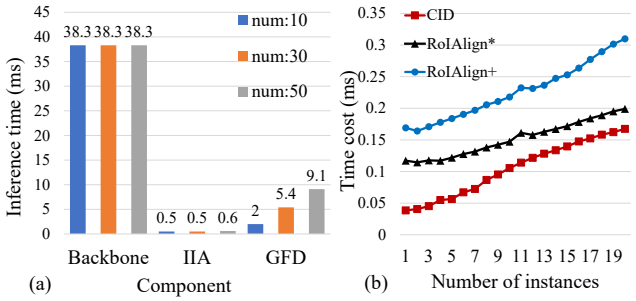

Figure 5. Efficiency analysis of CID. (a) shows inference time of each component *w.r.t.* different numbers of persons. (b) compares efficiency of CID with two variants of RoIAlign [7].

	Mem(G)	Params(M)	GFLOPs	Speed(fps)	AP
HrHRNet [3]	3.5	29.6	48.11	4.9	66.4
DEKR [6]	3.6	28.6	44.50	5.4	67.3
FCPose [16]	4.8	60.3	256.7	15.2	65.6
CID	3.8	29.4	43.17	21.0	68.9

Table 3. Efficiency comparison between CID and other methods. "Mem" refers to memory consumption during inference, which is tested on COCO `val` set with batchsize 1 and a single RTX 3090. Compared methods are tested with codes provided by authors.

Figure 6. Illustration of sampled pose estimation results of CID from (a) COCO [14], (b) CrowdPose [13], and (c) OCHuman [34].

decoupling are important for decoupling persons.

Analysis on embedding dimension. The embedding dimension in CID is an important hyper-parameter. Too small channel dimension is hard to encode cues of a large number of persons, while large dimension increases the memory and computation cost. Table 2 tests different embedding dimension from 8 to 64 and reports their performance. It indicates that smaller dimension corresponds to lower performance. Setting too large dimension, *e.g.*, 64, no longer substantially boosts the performance. We set the embedding dimension to 32 as a reasonable trade-off between accuracy and computational costs.

Runtime Analysis. We proceed to give a detailed analysis on the efficiency of CID. We first analyze the inference time of each component in our model on COCO `val` set with respect to different number of persons. Fig. 5 (a) shows that IIA and GFD only take a fraction of total time cost, even for cases with a large number of persons like 50. Fig. 5 (b) compares CID with the commonly used RoIAlign [7]. We report the time cost of generating instance features from global feature map with respect to different person numbers. We denotes the RoIAlign+ in Mask R-CNN as the green line, which outputs a feature map with size 14×14, then upsamples it to 56×56. We also denote RoIAlign$*$ as the black line, which directly outputs 56×56 sized feature map. It is clear that, CID achieves substantially better efficiency than RoIAlign+ and RoIAlign$*$.

Table 3 compares our method with recent works on parameter size and memory consumption during inference. Among those compared works, HrHRNet [3] follows the bottom-up pipeline. DEKR [6] and FCPose [16] are single-stage methods, hence are more efficiency than HrHRNet [3]. CID outperforms HrHRNet [3] in AP. Compared with both single-stage methods, our method consumes a comparable size of memory, and achieves a faster inference speed and substantially better performance.

Visualization to instance-aware feature maps and keypoint heatmaps of CID are demonstrated in Fig. 4. Fig. 6 further visualizes several pose estimation results on three datasets. It can be observed that, our method gets reliable and accuracy pose estimation even for challenging cases like heavy occlusion and person overlapping.

4.4. Comparison with Other Methods

We compare CID with recent works on COCO and two crowded scenes pose estimation benchmarks CrowdPose and OCHuman in Table 4, 5, and 6, respectively.

General Multi-Person Pose Estimation. Comparison with recent works on COCO are shown in Table 4. We compare three types of methods, including top-down methods: Mask R-CNN [7], bottom-up methods: OpenPose [2], AE [17], HGG [9], PifPaf [11] and HrHRNet [3], and recent single-stage methods: CenterNet [35], SPM [19], PointSet Anchor [31] and DEKR [6]. Compared with top-down methods, CID achieves better performance, outperforming Mask R-CNN by 5.8%. This indicates that our decoupling strategy is superior to box cropping. CID is also better than many bottom-up methods. For instance, we achieve 68.9% AP on COCO `test-dev` set, which is higher than AE by 6.1% and HigherHRNet by 2.5%. Compared with single-stage methods, CID also achieves superior performance, *i.e.*, outperforming DEKR by 1.6%. Benefit from its simple architecture and end-to-end trainable pipeline, CID also enjoys better inference speed. For instance, it achieves 21.0 FPS, faster than most of existing MPPE methods.

Crowded Scenes Pose Estimation. To test CID in more challenging scenarios, we compare it with recent works on crowded scenes pose estimation benchmarks CrowdPose and OCHuman. More severe person overlapping and occlusions in those datasets degrade the accuracy of person detection and keypoint localization, making existing methods get lower AP than on COCO.

We first evaluate on CrowdPose and summarize results in Table 5. Compared works include JC-SPPE [13], DEKR [6] and PINet [30]. Our method achieves 71.3% AP on CrowdPose `test` set, outperforming previous methods by a large margin, *e.g.*, CID outperforms DEKR by 5.6%. Comparison between Table 5 and Table 4 shows that CID gets more substantial advantages on CrowdPose.

Table 6 shows the comparison on OCHuman, a more challenging crowded scenes benchmarks. We report the performance under two evaluation protocols. The first is proposed in [24], which trains on OCHuman `val` set and tests on `test` set. Results are shown in the second column of Table 6, where CID achieves the best performance, *e.g.*, outperforming DEKR by 5.3%. We also report the results of

Method	Backbone	Input size	FPS	AP	AP^{50}	AP^{75}	AP^{M}	AP^{L}	AR
Top-down methods									
Mask R-CNN [7]	ResNet-50+FPN	800	14.2	63.1	87.3	68.7	57.8	71.4	-
SimpleBaseline^{+} [33]	ResNet-152	384x288	-	73.7	91.9	81.1	70.3	80.0	-
HRNet^{+} [26]	HRNet-W32	384x288	-	74.9	92.5	82.8	71.3	80.9	-
Bottom-up methods									
OpenPose* [2]	VGG-19	-	10.1	61.8	84.9	67.5	57.1	68.2	66.5
AE [17]	HourGlass	512	8.1	62.8	84.6	69.2	57.5	70.6	
HGG [9]	HourGlass	512	-	60.4	83.0	66.2	84.0	69.8	-
PifPaf [11]	ResNet-152	-	-	66.7	-	-	62.4	72.9	-
PersonLab [20]	ResNet-152	1401	5.0	66.5	88.0	72.6	62.4	72.3	71.0
HrHRNet [3]	HRNet-W32	512	4.9	66.4	87.5	72.8	61.2	74.2	-
SWAHR [15]	HrHRNet-W32	512	4.9	67.9	88.9	74.5	62.4	75.5	-
CenterAttention [1]	HrHRNet-W32	512	-	67.6	88.7	73.6	61.9	75.6	-
Single-stage methods									
CenterNet [35]	HourGlass	512	12.2	63.0	86.8	69.6	58.9	70.4	-
SPM* [19]	HourGlass	-	-	66.9	88.5	72.9	62.6	73.1	-
PointSet Anchor [31]	HRNet-W48	800	4.8	66.3	87.7	73.4	64.9	70.0	-
FCPose [16]	ResNet-101+FPN	800	15.2	65.6	87.9	72.6	62.1	72.3	-
DEKR [6]	HRNet-W32	512	5.1	67.3	87.9	74.1	61.5	76.1	72.4
Ours	HRNet-W32	512	21.0	**68.9**	**89.9**	**76.0**	**63.2**	**77.7**	**74.6**
Ours	HRNet-W48	640	12.5	**70.7**	**90.3**	**77.9**	**66.3**	**77.8**	**76.4**

Table 4. Comparison with recent works on COCO `test-dev` set. * denotes using refinement. $^{+}$ denotes using two separate models for person detection and pose estimation. FPS is measured on a single RTX 2080Ti.

Method	AP	AP^{50}	AP^{75}	AP^{E}	AP^{M}	AP^{H}
Top-down methods						
Mask R-CNN [7]	57.2	83.5	60.3	69.4	57.9	45.8
JC-SPPE [13]	66.0	84.2	71.5	75.5	66.3	57.4
Bottom-up methods						
OpenPose* [2]	-	-	-	62.7	48.7	32.3
HrHRNet-W48 [3]	65.9	86.4	70.6	73.3	66.5	57.9
CenterAttention [1]	67.6	87.7	72.7	75.8	68.1	58.9
Single-stage methods						
DEKR(HRNet-W32) [6]	65.7	85.7	70.4	73.0	66.4	57.5
PINet(HRNet-W32) [30]	68.9	88.7	74.7	75.4	69.6	61.5
Ours(HRNet-W32)	**71.3**	**90.6**	**76.6**	**77.4**	**72.1**	**63.9**
Ours(HRNet-W48)	**72.3**	**90.8**	**77.9**	**78.7**	**73.0**	**64.8**

Table 5. Comparison with recent works on CrowdPose `test` set. * denotes using refinement.

Method	OCHuman `val`	COCO `train`	
	`test`	`val`	`test`
Top-down methods			
Mask R-CNN [7]	20.2	-	-
JC-SPPE [13]	27.6	-	-
OPEC-Net [24]	29.1	-	-
RMPE [5]	-	38.8	30.7
SimpleBaseline [33]	-	41.0	33.3
Bottom-up methods			
AE [17]	-	32.1	29.5
HGG [9]	-	35.6	34.8
HrHRNet-W32 [3]	27.7	40.0	39.4
Single-stage methods			
SPM [19]	45.6	-	-
DEKR(HRNet-W32) [6]	52.2	37.9	36.5
Ours(HRNet-W32)	**57.5**	**44.9**	**44.0**
Ours(HRNet-W48)	**58.6**	**46.1**	**45.0**

Table 6. Comparison with recent works on OCHuman `val` and `test` set under two evaluation settings. The second column denotes the first evaluation protocol in [24] and the third and fourth columns denote the second protocol used in [9].

another evaluation protocol in [9] to train on COCO `train` set and test on OCHuman `val` and `test` set. The third and fourth columns of Table 6 show the comparison under this setting. Our method achieves 44.9% and 44.0% AP on the OCHuman `val` and `test` set, consistently outperforming competitors by a large margin.

Table 5 and Table 6 clearly demonstrate that CID achieves superior performance in crowded scenes. We hence could conclude that proposed contextual instance decoupling pipeline is superior to previous ones for MPPE.

5. Conclusion

Decoupling persons and estimating their keypoint locations in crowded scenes are challenging for MPPE. This paper proposes the CID to decouple persons into multiple instance-aware feature maps. CID hence adopts each feature maps to infer keypoints for a specific person. CID does not rely on bounding boxes, making it differentiable and more robust to detection errors than top-down methods. It alleviates the difficulties of keypoint grouping and long distance regression in bottom-up and single-stage methods. Another advantage is the capability of encoding more spatial contextual cues. Those characteristics make CID achieves substantially better performance than previous works on three MPPE benchmarks.

Acknowledgement This work is supported in part by The National Key Research and Development Program of China under Grant No. 2018YFE0118400, in part by Natural Science Foundation of China under Grant No. U20B2052, 61936011.

References

[1] Guillem Brasó, Nikita Kister, and Laura Leal-Taixé. The center of attention: Center-keypoint grouping via attention for multi-person pose estimation. In *ICCV*, 2021. 2, 8

[2] Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Realtime multi-person 2d pose estimation using part affinity fields. In *CVPR*, 2017. 1, 2, 7, 8

[3] 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 *CVPR*, 2020. 6, 7, 8

[4] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, 2009. 6

[5] Hao-Shu Fang, Shuqin Xie, Yu-Wing Tai, and Cewu Lu. Rmpe: Regional multi-person pose estimation. In *ICCV*, 2017. 8

[6] Zigang Geng, Ke Sun, Bin Xiao, Zhaoxiang Zhang, and Jingdong Wang. Bottom-up human pose estimation via disentangled keypoint regression. In *CVPR*, 2021. 1, 2, 6, 7, 8

[7] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *CVPR*, 2017. 1, 2, 6, 7, 8

[8] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *CVPR*, 2018. 3, 4, 5

[9] 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 *ECCV*, 2020. 1, 2, 6, 7, 8

[10] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *ICLR*, 2015. 6

[11] Sven Kreiss, Lorenzo Bertoni, and Alexandre Alahi. Pifpaf: Composite fields for human pose estimation. In *CVPR*, 2019. 1, 2, 7, 8

[12] Hei Law and Jia Deng. Cornernet: Detecting objects as paired keypoints. In *ECCV*, 2018. 5

[13] Jiefeng Li, Can Wang, Hao Zhu, Yihuan Mao, Hao-Shu Fang, and Cewu Lu. Crowdpose: Efficient crowded scenes pose estimation and a new benchmark. In *CVPR*, 2019. 2, 3, 5, 6, 7, 8

[14] 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 *ECCV*, 2014. 2, 5, 7

[15] Zhengxiong Luo, Zhicheng Wang, Yan Huang, Liang Wang, Tieniu Tan, and Erjin Zhou. Rethinking the heatmap regression for bottom-up human pose estimation. In *CVPR*, 2021. 8

[16] Weian Mao, Zhi Tian, Xinlong Wang, and Chunhua Shen. Fcpose: Fully convolutional multi-person pose estimation with dynamic instance-aware convolutions. In *CVPR*, 2021. 6, 7, 8

[17] Alejandro Newell, Zhiao Huang, and Jia Deng. Associative embedding: End-to-end learning for joint detection and grouping. In *NeurIPS*, 2017. 1, 2, 7, 8

[18] Alejandro Newell, Kaiyu Yang, and Jia Deng. Stacked hourglass networks for human pose estimation. In *ECCV*, 2016. 2

[19] Xuecheng Nie, Jiashi Feng, Jianfeng Zhang, and Shuicheng Yan. Single-stage multi-person pose machines. In *CVPR*, 2019. 1, 2, 7, 8

[20] 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 *ECCV*, 2018. 1, 2, 8

[21] 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 *CVPR*, 2017. 1, 2

[22] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. In *NeurIPS*, 2019. 6

[23] Leonid Pishchulin, Eldar Insafutdinov, Siyu Tang, Bjoern Andres, Mykhaylo Andriluka, Peter V Gehler, and Bernt Schiele. Deepcut: Joint subset partition and labeling for multi person pose estimation. In *CVPR*, 2016. 1, 2

[24] Lingteng Qiu, Xuanye Zhang, Yanran Li, Guanbin Li, Xiaojun Wu, Zixiang Xiong, Xiaoguang Han, and Shuguang Cui. Peeking into occluded joints: A novel framework for crowd pose estimation. In *ECCV*, 2020. 6, 7, 8

[25] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *CVPR*, 2016. 2

[26] Ke Sun, Bin Xiao, Dong Liu, and Jingdong Wang. Deep high-resolution representation learning for human pose estimation. In *CVPR*, 2019. 1, 2, 6, 8

[27] Xiao Sun, Bin Xiao, Fangyin Wei, Shuang Liang, and Yichen Wei. Integral human pose regression. In *ECCV*, 2018. 1

[28] Zhi Tian, Hao Chen, and Chunhua Shen. Directpose: Direct end-to-end multi-person pose estimation. *arXiv preprint arXiv:1911.07451*, 2019. 1, 2, 4

[29] Zhi Tian, Chunhua Shen, and Hao Chen. Conditional convolutions for instance segmentation. In *ECCV*, 2020. 4

[30] Dongkai Wang, Shiliang Zhang, and Gang Hua. Robust pose estimation in crowded scenes with direct pose-level inference. In *NeurIPS*, 2021. 1, 7, 8

[31] Fangyun Wei, Xiao Sun, Hongyang Li, Jingdong Wang, and Stephen Lin. Point-set anchors for object detection, instance segmentation and pose estimation. In *ECCV*, 2020. 7, 8

[32] Sanghyun Woo, Jongchan Park, Joon-Young Lee, and In So Kweon. Cbam: Convolutional block attention module. In *ECCV*, 2018. 3, 5

[33] Bin Xiao, Haiping Wu, and Yichen Wei. Simple baselines for human pose estimation and tracking. In *ECCV*, 2018. 1, 2, 8

[34] 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 *CVPR*, 2019. 2, 3, 5, 6, 7

[35] Xingyi Zhou, Dequan Wang, and Philipp Krähenbühl. Objects as points. *arXiv preprint arXiv:1904.07850*, 2019. 1, 2, 4, 5, 7, 8