Multi-Instance Pose Networks: Rethinking Top-Down Pose Estimation

Rawal Khirodkar[1*] Visesh Chari[2] Amit Agrawal[2] Ambrish Tyagi[2]
[1]Carnegie Mellon University [2]Amazon Lab 126
rkhirodk@cs.cmu.edu {viseshc, aaagrawa, ambrisht}@amazon.com

Abstract

A key assumption of top-down human pose estimation approaches is their expectation of having a single person/instance present in the input bounding box. This often leads to failures in crowded scenes with occlusions. We propose a novel solution to overcome the limitations of this fundamental assumption. Our Multi-Instance Pose Network (MIPNet) allows for predicting multiple 2D pose instances within a given bounding box. We introduce a Multi-Instance Modulation Block (MIMB) that can adaptively modulate channel-wise feature responses for each instance and is parameter efficient. We demonstrate the efficacy of our approach by evaluating on COCO, CrowdPose, and OCHuman datasets. Specifically, we achieve 70.0 *AP on CrowdPose and* 42.5 *AP on OCHuman test sets, a significant improvement of* 2.4 *AP and* 6.5 *AP over the prior art, respectively. When using ground truth bounding boxes for inference, MIPNet achieves an improvement of* 0.7 *AP on COCO,* 0.9 *AP on CrowdPose, and* 9.1 *AP on OCHuman validation sets compared to HRNet. Interestingly, when fewer, high confidence bounding boxes are used, HRNet's performance degrades (by* 5 *AP) on OCHuman, whereas MIPNet maintains a relatively stable performance (drop of* 1 *AP) for the same inputs.*

1. Introduction

Human pose estimation aims at localizing 2D human anatomical keypoints (*e.g.*, elbow, wrist, *etc.*) in a given image. Current human pose estimation methods can be categorized as *top-down* or *bottom-up* methods. Top-down methods [6, 13, 33, 40, 41, 43, 44] take as input an image region within a bounding box, generally the output of a human detector, and reduce the problem to the simpler task of *single human pose estimation*. Bottom-up methods [3, 22, 29, 32], in contrast, start by independently localizing keypoints in the entire image, followed by grouping them into 2D human pose instances.

The single human assumption made by top-down approaches limits the inference to a *single* configuration of human joints (a single instance) that can best explain the input. Top-down pose estimation approaches [6, 16, 30, 40, 44] are currently the best performers on datasets such as COCO [25], MPII [2]. However, when presented with inputs containing multiple humans like crowded or occluded instances, top-down methods are forced to select a single plausible configuration per human detection. In such cases, top-down methods may erroneously identify pose landmarks corresponding to the occluder (person in the front). See, for example, Fig. 1 (Middle). Therefore, on datasets such as CrowdPose [23] and OCHuman [48], which have a relatively higher proportion of occluded instances (Table 1), the performance of top-down methods suffer due to the single person assumption [8, 23, 48].

Figure 1: 2D pose estimation networks often fail in presence of heavy occlusion. (Left) Bounding boxes corresponding to two persons. (Middle) For both bounding boxes, HRNet predicts the pose for the front person and misses the occluded person. (Right) MIPNet allows multiple instances for each bounding box and recovers the pose of the occluded person.

In this paper, we rethink the architecture for top-down 2D pose estimators by predicting *multiple* pose instances for the input bounding box. The key idea of our proposed architecture is to allow the model to predict more than one pose instance for each bounding box. We demonstrate that

*Work done during an internship at Amazon Lab 126

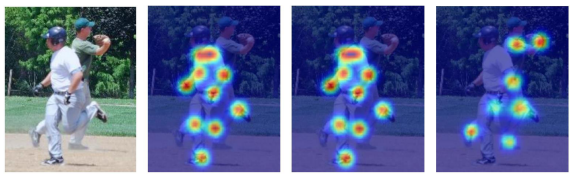

Figure 2: Heatmap predictions for a few keypoints from HRNet vs MIPNet. HRNet only focuses on the foreground person. MIPNet enables prediction of the multiple instances from the *same* input bounding box by varying λ during inference.

this conceptual change improves the performance of top-down methods, especially for images with crowding and heavy occlusion. A naïve approach to predict multiple instances per bounding box would be to add multiple prediction heads to an existing top-down network with a shared feature-extraction backbone. However, such an approach fails to learn different features corresponding to the various instances. A brute-force approach would then be to replicate the feature-extraction backbone, though at a cost of an N-fold increase in parameters, for N instances. In contrast, our approach enables predicting multiple instances for any existing top-down architecture with a small increase in the number of parameters ($< 3\%$) and inference time ($< 9\text{ms}, 16\%$). Technically, our approach can handle $N > 2$ instances. However, as shown in Figure 4, number of examples with 3+ annotated pose instances per ground truth bounding box in existing datasets is extremely small. Thus, similar to [35, 48], we primarily focus on the dominant occlusion scenario involving two persons.

To enable efficient training and inference of multiple instances in a given bounding box, we propose a novel Multi-Instance Modulation Block (MIMB). MIMB modulates the feature tensors based on a scalar *instance-selector*, λ, and allows the network to index on one of the N instances (Fig. 2). MIMB can be incorporated in any existing feature-extraction backbone, with a relatively simple (< 15 lines) code change (refer supplemental). At inference, for a given bounding box, we vary the instance-selector λ to generate multiple pose predictions (Fig. 3).

Since top-down approaches rely on the output from an object detector, they typically process a large number of bounding box hypotheses. For example, HRNet [40] uses more than $100K$ bounding boxes from Faster R-CNN [37] to predict 2D pose for ~ 6000 persons in the COCO `val` dataset. Many of these bounding boxes overlap and majority have low detection scores (< 0.4). This also adversely impacts the inference time, which increases linearly with the number of input bounding boxes. As shown in Fig. 5, using fewer, high confidence bounding boxes degrades the performance of HRNet from 37.8 to 32.8 AP on OCHuman, a degradation of 5 AP in performance. In contrast, MIPNet is robust and maintains a relatively stable performance for the same inputs (drop of 1 AP). Intuitively, our method can predict the 2D pose instance corresponding to a mis-detected bounding box based on predictions from its neighbors.

Dataset	IoU> 0.5	ΔAP_0	$\Delta AP_{0.9}$	ΔAP_{gt}
COCO	1.2K (1%)	0.0	+1.9	+0.7
CrowdPose	2.9K (15%)	+0.8	+2.3	+0.9
OCHuman	3.2K (68%)	+4.2	+8.2	+9.1

Table 1: MIPNet's relative improvement in AP compared to HRNet-W48 on the `val` set, using Faster R-CNN (AP_0: all, $AP_{0.9}$: high confidence) and ground truth (AP_{gt}) bounding boxes. For each dataset, the number (%) of instances with occlusion IoU > 0.5 is reported [35]. Datasets with more occlusions and crowding demonstrate higher gains.

Overall, MIPNet outperforms top-down methods and occlusion specific methods on various datasets as shown in Table 1. For challenging datasets such as CrowdPose and OCHuman, containing a larger proportion of cluttered scenes (with multiple overlapping people), MIPNet sets a new state-of-the-art achieving 70.0 AP and 42.5 AP respectively on the `test` set outperforming bottom-up methods. Our main contributions are

- We advance top-down 2D pose estimation methods by addressing limitations caused by the single person assumption during training and inference. Our approach achieves the state-of-the-art results on CrowdPose and OCHuman datasets.
- MIPNet allows predicting multiple pose instances for a given bounding box efficiently by modulating feature responses for each instance independently.
- The ability to predict multiple instances makes MIPNet resilient to bounding box confidence and allows it to deal with missing bounding boxes with minimal impact on performance.

2. Related Work

Biased benchmarks: Most human pose estimation benchmarks [1, 2, 12, 20, 25] do not uniformly represent possible poses and occlusions in the real world. Popular datasets such as COCO [25] and MPII [2] have less than 3% annotations with crowding at IoU of 0.3 [35]. More than 86% of annotations in COCO [25] have 5 or more keypoints visible [38]. These biases have seeped into our state-of-the-art data driven deep learning models [45], not only in the form of poor generalization to "in-the-tail" data but surprisingly in critical design decisions for network architectures. Recently, challenging datasets such as OCHuman [48] and CrowdPose [23] containing heavy occlusion have been proposed to capture these biases. These datasets demonstrate the failures of the state-of-art models under severe occlusions (Section 4.3). MIPNet shows a significant improvement in performance under such challenging conditions.

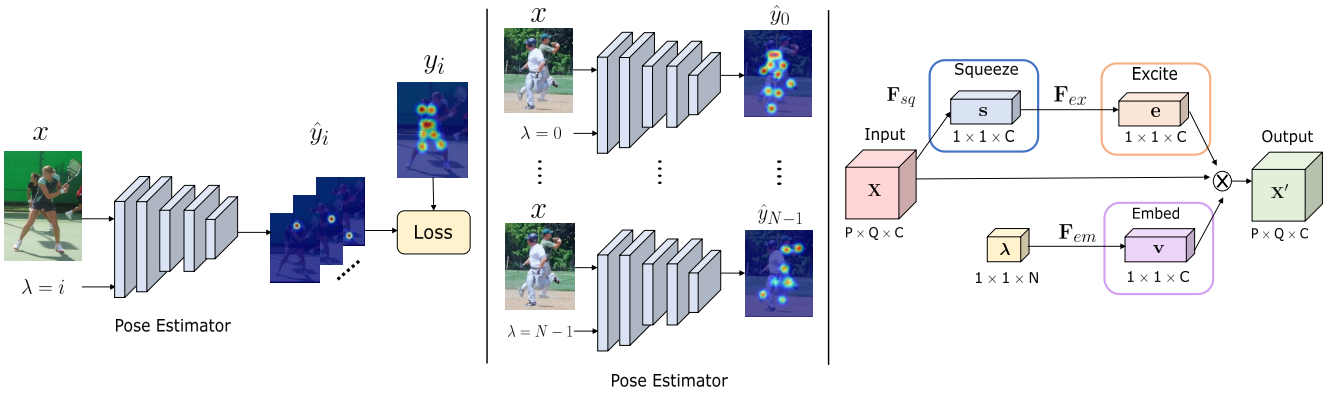

Figure 3: (Left) MIPNet is trained to predict the i^{th} instance from an input x by conditioning the network using $\lambda = i, \forall\, i = 0, \ldots, N-1$. (Middle) During inference, we obtain the N pose predictions by varying λ. (Right) MIMB uses squeeze, excitation and embed modules that enables λ to modulate the feature responses for each instance.

Top-down methods: Top-down methods [6, 11, 13, 16, 30, 33, 40, 44] detect the keypoints of a single person within a bounding box. These bounding boxes are usually generated by an object detector [7, 24, 26, 36, 37]. As top-down methods can normalize all the persons to approximately the same scale by cropping and resizing the bounding boxes, they are generally less sensitive to scale variations in images. Thus, state-of-the-art performances on various human pose estimation benchmarks are mostly achieved by top-down methods [40] in contrast to bottom-up methods [3, 8, 17, 18, 29, 34, 48]. However, these methods inherently assume a single person (instance) in the detection window and often fail under occlusions in multi-person cases. It is the ambiguity of the implicit bounding-box level representation that leads to this failure. MIPNet resolves this issue by predicting multiple instances within a single bounding box.

Occluded pose estimation: Many methods [42, 31, 39, 9] have made good progress in occluded person detection. Recent methods [23, 48, 35, 19] have also focused on occluded pose estimation. [23] uses a top-down model to make a multi-peak prediction and joint peaks are then grouped into persons using a graph model. [48] uses instance segmentation for occlusion reasoning. [35] use a graph neural network to refine pose proposals from a top-down model. [19] is a bottom-up method which uses a differentiable hierarchical graph grouping for joint association. In contrast, our approach is much simpler and does not require initial pose estimates, grouping or solving for joint association.

Lastly, in machine learning, many models have been trained to behave differently depending on a conditional input [5, 10, 21, 27, 28, 47]. Instead of training multiple models, our approach enables training a single network for predicting multiple outputs on the same input. Rather than duplicating the feature backbone, our novel MIMB block leads to a parameter efficient design. Our multi-instance pose network is fully supervised and not related to multiple instance learning [4, 46], which is a form of weakly-supervised learning paradigm where training instances are arranged in sets.

3. Method

Human pose estimation aims to detect the locations of K keypoints from an input image $x \in \mathbb{R}^{H \times W \times 3}$. Most top-down methods transform this problem to estimating K heatmaps, where each heatmap indicates the probability of the corresponding keypoint at any spatial location. Similar to [6, 30, 44] we define a convolutional pose estimator, P, for human keypoint detection. The bounding box at training and inference is scaled to $H \times W$ and is provided as an input to P. Let $y \in \mathbb{R}^{H' \times W' \times K}$ denote the K heatmaps corresponding to the ground truth keypoints for a given input x. The pose estimator transforms input x to a single set of predicted heatmaps, $\hat{y} \in \mathbb{R}^{H' \times W' \times K}$, such that $\hat{y} = P(x)$. P is trained to minimize the mean squared loss $\mathcal{L} = \texttt{MSE}(y, \hat{y})$.

3.1. Training Multi-Instance Pose Network

We propose to modify the top-down pose estimator P to predict multiple instances as follows. Our pose estimator P predicts N instances, $\hat{y}_0, \ldots, \hat{y}_{N-1}$ for an input x. This is achieved by conditioning the network P on a scalar *instance-selector* λ, $0 \leq \lambda \leq N-1$. P accepts both x and λ as input and predicts $\hat{y}_i = P(x, \lambda = i)$, where $i \in \{0, 1, \ldots, N-1\}$.

Let B_0 denote the ground truth bounding box used to crop the input x. Let B_i, $i \in \{1, ..n-1\}$, denote additional $n-1$ ground truth bounding boxes which overlap B_0, such that at least $k = 3$ keypoints from B_i fall within B_0. Thus, $B_0, \ldots, B_{n-1}$ represents the bounding boxes for n ground truth pose instances present in x. We denote the ground truth heatmaps corresponding to these n instances by $y_0, \ldots y_{n-1}$.

To define a loss, we need to assign the predicted pose instances to the ground truth heatmaps. The primary instance $\hat{y}_0 = P(x, \lambda = 0)$ is assigned to y_0, the pose instance corresponding to B_0. The next $N-1$ instances are assigned

to the remaining ground truth heatmaps ordered according to the distance of their corresponding bounding box from B_0. We train the network P to minimize the loss $\mathcal{L} = \frac{1}{N}\sum_{i=0}^{N-1}\mathcal{L}_i$, where,

$$\mathcal{L}_i = \begin{cases} \texttt{MSE}(y_i, P(x, \lambda = i)), & \forall\ 0 \leq i < \min(n, N), \\ \texttt{MSE}(y_0, P(x, \lambda = i)), & \forall\ \min(n, N) \leq i < N. \end{cases} \quad (1)$$

When $n <= N$, the available n ground truth pose instances are used to compute the loss for n predictions, and the loss for *residual* $N - n$ instances is computed using y_0. For example, when $n = 1$ and $N = 2$, both the predictions are encouraged to predict the heatmaps corresponding to the single ground truth instance present in x. In contrast, when $n > N$, only N ground truth pose instances (closest to B_0) are used to compute the loss.

In our experience, employing other heuristics such as not propagating the loss, *i.e., don't care* for residual instances resulted in less stable training. Additionally, a *don't care* based training scheme for residual instances resulted in significantly higher false positives, especially as we do not know the number of valid person instances per input at runtime. During inference, we vary λ to extract different pose predictions from the same input x as shown in Fig. 3.

3.2. Multi-Instance Modulation Block

In this section, we describe the Multi-Instance Modulation Block (MIMB) that can be easily introduced in any existing feature extraction backbone. The MIMBs allow a top-down pose estimator P to predict multiple instances from an input image x. Using MIMBs, P can now accept both x and the instance-selector λ as inputs. The design of MIMB is inspired by the squeeze excite block of [14]. Let $\mathbf{X} \in \mathbb{R}^{P \times Q \times C}$ be an intermediate feature map with C channels, such that $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_C]$. We use an instance-selector λ to modulate the channel-wise activations of the output of the excite module as shown in Fig. 3 (Right). The key insight of our design is that we can use the same set of convolutional filters to dynamically cater to different instances in the input. Compared to a brute force approach of replicating the feature backbone or assigning a fixed number of channels per instance, our design is parameter efficient.

Let $\mathbf{F}_{sq}$, $\mathbf{F}_{ex}$, $\mathbf{F}_{em}$ denote the *squeeze*, *excite*, and *embed* operations, respectively, within MIMB. We represent $\boldsymbol{\lambda}$ as the one hot representation of scalar λ. The feature map $\mathbf{X}$ is transformed to $\mathbf{X}' = [\mathbf{x}'_1, \mathbf{x}'_2, \ldots, \mathbf{x}'_C]$ as follows,

$$s_c = \mathbf{F}_{sq}(\mathbf{x}_c), \quad (2)$$

$$\mathbf{e} = \mathbf{F}_{ex}(\mathbf{s}), \quad (3)$$

$$\mathbf{v} = \mathbf{F}_{em}(\boldsymbol{\lambda}), \quad (4)$$

$$\mathbf{x}'_c = (v_c \times e_c)\mathbf{x}_c, \quad (5)$$

s.t. $\mathbf{s} = [s_1, \ldots, s_C]$, $\mathbf{v} = [v_1, \ldots, v_C]$ and $\mathbf{e} = [e_1, \ldots, e_C]$. $\mathbf{F}_{sq}$ *squeezes* the global spatial information into a channel descriptor using global average pooling. $\mathbf{F}_{ex}$ allows modeling for channel-wise interactions on the output of $\mathbf{F}_{sq}$. $\mathbf{F}_{ex}$ is implemented as a two layer, fully-connected, neural network. Following the output of the excite module, we modulate the channel-wise activations using the embedding of $\boldsymbol{\lambda}$ from another simple neural network $\mathbf{F}_{em}$. $\mathbf{F}_{em}$ has a similar design to $\mathbf{F}_{ex}$.

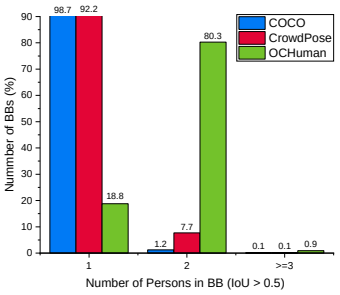

Figure 4: Percentage of examples with 1, 2 and 3+ pose instances per ground truth bounding box in various datasets.

During inference, we vary the instance-selector λ from 0 to $N-1$ to get N predictions and then apply OKS-NMS [40] after merging all predictions. Please refer supplemental for details. Figure 2 visualizes the predicted heatmaps from HRNet and MIPNet (using $N = 2$). Note that HRNet only outputs the heatmap corresponding to the foreground person while MIPNet predicts heatmaps for both persons using different values of λ at inference.

4. Experiments

We evaluate MIPNet on three datasets: *Common-Objects in Context*-COCO [25], *CrowdPose* [23] and *Occluded Humans*-OCHuman [48]. These datasets represent varying degrees of occlusion/crowding (see Table 1) and help illustrate the benefits of predicting multiple instances in top-down methods. We report standard metrics such as AP, AP^{50}, AP^{75}, AP^{M}, AP^{L}, AR, AP^{easy}, AP^{med} and AP^{hard} at various Object Keypoint Similarity as defined in [25, 23]. We report results using ground truth bounding boxes as well as bounding boxes obtained via YOLO [36] and Faster R-CNN [37] detectors.

We base MIPNet on recent state-of-the-art top-down architectures, namely, SimpleBaseline [44] and HRNet [40]. When comparing with HRNet, MIPNet employs a similar feature extraction backbone and adds MIMBs' at the output of the convolutional blocks at the end of stages 3 and 4 [40]. For comparisons with SimpleBaseline [44], two MIMB's are added to the last two ResNet blocks in the encoder.

Number of instances N: Trivially, $N = 1$ is equivalent to baseline top-down methods. By design, MIPNet supports predicting multiple instances. Empirically, on average we observed a small improvement of 0.3 AP, 0.5 AP using $N = 3$ and $N = 4$ on top of $N = 2$ respectively on the datasets. This is consistent with the fact that most datasets have very few examples with three or more ground-truth pose instances

Method	Arch	#Params	AP	AP^{50}	AP^{75}	AP^{M}	AP^{L}	AR
SBL†	R-50	34.0M	72.4	91.5	80.4	69.7	76.5	75.6
MIPNet†	R-50	35.0M (+2.8%)	**73.3 (+0.9)**	**93.3**	**81.2**	**70.6**	**77.6**	**76.7**
SBL†	R-101	53.0M	73.4	92.6	81.4	70.7	77.7	76.5
MIPNet†	R-101	54.0M (+1.7%)	**74.1 (+0.7)**	**93.3**	**82.3**	**71.3**	**78.6**	**77.4**
SBL†	R-152	68.6M	74.3	92.6	82.5	71.6	78.7	77.4
MIPNet†	R-152	69.6M (+1.4%)	**74.8 (+0.5)**	**93.3**	**82.4**	**71.7**	**79.4**	**78.2**
SBL⋆	R-50	34.0M	74.1	92.6	80.5	70.5	79.6	76.9
MIPNet⋆	R-50	35.0M (+0.4%)	**75.3 (+1.2)**	**93.4**	**82.4**	**72.0**	**80.4**	**78.4**
SBL⋆	R-101	35.0M	75.5	92.5	82.6	72.4	80.8	78.4
MIPNet⋆	R-101	54.0M (+0.3%)	**76.0 (+0.5)**	**93.4**	**83.5**	**72.6**	**81.1**	**79.1**
SBL⋆	R-152	68.6M	76.6	92.6	83.6	73.7	81.3	79.3
MIPNet⋆	R-152	69.6M (+2.8%)	**77.0 (+0.4)**	**93.5**	**84.3**	**73.7**	**81.9**	**80.0**
HRNet†	H-32	28.5M	76.5	93.5	83.7	73.9	80.8	79.3
MIPNet†	H-32	28.6M (+1.7%)	**77.6 (+1.1)**	**94.4**	**85.3**	**74.7**	**81.9**	**80.6**
HRNet†	H-48	63.6M	77.1	93.6	84.7	74.1	81.9	79.9
MIPNet†	H-48	63.7M (+1.4%)	**77.6 (+0.5)**	**94.4**	**85.4**	**74.6**	**82.1**	**80.6**
HRNet⋆	H-32	28.5M	77.7	93.6	84.7	74.8	82.5	80.4
MIPNet⋆	H-32	28.6M (+0.4%)	**78.5 (+0.8)**	**94.4**	**85.7**	**75.6**	**83.0**	**81.4**
HRNet⋆	H-48	63.6M	78.1	93.6	84.9	75.3	83.1	80.9
MIPNet⋆	H-48	63.7M (+0.3%)	**78.8 (+0.7)**	**94.4**	**85.7**	**75.5**	**83.7**	**81.6**

Table 2: MIPNet improves performance on COCO `val` set across various architectures and input sizes (using ground-truth bounding boxes). R-@ and H-@ stands for ResNet-@ and HRNet-W@ respectively. † and ⋆ denotes input resolution of 256×192 and 384×288 respectively. SBL refers to SimpleBaseline [44]. #Params are only of the pose estimation network, excluding bounding box computation.

per bounding box (Fig. 4). However, $N = 2$ provides a substantial improvement over $N = 1$ baseline as shown in our experiments. Note that since the MIMBs are added to the last few stages in our experiments, the increase in inference time due to predicting $N = 2$ instances is small (Table 3). For bigger HRNet-48 network with input resolution of 384×288, inference time increases by 8.2ms (16.7%). For smaller HRNet-32 network, increase in run-time is 4.7ms (11.9%). This is significantly better than replicating the backbone for each instance, which would lead to a 2x increase in inference time for $N = 2$. Please refer supplemental for more details.

4.1. COCO Dataset

Dataset: COCO contains $64K$ images and $270K$ persons labeled with 17 keypoints. For training we use the `train` set ($57K$ images, $150K$ persons) and for evaluation we use the `val` ($5K$ images, $6.3K$ persons) and the `test-dev` set ($20K$ images). The input bounding box is extended in either height or width to obtain a fixed aspect ratio of $4 : 3$. The detection box is then cropped from the image and is resized to a fixed size of either 256×192 or 384×288, depending on the experiment. Following [29], we use data augmentation with random rotation ($[-45°, 45°]$), random scale ($[0.65, 1.35]$), flipping, and half-body crops. Following [30, 40, 44], we use flipping and heatmap offset during inference.

Results: Table 2 compares the performance of MIPNet with SimpleBaseline (denoted as SBL) and HRNet using ground truth bounding boxes. MIPNet outperforms the baseline across various backbones and input sizes. Using ResNet-50 backbone, MIPNet improves the SimpleBaseline results by 0.9 AP for smaller input size and 1.2 AP for larger input size. Comparing with HRNet, MIPNet shows an improvement ranging from 0.7 to 1.1 AP on various architectures and input sizes. Note that MIPNet results in $< 3\%$ increase in parameters compared to the baselines.

Arch	Latency	AP	AP^{50}	AP^{75}	AP^{easy}	AP^{med}	AP^{hard}
HRNet-32†	27.5 ms	70.0	91.0	76.3	**78.8**	70.3	61.7
MIPNet†	30.9 ms	**71.2**	**91.9**	**77.4**	**78.8**	**71.5**	**63.8**
HRNet-48†	33.8 ms	71.3	91.1	77.5	80.5	71.4	62.5
MIPNet†	39.6 ms	**72.8**	**92.0**	**79.2**	**80.6**	**73.1**	**65.2**
HRNet-32⋆	39.4 ms	71.6	91.1	77.7	80.4	72.1	62.6
MIPNet⋆	44.1 ms	**73.0**	**91.8**	**79.3**	**80.7**	**73.3**	**65.5**
HRNet-48⋆	49.1 ms	72.8	**92.1**	78.7	**81.3**	73.3	64.0
MIPNet⋆	57.3 ms	**73.7**	91.9	**80.0**	80.7	**74.1**	**66.5**

Table 3: MIPNet outperforms HRNet on CrowdPose `val` set. † and ⋆ denote input resolution of 256×192 and 384×288, respectively. Average GPU latency is reported with batch size 24.

When using bounding boxes obtained from a person detector, as expected, MIPNet performs comparably to SBL and HRNet when using the same backbone (Table 5). Unsurprisingly, since most of the COCO bounding boxes contain a single person. The benefits of MIPNet are apparent on more challenging CrowdPose and OCHuman datasets (Sect. 4.2, 4.3).

4.2. CrowdPose Dataset

Dataset: CrowdPose contains $20K$ images and $80K$ persons labeled with 14 keypoints. CrowdPose has more crowded scenes as compared to COCO, but the index of

crowding is less compared to the OCHuman [48]. For training, we use the `train` set ($10K$ images, $35.4K$ persons) and for evaluation we use the `val` set ($2K$ images, $8K$ persons) and `test` set ($8K$ images, $29K$ persons).

Results: Table 3 compares the performance of MIPNet with HRNet when evaluated using ground-truth bounding boxes. MIPNet outperforms HRNet with improvements in AP ranging from 0.9 to 1.5 across different input sizes. As shown in Table 5, when evaluated using person detector bounding boxes, MIPNet improves SBL by 7.3 AP on the `test` set with an increase of less than 25 ms in inference time. For completeness, we also trained and evaluated HRNet on CrowdPose. MIPNet outperforms HRNet by 0.7 AP on the `test` set and 0.8 AP on the `val` set. MIPNet achieves state-of-the-art performance of 70.0 AP comparable to the two-stage method OPECNet [35] which refines initial pose estimates from AlphaPose+ [35]. We report additional results in the supplemental.

Method	Arch	AP	AP^{50}	AP^{75}	AP^{M}	AP^{L}	AR
SBL†	R-50	56.3	76.1	61.2	66.4	56.3	61.0
MIPNet†	R-50	**64.4 (+8.1)**	**86.0**	**70.4**	**66.8**	**64.4**	**72.3**
SBL†	R-101	60.5	77.2	66.6	**68.3**	60.5	64.7
MIPNet†	R-101	**68.2 (+7.7)**	**87.4**	**75.1**	67.0	**68.2**	**75.5**
SBL†	R-152	62.4	78.3	68.1	**68.3**	62.4	66.5
MIPNet†	R-152	**70.3 (+7.9)**	**88.6**	**77.9**	66.9	**70.2**	**77.0**
SBL⋆	R-50	55.8	74.8	60.4	64.7	55.9	60.7
MIPNet⋆	R-50	**65.3 (+9.5)**	**87.5**	**72.2**	**66.0**	**66.3**	**74.1**
SBL⋆	R-101	61.6	77.2	66.6	62.1	61.6	65.8
MIPNet⋆	R-101	**70.3 (+8.7)**	**88.4**	**77.1**	**64.1**	**70.4**	**77.7**
SBL⋆	R-152	64.2	78.3	69.1	66.5	64.2	68.1
MIPNet⋆	R-152	**72.4 (+8.2)**	**89.5**	**79.5**	**67.7**	**72.5**	**79.6**
HRNet†	H-32	63.1	79.4	69.0	64.2	63.1	67.3
MIPNet†	H-32	**72.5 (+9.4)**	**89.2**	**79.4**	**65.1**	**72.6**	**79.1**
HRNet†	H-48	64.5	79.4	70.1	65.1	64.5	68.5
MIPNet†	H-48	**72.2 (+7.7)**	**89.5**	**78.7**	**66.5**	**72.3**	**79.2**
HRNet⋆	H-32	63.7	78.4	69.0	64.3	63.7	67.6
MIPNet⋆	H-32	**72.7 (+9.0)**	**89.6**	**79.6**	**66.5**	**72.7**	**79.7**
HRNet⋆	H-48	65.0	78.4	70.3	**68.4**	65.0	68.8
MIPNet⋆	H-48	**74.1 (+9.1)**	**89.7**	**80.1**	**68.4**	**74.1**	**81.0**

Table 4: Comparisons on OCHuman `val` set with ground-truth bounding box evaluation after training on COCO `train` set. † and ⋆ denotes input resolution of 256×192 and 384×288 respectively. R-@ and H-@ stands for ResNet-@ and HRNet-W@ respectively. SBL refers to SimpleBaseline [44].

4.3. OCHuman Dataset

Dataset: OCHuman is focused on heavily occluded humans. It contains 4731 images and 8110 persons labeled with 17 keypoints. In OCHuman, on an average 67% of the bounding box area has overlap with other bounding boxes [48], compared to only 0.8% for COCO. Additionally, the number of examples with occlusion IoU > 0.5 is 68% for OCHuman, compared to 1% for COCO (Table 1). This makes the OCHuman dataset complex and challenging for human pose estimation under occlusion. The single person assumption made by existing top-down methods is not entirely applicable to examples in this dataset.

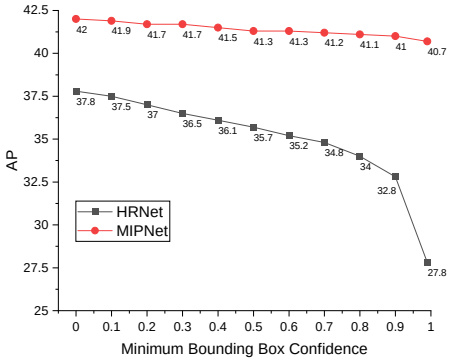

Figure 5: Unlike HRNet, MIPNet maintains a stable performance as a function of detector confidence for selecting input bounding boxes. Results are shown using HRNet-W48-384×288 evaluated on OCHuman `val` set.

Similar to [48], we use the `train` set of COCO for training. Note that we do not train on the OCHuman `train` set. For evaluation, we use the `val` set ($2,500$ images, $4,313$ persons) and the `test` set ($2,231$ images, $3,819$ persons).

Results: Table 4 compares the performance of MIPNet with SimpleBaseline and HRNet on OCHuman when evaluated with ground truth bounding boxes on the `val` set. MIPNet significantly outperforms SimpleBaseline with improvements in AP ranging from 7.7 to 10.5, across various architectures and input sizes. Similarly, for HRNet the performance gains between 7.7 to 9.4 AP are observed.

Current state-of-the-art results on OCHuman are reported by HGG [19] (bottom-up method, multi-scale testing) as shown in Table 5. In addition, we also evaluated MIPNet using person detector boxes on OCHuman with same backbones as baselines for a fair comparison. MIPNet with ResNet101 backbone and YOLO bounding boxes outperforms OPEC-Net by 5.9 AP on the `test` set. When using Faster R-CNN bounding boxes, MIPNet outperforms HRNet and HGG by 5.3 AP and 6.5 AP, respectively, on the `test` set. The improvements are significant and to the best of our knowledge, this is the first time a top-down method has outperformed the state-of-the-art bottom-up method using multi-scale testing on OCHuman.

Figure 8 shows qualitative results on several examples from OCHuman, highlighting the effectivness of MIPNet in recovering multiple poses under challenging conditions.

Robustness to Human Detector Outputs: The performance of top-down methods is often gated by the quality of human detection outputs. We analyze the robustness of HRNet and MIPNet with varying detector confidence on OCHuman in Fig. 5. As expected, HRNet performance degrades as low confidence bounding boxes are filtered out, leading to missed detections on occluded persons. Specifically, HRNet performance degrades from 37.8 AP (30637

Method	COCO		CrowdPose		OCHuman	
	val	test	val	test	val	test
Comparison with Top Down Methods, ResNet101 + YOLO-v3						
MaskRCNN [13]	-	64.8	-	57.2	-	20.2
AlphaPose [23]	-	70.1	-	61.0	-	-
JC-SPPE [23]	-	70.9	-	66.0	-	-
AlphaPose+ [35]	-	72.2	-	68.5	-	27.5
OPEC-Net [35]	-	73.9	-	**70.6**	-	29.1
SBL [44]	-	73.7	-	60.8	-	24.1
MIPNet (Ours)	**72.7**	**74.2**	**63.4**	68.1	**32.8**	**35.0**
Comparison with Top Down Methods, HRNet-W48-384 + Faster R-CNN						
HRNet [40]	**76.3**	75.5	68.0	69.3	37.8	37.2
MIPNet (Ours)	**76.3**	**75.7**	**68.8**	**70.0**	**42.0**	**42.5**
Comparison with Bottom Up Methods, Multi-scale [×2, ×1, ×0.5]						
AE [29]	–	65.5	–	–	40.0	32.8
HghrHRNet [8]	67.1	70.5	-	67.6	-	-
HghrHRNet+UDP [15]	–	70.5	-	68.2	-	-
HGG [19]	68.3	67.6	-	-	41.8	36.0
MIPNet (Ours, Top Down)	**76.3**	**75.7**	**68.8**	**70.0**	**42.0**	**42.5**

Table 5: Comparison with state-of-the-art methods using bounding boxes from a human detector on various datasets. Other numbers are reported from the respective publications.

bounding boxes) to 32.8 AP (6644 bounding boxes), when the detector confidence is varied from 0 to 0.9. Since HRNet is only able to provide a single output per bounding box, the average precision drops corresponding to misdetections on the occluded persons. In contrast, MIPNet maintains a relatively stable performance (drop of 1 AP) as shown in Fig. 5 for the same inputs. Since MIPNet can predict multiple instances, it can recover pose configurations for occluded persons despite misdetection of their corresponding bounding boxes. This is a desirable property afforded by the proposed MIPNet.

5. Discussions

Comparison to Two-Heads baseline: We compare MIPNet against the Two-Heads baseline which has a primary head ($\lambda = 0$) and a secondary head ($\lambda = 1$) in Table 6. To analyze the effect of head capacity in multi-instance prediction, we create two baselines: Two-Heads (*light*), and Two-Heads (*heavy*). MIPNet consistently outperforms the Two-Heads baseline on the OCHuman dataset. Please refer supplemental for more details.

Visualization with continuous λ: MIPNet's ability to predict multiple instances provides a useful tool to visualize how predictions can dynamically switch between various pose configurations. After training MIPNet using an one-hot representation of λ, during inference, we use a soft representation of $[\lambda, 1-\lambda]$ as instance-selector for the MIPNet. Fig. 6 shows how the predicted keypoints gradually shift from the foreground person to the other pose instance within the bounding box, as λ is varied from 0 to 1.

Method	#Params	COCO			OCHuman		
		AP	AP^{50}	AP^{75}	AP	AP^{50}	AP^{75}
HRNet	28.5M	76.5	93.5	83.7	63.1	79.4	69.0
Two-Heads (*light*)	28.6M	76.7	93.4	84.0	64.0	78.7	71.2
Two-Heads (*heavy*)	48.9M	77.1	94.1	**85.5**	69.8	84.5	74.9
MIPNet	28.6M	**77.6**	**94.4**	85.3	**72.5**	**89.2**	**79.4**

Table 6: Comparison with the Two-Heads baseline (*light*, *heavy*) and HRNet on the val sets using HRNet-W32 backbone with 256×192 input resolution and ground-truth bounding boxes.

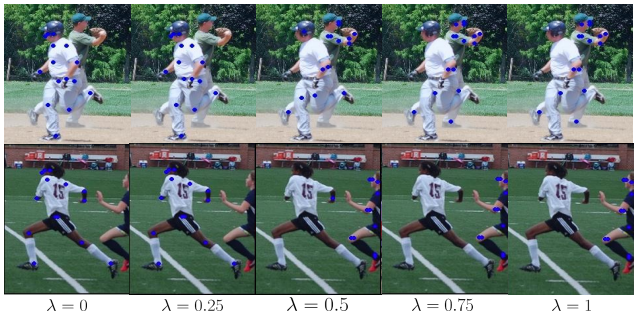

Figure 6: As λ is varied from 0 to 1 during inference, the keypoints (in blue) gradually shift from the foreground person to the other pose instance within the bounding box.

Figure 7: MIPNet fails in some cases with significant scale difference between multiple persons in the bounding box.

Limitations: In some cases, MIPNet can fail due to large difference in the scale of the various pose instances in a given bounding box, as shown in Figure 7.

6. Conclusion

Top-down 2D pose estimation methods make the key assumption of a single person within the input bounding box. While these methods have shown impressive results, the single person assumption limits their ability to perform well in crowded scenes with occlusions. Our proposed Multi-Instance Pose Network, MIPNet, enables top-down methods to predict multiple instances for a given input. MIPNet is efficient in terms of the number of additional network parameters and is stable with respect to the quality of the input bounding boxes. MIPNet achieves state-of-art results on challenging datasets with significant crowding and occlusions. We believe that the concept of predicting multiple instances is an important conceptual change and will inspire a new research direction for top-down methods.

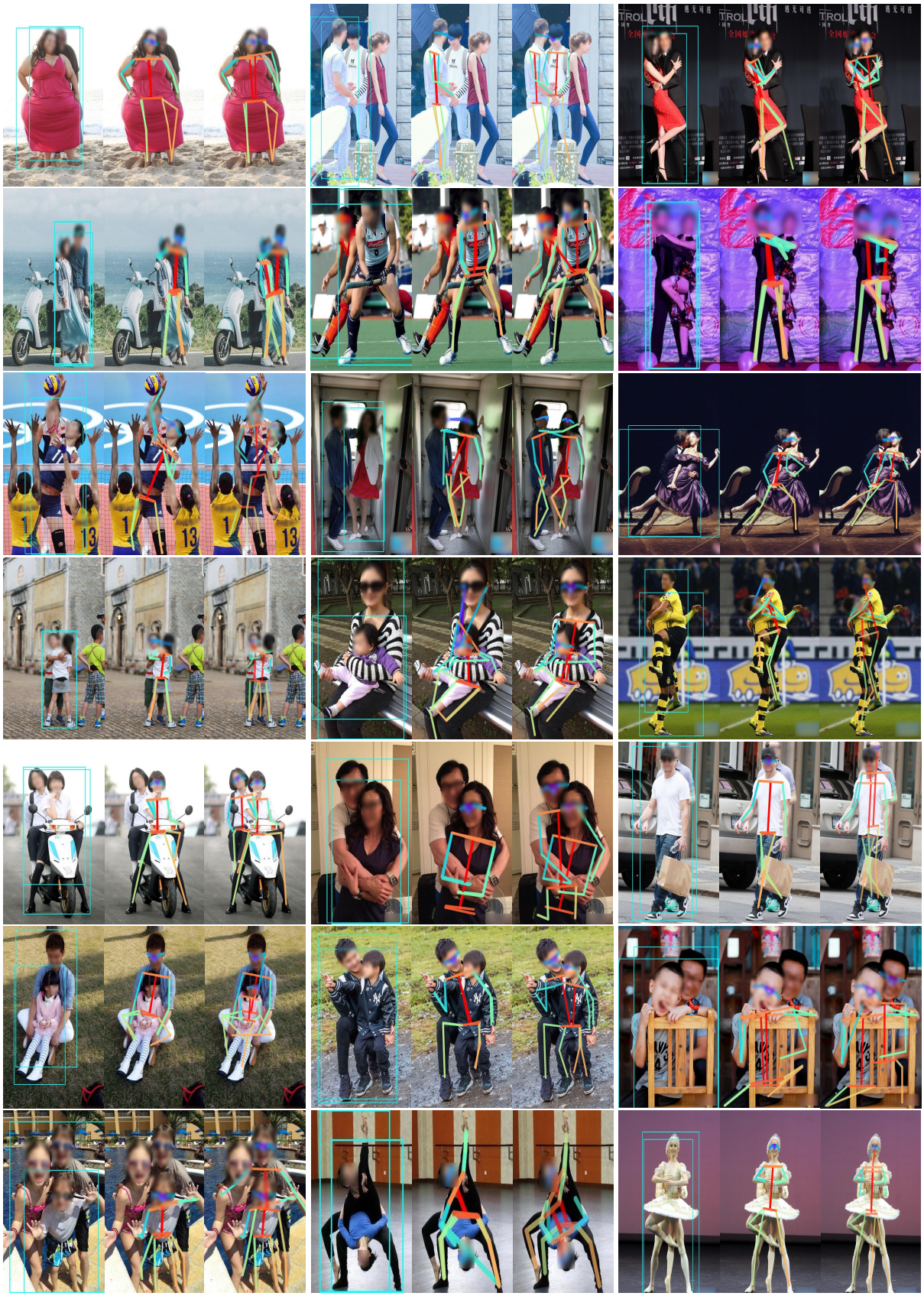

Figure 8: Qualitative results on OCHuman `val` set. Each image (left to right) shows input bounding boxes, HRNet predictions and MIPNet predictions. Due to occlusions, HRNet often misses the person in the background which is recovered by MIPNet. Please see additional results in supplemental.

References

[1] Mykhaylo Andriluka, Umar Iqbal, Eldar Insafutdinov, Leonid Pishchulin, Anton Milan, Juergen Gall, and Bernt Schiele. Posetrack: A benchmark for human pose estimation and tracking. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5167–5176, 2018. 2

[2] Mykhaylo Andriluka, Leonid Pishchulin, Peter Gehler, and Bernt Schiele. 2d human pose estimation: New benchmark and state of the art analysis. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2014. 1, 2

[3] Z Cao, T Simon, S Wei, and Y Sheikh. Realtime multi-person 2d pose estimation using part affinity fields. corr abs/1611.08050. *arXiv preprint arXiv:1611.08050*, 2016. 1, 3

[4] Marc-André Carbonneau, Veronika Cheplygina, Eric Granger, and Ghyslain Gagnon. Multiple instance learning: A survey of problem characteristics and applications. *Pattern Recognition*, 77:329–353, 2018. 3

[5] Rich Caruana. Multitask learning. *Machine learning*, 28(1):41–75, 1997. 3

[6] Yilun Chen, Zhicheng Wang, Yuxiang Peng, Zhiqiang Zhang, Gang Yu, and Jian Sun. Cascaded pyramid network for multi-person pose estimation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7103–7112, 2018. 1, 3

[7] Bowen Cheng, Yunchao Wei, Honghui Shi, Rogerio Feris, Jinjun Xiong, and Thomas Huang. Revisiting rcnn: On awakening the classification power of faster rcnn. In *Proceedings of the European conference on computer vision (ECCV)*, pages 453–468, 2018. 3

[8] 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. *arXiv preprint arXiv:1908.10357*, 2019. 1, 3, 7

[9] Xuangeng Chu, Anlin Zheng, Xiangyu Zhang, and Jian Sun. Detection in crowded scenes: One proposal, multiple predictions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12214–12223, 2020. 3

[10] Alexey Dosovitskiy and Josip Djolonga. You only train once: Loss-conditional training of deep networks. In *International Conference on Learning Representations*, 2019. 3

[11] Hao-Shu Fang, Shuqin Xie, Yu-Wing Tai, and Cewu Lu. Rmpe: Regional multi-person pose estimation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2334–2343, 2017. 3

[12] Rıza Alp Güler, Natalia Neverova, and Iasonas Kokkinos. Densepose: Dense human pose estimation in the wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7297–7306, 2018. 2

[13] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross B Girshick. Mask r-cnn. corr abs/1703.06870 (2017). *arXiv preprint arXiv:1703.06870*, 2017. 1, 3, 7

[14] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7132–7141, 2018. 4

[15] Junjie Huang, Zheng Zhu, Feng Guo, and Guan Huang. The devil is in the details: Delving into unbiased data processing for human pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5700–5709, 2020. 7

[16] Shaoli Huang, Mingming Gong, and Dacheng Tao. A coarse-fine network for keypoint localization. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3028–3037, 2017. 1, 3

[17] Eldar Insafutdinov, Leonid Pishchulin, Bjoern Andres, Mykhaylo Andriluka, and Bernt Schiele. Deepercut: A deeper, stronger, and faster multi-person pose estimation model. In *European Conference on Computer Vision*, pages 34–50. Springer, 2016. 3

[18] Umar Iqbal and Juergen Gall. Multi-person pose estimation with local joint-to-person associations. In *European Conference on Computer Vision*, pages 627–642. Springer, 2016. 3

[19] 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*, pages 718–734. Springer, 2020. 3, 6, 7

[20] Hanbyul Joo, Hao Liu, Lei Tan, Lin Gui, Bart Nabbe, Iain Matthews, Takeo Kanade, Shohei Nobuhara, and Yaser Sheikh. Panoptic studio: A massively multiview system for social motion capture. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3334–3342, 2015. 2

[21] Iasonas Kokkinos. Ubernet: Training a universal convolutional neural network for low-, mid-, and high-level vision using diverse datasets and limited memory. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6129–6138, 2017. 3

[22] Sven Kreiss, Lorenzo Bertoni, and Alexandre Alahi. Pifpaf: Composite fields for human pose estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 11977–11986, 2019. 1

[23] Jiefeng Li, Can Wang, Hao Zhu, Yihuan Mao, Hao-Shu Fang, and Cewu Lu. Crowdpose: Efficient crowded scenes pose estimation and a new benchmark. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 10863–10872, 2019. 1, 2, 3, 4, 7

[24] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2117–2125, 2017. 3

[25] 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 *European conference on computer vision*, pages 740–755. Springer, 2014. 1, 2, 4

[26] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C Berg. Ssd: Single shot multibox detector. In *European conference on computer vision*, pages 21–37. Springer, 2016. 3

[27] Kevis-Kokitsi Maninis, Ilija Radosavovic, and Iasonas Kokkinos. Attentive single-tasking of multiple tasks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1851–1860, 2019. 3

[28] Mehdi Mirza and Simon Osindero. Conditional generative adversarial nets. *arXiv preprint arXiv:1411.1784*, 2014. 3

[29] 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*, pages 2277–2287, 2017. 1, 3, 5, 7

[30] Alejandro Newell, Kaiyu Yang, and Jia Deng. Stacked hourglass networks for human pose estimation. In *European conference on computer vision*, pages 483–499. Springer, 2016. 1, 3, 5

[31] Wanli Ouyang and Xiaogang Wang. Single-pedestrian detection aided by multi-pedestrian detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3198–3205, 2013. 3

[32] George Papandreou, Tyler Zhu, Liang-Chieh Chen, Spyros Gidaris, Jonathan Tompson, and Kevin Murphy. Personlab: Person pose estimation and instance segmentation with a bottom-up, part-based, geometric embedding model. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 269–286, 2018. 1

[33] 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 *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4903–4911, 2017. 1, 3

[34] L Pishchulin, E Insafutdinov, S Tang, B Andres, M Andriluka, P Gehler, and Bb Schiele. Deepcut: Joint subset partition and labeling for multi person pose estimation.[arxiv], 2015. 3

[35] 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. *arXiv preprint arXiv:2003.10506*, 2020. 2, 3, 6, 7

[36] Joseph Redmon and Ali Farhadi. Yolov3: An incremental improvement. *arXiv preprint arXiv:1804.02767*, 2018. 3, 4

[37] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in neural information processing systems*, pages 91–99, 2015. 2, 3, 4

[38] Matteo Ruggero Ronchi and Pietro Perona. Supplementary materials for the iccv 2017 paper: Benchmarking and error diagnosis in multi-instance pose estimation. 2

[39] Russell Stewart, Mykhaylo Andriluka, and Andrew Y Ng. End-to-end people detection in crowded scenes. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2325–2333, 2016. 3

[40] Ke Sun, Bin Xiao, Dong Liu, and Jingdong Wang. Deep high-resolution representation learning for human pose estimation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5693–5703, 2019. 1, 2, 3, 4, 5, 7

[41] Xiao Sun, Jiaxiang Shang, Shuang Liang, and Yichen Wei. Compositional human pose regression. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2602–2611, 2017. 1

[42] Siyu Tang, Mykhaylo Andriluka, and Bernt Schiele. Detection and tracking of occluded people. *International Journal of Computer Vision*, 110(1):58–69, 2014. 3

[43] Jingdong Wang, Ke Sun, Tianheng Cheng, Borui Jiang, Chaorui Deng, Yang Zhao, Dong Liu, Yadong Mu, Mingkui Tan, Xinggang Wang, et al. Deep high-resolution representation learning for visual recognition. *IEEE transactions on pattern analysis and machine intelligence*, 2020. 1

[44] Bin Xiao, Haiping Wu, and Yichen Wei. Simple baselines for human pose estimation and tracking. In *Proceedings of the European conference on computer vision (ECCV)*, pages 466–481, 2018. 1, 3, 4, 5, 6, 7

[45] Makoto Yamada, Leonid Sigal, and Michalis Raptis. No bias left behind: Covariate shift adaptation for discriminative 3d pose estimation. In *European Conference on Computer Vision*, pages 674–687. Springer, 2012. 2

[46] Jun Yang. Review of multi-instance learning and its applications. *Technical report, School of Computer Science Carnegie Mellon University*, 2005. 3

[47] Amir R Zamir, Alexander Sax, William Shen, Leonidas J Guibas, Jitendra Malik, and Silvio Savarese. Taskonomy: Disentangling task transfer learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3712–3722, 2018. 3

[48] Song-Hai Zhang, Ruilong Li, Xin Dong, Paul Rosin, Zixi Cai, Xi Han, Dingcheng Yang, Haozhi Huang, and Shi-Min Hu. Pose2seg: Detection free human instance segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 889–898, 2019. 1, 2, 3, 4, 6