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

Rawal Khirodkar1 Visesh Chari2 Amit Agrawal2 Ambrish Tyagi2

1Carnegie Mellon University 2Amazon 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].

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...
this conceptual change improves the performance of top-down methods, especially for images with crowding and heavy occlusion. A naive approach to predict multiple instances per bounding box would be to add multiple predictions from the same input bounding box by varying $\lambda$ during inference.

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, $\lambda$, 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 $\lambda$ 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 $\sim 6000$ 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.

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.
During inference, we obtain the network is fully supervised and not related to multiple in-
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,

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

When $n \leq N$, 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 run-time. During inference, we vary $\lambda$ 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 $\lambda$ as inputs. The design of MIMB is inspired by the squeeze excite block of [14]. Let $X \in \mathbb{R}^{P \times Q \times C}$ be an intermediate feature map with $C$ channels, such that $X = [x_1, x_2, \ldots, x_C]$. We use an instance-selector $\lambda$ 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 $F_{sq}, F_{ex}, F_{cm}$ denote the squeeze, excite, and embed operations, respectively, within MIMB. We represent $\lambda$ as the one hot representation of scalar $\lambda$. The feature map $X$ is transformed to $X' = [x'_1, x'_2, \ldots, x'_C]$ as follows,

$$
s_c = F_{sq}(x_c),
$$

$$
e = F_{ex}(s),
$$

$$
v = F_{cm}(\lambda),
$$

$$
x'_c = (v_c \times e_c)x_c,
$$

s.t. $s = [s_1, \ldots, s_C], v = [v_1, \ldots, v_C]$ and $e = [e_1, \ldots, e_C]$. $F_{sq}$ squeezes the global spatial information into a channel descriptor using global average pooling. $F_{ex}$ allows modeling for channel-wise interactions on the output of $F_{sq}$. $F_{ex}$ 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 $\lambda$ from another simple neural network $F_{cm}$. $F_{cm}$ has a similar design to $F_{ex}$.

During inference, we vary the instance-selector $\lambda$ 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 $\lambda$ 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, $AP^{	ext{area}}$, $AP^{	ext{med}}$, and $AP^{	ext{hard}}$ 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.
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 \(\times\) 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 \(\times\) 192 or 384 \(\times\) 288, depending on the experiment. Following [29], we use data augmentation with random rotation \((-45^\circ, 45^\circ)\), 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-101 and HRNet-W32 as backbones, MIPNet improves 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.

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.

<table>
<thead>
<tr>
<th>Method</th>
<th>Arch</th>
<th>AP</th>
<th>AP50</th>
<th>AP75</th>
<th>AP40</th>
<th>AP50</th>
<th>AP75</th>
<th>AP40</th>
<th>AP50</th>
<th>AP75</th>
<th>AP40</th>
<th>AR</th>
</tr>
</thead>
<tbody>
<tr>
<td>SBL†</td>
<td>R-50</td>
<td>56.3</td>
<td>76.1</td>
<td>61.2</td>
<td>66.4</td>
<td>56.3</td>
<td>61.0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MIPNet†</td>
<td>R-50</td>
<td>64.4 (±8.1)</td>
<td>86.0</td>
<td>70.4</td>
<td>66.8</td>
<td>64.4</td>
<td>72.3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SBL†</td>
<td>R-101</td>
<td>60.5</td>
<td>77.2</td>
<td>66.6</td>
<td>68.3</td>
<td>60.5</td>
<td>64.7</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MIPNet†</td>
<td>R-101</td>
<td>68.2 (±7.7)</td>
<td>87.4</td>
<td>75.1</td>
<td>67.0</td>
<td>68.2</td>
<td>75.5</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SBL†</td>
<td>R-152</td>
<td>62.4</td>
<td>78.3</td>
<td>68.1</td>
<td>68.3</td>
<td>62.4</td>
<td>66.5</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MIPNet†</td>
<td>R-152</td>
<td>70.3 (±7.9)</td>
<td>88.6</td>
<td>77.9</td>
<td>66.9</td>
<td>70.2</td>
<td>77.0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SBL*</td>
<td>R-50</td>
<td>55.8</td>
<td>74.8</td>
<td>60.4</td>
<td>64.7</td>
<td>55.9</td>
<td>60.7</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MIPNet*</td>
<td>R-50</td>
<td>65.3 (±9.5)</td>
<td>87.5</td>
<td>72.2</td>
<td>66.0</td>
<td>66.3</td>
<td>74.1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SBL*</td>
<td>R-101</td>
<td>61.6</td>
<td>77.2</td>
<td>66.6</td>
<td>62.1</td>
<td>61.6</td>
<td>65.8</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MIPNet*</td>
<td>R-101</td>
<td>70.3 (±8.7)</td>
<td>88.4</td>
<td>77.1</td>
<td>64.1</td>
<td>70.4</td>
<td>77.7</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SBL*</td>
<td>R-152</td>
<td>64.2</td>
<td>78.3</td>
<td>69.1</td>
<td>66.5</td>
<td>64.2</td>
<td>68.1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MIPNet*</td>
<td>R-152</td>
<td>72.4 (±8.2)</td>
<td>89.5</td>
<td>79.5</td>
<td>67.7</td>
<td>72.5</td>
<td>79.6</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>HRNet†</td>
<td>H-32</td>
<td>63.1</td>
<td>79.4</td>
<td>69.0</td>
<td>64.2</td>
<td>63.1</td>
<td>67.3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MIPNet†</td>
<td>H-32</td>
<td>72.5 (±9.4)</td>
<td>89.2</td>
<td>79.4</td>
<td>65.1</td>
<td>72.6</td>
<td>79.1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>HRNet†</td>
<td>H-48</td>
<td>64.5</td>
<td>79.4</td>
<td>70.1</td>
<td>65.1</td>
<td>64.5</td>
<td>68.5</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MIPNet†</td>
<td>H-48</td>
<td>72.2 (±7.7)</td>
<td>89.5</td>
<td>78.7</td>
<td>66.5</td>
<td>72.3</td>
<td>79.2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>HRNet*</td>
<td>H-32</td>
<td>63.7</td>
<td>78.4</td>
<td>69.0</td>
<td>64.3</td>
<td>63.7</td>
<td>67.6</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MIPNet*</td>
<td>H-32</td>
<td>72.7 (±9.0)</td>
<td>89.6</td>
<td>79.6</td>
<td>66.5</td>
<td>72.7</td>
<td>79.7</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>HRNet*</td>
<td>H-48</td>
<td>65.0</td>
<td>78.4</td>
<td>70.3</td>
<td>68.4</td>
<td>65.0</td>
<td>68.8</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MIPNet*</td>
<td>H-48</td>
<td>74.1 (±9.1)</td>
<td>89.7</td>
<td>80.1</td>
<td>68.4</td>
<td>74.1</td>
<td>81.0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

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.

**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 5 shows qualitative results on several examples from OCHuman, highlighting the effectiveness 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...
## 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 $\lambda$: 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 $\lambda$, 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 $\lambda$ is varied from 0 to 1.

### 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.

### 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.

<table>
<thead>
<tr>
<th>Method</th>
<th>COCO val test</th>
<th>CrowdPose val test</th>
<th>OCHuman val test</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MaskRCNN [13]</td>
<td>- 64.8</td>
<td>- 57.2</td>
<td>- 20.2</td>
</tr>
<tr>
<td>AlphaPose [23]</td>
<td>- 70.1</td>
<td>- 61.0</td>
<td>-</td>
</tr>
<tr>
<td>JC-SPPE [23]</td>
<td>- 70.9</td>
<td>- 66.0</td>
<td>-</td>
</tr>
<tr>
<td>AlphaPose+ [35]</td>
<td>- 72.2</td>
<td>- 68.5</td>
<td>- 27.5</td>
</tr>
<tr>
<td>OPEC-Net [35]</td>
<td>- 73.9</td>
<td>- 70.6</td>
<td>- 29.1</td>
</tr>
<tr>
<td>SBL [44]</td>
<td>- 73.7</td>
<td>- 60.8</td>
<td>- 24.1</td>
</tr>
<tr>
<td>MIPNet (Ours)</td>
<td><strong>72.7</strong></td>
<td><strong>74.2</strong></td>
<td><strong>63.4</strong></td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>Method</th>
<th>#Params</th>
<th>COCO AP</th>
<th>COCO AP85</th>
<th>COCO AP85</th>
<th>OCHuman AP</th>
<th>OCHuman AP85</th>
<th>OCHuman AP85</th>
</tr>
</thead>
<tbody>
<tr>
<td>HRNet</td>
<td>28.5M</td>
<td>75.6</td>
<td>93.5</td>
<td>83.7</td>
<td>63.1</td>
<td>79.4</td>
<td>69.0</td>
</tr>
<tr>
<td>Two-Heads (light)</td>
<td>28.6M</td>
<td>76.7</td>
<td>93.4</td>
<td>84.0</td>
<td>64.0</td>
<td>78.7</td>
<td>71.2</td>
</tr>
<tr>
<td>Two-Heads (heavy)</td>
<td>48.9M</td>
<td>77.1</td>
<td>94.1</td>
<td><strong>85.5</strong></td>
<td>69.8</td>
<td>84.5</td>
<td>74.9</td>
</tr>
<tr>
<td>MIPNet</td>
<td>28.6M</td>
<td><strong>77.6</strong></td>
<td><strong>94.4</strong></td>
<td>85.3</td>
<td><strong>72.5</strong></td>
<td><strong>89.2</strong></td>
<td><strong>79.4</strong></td>
</tr>
</tbody>
</table>

![Figure 6: As $\lambda$ 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.](image1.png)

![Figure 7: MIPNet fails in some cases with significant scale difference between multiple persons in the bounding box.](image2.png)
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


