HigherHRNet: Scale-Aware Representation Learning for Bottom-Up Human Pose Estimation

Bowen Cheng[1], Bin Xiao[2], Jingdong Wang[2], Honghui Shi[1,3], Thomas S. Huang[1], Lei Zhang[2]

[1]UIUC, [2]Microsoft, [3]University of Oregon

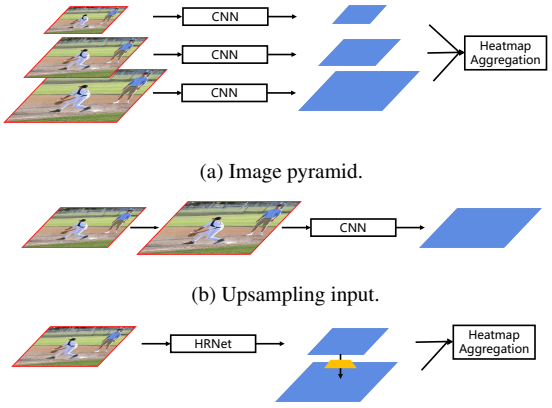
Abstract

Bottom-up human pose estimation methods have difficulties in predicting the correct pose for small persons due to challenges in scale variation. In this paper, we present ***HigherHRNet****: a novel bottom-up human pose estimation method for learning scale-aware representations using high-resolution feature pyramids. Equipped with multi-resolution supervision for training and multi-resolution aggregation for inference, the proposed approach is able to solve the scale variation challenge in* bottom-up multi-person *pose estimation and localize keypoints more precisely, especially for small person. The feature pyramid in HigherHRNet consists of feature map outputs from HRNet and upsampled higher-resolution outputs through a transposed convolution. HigherHRNet outperforms the previous best bottom-up method by* 2.5% *AP for medium person on COCO test-dev, showing its effectiveness in handling scale variation. Furthermore, HigherHRNet achieves new state-of-the-art result on COCO test-dev (*70.5% *AP) without using refinement or other post-processing techniques, surpassing all existing bottom-up methods. HigherHRNet even surpasses all top-down methods on CrowdPose test (*67.6% *AP), suggesting its robustness in crowded scene. The code and models are available at* https://github.com/HRNet/Higher-HRNet-Human-Pose-Estimation.

1. Introduction

2D human pose estimation aims at localizing human anatomical keypoints (*e.g.*, elbow, wrist, *etc.*) or parts. As a fundamental technique to human behavior understanding, it has received increasing attention in recent years.

Current human pose estimation methods can be categorized into *top-down* methods and *bottom-up* methods. Top-down methods [34, 9, 16, 42, 38, 40, 39, 16] take a dependency on person detector to detect person instances each with a bounding box and then reduce the problem to a simpler task of single person pose estimation. As top-down methods can normalize all the persons to approximately the same scale by cropping and resizing the detected person bounding boxes, they are generally less sensitive to the scale variance of persons. Thus, state-of-the-art performances on various multi-person human pose estimation benchmarks are mostly achieved by top-down methods. However, as such methods rely on a separate person detector and need to estimate pose for every person individually, they are normally computationally intensive and not truly end-to-end systems. By contrast, bottom-up methods [3, 30, 33, 22] start by localizing identity-free keypoints for all the persons in an input image through predicting heatmaps of different anatomical keypoints, followed by grouping them into person instances. This strategy effectively makes bottom-up methods faster and more capable of achieving real-time pose estimation. However, because bottom-up methods need to deal with scale variation, there still exists a large gap between the performances of bottom-up and top-down



(a) Image pyramid.

(b) Upsampling input.

(c) Our approach.

Figure 1. (a) Using image pyramid for heatmap prediction [33, 30]. (a) Generating higher resolution and spatially more accurate heatmaps by upsampling image. Recent work PersonLab [33] relies on enlarging input image size to generate high quality feature maps. (c) Our HigherHRNet uses high resolution feature pyramid.

methods, especially for small scale persons.

There are mainly two challenges in predicting keypoints of small persons. One is dealing with scale variation, *i.e.* to improve the performance of small person without sacrificing the performance of large persons. The other is generating a high-resolution heatmap with high quality for precise localizing keypoints of small persons. Previous bottom-up methods [3, 30, 33, 22] mainly focus on grouping keypoints and simply use a single resolution of feature map that is 1/4 of the input image resolution to predict the heatmap of keypoints. These methods neglect the challenge of scale variation and rely on image pyramid during inference (Figure 1 (a)). Feature pyramids are basic components for handling scale variation, however, smaller resolution feature maps in a top-down feature pyramid usually suffer from the second challenge. PersonLab [33] generates high-resolution heatmaps by increasing the input resolution (Figure 1 (b)). Although the performance of small persons increases consistently as input resolution, the performance of large persons begin decreasing when input resolution is too large. To solve these challenges, it is crucial to generate spatially more accurate and scale-aware heatmaps for bottom-up keypoint prediction in a natural and simple way without sacrificing computational cost.

In this paper, we propose a Scale-Aware High-Resolution Network (HigherHRNet) to address these challenges. HigherHRNet generates high-resolution heatmaps by a new high-resolution feature pyramid module. Unlike the traditional feature pyramid that starts from 1/32 resolution and uses bilinear upsampling with lateral connection to gradually increases feature map resolution to 1/4, high-resolution feature pyramid directly starts from 1/4 resolution which is the highest resolution feature in the backbone and generates even higher-resolution feature maps with deconvolution (Figure 1 (c)). We build the high-resolution feature pyramid on the 1/4 resolution path of HRNet [38, 40], to make it efficient. To make HigherHRNet capable of handling scale variation, we further propose a Multi-Resolution Supervision strategy to assign training target of different resolutions to the corresponding feature pyramid level. Finally, we introduce a simple Multi-Resolution Heatmap Aggregation strategy during inference to generate scale-aware high-resolution heatmaps.

We validate our method on the challenging COCO keypoint detection dataset [27] and demonstrate superior keypoint detection performance. Specifically, HigherHRNet achieves AP of 70.5% on COCO2017 test-dev *without any post processing*, outperforming all existing bottom-up methods by a large margin. Furthermore, we observe that most of the gain comes from medium person (there is no small person annotation for the keypoint detection task), HigherHRNet outperforms the previous best bottom-up method by 2.5% AP for medium persons without sacrificing the performance of large persons ($+0.3\%$ AP). This observation verifies HigherHRNet is indeed solving the scale variation challenge. We also provide a solid baseline for bottom-up methods on the new CrowdPose [24] dataset. Our HigherHRNet achieves AP of 67.6% on CrowdPose test, surpassing all existing methods. This result suggests bottom-up methods naturally have the advantages in the crowded scene.

To summarize our contributions:

- We attempt to address the scale variation challenge, which is rarely studied before in bottom-up multi-person pose estimation.
- We propose a HigherHRNet that generates high-resolution feature pyramid with multi-resolution supervision in the training stage and multi-resolution heatmap aggregation in the inference stage to predict scale-aware high-resolution heatmaps that are beneficial for small persons.
- We demonstrate the effectiveness of our HigherHRNet on the challenging COCO dataset. Our model outperforms all other bottom-up methods. We especially observe a large gain for medium persons.
- We achieve a new state-of-the-art result on the CrowdPose dataset, suggesting bottom-up methods are more robust to the crowded scene over top-down methods.

2. Related works

Top-down methods. Top-down methods [42, 38, 40, 34, 16, 18, 15, 9, 31] detect the keypoints of a single person within a person bounding box. The person bounding boxes are usually generated by an object detector [36, 26, 14, 13]. Mask R-CNN [16] directly adds a keypoint detection branch on Faster R-CNN [36] and reuses features after ROIPooling. G-RMI [34] and the following methods further break top-down methods into two steps and use separate models for person detection and pose estimation.

Bottom-up methods. Bottom-up methods [35, 19, 20, 3, 30] detect identity-free body joints for all the persons in an image and then group them into individuals. OpenPose [3] uses a two-branch multi-stage netork with one branch for heatmap prediction and one branch for grouping. OpenPose uses a grouping method named part affinity field which learns a 2D vector field linking two keypoints. Grouping is done by calculating line integral between two keypoints and group the pair with the largest integral. Newell *et al.* [30] use stacked hourglass network [31] for both heatmap prediction and grouping. Grouping is done by a method named associate embedding, which assigns each keypoint with a "tag" (a vector representation) and groups keypoints based on the l_2 distance between tag vectors. PersonLab [33] uses

dilated ResNet [17] and groups keypoints by directly learning a 2D offset field for each pair of keypoints. PifPaf [22] uses a Part Intensity Field (PIF) to localize body parts and a Part Association Field (PAF) to associate body parts with each other to form full human poses.

Feature pyramid. Pyramidal representation has been widely adopted in recent object detection and segmentation frameworks to handle scale variation. SSD [29] and MS-CNN [2] predict objects at multiple layers of the network without merging features. Feature pyramid networks [26] extend the backbone model with a top-down pathway that gradually recovers feature resolution from 1/32 to 1/4, using bilinear upsampling and lateral connection. The motivation in common is to let features from different pyramid level to predict instances of different scales. However, this pyramidal representation is less explored in bottom-up multi-person pose estimation. In this work, we design a high-resolution feature pyramid that extend the pyramid to a different direction, starting from 1/4 resolution feature and generate pyramid of features with higher resolution.

High resolution feature maps. There are mainly 4 methods to generate high resolution feature maps. (1) Encoder-decoder [31, 16, 9, 37, 1, 25, 41, 10] captures the context information in the encoder path and recover high resolution features in the decoder path. The decoder usually contains a sequence of bilinear upsample operations with skip connections from encoder features with the same resolution. (2) Dilated convolution [44, 5, 6, 7, 8, 4, 28, 43, 11, 12] (*a.k.a.* "atrous" convolution) is used to remove several stride convolutions/max poolings to preserve feature map resolution. Dilated convolution prevents losing spatial information but introduces more computational cost. (3) Deconvolution (transposed convolution) [42] is used in sequence at the end of a network to efficiently increase feature map resolution. SimpleBaseline [42] demonstrates that deconvolution can generate high quality feature maps for heatmap prediction. (4) Recently, a High-Resolution Network (HRNet) [38, 40] is proposed as an efficient way to keep a high resolution pass throughout the network. HRNet [38, 40] consists of multiple branches with different resolutions. Lower resolution branches capture contextual information and higher resolution branches preserve spatial information. With multi-scale fusions between branches, HRNet [38, 40] can generate high resolution feature maps with rich semantic.

We adopt HRNet [38, 40] as our base network to generate high-quality feature maps. And we add a deconvolution module to generate higher resolution feature maps to predict heatmaps. The resulting model is named "Scale-Aware High-Resolution Network" (HigherHRNet). As both HRNet [38, 40, 40] and deconvolution are efficient, HigherHRNet is an efficient model for generating higher resolution feature maps for heatmap prediction.

3. Higher-Resolution Network

In this section, we introduce our proposed Scale-Aware High-Resolution Representation Learning using the HigherHRNet. Figure 2 illustrates the overall architecture of our method. We will firstly give a brief overview on the proposed HigherHRNet and then describe its components in details.

3.1. HigherHRNet

HRNet. HigherHRNet uses HRNet [38, 40] (shown in Figure 2) as backbone. HRNet [38, 40] starts with a high-resolution branch in the first stage. In every following stage, a new branch is added to current branches in parallel with $\frac{1}{2}$ of the lowest resolution in current branches. As the network has more stages, it will have more parallel branches with different resolutions and resolutions from previous stages are all preserved in later stages. An example network structure, containing 3 parallel branches, is illustrated in Figure 2.

We instantiate the backbone using a similar manner as HRNet [38, 40]. The network starts from a stem that consists of two strided 3×3 convolutions decreasing the resolution to $1/4$. The 1st stage contains 4 residual units where each unit is formed by a bottleneck with width (number of channels) 64, followed by one 3×3 convolution reducing the width of feature maps to C. The 2nd, 3rd, 4th stages contain 1, 4, and 3 multi-resolution blocks, respectively. The widths of the convolutions of the four resolutions are C, $2C$, $4C$, and $8C$, respectively. Each branch in the multi-resolution group convolution has 4 residual units and each unit has two 3×3 convolutions in each resolution. We experiment with two networks with different capacity by setting C to 32 and 48 respectively.

HRNet [38, 40] was originally designed for top-down pose estimation. In this work, we adopt HRNet [38, 40] to a bottom-up method by adding a 1×1 convolution to predict heatmaps and tagmaps similar to [30]. We only use the highest resolution ($\frac{1}{4}$ of the input image) feature maps for prediction. Following [30], we use a scalar tag for each keypoint.

HigherHRNet. Resolution of the heatmap is important for predicting keypoints for small persons. Most existing human pose estimation methods predict Gaussian-smoothed heatmaps by preparing the ground truth headmaps with an unnormalized Gaussian kernel applyed to each keypoint location. Adding this Gaussian kernel helps training networks as CNNs tend to output spatially smooth responses as a nature of convolution operations. However, applying a Gaussian kernel also introduces confusion in precise localization of keypoints, especially for keypoints belonging to small persons. A trivial solution to reduce this confusion is to reduce the standard deviation of the Gaus-

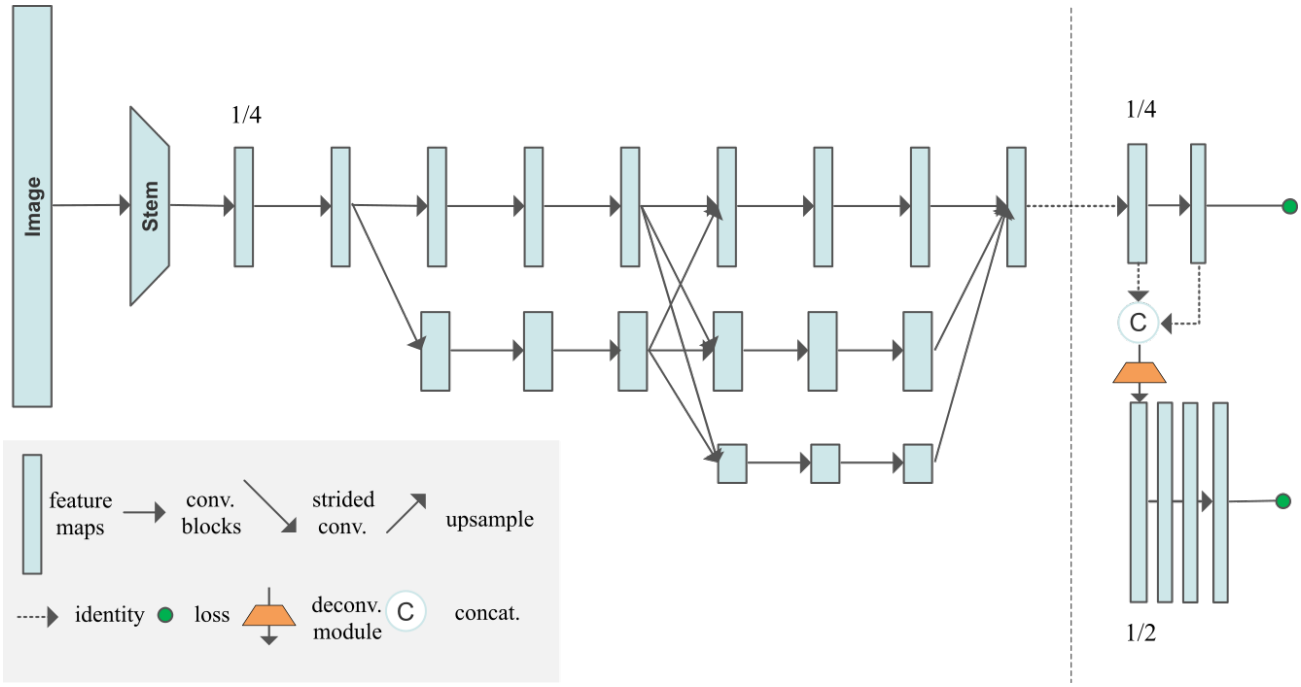


Figure 2. An illustration of HigherHRNet. The network uses HRNet [38, 40] as backbone, followed by one or more deconvolution modules to generate multi-resolution and high-resolution heatmaps. Multi-resolution supervision is used for training. More details are given in Section 3.

sian kernel. However, we empirically find that it makes optimization harder and leads to even worse results.

Instead of reducing standard deviation, we solve this problem by predicting heatmaps at higher resolution with standard deviation unchanged at different resolutions. Bottom-up methods usually predict heatmaps at resolution $\frac{1}{4}$ of the input image. Yet we find this resolution is not high enough for predicting accurate heatmaps. Inspired by [42], which shows that deconvolution can be used to effectively generate high quality and high resolution feature maps, we build HigherHRNet on top of the highest resolution feature maps in HRNet as shown in Figure 2 by adding a deconvolution module as discussed in Section 3.3.

The deconvolution module takes as input both features and predicted heatmaps from HRNet and generates new feature maps that are 2 times larger in resolution than the input feature maps. A feature pyramid with two resolutions is thus generated by the deconvolution module together with the feature maps from HRNet. The deconvolution module also predicts heatmaps by adding an extra 1×1 convolution. We follow Section 3.4 to train heatmap predictors at different resolutions and use a heatmap aggregation strategy as described in (Section 3.5) for inference.

More deconvolution modules can be added if larger resolution is desired. We find the number of deconvolution modules is dependent on the distribution of person scales of the dataset. Generally speaking, a dataset containing smaller persons requires larger resolution feature maps for prediction and vice versa. In experiments, we find adding a single deconvolution module achieves the best performance on the COCO dataset.

3.2. Grouping.

Recent works [30, 23] have shown that grouping can be solved with high accuracy by a simple method using associative embedding [30]. As an evidence, experimental results in [30] show that using the ground truth detections with the predicted tags improves AP from 59.2 to 94.0 on a held-out set of 500 training images of the COCO keypoint detection dataset [27]. We follow [30] to use associative embedding for keypoint grouping. The grouping process clusters identity-free keypoints into individuals by grouping keypoints whose tags have small l_2 distance.

3.3. Deconvolution Module

We propose a simple deconvolution module for generating high quality feature maps whose resolution is two times higher than the input feature maps. Following [42], we use a 4×4 deconvolution (*a.k.a.* transposed convolution) followed by BatchNorm and ReLU to learn to upsample the input feature maps. Optionally, we could further add several Basic Residual Blocks [17] after deconvolution to refine the upsampled feature maps. We add 4 Residual Blocks in HigherHRNet.

Different from [42], the input to our deconvolution module is the concatenation of the feature maps and the pre-

dicted heatmaps from either HRNet or previous deconvolution modules. And the output feature maps of each deconvolution module are also used to predict heatmaps in a multi-scale fashion.

3.4. Multi-Resolution Supervision

Unlike other bottom-up methods [30, 33, 3] that only apply supervision to the largest resolution heatmaps, we introduce a multi-resolution supervision during training to handle scale variation. We transform ground truth keypoint locations to locations on the heatmaps of all resolutions to generate ground truth heatmaps with different resolutions. Then we apply a Gaussian kernel *with the same standard deviation* (we use standard deviation $= 2$ by default) to all these ground truth heatmaps. We find it important not to scale standard deviation of the Gaussian kernel. This is because different resolution of feature pyramid is suitable to predict keypoints of different scales. On higher-resolution feature maps, a relatively small standard deviation (compared to the resolution of the feature map) is desired to more precisely localize keypoints of small persons.

At each prediction scale in HigherHRNet, we calculate the mean squared error between the predicted heatmaps of that scale and its associated ground truth heatmaps. The final loss for heatmaps is the sum of mean squared errors for all resolutions.

It is worth highlighting that we do not assign different scale of persons to different levels in the feature pyramid, due to the following reasons. First, the heuristic used for assigning training target depends on both the dataset and network architecture. It is hard to transform the heuristic for FPN [26] to HigherHRNet as both the dataset (scale distribution of person *v.s.* all objects) and architecture (HigherHRNet only has 2 levels of pyramid while FPN has 4) change. Second, ground truth keypoint targets interact with each other since we apply the Gaussian kernel. Thus, it is very hard to decouple keypoints by simply setting ignored regions. We believe model has the ability to automatically focus on specific scales in different levels of the feature pyramid.

Tagmaps are trained differently from heatmaps in HigherHRNet. We only predict tagmaps at the lowest resolution, instead of using all resolutions. This is because learning tagmaps requires global reasoning and it is more suitable to predict tagmaps in lower resolution. Empirically, we also find higher resolutions do not learn to predict tagmaps well and even do not converge. Thus, we follow [30] to train the tagmaps on feature maps at $\frac{1}{4}$ resolution of input image.

3.5. Heatmap Aggregation for Inference

We propose a heatmap aggregation strategy during inference. We use bilinear interpolation to upsample all the predicted heatmaps with different resolutions to the resolution of the input image and average the heatmaps from all scales for final prediction. This strategy is quite different from previous methods [3, 30, 33] which only use heatmaps from a single scale or single stage for prediction.

The reason that we use heatmap aggregation is to enable scale-aware pose estimation. For example, the COCO Keypoint dataset [27] contains persons of large scale variance from 32^2 pixels to more than 128^2 pixels. Top-down methods [34, 9, 42] solve this problem by normalizing person regions approximately into a single scale. However, bottom-up methods need to be aware of scales to detect keypoints from all scales. We find heatmaps from different scales in HigherHRNet captures keypoints with different scales better. For example, keypoints for small persons missed in lower-resolution heatmap can be recovered in the higher-resolution heatmap. Thus, averaging predicted heatmaps from different resolutions makes HigherHRNet a scale-aware pose estimator.

4. Experiments

4.1. COCO Keypoint Detection

Dataset. The COCO dataset [27] contains over $200,000$ images and $250,000$ person instances labeled with 17 keypoints. COCO is divided into *train*/*val*/*test-dev* sets with 57k, 5k and 20k images respectively. All the experiments in this paper are trained only on the *train* set. We report results on the *val* set for ablation studies and compare with other state-of-the-art methods on the *test-dev* set.

Evaluation metric. The standard evaluation metric is based on Object Keypoint Similarity (OKS): $\text{OKS} = \frac{\sum_i \exp(-d_i^2/2s^2k_i^2)\delta(v_i>0)}{\sum_i \delta(v_i>0)}$. Here d_i is the Euclidean distance between a detected keypoint and its corresponding ground truth, v_i is the visibility flag of the ground truth, s is the object scale, and k_i is a per-keypoint constant that controls falloff. We report standard average precision and recall scores[1]: AP^{50} (AP at $\text{OKS} = 0.50$), AP^{75}, AP (the mean of AP scores at $\text{OKS} = 0.50, 0.55, \ldots, 0.90, 0.95$), AP^M for medium objects, AP^L for large objects, and AR (the mean of recalls at $\text{OKS} = 0.50, 0.55, \ldots, 0.90, 0.95$).

Training. Following [30], we use data augmentation with random rotation ($[-30°, 30°]$), random scale ($[0.75, 1.5]$), random translation ($[-40, 40]$) to crop an input image patch of size 512×512 as well as random flip. As mentioned in Section 3.4, we generate two ground truth heatmaps with resolutions 128×128 and 256×256 respectively.

We use the Adam optimizer [21]. The base learning rate is set to $1e-3$, and dropped to $1e-4$ and $1e-5$ at the $200th$ and $260th$ epochs respectively. We train the model for a total of 300 epochs. To balance the heatmap loss and the grouping loss, we set the weight to 1 and $1e-3$ respectively for the two losses.

[1] http://cocodataset.org/#keypoints-eval

Method	Backbone	Input size	#Params	GFLOPs	AP	AP^{50}	AP^{75}	AP^{M}	AP^{L}
w/o multi-scale test									
OpenPose [3]†	-	-	-	-	61.8	84.9	67.5	57.1	68.2
Hourglass [30]	Hourglass	512	277.8M	206.9	56.6	81.8	61.8	49.8	67.0
PersonLab [33]	ResNet-152	1401	68.7M	405.5	66.5	88.0	72.6	62.4	72.3
PifPaf [22]	-	-	-	-	66.7	-	-	62.4	72.9
Bottom-up HRNet‡	HRNet-W32	512	28.5M	38.9	64.1	86.3	70.4	57.4	73.9
HigherHRNet (Ours)	HRNet-W32	512	28.6M	47.9	66.4	87.5	72.8	61.2	74.2
HigherHRNet (Ours)	HRNet-W48	640	63.8M	154.3	**68.4**	**88.2**	**75.1**	**64.4**	**74.2**
w/ multi-scale test									
Hourglass [30]	Hourglass	512	277.8M	206.9	63.0	85.7	68.9	58.0	70.4
Hourglass [30]†	Hourglass	512	277.8M	206.9	65.5	86.8	72.3	60.6	72.6
PersonLab [33]	ResNet-152	1401	68.7M	405.5	68.7	89.0	75.4	64.1	75.5
HigherHRNet (Ours)	HRNet-W48	640	63.8M	154.3	**70.5**	**89.3**	**77.2**	**66.6**	**75.8**

† Indicates using refinement.
‡ Our implementation, not reported in [38, 40]

Table 1. Comparisons with bottom-up methods on the **COCO2017 test-dev** set. All GFLOPs are calculated at single-scale. For PersonLab [33], we only calculate its backbone's #Params and GFLOPs. Top: w/o multi-scale test. Bottom: w/ multi-scale test. *It is worth noting that our results are achieved without refinement.*

Method	AP	AP^{50}	AP^{75}	AP^{M}	AP^{L}	AR
Top-down methods						
Mask-RCNN [16]	63.1	87.3	68.7	57.8	71.4	-
G-RMI [34]	64.9	85.5	71.3	62.3	70.0	69.7
Integral Pose Regression [39]	67.8	88.2	74.8	63.9	74.0	-
G-RMI + extra data [34]	68.5	87.1	75.5	65.8	73.3	73.3
CPN [9]	72.1	91.4	80.0	68.7	77.2	78.5
RMPE [15]	72.3	89.2	79.1	68.0	78.6	-
CFN [18]	72.6	86.1	69.7	78.3	64.1	-
CPN (ensemble) [9]	73.0	91.7	80.9	69.5	78.1	79.0
SimpleBaseline [42]	73.7	91.9	81.1	70.3	80.0	79.0
HRNet-W48 [38, 40]	75.5	92.5	83.3	71.9	81.5	80.5
HRNet-W48 + extra data [38, 40]	77.0	92.7	84.5	73.4	83.1	82.0
Bottom-up methods						
OpenPose* [3]	61.8	84.9	67.5	57.1	68.2	66.5
Hourglass*+ [30]	65.5	86.8	72.3	60.6	72.6	70.2
PifPaf [22]	66.7	-	-	62.4	72.9	-
SPM [32]	66.9	88.5	72.9	62.6	73.1	-
PersonLab+ [33]	68.7	89.0	75.4	64.1	75.5	75.4
Ours: HigherHRNet-W48+	70.5	89.3	77.2	66.6	75.8	74.9

Table 2. Comparisons with both top-down and bottom-up methods on **COCO2017 test-dev** dataset. * means using refinement. + means using multi-scale test.

Method	Feat. stride/resolution	AP	AP^{M}	AP^{L}
HRNet	4/128	64.4	57.1	75.6
HigherHRNet	2/256	66.9	61.0	75.7
HigherHRNet	1/512	66.5	61.1	74.9

Table 3. Ablation study of HRNet *vs*. HigherHRNet on **COCO2017 val** dataset. Using one deconvolution module for HigherHRNet performs best on the COCO dataset.

Testing. We first resize the short side of the input image to 512 and keep the aspect ratio. Heatmap aggregation is done by resizing all the predicted heatmaps to the size of input image and taking the average. Following [30], flip testing is used for all the experiments. All reported numbers have been obtained with single model without ensembling.

Results on COCO2017 test-dev. Table 1 summarizes the results on COCO2017 test-dev dataset. From the results, we can see that using HRNet [38, 40] itself already serves as a simple and strong baseline for bottom-up methods (64.1 AP). Our baseline method of HRNet with only single scale test outperforms Hourglass [30] using multi-scale test, while HRNet has much less parameters and computation in terms of FLOPs. Equipped with light-weight deconvolution modules, our proposed HigherHRNet (66.4 AP) outperforms HRNet by +2.3 AP with only marginal increase in parameters (+0.4%) and FLOPs (+23.1%). HigherHRNet is comparable with PersonLab [33] but with only 50% parameters and 11% FLOPs. If we further use multi-scale test, our HigherHRNet achieves 70.5 AP, outperforming all existing bottom-up methods by a large margin. We do not use any post processing like refining with top-down methods in [3, 30].

Table 2 lists both bottom-up and top-down methods on the COCO2017 test-dev dataset. HigherHRNet further closes the performance gap between bottom-up and top-down methods.

4.2. Ablation Experiments

We perform a number of ablation experiments to analyze Scale-Aware High-Resolution Network (HigherHRNet) on the COCO2017 [27] val dataset.

HRNet *vs*. HigherHRNet. We perform ablation study comparing HRNet and HigherHRNet. For HigherHRNet, deconvolution module without extra residual blocks is used, and heatmaps aggregation is used for inference. Results are shown in Table 3. A simple bottom-up baseline by using HRNet with a feature stride of 4 achieves AP = 64.4. By adding one deconvolution module, our HigherHRNet with a feature stride of 2 outperforms HRNet by a large margin of +2.5 AP (achieving 66.9 AP). Furthermore, the main

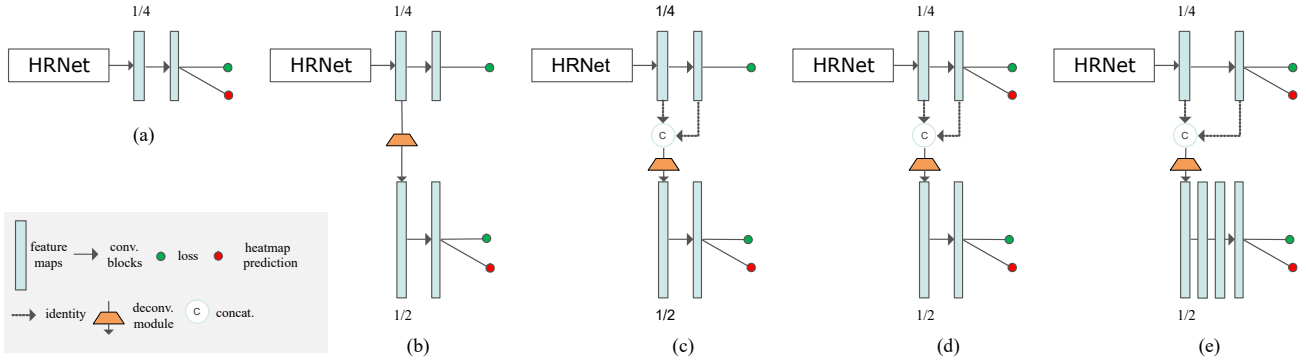


Figure 3. (a) Baseline method using HRNet [38, 40] as backbone. (b) HigherHRNet with multi-resolution supervision (MRS). (c) HigherHRNet with MRS and feature concatenation. (d) HigherHRNet with MRS and feature concatenation. (e) HigherHRNet with MRS, feature concatennation and extra residual blocks. For (d) and (e), heatmap aggregation is used.

	Network	w/ MRS	feature concat.	w/ heatmap aggregation	extra res. blocks	AP	AP^M	AP^L
(a)	HRNet					64.4	57.1	75.6
(b)	HigherHRNet	✓				66.0	60.7	74.2
(c)	HigherHRNet	✓	✓			66.3	60.8	74.0
(d)	HigherHRNet	✓	✓	✓		66.9	61.0	75.7
(e)	HigherHRNet	✓	✓	✓	✓	67.1	61.5	76.1

Table 4. Ablation study of HigherHRNet's components on **COCO2017 val** dataset. MSR: multi-resolution supervision. feature concat.: feature concatenation. res. blocks: residual blocks.

improvement comes from medium persons, where AP^M is improved from 57.1 for HRNet to 61.0 for HigherHRNet.

These results show that HigherHRNet performs much better with small scales thanks to its higher resolution heatmaps. We also find the AP for large person pose does no drop. This is mainly because we also use smaller resolution heatmaps for prediction. It demonstrates that 1) making prediction at higher resolution is beneficial to bottom-up pose estimation and 2) scale-aware prediction is important.

If we add a sequence of two deconvolution modules after HRNet to generate feature maps that is of the same resolution as the input image, we observe that the performance decreases to 66.5 AP from 66.9 AP for adding only one deconvolution module. The improvement for medium person is marginal (+0.1 AP) but there is a large drop in the performance of large person (−0.8 AP). We hypothesize this is because the misalignment between feature map scale and object scales. Larger resolution feature maps (feature stride $= 1$) are good for detecting keypoints from even smaller persons but the small persons in COCO are not considered for pose estimation. Therefore, we only use one deconvolution module by default for the COCO dataset. But we would like to point out that the number of cascaded deconvolution modules should be dependent on datasets and we will validate this on more datasets in our future work.

HigherHRNet gain breakdown. To better understand the gain of the proposed components, we perform detailed ablation studies on each individual component. Figure 3 illustrates all the architectures of our experiments. Results are shown in Table 4.

Effect of deconvolution module. We perform ablation study on the effect of adding deconvolution module to generate higher resolution heatmaps. For a fair comparison, we only use the highest resolution feature maps to generate heatmaps for prediction (Figure 3 (b)). HRNet (Figure 3 (a)) achieves a baseline of 64.4 AP. By adding one deconvolution module, the model achieves 66.0 AP which is 1.6 AP better than the baseline. This improvement is completely due to predicting on larger feature maps with higher quality. The result verifies our claim that it is important to predict on higher resolution feature maps for bottom-up pose estimation.

Effect of feature concatenation. We concatenate feature maps with predicted heatmaps from HRNet as input to the deconvolution module (Figure 3 (c)) and the performance is further improved to 66.3 AP. We also observe there is a large gain in medium persons while the performance for large persons decreases. Comparing method (a) and (c), the gain of predicting heatmaps at higher resolution mainly comes from medium persons ($+3.7AP^M$). Moreover, the drop in large persons (−1.6 AP) justifies our claim that different different resolutions of feature maps are sensitive to different scales of persons.

Effect of heatmap aggregation. We further use all resolutions of heatmaps following the heatmap aggregation strategy for inference (Figure 3 (d)). Compared with Fig-

Training size	AP	AP^M	AP^L
512	67.1	61.5	76.1
640	68.5	64.3	75.3
768	68.5	64.9	73.8

Table 5. Ablation study of HigherHRNet with different training image size on **COCO2017 val** dataset.

Backbone	#Params	GFLOPs	AP	AP^M	AP^L
HRNet-W32	28.6	47.8	68.5	64.3	75.3
HRNet-W40	44.5	110.7	69.2	64.9	75.9
HRNet-W48	63.8	154.3	69.9	65.4	76.4

Table 6. Ablation study of HigherHRNet with different backbone on **COCO2017 val** dataset.

ure 3 (c) (66.3 AP) that only uses the highest resolution heatmaps for inference, applying heatmap aggregation strategy achieves 66.9 AP. Comparing method (d) and (e), the gain of heatmap aggregation comes from large person (+1.7 AP). And the performance of large person is even marginally better than predicting at lower resolution (method (a)). It means that predicting heatmaps using heatmap aggregation strategy is truly scale-aware.

Effect of extra residual blocks. We add 4 residual blocks in the deconvolution module and our best model achieves 67.1 AP. Adding residual blocks can further refine the feature maps and it increases AP for both medium and large persons equally.

Training with larger image size. A natural question is can training with larger input size further improve performance? To answer this question, we train HigherHRNet with 640×640 and 768×768 and results are shown in Table 5, all three models are tested using the training image size. We find that by increasing training image size to 640, there is a significant gain of 1.4 AP. Most of the gain comes from medium person while the performance of large person degrades slightly. When we further change the training image size to 768, the overall AP does not change anymore. We observe a marginal improvement in medium person along with large degradation in large person.

Larger backbone. In previous experiments, we use HRNet-W32 (1/4 resolution feature map has 32 channels) as backbone. We perform experiments with larger backbones HRNet-W40 and HRNet-W48. Results are shown in Table 6. We find using larger backbone consistently improves performance for both medium and large person.

4.3. CrowdPose

The CrowdPose [24] dataset consists of 20,000 images, containing about 80,000 persons. The training, validation and testing subset are split in proportional to 5:1:4. CrowdPose has more crowded scenes than the COCO keypoint dataset, posing more challenges to pose estimation methods. The evaluation metric is the same as COCO [27].

Method	AP	AP^{50}	AP^{75}	AP^E	AP^M	AP^H
Top-down methods						
Mask-RCNN [16]	57.2	83.5	60.3	69.4	57.9	45.8
AlphaPose [15]	61.0	81.3	66.0	71.2	61.4	51.1
Top-down with refinement						
SPPE [24]	66.0	84.2	71.5	75.5	66.3	57.4
Bottom-up methods						
OpenPose [3]	-	-	-	62.7	48.7	32.3
Ours: HigherHRNet-W48	65.9	86.4	70.6	73.3	66.5	57.9
Ours: HigherHRNet-W48$^+$	67.6	87.4	72.6	75.8	68.1	58.9

Table 7. Comparisons with both top-down and bottom-up methods on **CrowdPose test** dataset. Superscripts E, M, H of AP stand for easy, medium and hard. $^+$ means using multi-scale test.

The strong assumption of top-down methods that each person detection only contains a single person in the center, is no more valid in crowded scene. As shown in Table 7, top-down methods [16, 15] that perform well on COCO fail on the CrowdPose dataset.

On the other hand, bottom-up methods naturally have the advantage in crowded scene. To validate the robustness of HigherHRNet in crowded scene, as well as setting up a strong baseline for bottom-up methods. We train our best HigherHRNet-W48 model on the CrowdPose *train and val set* and report performance on the *test* set. All training parameters follow COCO exactly and we use a crop size of 640×640 for both training and testing.

Results are shown in Table 7. Our HigherHRNet outperforms naïve top-down methods by a large margin of 6.6 AP. HigherHRNet also outperforms the previous best method [24] (which performs a global refinement of top-down method [15]) by a healthy margin of 1.6 AP and most of the gain comes from AP^M (+1.8 AP) and AP^H (+1.5 AP), which contains images with the most crowd. Even without multi-scale test, HigherHRNet outperforms SPPE [24] by 0.5 in AP^H.

5. Conclusion

We have presented a Scale-Aware High-Resolution Network (HigherHRNet) to solve the scale variation challenge in the bottom-up multi-person pose estimation problem, especially for precisely localizing keypoints of small persons. We find multi-scale image pyramid and larger input size partially solve the problem, but these methods suffer from high computational cost. To solve the problem, we propose an efficient high-resolution feature pyramid based on HRNet and train it with multi-resolution supervision. During the inference, HigherHRNet with multi-resolution heatmap aggregation is capable of efficiently generating muilt- and higher-resolution heatmaps for more accurate human pose estimation. HigherHRNet outperforms all existing bottom-up methods by a large margin on the challenging COCO dataset, especially for small persons.

References

[1] Vijay Badrinarayanan, Alex Kendall, and Roberto Cipolla. Segnet: A deep convolutional encoder-decoder architecture for image segmentation. *IEEE TPAMI*, 2017. 3

[2] Zhaowei Cai, Quanfu Fan, Rogerio S Feris, and Nuno Vasconcelos. A unified multi-scale deep convolutional neural network for fast object detection. In *ECCV*, 2016. 3

[3] 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, 5, 6, 8

[4] Liang-Chieh Chen, Maxwell D. Collins, Yukun Zhu, George Papandreou, Barret Zoph, Florian Schroff, Hartwig Adam, and Jonathon Shlens. Searching for efficient multi-scale architectures for dense image prediction. In *NeurIPS*, 2018. 3

[5] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. Semantic image segmentation with deep convolutional nets and fully connected crfs. 2015. 3

[6] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE TPAMI*, 2018. 3

[7] Liang-Chieh Chen, George Papandreou, Florian Schroff, and Hartwig Adam. Rethinking atrous convolution for semantic image segmentation. *arXiv preprint arXiv:1706.05587*, 2017. 3

[8] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *ECCV*, 2018. 3

[9] Yilun Chen, Zhicheng Wang, Yuxiang Peng, Zhiqiang Zhang, Gang Yu, and Jian Sun. Cascaded pyramid network for multi-person pose estimation. In *CVPR*, 2018. 1, 2, 3, 5, 6

[10] Bowen Cheng, Liang-Chieh Chen, Yunchao Wei, Yukun Zhu, Zilong Huang, Jinjun Xiong, Thomas S Huang, Wen-Mei Hwu, and Honghui Shi. Spgnet: Semantic prediction guidance for scene parsing. In *ICCV*, 2019. 3

[11] Bowen Cheng, Maxwell D Collins, Yukun Zhu, Ting Liu, Thomas S Huang, Hartwig Adam, and Liang-Chieh Chen. Panoptic-deeplab. *arXiv preprint arXiv:1910.04751*, 2019. 3

[12] Bowen Cheng, Maxwell D Collins, Yukun Zhu, Ting Liu, Thomas S Huang, Hartwig Adam, and Liang-Chieh Chen. Panoptic-deeplab: A simple, strong, and fast baseline for bottom-up panoptic segmentation. In *CVPR*, 2020. 3

[13] Bowen Cheng, Yunchao Wei, Honghui Shi, Rogerio Feris, Jinjun Xiong, and Thomas Huang. Decoupled classification refinement: Hard false positive suppression for object detection. *arXiv preprint arXiv:1810.04002*, 2018. 2

[14] Bowen Cheng, Yunchao Wei, Honghui Shi, Rogerio Feris, Jinjun Xiong, and Thomas Huang. Revisiting rcnn: On awakening the classification power of faster rcnn. In *ECCV*, 2018. 2

[15] Hao-Shu Fang, Shuqin Xie, Yu-Wing Tai, and Cewu Lu. Rmpe: Regional multi-person pose estimation. In *ICCV*, 2017. 2, 6, 8

[16] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *ICCV*, 2017. 1, 2, 3, 6, 8

[17] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 3, 4

[18] Shaoli Huang, Mingming Gong, and Dacheng Tao. A coarse-fine network for keypoint localization. In *ICCV*, 2017. 2, 6

[19] Eldar Insafutdinov, Leonid Pishchulin, Bjoern Andres, Mykhaylo Andriluka, and Bernt Schiele. Deepercut: A deeper, stronger, and faster multi-person pose estimation model. In *ECCV*, 2016. 2

[20] Umar Iqbal and Juergen Gall. Multi-person pose estimation with local joint-to-person associations. In *ECCV*, 2016. 2

[21] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 5

[22] Sven Kreiss, Lorenzo Bertoni, and Alexandre Alahi. Pifpaf: Composite fields for human pose estimation. In *CVPR*, 2019. 1, 2, 3, 6

[23] Hei Law and Jia Deng. Cornernet: Detecting objects as paired keypoints. In *ECCV*, 2018. 4

[24] 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, 8

[25] Guosheng Lin, Anton Milan, Chunhua Shen, and Ian Reid. Refinenet: Multi-path refinement networks for high-resolution semantic segmentation. In *CVPR*, 2017. 3

[26] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *CVPR*, 2017. 2, 3, 5

[27] 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, 4, 5, 6, 8

[28] Chenxi Liu, Liang-Chieh Chen, Florian Schroff, Hartwig Adam, Wei Hua, Alan Yuille, and Li Fei-Fei. Auto-deeplab: Hierarchical neural architecture search for semantic image segmentation. In *CVPR*, 2019. 3

[29] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C Berg. Ssd: Single shot multibox detector. In *ECCV*, 2016. 3

[30] Alejandro Newell, Zhiao Huang, and Jia Deng. Associative embedding: End-to-end learning for joint detection and grouping. In *NeurIPS*. 2017. 1, 2, 3, 4, 5, 6

[31] Alejandro Newell, Kaiyu Yang, and Jia Deng. Stacked hourglass networks for human pose estimation. In *ECCV*, 2016. 2, 3

[32] Xuecheng Nie, Jiashi Feng, Jianfeng Zhang, and Shuicheng Yan. Single-stage multi-person pose machines. In *ICCV*, 2019. 6

[33] George Papandreou, Tyler Zhu, Liang chieh Chen, Spyros Gidaris, Jonathan Tompson, and Kevin Murphy. Personlab: Person pose estimation and instance segmentation with a part-based geometric embedding model. In *ECCV*, 2018. 1, 2, 5, 6

[34] 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, 5, 6

[35] 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. 2

[36] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *NeurIPS*, 2015. 2

[37] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *MICCAI*, 2015. 3

[38] Ke Sun, Bin Xiao, Dong Liu, and Jingdong Wang. Deep high-resolution representation learning for human pose estimation. In *CVPR*, 2019. 1, 2, 3, 4, 6, 7

[39] Xiao Sun, Bin Xiao, Fangyin Wei, Shuang Liang, and Yichen Wei. Integral human pose regression. In *ECCV*, 2018. 1, 6

[40] Jingdong Wang, Ke Sun, Tianheng Cheng, Borui Jiang, Chaorui Deng, Yang Zhao, Dong Liu, Yadong Mu, Mingkui Tan, Xinggang Wang, Wenyu Liu, and Bin Xiao. Deep high-resolution representation learning for visual recognition. *CoRR*, abs/1908.07919, 2019. 1, 2, 3, 4, 6, 7

[41] Zbigniew Wojna, Vittorio Ferrari, Sergio Guadarrama, Nathan Silberman, Liang-Chieh Chen, Alireza Fathi, and Jasper Uijlings. The devil is in the decoder. In *BMVC*, 2017. 3

[42] Bin Xiao, Haiping Wu, and Yichen Wei. Simple baselines for human pose estimation and tracking. In *ECCV*, 2018. 1, 2, 3, 4, 5, 6

[43] Tien-Ju Yang, Maxwell D Collins, Yukun Zhu, Jyh-Jing Hwang, Ting Liu, Xiao Zhang, Vivienne Sze, George Papandreou, and Liang-Chieh Chen. Deeperlab: Single-shot image parser. *arXiv preprint arXiv:1902.05093*, 2019. 3

[44] Fisher Yu and Vladlen Koltun. Multi-scale context aggregation by dilated convolutions. *arXiv preprint arXiv:1511.07122*, 2015. 3