Joint COCO and Mapillary Workshop at ICCV 2019: COCO Keypoint Detection Challenge Track

Technical Report: HigherHRNet

Bowen Cheng[1], Bin Xiao[2], Jingdong Wang[3], Honghui Shi[1,4], Thomas S. Huang[1], Lei Zhang[3]

[1]UIUC, [2]ByteDance, [3]Microsoft, [4]University of Oregon

1. Introduction

In this paper, we present a Higher-Resolution Network (HigherHRNet) for generating spatially more accurate and scale-aware heatmaps for *bottom-up* multi-person pose estimation. HigherHRNet is an extention of High-Resolution Network (HRNet) [14, 16], by simply adding one or more deconvolution modules. Furthermore, HigherHRNet is naturally equipped with high quality multi-resolution heatmaps that can be used for heatmap aggegation with small computation overhead. We demonstrate superior keypoint detection performance on the COCO keypoint detection dataset [9]. Specifically, HigherHRNet achieves AP of 70.5 on COCO2017 test-dev, outperforming all existing bottom-up methods by a large margin. The code will be released at https://github.com/HRNet/Higher-HRNet-Human-Pose-Estimation.

2. Higher-Resolution Network

In this section, we introduce our proposed Higher-Resolution Network (HigherHRNet). Figure 1 illustrates the overall architecture of our method.

Backbone. HigherHRNet uses HRNet [14, 16] as backbone. We first adopt HRNet [14, 16] to a bottom-up method by adding a 1×1 convolution to predict heatmaps and tagmaps similar to [11]. Then, HigherHRNet is constructed by adding a deconvolution module to HRNet to generate higher resolution and multi-scale heatmaps.

Grouping. We follow [11] 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.

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 [17], 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 [5] after deconvolution to refine the upsampled feature maps. We add 4 Residual Blocks in HigherHRNet.

Multi-Resolution Supervision. We introduce a multi-resolution supervision during training. 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 $std = 2$ by default) to all these ground truth heatmaps. The final loss for heatmaps is the sum of mean squared errors for all resolutions. For Tagmaps, we only predict tagmaps at the lowest resolution, instead of using all resolutions.

Heatmap Aggregation for Inference. 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 [1, 11, 12] which only use heatmaps from a single scale or single stage for prediction.

PoseFix [10]. Our final result submitted for the COCO 2019 Keypoint Detection Challenge is further refined by PoseFix [10].

3. Experiments

3.1. COCO Keypoint Detection

Dataset. The COCO dataset [9] 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.

Training. Following [11], we use data augmentation with random rotation ($[-30°, 30°]$), random scale ($[0.75, 1.25]$), random translation ($[-40, 40]$) to crop an input image of size 512×512 as well as random flip. We generate two

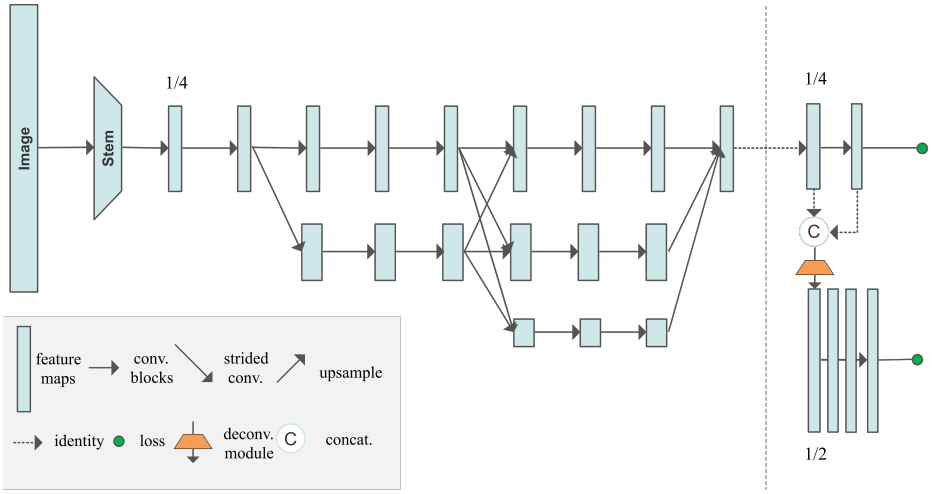

Figure 1: An illustration of HigherHRNet. The network uses HRNet [14] as backbone, followed by one or more deconvolution modules to generate multi-resolution and high-resolution heatmaps. Multi-resolution supervision is used for training. More details are given in Section 2.

Method	Backbone	Input size	#Params	GFLOPs	AP	AP^{50}	AP^{75}	AP^{M}	AP^{L}
w/o multi-scale test									
OpenPose*[1]	–	–	–	–	61.8	84.9	67.5	57.1	68.2
Hourglass [11]	Hourglass	512	277.8M	206.9	56.6	81.8	61.8	49.8	67.0
PersonLab [12]	ResNet-101	1401	68.7M	405.5	66.5	88.0	72.6	62.4	72.3
PifPaf [8]	–	–	–	–	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
Ours	HRNet-w32	512	28.6M	44.6	66.4	87.5	72.8	61.2	74.2
Ours	HRNet-w48	640	63.8M	154.3	**68.4**	**88.2**	**75.1**	**64.4**	**74.2**
w/ multi-scale test									
Hourglass [11]	Hourglass	512	277.8M	206.9	63.0	85.7	68.9	58.0	70.4
Hourglass*[11]	Hourglass	512	277.8M	206.9	65.5	86.8	72.3	60.6	72.6
PersonLab [12]	ResNet-101	1401	68.7M	405.5	68.7	89.0	75.4	64.1	75.5
Ours	HRNet-w48	640	63.8M	154.3	**70.5**	**89.3**	**77.2**	**66.6**	**75.8**
Ours + PoseFix [10]	HRNet-w48	640	63.8M	154.3	*72.1*	*89.5*	*78.4*	*68.1*	*77.5*

* Indicates using refinement.
† Our implementation, not reported in [14]

Table 1: Comparisons with bottom-up methods on the COCO2017 test-dev set. All GFLOPs are calculated at single-scale. For PersonLab [12], 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.* Ours + PoseFix [10] is our final result submitted to the COCO 2019 Keypoint Detection Challenge.

ground truth heatmaps with resolutions 128×128 and 256×256 respectively.

We use the Adam optimizer [7]. The base learning rate is set to $1e-3$, and dropped to $1e-4$ and $1e-5$ at the $200th$ and $260th$ epochs respectively. The total epochs is 300. To balance the heatmap loss and the grouping loss, we set the weight to 1 and $1e-3$ respectively for the two losses.

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 [11], 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 [14] itself already serves as a simple and strong baseline for bottom-up methods (64.1 AP). Equipped with light-weight deconvolution modules, our proposed HigherHRNet (66.4 AP) outperforms HRNet by +2.3 AP with only marginal increase in parameters (+0.1M) and FLOPs (+14.7%). If we increase the width (increases channels from 32 to 48), HigherHRNet outperforms all bottom-up methods by a large margin in both single-scale and multi-scale settings. HigherHRNet sets a new state-of-the-art result of 70.5 AP for bottom-up method.

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-

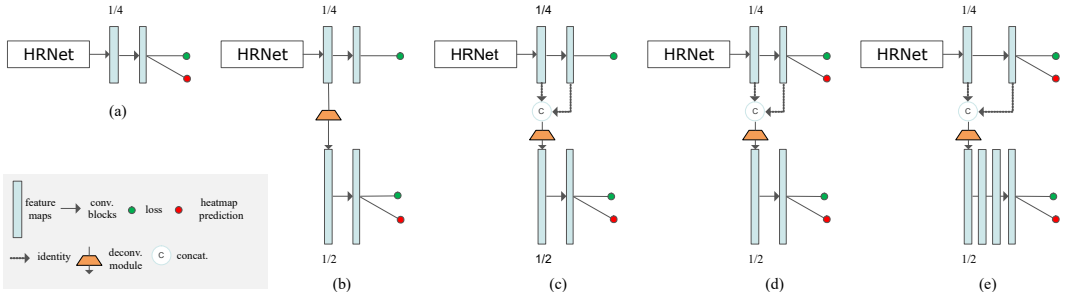

Figure 2: (a) Baseline method using HRNet [14] as backbone. (b) HigherHRNet with multi-resolution supervision (MRS). (c) HigherHRNet with MRS and feature concatenation. (d) HigherHRNet with MRS and feature concatenation. (e) HigherHRNet with MRS, feature concatennation and extra residual blocks. For (d) and (e), heatmap aggregation is used.

Method	AP	AP^{50}	AP^{75}	AP^{M}	AP^{L}	AR
Top-down methods						
Mask-RCNN [4]	63.1	87.3	68.7	57.8	71.4	–
G-RMI [13]	64.9	85.5	71.3	62.3	70.0	69.7
Integral Pose Regression [15]	67.8	88.2	74.8	63.9	74.0	–
G-RMI + extra data [13]	68.5	87.1	75.5	65.8	73.3	73.3
CPN [2]	72.1	91.4	80.0	68.7	77.2	78.5
RMPE [3]	72.3	89.2	79.1	68.0	78.6	–
CFN [6]	72.6	86.1	69.7	78.3	64.1	–
CPN (ensemble) [2]	73.0	91.7	80.9	69.5	78.1	79.0
SimpleBaseline [17]	73.7	91.9	81.1	70.3	80.0	79.0
HRNet-W48 [14]	75.5	92.5	83.3	71.9	81.5	80.5
HRNet-W48 + extra data [14]	77.0	92.7	84.5	73.4	83.1	82.0
Bottom-up methods						
OpenPose* [1]	61.8	84.9	67.5	57.1	68.2	66.5
Hourglass+AE*+ [11]	65.5	86.8	72.3	60.6	72.6	70.2
PifPaf [8]	66.7	-	-	62.4	72.9	-
PersonLab+ [12]	68.7	89.0	75.4	64.1	75.5	75.4
Ours: HigherHRNet-W48+AE+	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*. HigherRNet on COCO2017 val dataset. Using one deconvolution module for HigherHRNet performs best on the COCO dataset.

down methods.

3.2. Ablation Experiments

We perform a number of ablation experiments to analyze Higher-Resolution Network (HigherHRNet) on the COCO2017 [9] val dataset.

HRNet *vs*. HigherHRNet. We perform ablation study comparing HRNet and HigherHRNet. Results are shown in Table 3. A simple bottom-up baseline by using HRNet 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 improvement comes from medium persons, where AP^{M} is improved from 57.1 for HRNet to 61.0 for HigherHRNet.

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.

HigherHRNet gain breakdown. To better understand the gain of the proposed components, we perform detailed ablation studies on each individual component. Figure 2 illustrates all the architectures of our experiments. Results are shown in Table 4.

Effect of deconvolution module. For a fair comparison, we only use the highest resolution feature maps to generate heatmaps for prediction (Figure 2 (b)). HRNet (Figure 2 (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.

Effect of feature concatenation. We concatenate feature maps with predicted heatmaps from HRNet as input to the deconvolution module (Figure 2 (c)) and the performance is further improved to 66.3 AP.

Effect of heatmap aggregation. We further use all resolutions of heatmaps following the heatmap aggregation strategy for inference (Figure 2 (d)). Compared with Figure 2 (c) (66.3 AP) that only uses the highest resolution heatmaps, applying heatmap aggregation strategy achieves 66.9 AP.

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

	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. MSR: multi-resolution supervision. feature concat.: feature concatenation. res. blocks: residual blocks.

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 HigherRNet with different training image size.

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 HigherRNet with different training image size.

persons equally.

Training with larger image size. 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.

References

[1] 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, 3

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

[3] Hao-Shu Fang, Shuqin Xie, Yu-Wing Tai, and Cewu Lu. Rmpe: Regional multi-person pose estimation. In *ICCV*, 2017. 3

[4] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *ICCV*, 2017. 3

[5] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 1

[6] Shaoli Huang, Mingming Gong, and Dacheng Tao. A coarse-fine network for keypoint localization. In *ICCV*, 2017. 3

[7] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 2

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

[9] 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. 1, 3

[10] Gyeongsik Moon, Ju Yong Chang, and Kyoung Mu Lee. Posefix: Model-agnostic general human pose refinement network. In *CVPR*, 2019. 1, 2

[11] Alejandro Newell, Zhiao Huang, and Jia Deng. Associative embedding: End-to-end learning for joint detection and grouping. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *NIPS*. 2017. 1, 2, 3

[12] 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, 3

[13] 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. 3

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

[15] Xiao Sun, Bin Xiao, Fangyin Wei, Shuang Liang, and Yichen Wei. Integral human pose regression. In *ECCV*, 2018. 3

[16] 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. *arXiv preprint arXiv:1908.07919*, 2019. 1

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