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

Bowen Cheng, Bin Xiao, Jingdong Wang, Honghui Shi, Thomas S. Huang, Lei Zhang

Abstract

Bottom-up human pose estimation methods have difficulties in predicting the correct pose for small persons due to challenges in scale variation. This paper 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.

Keywords：Pose Estimation, HigherHRNet.

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 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 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 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.

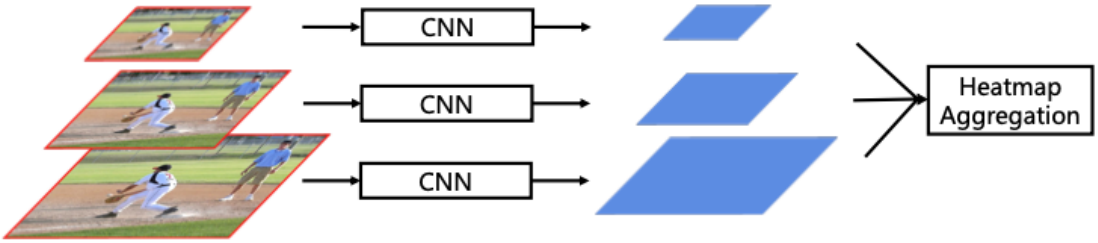

Figure 1: Image pyramid

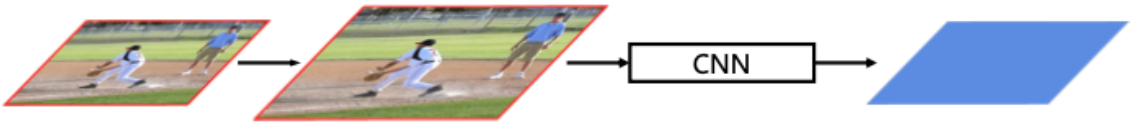

Figure 2: Upsampling input.

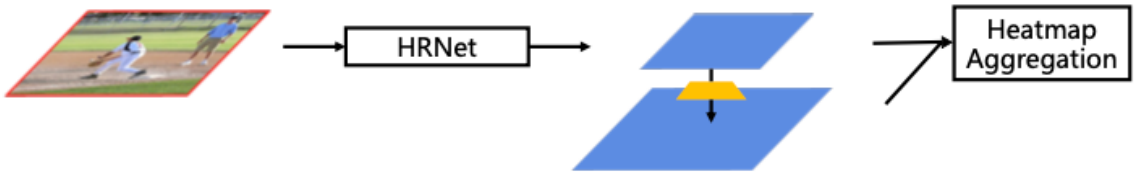

Figure 3: Author's approach

This paper 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 3). Build the high-resolution feature pyramid on the 1/4 resolution path of HRNet[1][2], to make it efficient. To make HigherHRNet capable of handling scale variation, further propose a Multi-Resolution Supervision strategy to assign training target of different resolutions to the corresponding feature pyramid level. Finally, introduce a simple Multi-Resolution Heatmap Aggregation strategy during inference to generate scale-aware high-resolution heatmaps.

2 Related works

2.1 Top-down methods

Top-down methods detect the keypoints of a single person within a person bounding box. The person bounding boxes are usually generated by an object detector . Mask R-CNN[3] directly adds a keypoint detection branch on Faster R-CNN[4] and reuses features after ROIPooling. G-RMI[5] and the following methods further break top- down methods into two steps and use separate models for person detection and pose estimation.

2.2 Bottom-up methods

Bottom-up methods detect identity-free body joints for all the persons in an image and then group them into individuals. OpenPose[6] uses a two-branch multi-stage netork with one branch for heatmap prediction and one branch for grouping. Open- Pose 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.[7] use stacked hourglass network[8] 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 l2 distance between tag vectors. PersonLab[9] uses dilated ResNet[10] and groups keypoints by directly learning a 2D offset field for each pair of keypoints. PifPaf[11] 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.

2.3 Feature pyramid

Pyramidal representation has been widely adopted in recent object detection and segmentation frameworks to handle scale variation. SSD[12] and MS-CNN[13] predict objects at multiple layers of the network without merging features. Feature pyramid networks[14] 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.

2.4 High resolution feature maps

There are mainly 4 methods to generate high resolution feature maps. (1) Encoder-decoder 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 (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)[15] is used in sequence at the end of a network to efficiently increase feature map resolution. SimpleBaseline[15] demonstrates that deconvolution can generate high quality feature maps for heatmap prediction. (4) Recently, a High-Resolution Network (HRNet)[1][2] is proposed as an efficient way to keep a high resolution pass throughout the network. HRNet[1][2] 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[1][2] can generate high resolution feature maps with rich semantic.

3 Method

3.1 Overview

The network uses HRNet[1][2] as backbone, followed by one or more deconvolution modules to generate multi-resolution and high-resolution heatmaps. Multi-resolution supervision is used for training.

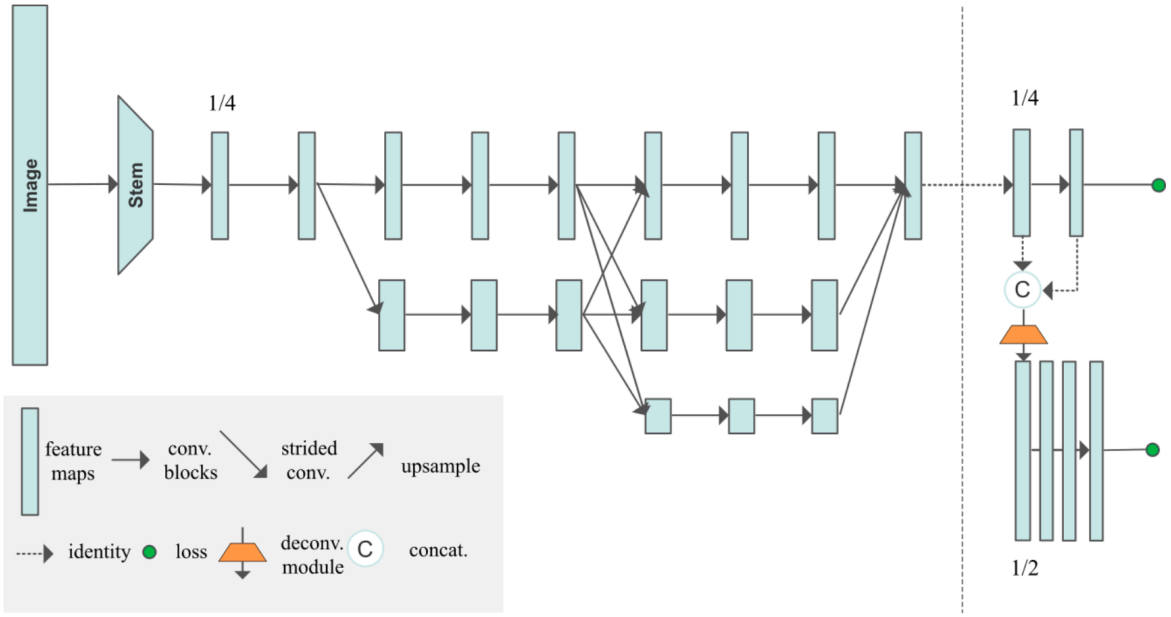

Figure 4: An illustration of HigherHRNet

3.2 HRNet

The author instantiate the backbone using a similar manner as HRNet[1][2] . 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. experiment with two networks with different capacity by setting C to 32 and 48 respectively. HRNet[1][2] was originally designed for top-down pose estimation. In this work, we adopt HRNet[1][2] to a bottom-up method by adding a 1 × 1 convolution to predict heatmaps and tagmaps similar to[7] . Only use the highest resolution (41 of the input image) feature maps for prediction. Following[7] , use a scalar tag for each keypoint.

3.3 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 applied 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 key-points belonging to small persons. A trivial solution to reduce this confusion is to reduce the standard deviation

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

Instead of reducing standard deviation, the author solve this problem by predicting heatmaps at higher resolution with standard deviation unchanged at different resolutions. Bottom-up methods usually predict heatmaps at resolution 14 of the input image. Yet find this resolution is not high enough for predicting accurate heatmaps. Inspired by[15] , which shows that deconvolution can be used to effectively generate high quality and high resolution feature maps, build HigherHRNet on top of the highest resolution feature maps in HRNet as shown in Figure 2 by adding a deconvolution module.

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. To train heatmap predictors at different resolutions and use a heatmap aggregation strategy.

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.

3.4 Grouping

Recent works have shown that grouping can be solved with high accuracy by a simple method using associative embedding[7]. As an evidence, experimental results in[7] 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[16]. Follow[7] to use associative embedding for keypoint grouping. The grouping process clusters identity-free keypoints into individuals by grouping keypoints whose tags have small l2 distance.

3.5 Deconvolution Module

The author propose a simple deconvolution module for generating high quality feature maps whose resolution is two times higher than the input feature maps. Following[15] , 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[10] after deconvolution to refine the upsampled feature maps. We add 4 Residual Blocks in HigherHRNet.

Different from[15] , the input to the author's deconvolution module is the concatenation of the feature maps and the predicted 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.

3.6 Multi-Resolution Supervision

Unlike other bottom-up methods that only apply supervision to the largest resolution heatmaps, the author introduce a multi-resolution supervision during training to handle scale variation. Transform ground truth key-

point locations to locations on the heatmaps of all resolutions to generate ground truth heatmaps with different resolutions. Then apply a Gaussian kernel with the same standard deviation (the author use standard deviation = 2 by default) to all these ground truth heatmaps. 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, 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[14] 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. Tagmaps are trained differently from heatmaps in HigherHRNet. 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, the author also find higher resolutions do not learn to predict tagmaps well and even do not converge. Thus, follow[7] to train the tagmaps on feature maps at 14 resolution of input image.

3.7 Heatmap Aggregation for Inference

The author propose a heatmap aggregation strategy during inference. 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 which only use heatmaps from a single scale or single stage for prediction.

The reason that use heatmap aggregation is to enable scale-aware pose estimation. For example, the COCO Keypoint dataset[16] contains persons of large scale variance from 322 pixels to more than 1282 pixels. Top-down methods 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.

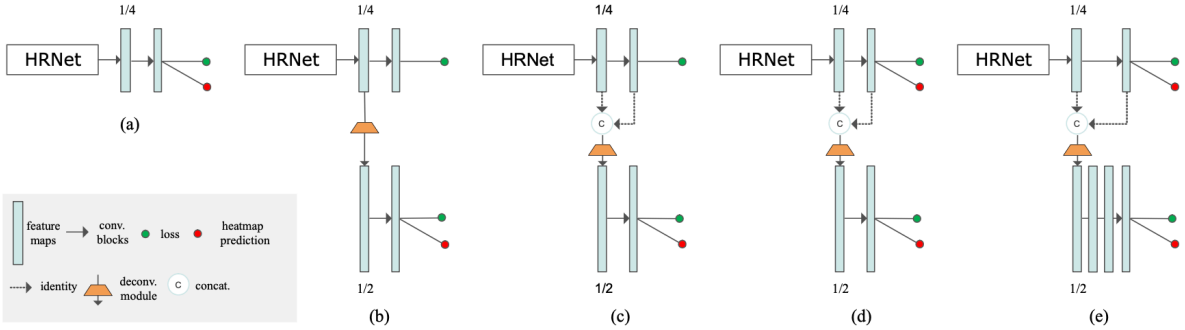

Figure 5: (a) Baseline method using HRNet as backbone. (b) HigherHRNet with multi-resolution supervision (MRS). (c) High- erHRNet with MRS and feature concatenation. (d) HigherHRNet with MRS and feature concatenation. (e) HigherHRNet with MRS, feature concatennation and extra residual blocks. For (d) and (e), heatmap aggregation is used.

4 Implementation details

4.1 Comparing with released source codes

Compared with the open source code, there are two main changes here. First, according to the conclusion drawn by the author, a more accurate heat map can be obtained by using a high-resolution feature map, and the deconvolution operation can improve the resolution of the feature map, so based on the author's code, a deconvolution module is added to obtain higher resolution feature maps. Second, add several Basic Residual Blocks after deconvolution to refine the upsampled feature maps. The number of Basic Residual Blocks in the author's source code is 4. If the network structure is changed, the number needs to be changed according to the actual situation.

4.2 Experimental environment setup

Hardware Environment: Intel 6700HQ, NVIDIA GT970M

Driver Version: CUDA 10.1, CUDNN 7.6.5

Software Environment: Anaconda3.9, Python3.6

Python Environment: Torch1.5.1, Torchvision0.6.1, EasyDict1.7, Opencv-Python4.6.0.66, Pandas1.1.5, Tqdm4.64.1, Cython, Scipy, Pyyaml, Json_Tricks, Scikit-Image, TensorboardX2.5.1, Yacs, Cffi, Munkres

DataSets: COCOAPI, CrowdPoseAPI

4.3 Main contributions

According to the author's methodology and open source code, my Implementation includes two points. First, deploy the author's code, configure the system environment and datasets required for the code to run, and execute the author's code to view the effect of the method. Second, make slight changes to the author's code to add my own ideas.Added a deconvolution module to the author's network structure, and changed the number of Basic Residual Blocks from 4 to 8.

5 Results and analysis

Figure 6 shows the experimental results of the method proposed by the authors. The author 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. During the inference, HigherHRNet with multi-resolution heatmap aggregation is capable of efficiently generating higher-resolution heatmaps for more accurate human pose estimation. HigherHRNet Performs well on the challenging COCO dataset, especially for small persons.

Figure 6: Experimental results

6 Conclusion and future work

Through the lectures of experts in various fields in the class, I have some understanding of the cutting-edge technologies in various fields of computer, and I have a strong interest in several fields. The reproduction work of the thesis has greatly improved my hands-on ability. Since I have not learned the relevant knowledge of machine learning or deep learning before, it took me a lot of time just to make the author's code run smoothly. According to the tutorials on the Internet, I realized the construction of the system development environment, and have a basic understanding of the development environment. Although I have some ideas to make some changes to the author's method, due to the huge project and my computer configuration, the training test cannot be completed. I hope that I can realize my own ideas with a suitable computer configuration in the future.

Finally, I would like to thank the college for inviting experts in various fields to give lectures. I feel that I have gained a lot.

References

[1] SUN K, XIAO B, LIU D, et al. Deep high-resolution representation learning for human pose estimation [C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2019: 5693-5703.

[2] WANG J, SUN K, CHENG T, et al. Deep high-resolution representation learning for visual recognition [J]. IEEE transactions on pattern analysis and machine intelligence, 2020, 43(10): 3349-3364.

[3] HE K, GKIOXARI G, DOLLÁR P, et al. Mask r-cnn[C]//Proceedings of the IEEE international conference on computer vision. 2017: 2961-2969.

[4] REN S, HE K, GIRSHICK R, et al. Faster r-cnn: Towards real-time object detection with region proposal networks[J]. Advances in neural information processing systems, 2015, 28.

[5] PAPANDREOU G, ZHU T, KANAZAWA N, et al. Towards accurate multi-person pose estimation in the wild[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2017: 4903-4911.

[6] CAO Z, SIMON T, WEI S E, et al. Realtime multi-person 2d pose estimation using part affinity fields[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2017: 7291-7299.

[7] NEWELL A, HUANG Z, DENG J. Associative embedding: End-to-end learning for joint detection and grouping[J]. Advances in neural information processing systems, 2017, 30.

[8] NEWELL A, YANG K, DENG J. Stacked hourglass networks for human pose estimation[C]//European conference on computer vision. 2016: 483-499.

[9] PAPANDREOU G, ZHU T, CHEN L C, et al. Personlab: Person pose estimation and instance segmentation with a part-based geometric embedding model[J]., 2018.

[10] HE K, ZHANG X, REN S, et al. Deep residual learning for image recognition. CVPR. 2016[J]. arXiv preprint arXiv:1512.03385, 2016.

[11] KREISS S, BERTONI L, ALAHI A. Pifpaf: Composite fields for human pose estimation[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2019: 11977-11986.

[12] LIU W, ANGUELOV D, ERHAN D, et al. Ssd: Single shot multibox detector[C]//European conference on computer vision. 2016: 21-37.

[13] CAI Z, FAN Q, FERIS R S, et al. A unified multi-scale deep convolutional neural network for fast object detection[C]//European conference on computer vision. 2016: 354-370.

[14] LIN T Y, DOLLÁR P, GIRSHICK R, et al. Feature pyramid networks for object detection[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2017: 2117-2125.

[15] XIAO B, WU H, WEI Y. Simple baselines for human pose estimation and tracking[C]//Proceedings of the European conference on computer vision (ECCV). 2018: 466-481.

[16] LIN T Y, MAIRE M, BELONGIE S, et al. Microsoft coco: Common objects in context[C]//European conference on computer vision. 2014: 740-755.