Bidirectional Pyramid Networks for Semantic Segmentation Supplementary Material

Dong Nie[1], Jia Xue*[2], and Xiaofeng Ren[1]

[1] Amap, Alibaba Group {dong.nie,x.ren}@alibaba-inc.com
[2] Rutgers University jia.xue@rutgers.edu

This supplementary material provides details and additional visual results that could not be included in the paper submission due to space limitation. In Sec. 1, we first provide more implementation details, such as model architecture and the loss functions we use to train the networks. Then, we illustrate the uses of bidirectional pyramid network (in Sec. 2 in the paper) and unary attention and pairwise attention (in Sec. 3 in the paper) by visualizing feature maps of an example. Sec. 4 provides more experimental details for prostate MRI segmentation. Sec. 5 discuss additional state-of-the-art works. In Sec. 6, we show more visual results on Cityscapes, CamVid, PASCAL Context.

1 More Implementation Details

1.1 Architecture Details

Our network starts from a a stem that consists of one strided 3×3 convolution and a pooling layer to decrease the resolution to $1/4$ of the input image. Then, the first step upward contains 1 residual module in the subsampling pathway (i.e., feature level in L1), with a bottleneck of width C, one strided 3×3 convolution to achieve larger receptive field (lower resolution, i.e., L2). The 2nd, 3rd, 4th steps behave the same way as the 1st step, as shown in Fig.1 in the paper. The widths (number of channels) of the convolutions of the five resolutions are C, 2C, 4C, 8C and 16C, respectively. In the highest layer (i.e., L5, with lowest resolution), we apply a pyramid pooling module to gain even larger receptive fields. The bidirectional information flow is implemented by bilinear interpolation (for upsampling) and the AMA fusion strategy. To save computational cost, we reduce the resolution of the context aggregation module to $1/2$ (or $1/4$) and all operations of unary-pairwise attention module is working on this lower resolution. The APNB module follows exactly the same design in [1].

1.2 Loss Functions for Training

Ohem loss [2] is widely used to train semantic segmentation networks because of its ability to handle hard samples. Dice loss [3] is often adopted as a segmentation

* Work done during internship at Amap

loss function to solve the category imbalance issues. We use both ohem loss and Dice loss to train our segmentation network, as shown in Eq. 1.

$$L_{Hyb} = L_{ohem} + \lambda L_{Dice} \tag{1}$$

where λ is a non-negative weighting coefficient, and it is set to 1.6 in all our experiments.

Besides, We utilize auxiliary loss functions to supervise the training of our proposed method. We use the principal loss function (OHEM loss and Dice loss, jointly) to supervise the output of the whole BPNet. Moreover, we add two (or three, for BPNet-S4) auxiliary loss functions to supervise the output of the intermediate stages at the lowest layer (similar to deep supervision [4].

2 Exploration of Bidirectional Pyramid Structure

The bidirectional pyramid network is the core of our work. We have quantitatively demonstrate the effectiveness of our proposed structure in the paper. Top-down and bottom-up information flows are shown to both provide performance gains. Here we show a qualitative analysis how the information flows help the segmentation task.

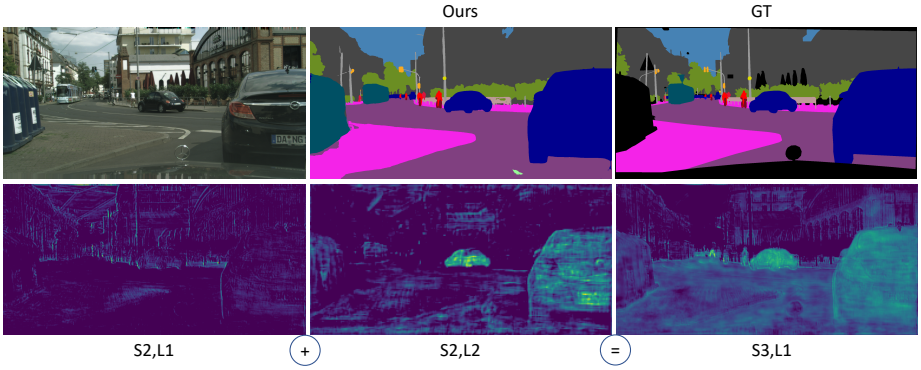

Fig. 1. Visualization of a top-down information flow. "S2,L1" is the 2nd stage feature map at feature level 1, which contains many fine details (e.g., edges) and some unrelated noises; "S2,L2" is the 2nd stage feature map at feature level 2, which captures the semantic shape of the cars. Fusing these two information sources, "S3,L1" retains the necessary fine details but is more semantically meaningful.

We visualize an example of top-down information flow in Fig. 1. From the example, we can see that top-down information flow enhances the later-stage feature maps to capture more semantically meaningful cues while retaining fine details. With such (bidirectional) information flow going in every step of the network processing, the representation for the input image is both rich in semantic cues and in high-resolution details, as shown in Fig. 2.

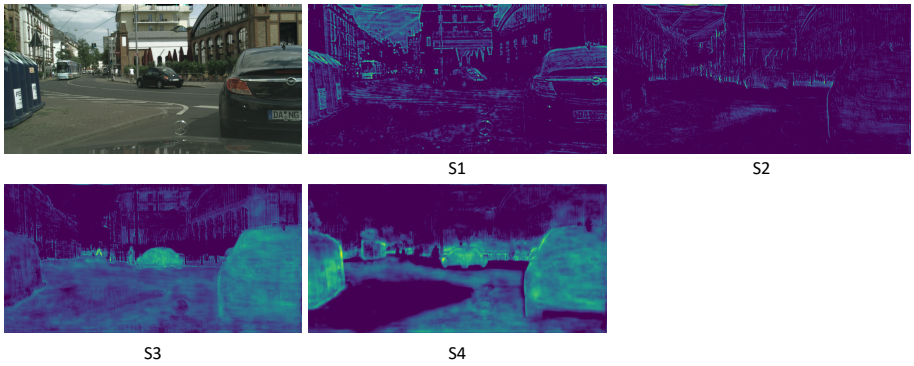

Fig. 2. Visualization of feature maps in different stages (i.e., from S1 to S4) at feature level 1. The necessary detailed information is always maintained via this high-resolution branch, and the semantic information is largely enhanced by the bidirectional information flow in the pyramid.

3 Exploration of Unary-Pairwise Attention

We conduct quantitative ablation studies to understand the impact of unary-pairwise attention. As indicated in the main paper, the unary-pairwise attention can work better than either unary attention or pairwise attention by themselves. Here we further explore what unary-pairwise attention learns, and why it works better than individual attention mechanisms.

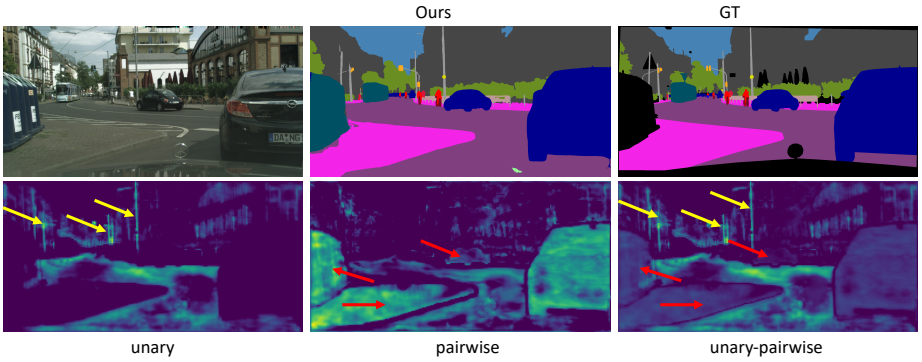

Fig. 3. Visualization of the unary-pairwise attention maps. Unary attention maps can preserve thin structures, and the pairwise attention maps can help capture the concept of large objects. With parallel pairwise-attention mechanism, we can boost the segmentation for both fine details and large objects.

In Fig. 3, we visualize the feature map from the unary-attention block, pairwise attention block and unary-pairwise attention block , respectively. It can be observed that uanry attention contributes more to edges or details, while pairwise attention focuses more on large objects. With a parallel combination (which works better than sequential combination), we can gain the advantages of both attention mechanism.

4 Details of Experiments on Medical Image Data: Prostate Segmentation

The detailed comparison results on prostate MRI dataset are shown in Table 1.

Table 1. Semantic segmentation results on prostate MRI.

Dataset	training time	test time	mIoU
3D-UNet [5]	3d	27s	82.1
3DVNet [3]	3d	29s	84.6
3DUNet++ [6]	3d	31s	87.2
nnUNet [7]	5d	45s	92.3
BPNet-S4	17h	1s	91.1

We also visualize the segmentations of the prostate for two typical subjects with their 3D renderings in Fig. 4.

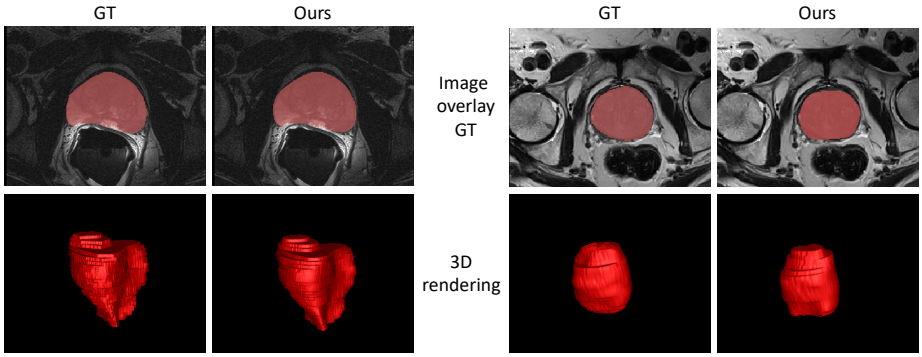

Fig. 4. Qualitative visualizations of our BPNet-S4 on the prostate MRI val set. The first row shows segmentation results of two typical subjects by experts' manual segmentation (GT) and automatic segmentation (Our BPNet-S4). The second row shows the corresponding 3D renderings of the segmentation results.

5 Additional Discussion on Related Work

5.1 State-of-the-art real-time segmentation methods

As shown in the main paper, our BPNet-S3-W32 network can achieve test mIoU 76.3 with 5.1M parameters and 36 fps (on Nvidia 2080 Ti) on the Cityscapes test set. Compared with many recent state-of-the-art real-time segmentation methods [8–14], our BPNet-S3-W32 achieves the best accuracy with a comparable or smaller model size. The inference speed of our network is also high, comparable to BiseNet [11]. We believe that our BPNet-S3-W32 provides a competitive and general solution to real-time semantic segmentation.

It is worth noting the design strategy of our real-time BPNet-S3-W32 is different from many existing real-time segmentation methods, which often downsample feature maps by a large factor at the beginning stage, in order to save computational cost. Our BPNet-S3-W32 keeps a high-resolution pathway all the way at the bottom of the pyramid, so that we can capture detailed features. To reduce computational cost, comapre to our heavyweight networks, we reduces the number of channels overall the networks and shorten the stages from four to three. The systematic information communication in the bidirectional pyramid network and the unary-pairwise attention mechanism could remedy the loss of feature representation ability due to the shrinkage of the networks.

5.2 State-of-the-art segmentation methods with external training

In the main paper, we have compared to many state-of-the-art segmentation methods in terms of accuracy. Most of these compared methods use the Imagenet dataset to pretrain a strong backbone, but typically not use other external datasets, for instance, Cityscapes coarse set and Mapillary dataset. We note that some proposed segmentation methods do use external datasets to achieve state-of-the-art performance on Cityscapes [15–21]. In Gated-SCNN [20], a shape branch is constructed to support the segmentation branch achieving better boundary accuracy. The Gated-SCNN can achieve 80.8% mIoU on Cityscapes val set with multi-scale inference (Our BPNet-S4 can achieve 81.2% with multi-scale inference on the Cityscapes val set). Using extra Mapillary data, they achieve a 82.8% mIoU on Cityscapes test set. Our BPNet-S4 achieves a 81.9% on the test set without using any external dataset (Imagenet or Cityscapes coarse set). By using even more external data in training, better performance can be obtained. Zhu et al. [21] proposes an algorithm to use more data to train the segmentation network on Cityscapes. Besides ImageNet and Cityscapes coarse set, their network also uses Cityscapes video set to help train the segmentation network, their final accuracy is 83.5% on Cityscapes test set. It is conceivable that our method can also benefit from using extra training data. For example, using Cityscapes train set only, our model can achieve 81.0% on the test set with multi-scale inference; using train+val set for training, 81.9% is achieved on the test set. Thus, we believe our method can further improve performance if using external training data, such as Cityscapes coarse set and Mapillary data.

5.3 Details about Difference from HRNet.

We elaborate the details of difference from HRNet [22]. HRNet as an example to explain the difference. (1) Feature fusion design is different. HRNet aims for high-resolution representation and accuracy. In each stage, features of all scales are fully connected and fused with '+' op, followed by a heavy translation block (4 RBs). In contrast, BPNet aims for a balance between scales and the efficiency of information flow, using AMA to fuse features in a pyramidal scheme. Note, in BPNet, each fusion just involves two information flow, with one providing the resolution and the other sharing the semantics. (2) As a result, BPNet is much more efficient for both the lightweight and heavy variants, 2× to 3× more efficient than HRNet. Testing on 2080 Ti, BPNet-S3-W32 has mIoU 77.2 at 36.5 fps, while HRNet-v2-small-v2 (released model) has mIoU 76.2 at 19.9 fps. Our heavier model BPNet-S4 is at 14.8 fps, while HRNet-v2-W48 is at 5.1 fps. (3) Moreover, BPNet does not use pretraining, which increases its flexibility. HRNet is trained with ImageNet pretraining.

6 More Visual Results

Here we show additinoal visual results on Cityscapes, CamVid, PASCAL-Context and prostate MRI, to help illustrate the reuslts of our method.

As shown in Fig. 5, our BPNet-S4 is able to perform accurate prediction on challenging urban scenes, producing results very similar to the ground-truth.

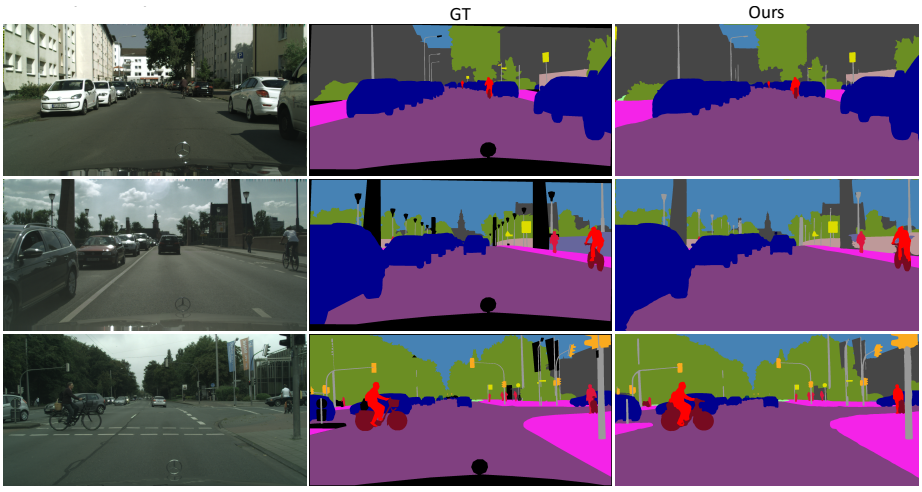

Fig. 5. Qualitative visualization of our BPNet-S4 on the Cityscapes val set. Best viewed in color.

We also show visual results of our BPNet-S4 on CamVid and PASCAL-Context in Fig. 6 and Fig. 7, respectively. In those examples, our BPNet-S4 achieves accurate results, which are close to the ground-truth.

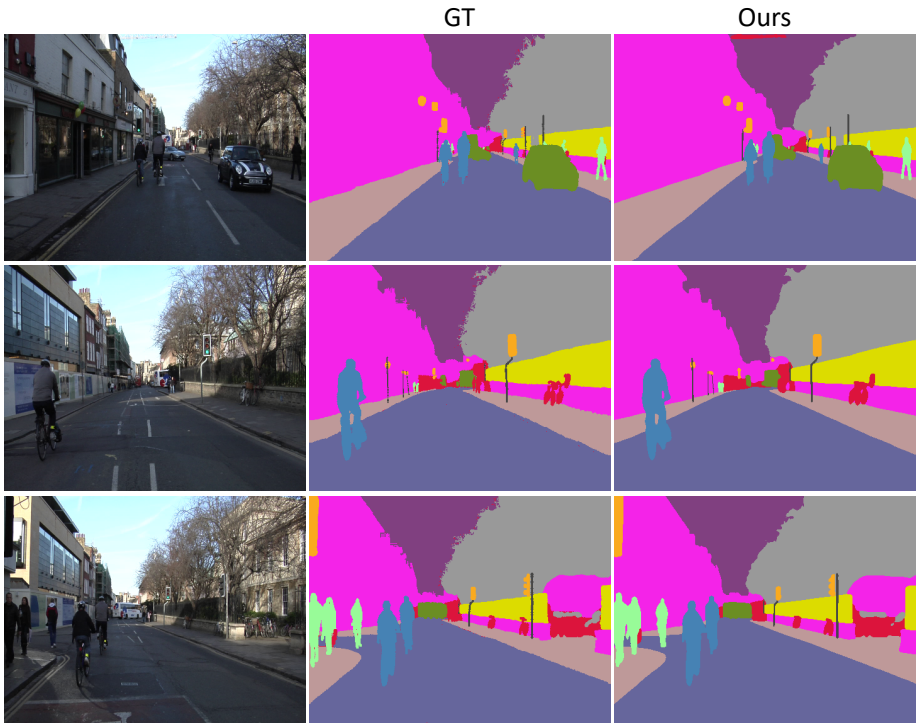

Fig. 6. Qualitative visualizations of our BPNet-S3 on the CamVid val set. Best viewed in color.

References

1. Zhu, Z., Xu, M., Bai, S., Huang, T., Bai, X.: Asymmetric non-local neural networks for semantic segmentation. In: Proceedings of the IEEE International Conference on Computer Vision. (2019) 593–602
2. Shrivastava, A., Gupta, A., Girshick, R.: Training region-based object detectors with online hard example mining. In: Proceedings of the IEEE conference on computer vision and pattern recognition. (2016) 761–769
3. Milletari, F., Navab, N., Ahmadi, S.A.: V-net: Fully convolutional neural networks for volumetric medical image segmentation. In: 2016 Fourth International Conference on 3D Vision (3DV), IEEE (2016) 565–571
4. Lee, C.Y., Xie, S., Gallagher, P., Zhang, Z., Tu, Z.: Deeply-supervised nets. In: Artificial intelligence and statistics. (2015) 562–570

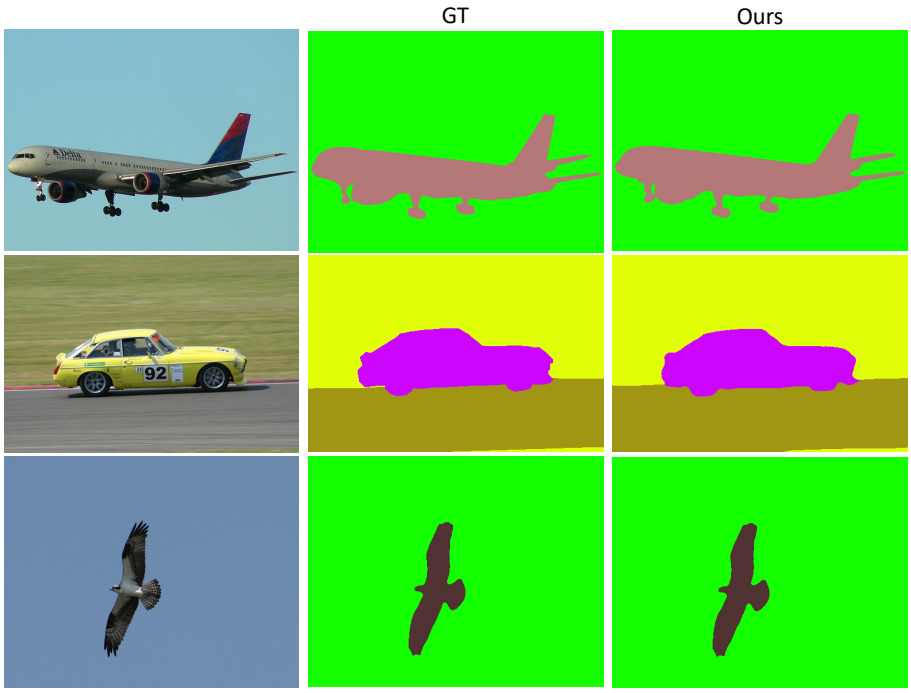

Fig. 7. Qualitative visualizations of our BPNet-S4 on the Pascal-Context val set. Best viewed in color.

5. Çiçek, Ö., Abdulkadir, A., Lienkamp, S.S., Brox, T., Ronneberger, O.: 3d u-net: learning dense volumetric segmentation from sparse annotation. In: International conference on medical image computing and computer-assisted intervention, Springer (2016) 424–432
6. Zhou, Z., Siddiquee, M.M.R., Tajbakhsh, N., Liang, J.: Unet++: A nested u-net architecture for medical image segmentation. In: Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support. Springer (2018) 3–11
7. Isensee, F., Petersen, J., Klein, A., Zimmerer, D., Jaeger, P.F., Kohl, S., Wasserthal, J., Koehler, G., Norajitra, T., Wirkert, S., et al.: nnu-net: Self-adapting framework for u-net-based medical image segmentation. arXiv preprint arXiv:1809.10486 (2018)
8. Nekrasov, V., Shen, C., Reid, I.: Light-weight refinenet for real-time semantic segmentation. arXiv preprint arXiv:1810.03272 (2018)
9. Li, H., Xiong, P., Fan, H., Sun, J.: Dfanet: Deep feature aggregation for real-time semantic segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (2019) 9522–9531
10. Zhao, H., Qi, X., Shen, X., Shi, J., Jia, J.: Icnet for real-time semantic segmentation on high-resolution images. In: Proceedings of the European Conference on Computer Vision (ECCV). (2018) 405–420

11. Yu, C., Wang, J., Peng, C., Gao, C., Yu, G., Sang, N.: Bisenet: Bilateral segmentation network for real-time semantic segmentation. In: Proceedings of the European conference on computer vision (ECCV). (2018) 325–341
12. Paszke, A., Chaurasia, A., Kim, S., Culurciello, E.: Enet: A deep neural network architecture for real-time semantic segmentation. arXiv preprint arXiv:1606.02147 (2016)
13. Wu, Z., Shen, C., Hengel, A.v.d.: Real-time semantic image segmentation via spatial sparsity. arXiv preprint arXiv:1712.00213 (2017)
14. Pohlen, T., Hermans, A., Mathias, M., Leibe, B.: Full-resolution residual networks for semantic segmentation in street scenes. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (2017) 4151–4160
15. Chen, L.C., Collins, M., Zhu, Y., Papandreou, G., Zoph, B., Schroff, F., Adam, H., Shlens, J.: Searching for efficient multi-scale architectures for dense image prediction. In: Advances in neural information processing systems. (2018) 8699–8710
16. Zhuang, Y., Yang, F., Tao, L., Ma, C., Zhang, Z., Li, Y., Jia, H., Xie, X., Gao, W.: Dense relation network: Learning consistent and context-aware representation for semantic image segmentation. In: 2018 25th IEEE international conference on image processing (ICIP), IEEE (2018) 3698–3702
17. Rota Bulò, S., Porzi, L., Kontschieder, P.: In-place activated batchnorm for memory-optimized training of dnns. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (2018) 5639–5647
18. Valada, A., Mohan, R., Burgard, W.: Self-supervised model adaptation for multimodal semantic segmentation. (International Journal of Computer Vision) 1–47
19. Li, X., Zhang, L., You, A., Yang, M., Yang, K., Tong, Y.: Global aggregation then local distribution in fully convolutional networks. arXiv preprint arXiv:1909.07229 (2019)
20. Takikawa, T., Acuna, D., Jampani, V., Fidler, S.: Gated-scnn: Gated shape cnns for semantic segmentation. In: Proceedings of the IEEE International Conference on Computer Vision. (2019) 5229–5238
21. Zhu, Y., Sapra, K., Reda, F.A., Shih, K.J., Newsam, S., Tao, A., Catanzaro, B.: Improving semantic segmentation via video propagation and label relaxation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (2019) 8856–8865
22. Sun, K., Zhao, Y., Jiang, B., Cheng, T., Xiao, B., Liu, D., Mu, Y., Wang, X., Liu, W., Wang, J.: High-resolution representations for labeling pixels and regions. arXiv preprint arXiv:1904.04514 (2019)