Panoptic-DeepLab: A Simple, Strong, and Fast Baseline for Bottom-Up Panoptic Segmentation – Supplementary Material –

Bowen Cheng[1,2], Maxwell D. Collins[2], Yukun Zhu[2], Ting Liu[2],
Thomas S. Huang[1], Hartwig Adam[2], Liang-Chieh Chen[2]

[1]UIUC [2]Google Research

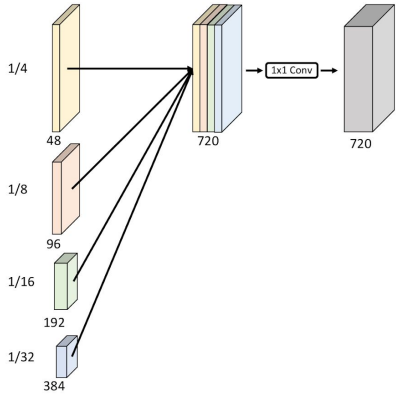


Figure 1. Semantic segmentation head proposed in HRNet [11].

1. HRNet Variant

We introduce our modifications to the HRNet [10, 11] that are used in our ensemble model [2] for Mapillary Vistas. All hyper-parameters for training HRNet variants are the same as Xception, except that the learning rate is set to $7.5e-4$.

1.1. HRNet

The original segmentation head for HRNet is shown in Fig. 1. Features from all four resolutions are first upsampled to the 1/4 resolution and concatenated, followed by another 1×1 convolution to fuse features.

To pre-train the HRNet on ImageNet [4], Wang *et al*. [11] designed a specific image classification head which gradually downsamples the feature maps, as shown in Fig. 2 (a). Specifically, a bottleneck residual module [6] is applied to every output resolution to increase the channels. The feature map from the finest spatial resolution (*i.e*., 1/4 resolution) is then downsampled by sequentially using a 3×3 convolution with stride 2. At the final 1/32 resolution feature map, a global average pooling and a fully connected layer are attached for ImageNet classification.

1.2. HRNet+

After pre-training on ImageNet, Wang *et al*. [11] removed the image classification head. However, we observe that the classification head takes around 20% of the total parameters, which is a waste of information if discarded. Therefore, we propose to keep this classification head in our modified HRNet+ (Fig. 2 (b)). Starting from the image classification HRNet, we replace the final global average pooling and linear classifier with an ASPP module, and build a similar decoder as shown in Fig. 2 of main paper with some differences that the output stride of encoder is now 32 instead of 16 and we introduce one more encoder feature map of stride 16 to the decoder by first projecting its channels to 96.

1.3. HRNet-Wider+

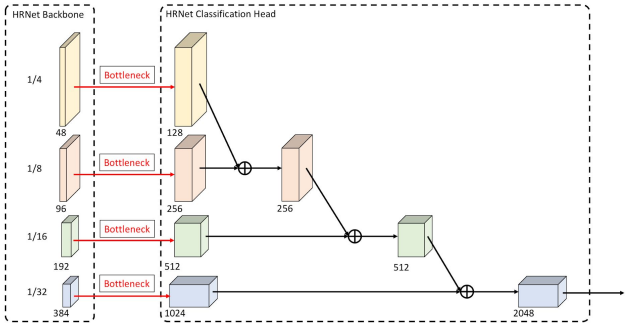
We additionally propose HRNet-Wider+ (Fig. 2 (c)) that replaces the basic residual module [6] with the Xception module [3], significantly reducing the model parameters and computation FLOPs at the cost of marginal degradation in performance. Additionally, we employ the number of channels $\{64, 256, 384, 384\}$ for each resolution (instead of $\{48, 96, 192, 384\}$).

1.4. Atrous HRNet

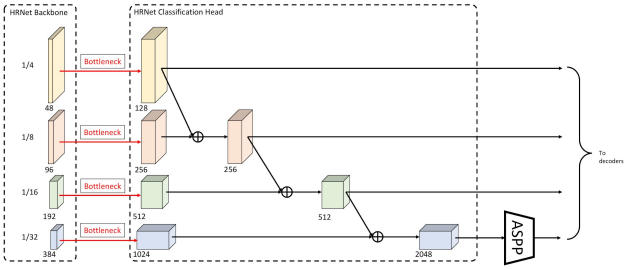
Another modification of HRNet that we have explored is referred to as HRNet+ (Atrous), where we remove all the downsampling operations that generate 1/32 resolution feature maps and apply atrous convolution with rate equal to 2 in that branch. This modification increases the computation FLOPs but does not improve the performance compared to its HRNet+ counterpart.

Decoder	Backbone	Input Size	PQ (%)	AP (%)	mIoU (%)	Speed (ms)	Params (M)	M-Adds (B)
DeepLabV3+ [1]	Xception-71	1025×2049	62.5	34.5	80.2	176	46.61	553.41
Panoptic-DeepLab	Xception-71	1025×2049	63.0	35.3	80.5	175	46.72	547.49

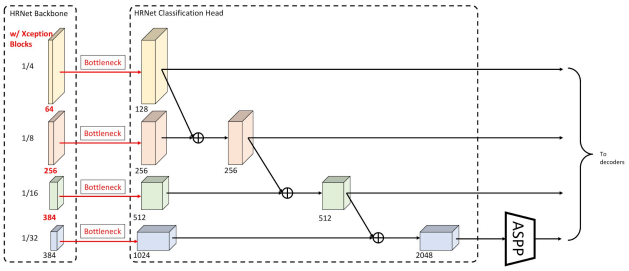
Table 1. Comparison between the decoder design of DeepLabV3+ [1] and Panoptic-DeepLab on Cityscapes validation set.



(a) Image classification head proposed in HRNet [11], which is discarded after pre-training on ImageNet.



(b) Our proposed HRNet+, which keeps the image classification head and attaches the ASPP module as well as the decoder module for segmentation tasks.



(c) Our proposed HRNet-Wider+, which reduces the model parameters and computations by adopting the Xception module.

Figure 2. Demonstration of our proposed variants of HRNet [11].

2. Auto-DeepLab Variant

We make a simple modification to the Auto-DeepLab [9] in Fig. 3 by removing the stride in the convolution that generates the 1/32 feature map in order to keep high spatial resolution within the network backbone. We find this modification improves 1% PQ on Mapillary Vistas validation set.

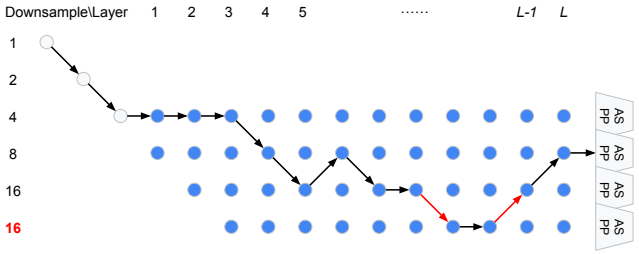


Figure 3. Our proposed Auto-DeepLab+, which keeps the high spatial resolution of feature maps by removing the last stride, *i.e.*, no spatial resolution changes marked in the red arrows.

3. Comparison with DeepLabV3+ decoder

As mentioned in the main paper that the decoder of Panoptic-DeepLab is slightly different from the one in DeepLabv3+ [1]. Herein, we compare their performance on Cityscapes validation set, as shown in Tab. 1. Panoptic-DeepLab outperforms DeepLabv3+ by 0.5% PQ, 0.8% AP, and 0.3% mIOU, showing more improvement in the instance segmentation task. Additionally, Panoptic-DeepLab is slightly faster than DeepLabv3+ at the cost of extra marginal parameters.

4. Comparison with different instance scores

Instance score	PQ (%)	AP (%)	mIoU (%)
Score(Objectness)	63.0	28.9	80.5
Score(Class)	63.0	35.1	80.5
Score(Objectness) x Score(Class)	63.0	**35.3**	80.5

Table 2. Ablation study on using different confidence scores. Note the choice of confidence scores only affects AP.

In Tab. 2, we experiment with different confidence scores when evaluating instance segmentation results. We found that using Score(Objectness) alone leads to 28.9% AP, Score(Class) alone produces 35.1% AP, while employing Score(Objectness) $\times$ Score(Class) generates the best result (35.3% AP). We would like to highlight that the choice of different confidence score does not affect our final mIoU and PQ results, since our Panoptic-DeepLab does not produce overlapping predictions and therefore does not require a confidence score to rank predictions (or to resolve the conflict among overlapping predictions) like Panoptic-FPN [7]. Confidence score is only used in computing AP to rank instance mask predictions. Since it is only used for the purpose of ranking, the confidence score does not necessarily

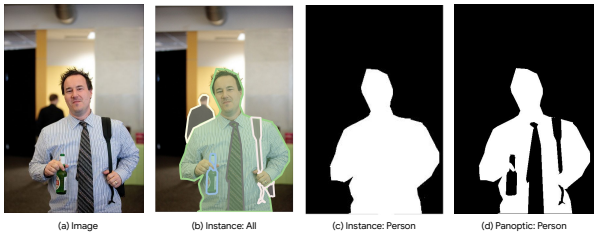


Figure 4. Illustration of the difference between instance and panoptic annotation on COCO.

need to be a probability.

5. Instance and Panoptic Annotation

Fig. 4 shows an example to illustrate the difference between instance annotation and panoptic annotation on the COCO dataset. Instance annotation, unlike panoptic annotation, allows overlapping groundtruth masks. For example, the 'person' mask ignores the existence of the 'tie' and 'bottle' masks in the instance annotation, while the 'person' mask has occlusions caused by other instances in the panoptic annotation.

We notice that all top-down methods based on Mask R-CNN [5] use the *instance annotation* [8, 7, 12] when trained on COCO, while bottom-up methods [13] including our Panoptic-DeepLab use the *panoptic annotation* on all datasets.

6. More Visualization

We provide more visualization results of our Panoptic-DeepLab in Fig. 5, Fig. 6, and Fig. 7.

References

[1] 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. 2

[2] Bowen Cheng, Maxwell D Collins, Yukun Zhu, Ting Liu, Thomas S Huang, Hartwig Adam, and Liang-Chieh Chen. Panoptic-deeplab. In *ICCV COCO + Mapillary Joint Recognition Challenge Workshop*, 2019. 1

[3] François Chollet. Xception: Deep learning with depthwise separable convolutions. In *CVPR*, 2017. 1

[4] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, 2009. 1

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

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

[7] Alexander Kirillov, Ross Girshick, Kaiming He, and Piotr Dollár. Panoptic feature pyramid networks. In *CVPR*, 2019. 2, 3

[8] Yanwei Li, Xinze Chen, Zheng Zhu, Lingxi Xie, Guan Huang, Dalong Du, and Xingang Wang. Attention-guided unified network for panoptic segmentation. In *CVPR*, 2019. 3

[9] 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. 2

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

[11] 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. *arXiv:1908.07919*, 2019. 1, 2

[12] Yuwen Xiong, Renjie Liao, Hengshuang Zhao, Rui Hu, Min Bai, Ersin Yumer, and Raquel Urtasun. Upsnet: A unified panoptic segmentation network. In *CVPR*, 2019. 3

[13] 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:1902.05093*, 2019. 3

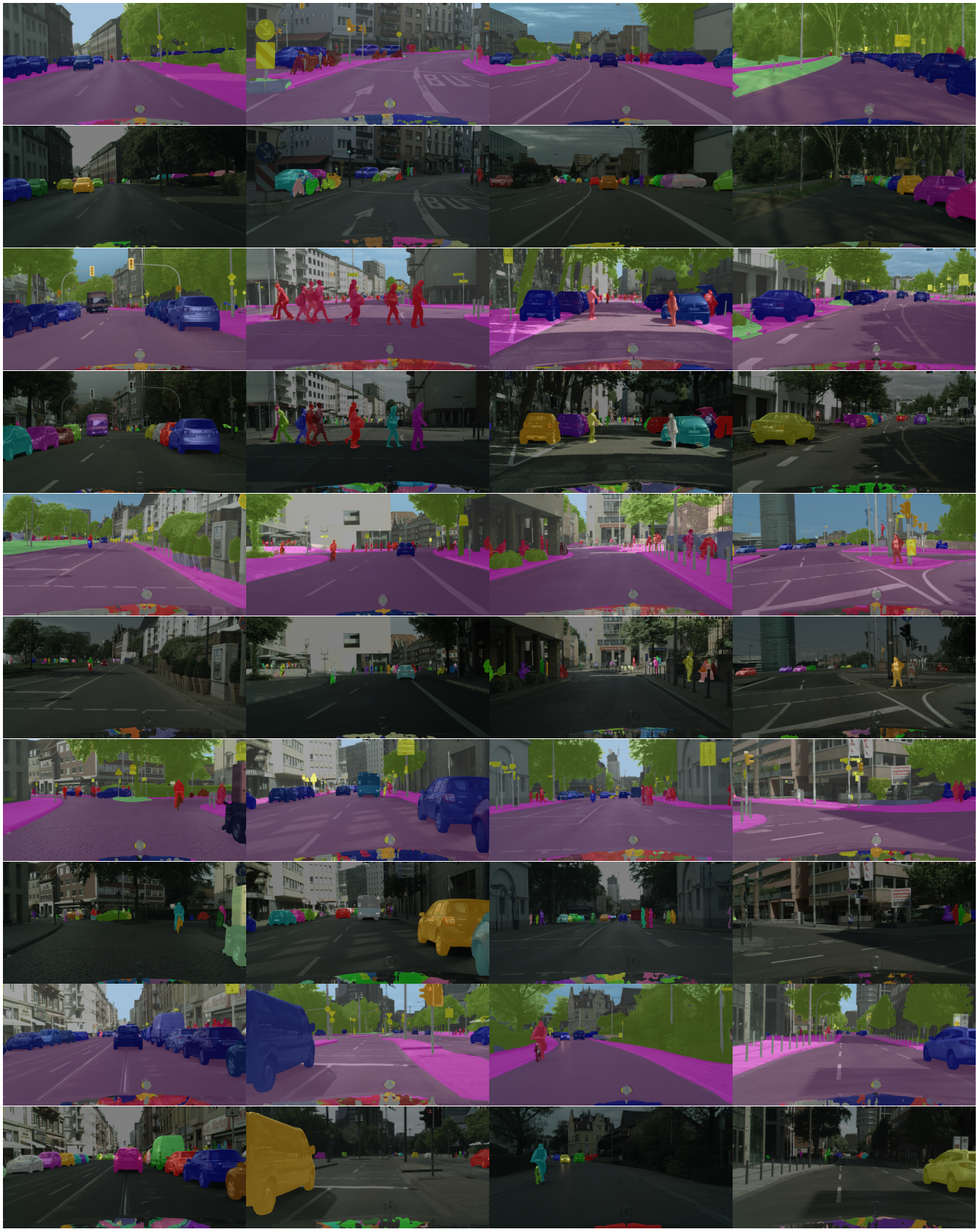


Figure 5. Visualization of Panoptic-DeepLab with Xception-71 on Cityscapes *val* set. Only single scale inference is used and the model achieves 63.0% PQ. The first row is panoptic prediction and the second row is instance prediction.

Figure 6. Visualization of Panoptic-DeepLab with Xception-71 on Mapillary Vistas *val* set. Only single scale inference is used and the model achieves 37.7% PQ. The first row is panoptic prediction and the second row is instance prediction.

Figure 7. Visualization of Panoptic-DeepLab with Xception-71 on COCO *val* set. Only single scale inference is used and the model achieves 39.7% PQ. The first row is panoptic prediction and the second row is instance prediction.