Supplementary Material for
Unifying Training and Inference for Panoptic Segmentation

Qizhu Li Xiaojuan Qi* Philip H.S. Torr
University of Oxford
{qizhu.li, xiaojuan.qi, philip.torr}@eng.ox.ac.uk

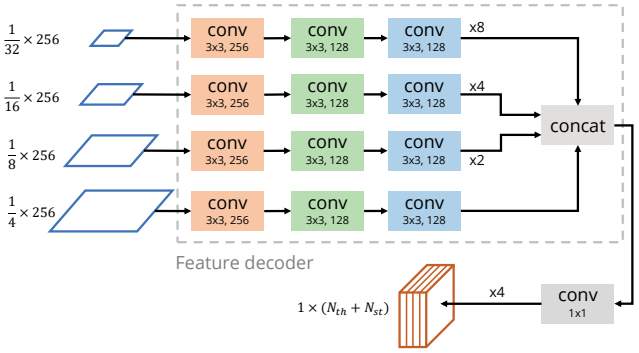

Figure A. Semantic segmentation submodule. Each 3×3 convolution block consists of a deformable convolution (with the indicated number of output channels), a Group Normalisation operation, and a ReLU activation. Weights are **shared** across 3×3 convolution blocks with the same colour code.

Appendices

A. Architecture and design

A.1 Semantic segmentation submodule

Our semantic segmentation submodule is modified from [9], by performing Group Normalisation [8] after each 3×3 convolution. We illustrate the pipeline in Fig. A. Note that the architecture of the *feature decoder* inside this submodule is also adopted by our dense instance affinity head to extract affinity features $\boldsymbol{Q}$. This submodule is supervised by a cross-entropy loss, unless otherwise stated.

A.2 Object detection submodule

In our experiments, we use the standard box head from Faster-RCNN [7] and optionally the mask head from Mask-RCNN [2] for this submodule, following [9, 3]. For the mask head, we use the Lovasz Hinge loss to replace the binary cross entropy loss. Thanks to the modular design of our network, it is easy to substitute it with any other detector architecture.

*Xiaojuan Qi is now with the University of Hong Kong.

Dataset	Variant B PQ	Variant B SQ	Variant B RQ	Variant C PQ	Variant C SQ	Variant C RQ
Cityscapes	**61.4**	**81.8**	**74.7**	60.3	80.8	73.5
COCO	42.7	79.4	52.2	**43.4**	**79.6**	**53.0**

Table A. Ablation study on two design variants for the dynamic potential head. On Cityscapes, variant B outperforms variant C, whereas on COCO, variant C achieves higher accuracies.

		Classified as th.	Classified as st.
GT	th.	95.1	4.9
	st.	0.0	100.0

(1) Cityscapes

		Classified as th.	Classified as st.
GT	th.	90.1	9.9
	st.	4.8	95.2

(2) COCO

Table B. Confusion matrices between "thing" and "stuff" for semantic segmentation submodule outputs on Cityscapes and COCO validation sets. All numbers are percentages, normalised row-wise.

A.3 Dynamic potential head

We refer to the design variant B and C presented in Sec. 3.4.1 (Fig. 4). At first glance, variant B, which multiplies semantic segmentation probabilities $V_i(c_k)$ with mask scores $M_i(k)$, appears to be a more appropriate method than variant C which sums probabilities instead. The output of variant B is high only when both inputs are unanimously high. This can filter out spurious misclassifications from either input, and improve robustness towards false positive predictions. Indeed, on Cityscapes, we observe that variant B achieves a 1.1 PQ lead over the variant C counterpart (first row of Table A).

However, on COCO, we notice a high tendency for the semantic segmentation submodule to mistake "things" for "stuff" (Table. B2). The multiplicative action of variant B can systematically and substantially weaken the panoptic logits for "thing" classes, relative to the unattenuated panoptic logits of "stuff" classes. This can be undesirable for models whose semantic segmentation submodule is already prone to misclassifying "things" as "stuff". On the other hand,

Dets. for training	PQ	SQ	RQ	IoU	AP_{box}
Ground truths	58.6	80.0	72.0	77.8	36.8
Predictions	**59.0**	**80.1**	**72.4**	77.8	**38.1**

Table C. Comparison between two different training strategies. The top row uses ground truth detections to train the panoptic segmentation submodule, whereas the bottom row uses the ones predicted by the network on-the-fly. Results are reported on the Cityscapes validation set.

the opposite is true for variant C, as summation strengthens panoptic logits of "things" in comparison to unmodified "stuff" scores. This led us to use variant C for COCO, and we observe a 0.7 PQ improvement in comparison to B (second row of Table A).

A.4 Training with predicted detections

In contrast with the practice in [9], we argue that, during training, the dynamic potential head should use predicted detections instead of ground truth ones to construct its output $\mathbf{\Psi}$. This allows the network to learn from realistic examples, and build up its robustness towards imperfections in detection localisation and scoring. To test our hypothesis, we carried out an ablation study on Cityscapes using our mask-free model. When training with ground truth boxes, a uniform score of 1.0 is used for their confidence scores. Results are shown in Table C. As expected, training with predicted detections yields performance improvements across all panoptic metrics, including a 0.4 increase in PQ. A large boost in observed for AP_{box} ($+1.3$), because training with predicted boxes allows gradients from the panoptic segmentation submodule to flow to the object detection submodule, giving it more fine-grained supervision. IoU has not changed, as this ablation setting does not affect the semantic segmentation module.

B. Implementation details

Cityscapes training. We run our experiments on four V100-32GB GPUs. This allows us to load each GPU with eight image crops and obtain an effective batch size of 32. The large number of crops per GPU enables us to use a Lovasz Softmax loss [1] instead of a cross entropy loss for supervising semantic segmentation, which we found to be effective. Following [3], we use a base learning rate of 0.01, a weight decay of 0.0001, and train for a total of 65k iterations. The learning rate is reduced by 10 folds after the first 40k iterations, and once more after another 15k iterations. Additionally, we adopt a "warm-up" period at the start of training – linearly increasing the learning rate from a third of the base rate to the full rate in 500 iterations, which helps stabilise the training.

We augment input images on-the-fly during training to reduce the network's tendency to overfit. Our augmentation pipeline resizes the input image by a random factor between 0.5 and 2, takes a random 512×1024 crop, and applies a horizontal flip with 50% chance. On top of these techqniues, we also apply image relighting, randomly adjusting the brightness, contrast, hue, and saturation of the image by a small amount, as used in [3].

COCO training. On COCO, as the dataset is larger than Cityscapes, less overfitting is observed. Therefore, in terms of data augmentation techniques, we only apply resizing where the shorter size is resized to 800 and the longer size is kept under 1333, and random horizontal flipping with 0.5 probability.

Miscellaneous. We use ImageNet pretrained ResNet-50 to initialise all experiments. The batch normalisation statistics are kept unchanged, though further performance gains are likely if they are finetuned on the target dataset. When a normalisation step is used in either the semantic or panoptic submodules, we use the Group Normalisation operation [8], as it is less sensitive to batch sizes.

Inference. We conduct single-scale inference for all experiments, letting the network process and make predictions on full-resolution images in a single forward pass. Note that only detection predictions whose confidence scores are more than a threshold are fed into the dynamic potential head during inference, to minimise unnecessary computation. This cut-off is 0.5 and 0.75 for Cityscapes and COCO respectively.

C. Evaluation of "stuff"

The PQ metrics effectively treats "stuff" classes as image-wide instances – making all "stuff" segments undergo the same matching procedure with ground truth segments as "thing" segments. While this approach has its merits including a unified evaluation logic and a simplified PQ implementation, it should be noted that matching "stuff" predictions to ground truth is not strictly necessary, since at most one "stuff" instance for each "stuff" class is present per image.

Furthermore, this approach towards "stuff" is neither robust nor fair as a measure for "stuff" segmentation quality, and arguably encourages post-processing of panoptic predictions. Under the PQ formulation, misclassifying even a single pixel into a "stuff' class absent in the ground truth will increment false positive detections by one, and such mistakes – exacerbated by the relatively small number of ground truth "stuff" segments in a dataset – attract a large penalty on the "stuff" RQ, even though the practical impact on perceptual quality is minimal. This also contrasts in spirit with the mean IoU metric widely adopted to measure semantic segmentation quality, as the mean IoU accumulates

Table D. Comparison of various evaluation metrics for "stuff", before and after small stuff areas are set to "void" on Cityscapes validation set. Note that the IoU^{st} here is computed from the final panoptic segmentation, by combining instances of the same semantic class. This is different from the IoU metrics reported in Table 1 and 3, which measure the quality of the semantic segmentation input to the heuristic merger / our panoptic segmentation submodule.

Model	Trim stuff	PQ^{st}	SQ^{st}	RQ^{st}	IoU^{st}
Pan. FPN [3]*		59.9	79.3	72.9	74.7
Pan. FPN [3]*	✓	62.0	79.6	75.5	74.5
		+2.1	+0.3	+2.6	-0.2
UPSNet [9]†		60.5	79.8	73.6	75.8
UPSNet [9]†	✓	62.8	80.0	76.3	75.7
		+2.3	+0.2	+2.7	-0.1
Ours		64.2	81.4	77.1	78.3
Ours	✓	66.3	81.8	79.4	78.2
		+2.1	+0.4	+2.3	-0.1

* Results obtained from our re-implementation of Panoptic FPN.
† Results obtained by running the public inference script of [9].

intersection and union counts over the whole dataset and is minimally affected by individual pixels.

On the other hand, CNN-based semantic segmentation models are typically prone to produce spurious misclassifications, as they usually do not explicitly enforce smoothness. As a result, recent panoptic segmentation works [4, 6, 5, 3, 9, 10] collectively resort to setting small "stuff" segments to "void" in the final panoptic segmentation. Therefore, to foster meaningful comparison with other state-of-the-art panoptic segmentation approaches, unless specified otherwise, we also carry out this strategy as part of evaluation.

Effects of trimming small stuff segments on evaluation metrics. On Cityscapes validation set, we test our full model, our re-implemented Panoptic FPN [3], and the released UPSNet model [9] with and without trimming off small "stuff" regions, to quantitatively assess the impact of this step on state-of-the-art models. The findings are reported in Table D.

The results show that PQ and RQ are very sensitive to such operations, as removing small stuff segments consistently results in an increase of approximately 2 points for "stuff" PQ, and 2.5 points for "stuff" RQ. This can be largely attributed to the reduced number of false positive stuff segments. On the other hand, the "stuff" IoU metric is insensitive to such modifications, as in all three cases, it suffers a slight decrease of 0.1 or 0.2 points. This prompts us to believe that "stuff" IoU is a better metric for capturing "stuff" segmentation quality than the "thing"-centric PQ family.

D. Detailed validation set results

We report the detailed results of our models on the Cityscapes and COCO validation sets in Table E. In addition to the metrics reported in the main paper, this table also includes breakdowns of SQ and RQ by "stuff" and "thing".

E. Visualisation of learnt instance affinities

Additional visualisations of some predicted instance affinities are provided in Fig. B. Note that these instance affinities are extracted from our mask-free model. Interestingly, the model has learnt to resolve cars regions covered by multiple car bounding boxes – a problem difficult for methods only using boxes as localisation cues – by creating strong instance affinities to the bottoms and tyres of cars. The model has found that these regions of cars are normally not covered by multiple bounding boxes, and therefore it is most helpful for instance discrimination by associating uncertain pixels with these regions.

F. Qualitative results

We show more qualitative results in Fig. C and D, and comparisons to previous state-of-the-art methods [3, 9].

References

[1] Maxim Berman, Amal Rannen Triki, and Matthew B Blaschko. The lovász-softmax loss: A tractable surrogate for the optimization of the intersection-over-union measure in neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4413–4421, 2018. 2

[2] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *ICCV*, pages 2961–2969, 2017. 1

[3] Alexander Kirillov, Ross Girshick, Kaiming He, and Piotr Dollár. Panoptic feature pyramid networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6399–6408, 2019. 1, 2, 3

[4] Alexander Kirillov, Kaiming He, Ross Girshick, Carsten Rother, and Piotr Dollár. Panoptic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 9404–9413, 2019. 3, 5

[5] Jie Li, Allan Raventos, Arjun Bhargava, Takaaki Tagawa, and Adrien Gaidon. Learning to fuse things and stuff. *arXiv preprint arXiv:1812.01192*, 2018. 3

[6] Yanwei Li, Xinze Chen, Zheng Zhu, Lingxi Xie, Guan Huang, Dalong Du, and Xingang Wang. Attention-guided unified network for panoptic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7026–7035, 2019. 3

[7] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in neural information processing systems*, pages 91–99, 2015. 1

[8] Yuxin Wu and Kaiming He. Group normalization. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 3–19, 2018. 1, 2

Dataset	Method	PQ all	PQ th.	PQ st.	SQ all	SQ th.	SQ st.	RQ all	RQ th.	RQ st.	IoU all	IoU th.	IoU st.	AP mask	AP box
Cityscapes	Ours (w/o mask)	59.0	50.2	65.3	80.1	78.4	81.2	72.4	63.9	78.6	77.8	78.7	77.2	–	38.1
Cityscapes	Ours (w/ mask)	61.4	54.7	66.3	81.1	80.0	81.8	74.7	68.2	79.4	79.5	81.0	78.4	33.7	38.8
COCO	Ours (w/ mask)	43.4	48.6	35.5	79.6	80.0	78.9	53.0	59.2	43.8	53.7	60.4	43.6	36.4	40.5

Table E. Full panoptic segmentation results on Cityscapes validation set and COCO validation set. All models are ResNet-50 based, and tested with a *single-scale* inference scheme, without test-time augmentation.

(a) Image with boxes (b) Ground truth (c) Instance affinities (d) Panoptic prediction

Figure B. Additional examples of instance affinities. In (c), we show the instance affinities – overlaid on input images to aid visualisation – of the cross-marked pixels in (a). These affinities and predictions are predicted by our mask-free models which use only bounding boxes. They can be seen to help segment full objects when bounding box localisation is poor (Row 1), and attribute pixels within multiple bounding boxes to the correct instances (Row 2 to 5). For Row 4, our proposed method is able to overcome a false positive detection, as the dynamic potential is robust towards false detections. For Row 5, the cross-marked pixel is on the wing mirror of the closest silver car, and our fine-grained instance affinity is able to attribute the mirror to the correct car, while the ground truth has failed to correctly label as such.

(a) Image (b) Ground truth (c) Heuristic Fusion [4] (d) Ours

Figure C. Qualitative results on Cityscapes. Column (c) and (d) are produced by the same model under different inference strategies – either by heuristic merger [4] or with our proposed panoptic segmentation submodule. Row 1 to 3 shows that our model are able to revise erroneous cues and resolve conflicts between overlapping object masks. Row 4 and 5 demonstrate our network's ability to segment outside boxes, when boxes do not cover the full extent of an object.

[9] Yuwen Xiong, Renjie Liao, Hengshuang Zhao, Rui Hu, Min Bai, Ersin Yumer, and Raquel Urtasun. Upsnet: A unified panoptic segmentation network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8818–8826, 2019. 1, 2, 3, 6

[10] 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

(a) Image (b) Ground truth (c) UPSNet [9] (d) Ours

Figure D. Qualitative results on COCO. Column (c) is produced by running the publicly available inference script of [9]. With our parametrised panoptic segmentation submodule, we are able to produce more coherent, accurate, and visually appealing predictions than the parameter-free approach of [9].