Exploring Cross-Image Pixel Contrast for Semantic Segmentation
Supplemental Material

Wenguan Wang[1*], Tianfei Zhou[1*], Fisher Yu[1], Jifeng Dai[2], Ender Konukoglu[1], Luc Van Gool[1]
[1] Computer Vision Lab, ETH Zurich [2] SenseTime Research
https://github.com/tfzhou/ContrastiveSeg

In this document, we first present additional quantitative results on Cityscapes `val` (*cf.* §A). Then, we conduct an empirical analysis of our contrastive loss against other semantic segmentation loss designs (*cf.* §B). Last, we provide more qualitative semantic segmentation results on Cityscapes `val` [7], PASCAL-Context `test` [12], COCO-Stuff `test` [3], and CamVid `test` [2] (*cf.* §C).

A. Additional Quantitative Result

Table 1 provides comparison results with representative approaches on Cityscapes `val` [7] in terms of mIoU and training speed. We train our models on Cityscapes `train` for 80,000 iterations with a mini-batch size of 8. We find that, by equipping with cross-image pixel contrast, the performance of baseline models enjoy consistently improvements (**1.2/1.1/0.8** points gain over DeepLabV3, HRNetV2 and OCR, respectively). We also carry out extra experiments over lightweight backbones (*i.e.*, MobileNet V1/V2/V3) for DeepLab V3. All the models are trained for 80,000 iterations with a mini-batch size of 16. As seen, our method also at- tains consistent improvements on lightweight backbones. In addition, the contrastive loss computation brings negligible training speed decrease, and does not incur any additional overhead during inference.

Model	Backbone	sec./iter.	mIoU (%)
SegSort$_{19}$ [8]	D-ResNet-101	-	78.2
AAF$_{18}$ [9]	D-ResNet-101	-	79.2
DeepLabV3+$_{18}$ [5]	D-Xception-71	-	79.6
PSPNet$_{17}$ [16]	D-ResNet-101	-	79.7
Auto-DeepLab-L$_{19}$ [11]	-	-	80.3
HANet$_{20}$ [6]	D-ResNet-101	-	80.3
SpyGR$_{20}$ [10]	D-ResNet-101	-	80.5
ACF$_{19}$ [15]	D-ResNet-101	-	81.5
DeepLabV3$_{17}$	MobileNet-V1	0.19	70.8
DeepLabV3+ **Ours**	MobileNet-V1	0.31	**72.1** (+1.3)
DeepLabV3$_{17}$	MobileNet-V2	0.21	71.3
DeepLabV3+ **Ours**	MobileNet-V2	0.35	**72.3** (+1.0)
DeepLabV3$_{17}$	MobileNet-V3	0.21	70.7
DeepLabV3+ **Ours**	MobileNet-V3	0.31	**71.9** (+1.2)
DeepLabV3$_{17}$ [4]	D-ResNet-101	1.18	78.5
DeepLabV3+ **Ours**	D-ResNet-101	1.37	**79.7** (+1.2)
HRNetV2$_{20}$ [13]	HRNetV2-W48	1.67	81.1
HRNetV2+ **Ours**	HRNetV2-W48	1.87	**82.2** (+1.1)
OCR$_{20}$ [14]	D-ResNet-101	1.29	80.6
OCR+ **Ours**	D-ResNet-101	1.41	**81.2** (+0.6)
OCR$_{20}$ [14]	HRNetV2-W48	1.75	81.6
OCR+ **Ours**	HRNetV2-W48	1.90	**82.4** (+0.8)

Table 1: **Quantitative semantic segmentation results** on Cityscapes `val` [7]. D-ResNet-101 = Dilated-ResNet-101. D-Xception-71 = Dilated-Xception-71. See §A for more details.

B. Comparison to Other Losses

We further study the effectiveness of our contrastive loss against representative semantic segmentation losses, including Cross-Entropy (CE) Loss, AAF loss [9], Lovász Loss [1], and RMI Loss [17].

For fair comparison, we examine each loss using HRNetV2 [13] as the base segmentation network, and train the loss jointly with CE on Cityscapes `train` for 40,000 iterations with a mini-batch size of 8. The results are reported in Table 2. We observe that all structure-aware losses outperform the standard CE loss. Notably, our contrastive loss achieves the best performance, outperforming the second-best Lovász loss by **0.7%**, and the pairwise losses, *i.e.*, RMI and AAF, by **1.2%** and **2.3%**, respectively.

Additionally, Table 2 reports results of each loss in combination with our contrastive loss. From a perspective of metric learning, the CE loss can be viewed as a pixel-wise *unary* loss that penalizes each pixel independently and ignores dependencies between pixels, while AAF is a *pairwise* loss, which models the pairwise relations between spatially adjacent pixels. Moreover, the RMI and Lovász losses are *higher-order* losses: the former one accounts for region-level mutual information, and the latter one directly optimizes the intersection-over-union score over the pixel clique level. However, all these existing loss designs are defined within individual images, capturing *local* context/pixel relations only. Our contrastive loss, as it explores *pairwise* pixel-to-pixel dependencies, is also a

*The first two authors contribute equally to this work.

Loss	Type	Context	Backbone	mIoU (%)
Cross-Entropy Loss	unary	local	HRNetV2-W48	78.1
+AAF Loss [9]	pairwise	local	HRNetV2-W48	78.7
+RMI Loss [17]	higher-order	local	HRNetV2-W48	79.8
+Lovász Loss [1]	higher-order	local	HRNetV2-W48	80.3
+Contrastive Loss (**Ours**)	pairwise	global	HRNetV2-W48	81.0
+AAF [9] + Contrastive	-	-	HRNetV2-W48	81.0
+RMI [17] + Contrastive	-	-	HRNetV2-W48	81.3
+Lovász [1] + Contrastive	-	-	HRNetV2-W48	81.5

Table 2: **Comparison of different loss designs** on Cityscapes `val` [7]. See §B for more details.

pairwise loss. But it is computed over the whole training dataset, addressing the *global* context over the whole data space. Therefore, AAF can be viewed as a specific case of our contrastive loss, and additionally considering AAF does not bring any performance improvement. For other losses, our contrastive loss are complementary to them (global *vs.* local, pairwise *vs.* higher-order) and thus enables further performance uplifting. This suggests that designing a higher-order, global loss for semantic segmentation is a promising direction.

C. More Qualitative Result

We provide additional qualitative improvements of HRNetV2+Ours over HRNetV2 [13] on four benchmarks, including Cityscapes `val` [7] in Fig. 1, PASCAL-Context `test` [12] in Fig. 2, COCO-Stuff `test` [3] in Fig. 3, and CamVid `test` [2] in Fig. 4. The improved regions are marked by dashed boxes. As can be seen, our approach is able to produce great improvements on those hard regions, *e.g.*, small objects, cluttered background.

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 *CVPR*, 2018. 1, 2

[2] Gabriel J Brostow, Julien Fauqueur, and Roberto Cipolla. Semantic object classes in video: A high-definition ground truth database. *PRL*, 30(2):88–97, 2009. 1, 2, 4

[3] Holger Caesar, Jasper Uijlings, and Vittorio Ferrari. Cocostuff: Thing and stuff classes in context. In *CVPR*, 2018. 1, 2, 4

[4] Liang-Chieh Chen, George Papandreou, Florian Schroff, and Hartwig Adam. Rethinking atrous convolution for semantic image segmentation. *arXiv preprint arXiv:1706.05587*, 2017. 1

[5] 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. 1

[6] Sungha Choi, Joanne T Kim, and Jaegul Choo. Cars can't fly up in the sky: Improving urban-scene segmentation via height-driven attention networks. In *CVPR*, 2020. 1

[7] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *CVPR*, 2016. 1, 2, 3

[8] Jyh-Jing Hwang, Stella X Yu, Jianbo Shi, Maxwell D Collins, Tien-Ju Yang, Xiao Zhang, and Liang-Chieh Chen. Segsort: Segmentation by discriminative sorting of segments. In *ICCV*, 2019. 1

[9] Tsung-Wei Ke, Jyh-Jing Hwang, Ziwei Liu, and Stella X Yu. Adaptive affinity fields for semantic segmentation. In *ECCV*, 2018. 1, 2

[10] Xia Li, Yibo Yang, Qijie Zhao, Tiancheng Shen, Zhouchen Lin, and Hong Liu. Spatial pyramid based graph reasoning for semantic segmentation. In *CVPR*, 2020. 1

[11] Chenxi Liu, Liang-Chieh Chen, Florian Schroff, Hartwig Adam, Wei Hua, Alan L Yuille, and Li Fei-Fei. Auto-deeplab: Hierarchical neural architecture search for semantic image segmentation. In *CVPR*, 2019. 1

[12] Roozbeh Mottaghi, Xianjie Chen, Xiaobai Liu, Nam-Gyu Cho, Seong-Whan Lee, Sanja Fidler, Raquel Urtasun, and Alan Yuille. The role of context for object detection and semantic segmentation in the wild. In *CVPR*, 2014. 1, 2, 4

[13] 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. *IEEE TPAMI*, 2020. 1, 2, 3, 4

[14] Yuhui Yuan, Xilin Chen, and Jingdong Wang. Object-contextual representations for semantic segmentation. In *ECCV*, 2020. 1

[15] Fan Zhang, Yanqin Chen, Zhihang Li, Zhibin Hong, Jingtuo Liu, Feifei Ma, Junyu Han, and Errui Ding. Acfnet: Attentional class feature network for semantic segmentation. In *ICCV*, 2019. 1

[16] Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia. Pyramid scene parsing network. In *CVPR*, 2017. 1

[17] Shuai Zhao, Yang Wang, Zheng Yang, and Deng Cai. Region mutual information loss for semantic segmentation. In *NeurIPS*, 2019. 1, 2

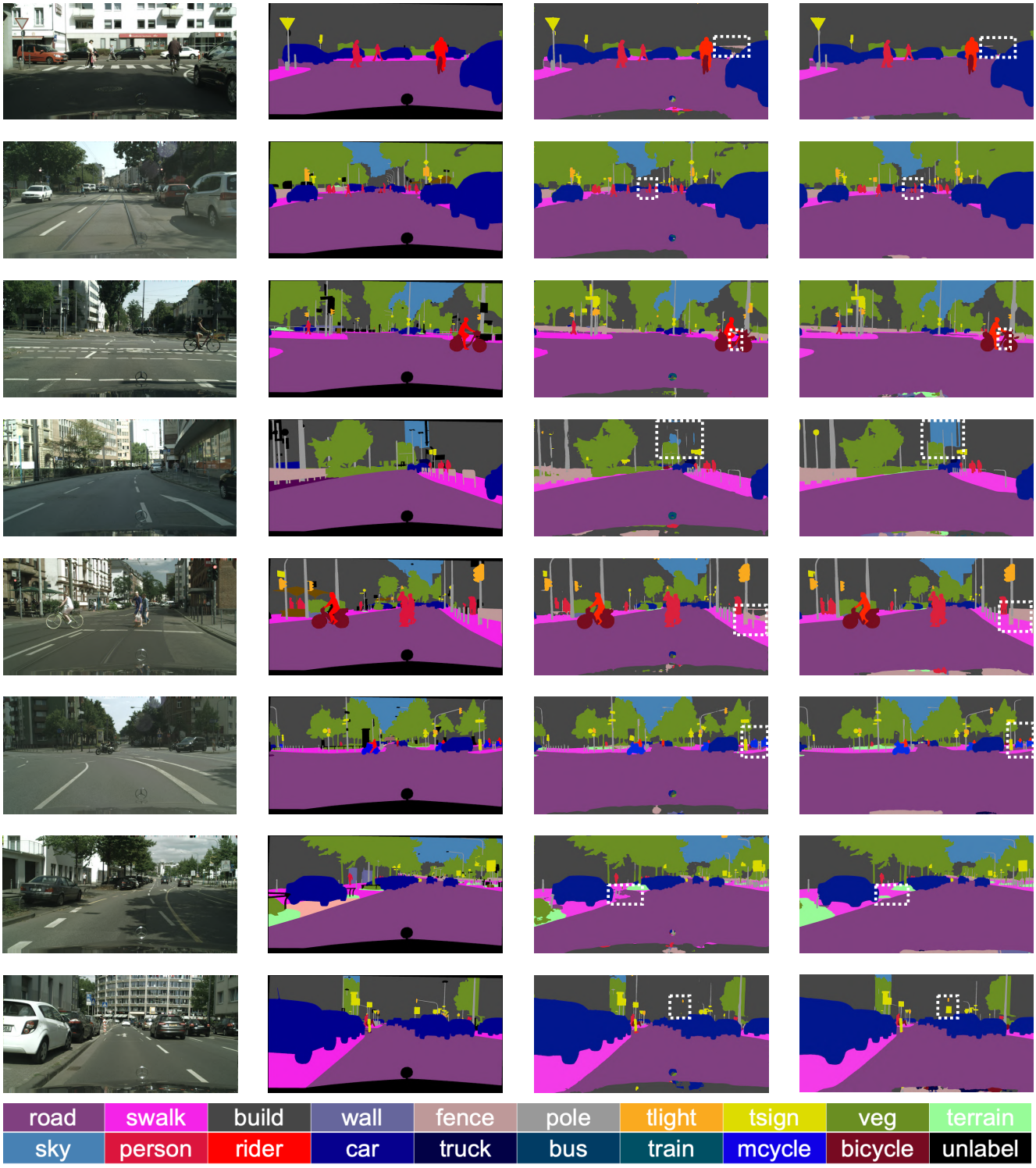

Figure 1: **Qualitative semantic segmentation results** on Cityscapes `val` [7]. From left to right: input images, ground-truths, results of HRNetV2 [13], results of HRNetV2+Ours. The improved regions are marked by white dashed boxes.

Figure 2: **Qualitative semantic segmentation results** on PASCAL-Context `test` [12]. From left to right: input images, ground-truths, results of HRNetV2 [13], results of HRNetV2+Ours. The improved regions are marked by black dashed boxes.

Figure 3: **Qualitative semantic segmentation results** on COCO-Stuff `test` [3]. From left to right: input images, ground-truths, results of HRNetV2 [13], results of HRNetV2+Ours. The improved regions are marked by black dashed boxes.

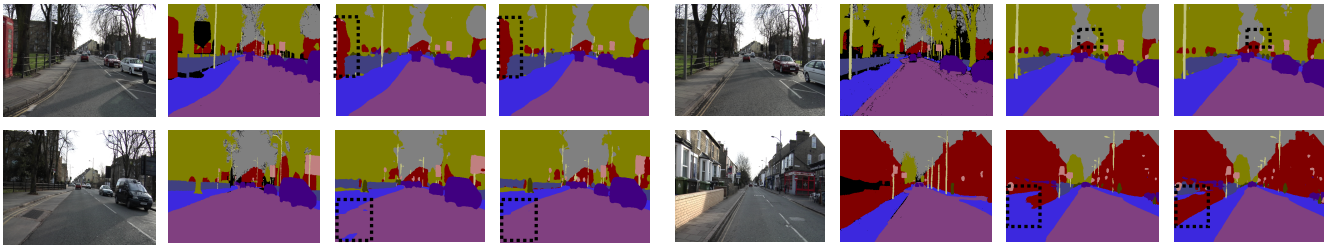

Figure 4: **Qualitative semantic segmentation results** on CamVid `test` [2]. From left to right: input images, ground-truths, results of HRNetV2 [13], results of HRNetV2+Ours. The improved regions are marked by black dashed boxes.