Tree Energy Loss: Towards Sparsely Annotated Semantic Segmentation
Supplenmentary Material

Zhiyuan Liang[1] Tiancai Wang[2] Xiangyu Zhang[2]
Jian Sun[2] Jianbing Shen[3*]
[1]Beijing Institute of Technology [2]MEGVII Technology [3]SKL-IOTSC, University of Macau

1. Failure cases

Fig. 1 shows some failure cases of our method. The predictive error occurs due to the color mutation within one object region or the color similarity between two semantic objects. Since our approach partially relies on the low-level color prior to generate the pseudo labels, so it may make wrong predictions under point-wise supervision. Fortunately, the predictive error can be reduced by introducing more complete annotations (i.e., scribble annotation).

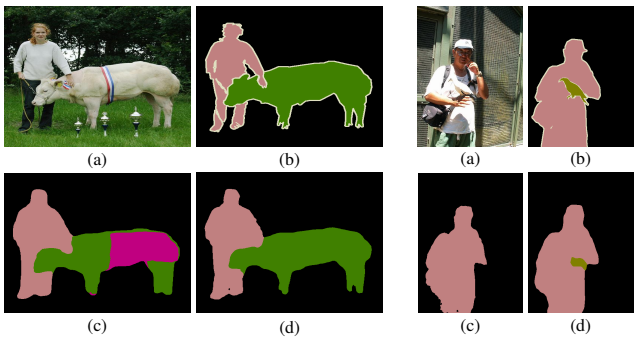

Figure 1. Failure cases of the proposed method on Pascal VOC 2012 dataset [2]. (a) input image. (b) ground truth. (c) result with point-wise supervision. (d) result with scribble-wise supervision.

2. Computational costs

The proposed method can be plugged into most existing semantic segmentation frameworks. During inference time, the auxiliary branch is removed to avoid extra memory costs. During training, the GPU loads on Pascal VOC 2012 dataset are reported in Tab. 1. The DeeplabV3+ and the LTF [3] are selected as the baselines. The resolution of the input image is 512×512 and the batch size is 16. Leveraging the low-level information almost introduces no extra memory cost. Overall, introducing both low-level and high-level information requires 1.93 and 1.78 GB GPU loads for DeeplabV3+ and LTF models, respectively.

Low-level	High-level	DeeplabV3+	LTF
		32.85	46.65
✓		32.86 (+0.01)	46.67 (+0.02)
✓	✓	34.78 (+1.93)	48.43 (+1.78)

Table 1. The memory cost (GB) during training, leveraging various levels of structural information.

3. Visualizations of block-supervised settings

We synthesize the block-wise annotations for ADE20k [4] and Cityscapes [1] datasets. Given the full annotations of the image, we discard the annotations from the semantic boundary to the interior region until the ratio of the rest annotations reaches the preset threshold. The block-wise annotations are generated at 3 levels, including 10%, 20%, and 50% of full annotations. The visualizations of block-wise annotation for the two datasets are respectively illustrated in Fig. 2 and Fig. 3. Moreover, qualitative results of our method on the ADE20k and the Cityscapes datasets are illustrated in Fig. 4 and Fig. 5, respectively. Here, HRNet is selected as the segmentation model. The proposed approach can be used to train the segmentation model with the annotations of different sparsity.

*Corresponding author: *Jianbing Shen*.

References

[1] 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] Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *IJCV*, 88(2):303–338, 2010. 1

[3] Lin Song, Yanwei Li, Zeming Li, Gang Yu, Hongbin Sun, Jian Sun, and Nanning Zheng. Learnable tree filter for structure-preserving feature transform. In *NeurIPS*, 2019. 1

[4] Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Scene parsing through ade20k dataset. In *CVPR*, 2017. 1

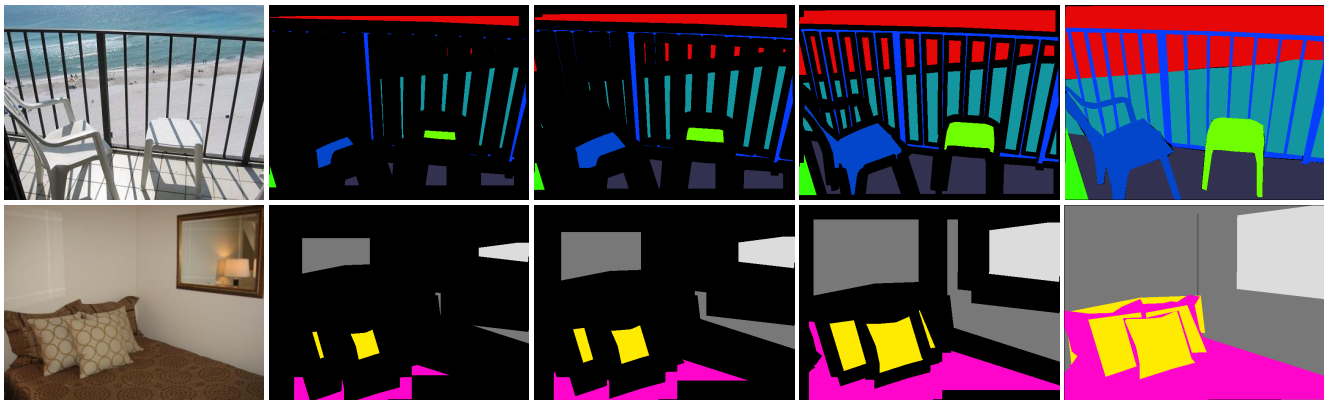

image ratio 10% ratio 20% ratio 50% ground truth

Figure 2. The block-wise annotations for ADE20k dataset.

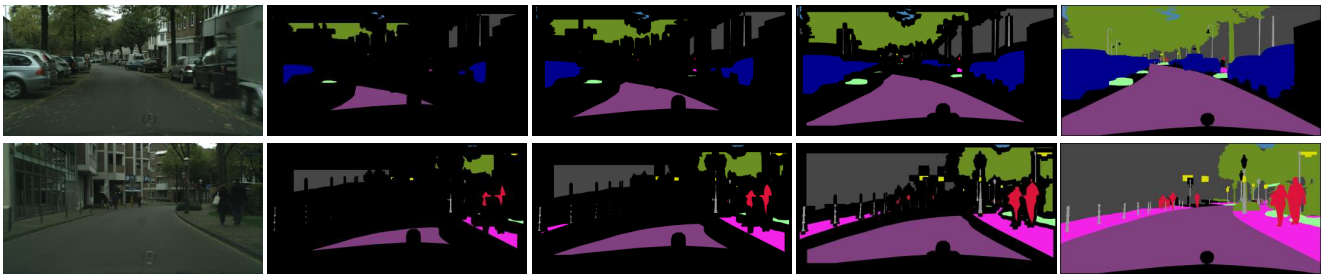

image ratio 10% ratio 20% ratio 50% ground truth

Figure 3. The block-wise annotations for Cityscapes dataset.

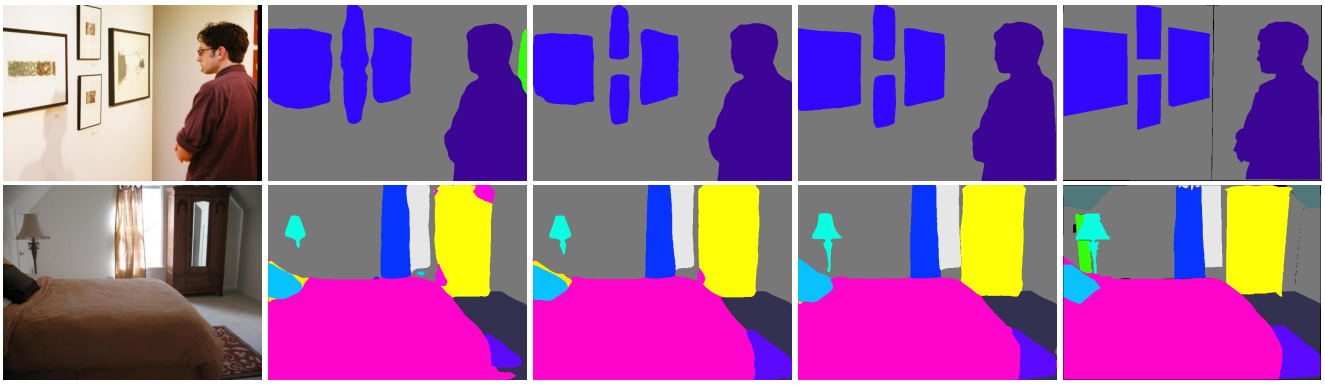

image ratio 10% ratio 20% ratio 50% ground truth

Figure 4. Qualitative results for the proposed TEL on ADE20k dataset.

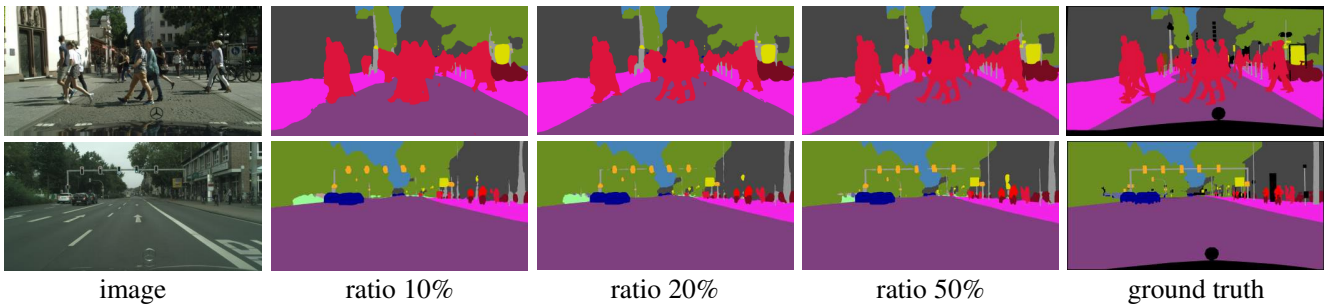

image ratio 10% ratio 20% ratio 50% ground truth

Figure 5. Qualitative results for the proposed TEL on Cityscapes dataset.