Scribble-Supervised Semantic Segmentation by Uncertainty Reduction on Neural Representation and Self-Supervision on Neural Eigenspace
Supplementary Material

Zhiyi Pan[1] Peng Jiang[1*] Yunhai Wang[1] Changhe Tu[1] Anthony G. Cohn[2,1]
[1]Shandong University, China [2]University of Leeds, UK
{panzhiyi1996, sdujump, cloudseawang, changhe.tu}@gmail.com a.g.cohn@leeds.ac.uk

1. Computing Matrix

For self-supervision on eigenspace, we compute the KL divergence distance between $P(t_\phi(x))$ and $T_\phi(P(x))$, where t_ϕ and T_ϕ are corresponding transform operations on x and $P(x)$. Though $T_\phi(P(x))$ has a complicated form, it can be defined as the multiplication of the original $P(x)$ with the predefined computing matrices to facilitate computation. In Fig. 1, we visualize computing matrices for horizontal flip and vertical translation when using soft eigenspace self-supervision. The formulation can be defined as,

$$T_\phi(P(x)) = T_\phi r \cdot P(x) \cdot T_\phi c, \tag{1}$$

where $T_\phi r$ and $T_\phi c$ are predefined computing matrices for transform ϕ. The detail definitions of predefined computing matrices are given below.

1.1. Horizontal Flip

Assuming that the size of x is $M \times N$, then the size of $P(x)$ is $MN \times MN$.

$$T^{i,j}r = \begin{cases} 1 & if\ i+j = kN\ and \\ & |i-j| < N\ , k = 1, 2.., M\ . \\ 0 & otherwise \end{cases} \tag{2}$$

Tc share the same definition as Tr. Tr and Tc are also of the size $MN \times MN$.

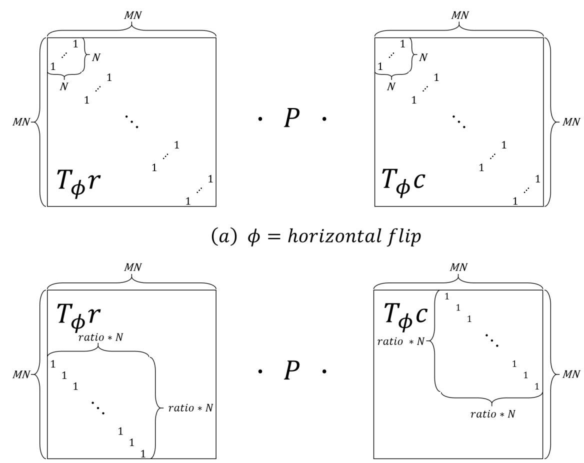
1.2. Vertical Translation

Assuming that the size of x is $M \times N$, take the ratio of vertical translation as η. T_r and T_c are defined as

$$T^{i,j}r = \begin{cases} 1 & if\ j-i = \eta MN \\ 0 & otherwise \end{cases}, \tag{3}$$

$$T^{i,j}c = \begin{cases} 1 & if\ i-j = \eta MN \\ 0 & otherwise \end{cases}. \tag{4}$$

*Corresponding Author



$(a)\ \phi = horizontal\ flip$

$(b)\ \phi = vertical\ translation$

Figure 1. Visualization of predefined computing matrices for self-supervision on eigenspace.

2. More Detailed Experimental Results

2.1. Hyper-parameters

The proposed method has 100 epochs of training, with the first 50 epochs have no self-supervised loss. For every step, sixteen images (batch size) are randomly selected to train the network with Adam [2] optimizer, Sync-BatchNorm [1] and learning rate as $1e$-3 for the first 50 epochs and $1e$-4 for the rest. The total loss in our work is defined as:

$$L = \sum_{p \in \Omega_\mathcal{L}} c(s(x)_p, y_p) + \omega_1 E_{\Omega_{-\mathcal{B}}} + \omega_2 * ss_P(x, \phi), \tag{5}$$

where the weights γ, ω_1 and ω_2 are set to be 0.01, 1 and 30, respectively. Moreover, in common with other approaches to semantic segmentation [3], data augmentation is performed during training.

category	bkg	plane	bike	bird	boat	bottle	bus	car	cat	chair	cow
mIoU	93.3	83.4	35.7	85.3	66.8	76.1	89.8	86.6	91.5	42.0	89.9
category	table	dog	horse	mbike	person	plant	sheep	sofa	train	monitor	mean
mIoU	59.3	89.0	85.2	81.0	85.0	66.2	86.0	49.9	83.4	71.7	76.1

Table 1. Detail mIoU scores on the validation set of *scribblesup*.

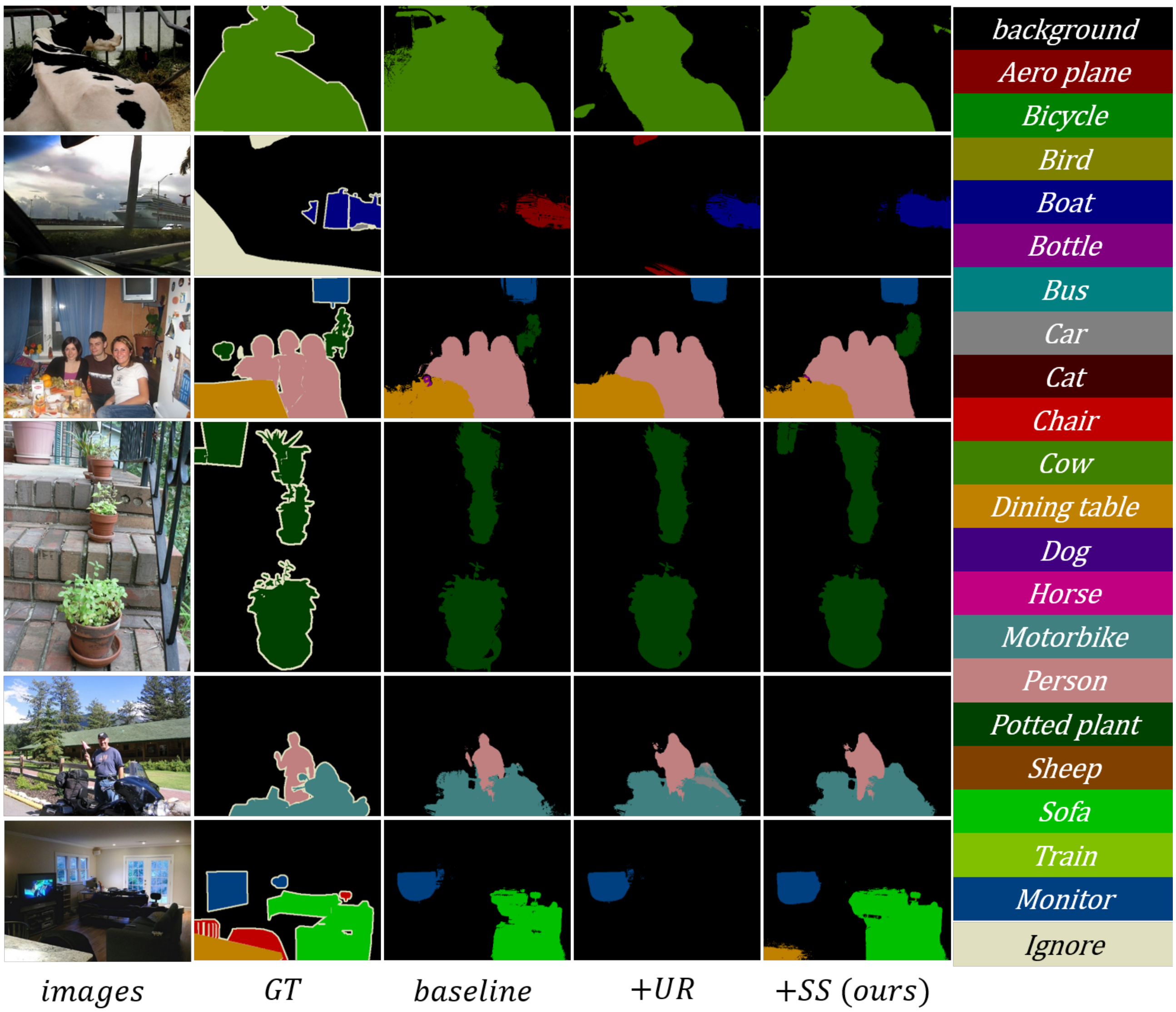


Figure 2. Visual results on the validation set of *scribblesupp*

2.2. Quantitative Results

In this part, we report the detailed mIoU score on every category on the validation set of *scribblesup* in Tab. 1. Our method performs well on most categories but still has room for improvement on categories with similar appearance (*e.g.* chair and sofa) and complex structure (*e.g.* bike and plant).

2.3. Qualitative Results

We show more visual comparison in Fig. 2. With the proposed uncertainty reduction (UR) and self-supervision on eigenspace (SS), the results are gradually refined, and the complete method shows significant improvement over the baseline.

References

[1] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International Conference on Machine Learning*, pages 448–456, 2015.

[2] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[3] Hengshuang Zhao, Yi Zhang, Shu Liu, Jianping Shi, Chen Change Loy, Dahua Lin, and Jiaya Jia. Psanet: Point-wise spatial attention network for scene parsing. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 267–283, 2018.