SegRCDB: Semantic Segmentation via Formula-Driven Supervised Learning Supplemental Material

Risa Shinoda[1], Ryo Hayamizu[1], Kodai Nakashima[1], Nakamasa Inoue[2], Rio Yokota[2], Hirokatsu Kataoka[1]
[1]National Institute of Advanced Industrial Science and Technology (AIST)
[2]Tokyo Institute of Technology

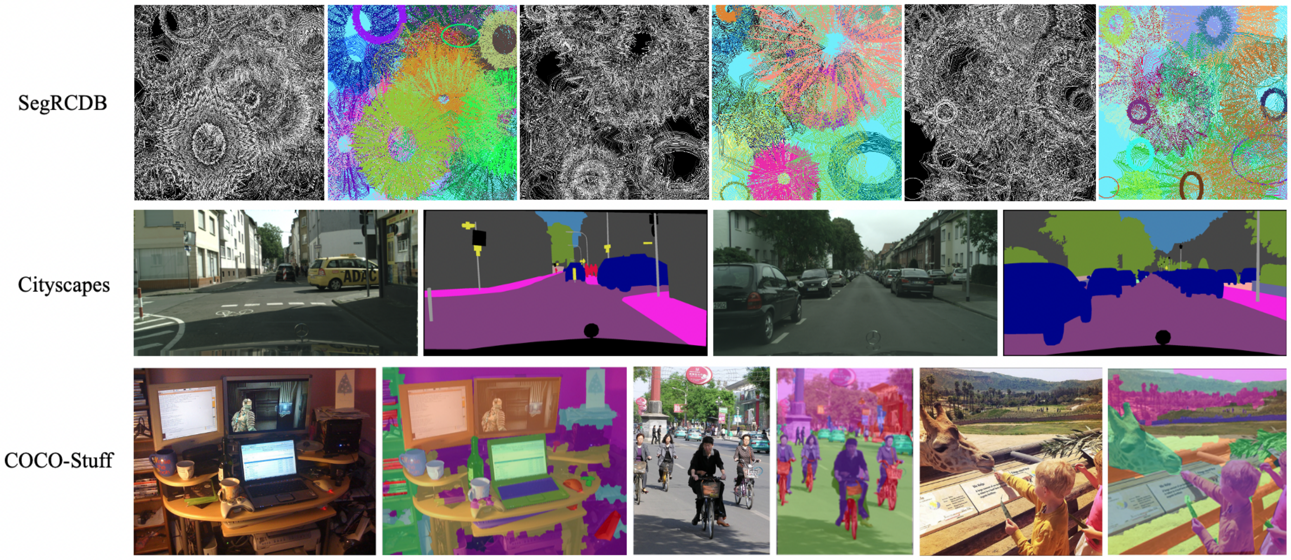

Figure 1: **Example images of semantic segmentation datasets.** The first row is the visualization of SegRCDB. The second row and third row are the visualization of Cityscapes [3] and COCO-Stuff [1].

1. Dataset details

We list the example images of SegRCDB and other datasets in Fig 1. SegRCDB generation code is available at: https://github.com/dahlian00/SegRCDB.

2. Experiment Details

Throughout this experiment, in the augmentation process, we adopt MMSegmentation [2] official settings. We use the following augmentations: resize, random crop, random flip, photometric distortion, normalize, and padding.

As the model architecture, we utilize UPerNet [6] as our base model with a Swin Transformer base [5] backbone. Throughout the paper, we adopt cross entropy loss with a loss weight of 1.0 for UPerNet head training.

2.1. Investigation (Main paper Sec 5.2)

Settings. We set 300 epochs for pre-training and 60 epochs for fine-tuning. In investigation (F1) - (F6), all parameters follow MMSegmentation official settings, with batch size set to 32. In (F7), fine-tuning is set to 120 epochs to allow for convergence. SegRCDB parameters in (F7) are set to the best result obtained from the investigation.

Table 1 compares the base parameters in the investigation with the parameter combinations that gave the best results. Table 2 shows the result using the base parameters and best parameters of SegRCDB with 20k images fine-tuned on ADE-20k [7]. Fine-tuning epoch is set to 60 epochs. This confirms that SegRCDB has acquired more effective image representations due to the investigation, with a 20.51 mIoU improvement over the baseline.

Table 1: SegRCDB parameter

	Base line	Best
Number of instances (N)	1	32
Mask type	$\boldsymbol{m}^1$	$\boldsymbol{m}^1$
Line width (d)	1.0	1.0
Number of polygons (K)	$\{1, 2...50\}$	$\{1, 2...25\}$
Occlusion (r)	512	400
Colors	grayscale	grayscale
Number of categories (C)	255	255

Table 2: Investigation result at ADE-20k-val. SegRCDB contains 20k training images. Fine-tuning is set to 60 epochs.

	Base line	Best
mIoU	19.21	39.72

2.2. Supervised pre-training for semantic segmentation datasets. (Main paper Sec 5.3)

Settings. For the pre-training of natural image datasets, we follow the MMSegmentation official settings, with the backbone learning rate set to 0.00024 and a batch size of 64. SegRCDB can converge even when using a higher learning rate; therefore, we set a higher backbone learning rate of 0.0006. In the fine-tuning phase, the learning rate is set to 0.0005, and the batch size to 16 for all datasets.

2.3. Backbone pre-training for semantic segmentation. (Main paper Sec 5.3)

Settings. For ImageNet-1k, RCDB-1k [4], and ExFractalDB-1k [4] training, we followed the official settings of Swin Transformer [5], with the exception of the augmentation process, where we adopt MMSegmentation's official settings. We use the following augmentations: resize, random crop, random flip, photometric distortion, normalize, and padding. SegRCDB pre-training settings are the same as Sec 2.2. Fine-tuning is 150 epochs long for all datasets.

In the fine-tuning phase using ImageNet backbone, the learning rate is set to 0.00006 following MMSegmentation official settings. The backbone learning rate of RCDB and ExFractalDB is set to 0.001 to allow for convergence. We also tested a learning rate of 0.001 when using a pre-trained ImageNet backbone for fairness. Table 3 shows the results of fine-tuning ImageNet with a backbone learning rate of 0.001. The fine-tuning from ImageNet shows lower mIoU compared to the official setting (ADE-20k: 43.60 – 46.37, Cityscapes: 71.85 –75.26). For this reason, in the main paper, we show follow official learning rates for ImageNet's fine-tuning phase, for the fairness comparison.

3. Visualization

Figure 2 and Figure 3 show the visualizations of fine-tunings on Cityscapes and ADE-20k. As can be seen, models pre-trained on SegRCDB can still annotate details comparable to pre-trained models on other real image datasets, despite not having seen any natural images during training.

References

[1] Holger Caesar, Jasper Uijlings, and Vittorio Ferrari. Coco-stuff: Thing and stuff classes in context. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1209–1218, 2018. 1

[2] MMSegmentation Contributors. MMSegmentation: Openmmlab semantic segmentation toolbox and benchmark. https://github.com/open-mmlab/mmsegmentation, 2020. 1

[3] 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 *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3213–3223, 2016. 1

[4] Hirokatsu Kataoka, Ryo Hayamizu, Ryosuke Yamada, Kodai Nakashima, Sora Takashima, Xinyu Zhang, Edgar Josafat Martinez-Noriega, Nakamasa Inoue, and Rio Yokota. Replacing labeled real-image datasets with auto-generated contours. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 21232–21241, 2022. 2

[5] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 9992–10002, 2021. 1, 2

[6] Tete Xiao, Yingcheng Liu, Bolei Zhou, Yuning Jiang, and Jian Sun. Unified perceptual parsing for scene understanding. In *European Conference on Computer Vision (ECCV)*, page 432–448. Springer, 2018. 1

[7] Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Scene parsing through ade20k dataset. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5122–5130, 2017. 1

Table 3: Comparison with backbone pre-training. This represents the result of fine-tuning ImageNet with a learning rate of 0.001 as other classification datasets.

		ADE-20k		Cityscapes	
Pre-training	#Img	mIoU	mAcc	mIoU	mAcc
Scratch	-	31.40	41.02	54.65	62.89
ImageNet	1.28M	43.60	54.66	71.85	80.71
ExFractalDB	1M	41.10	52.05	68.93	77.96
RCDB	1M	38.50	49.43	66.57	75.88
SegRCDB	118k	**43.85**	**54.98**	**73.06**	**81.59**

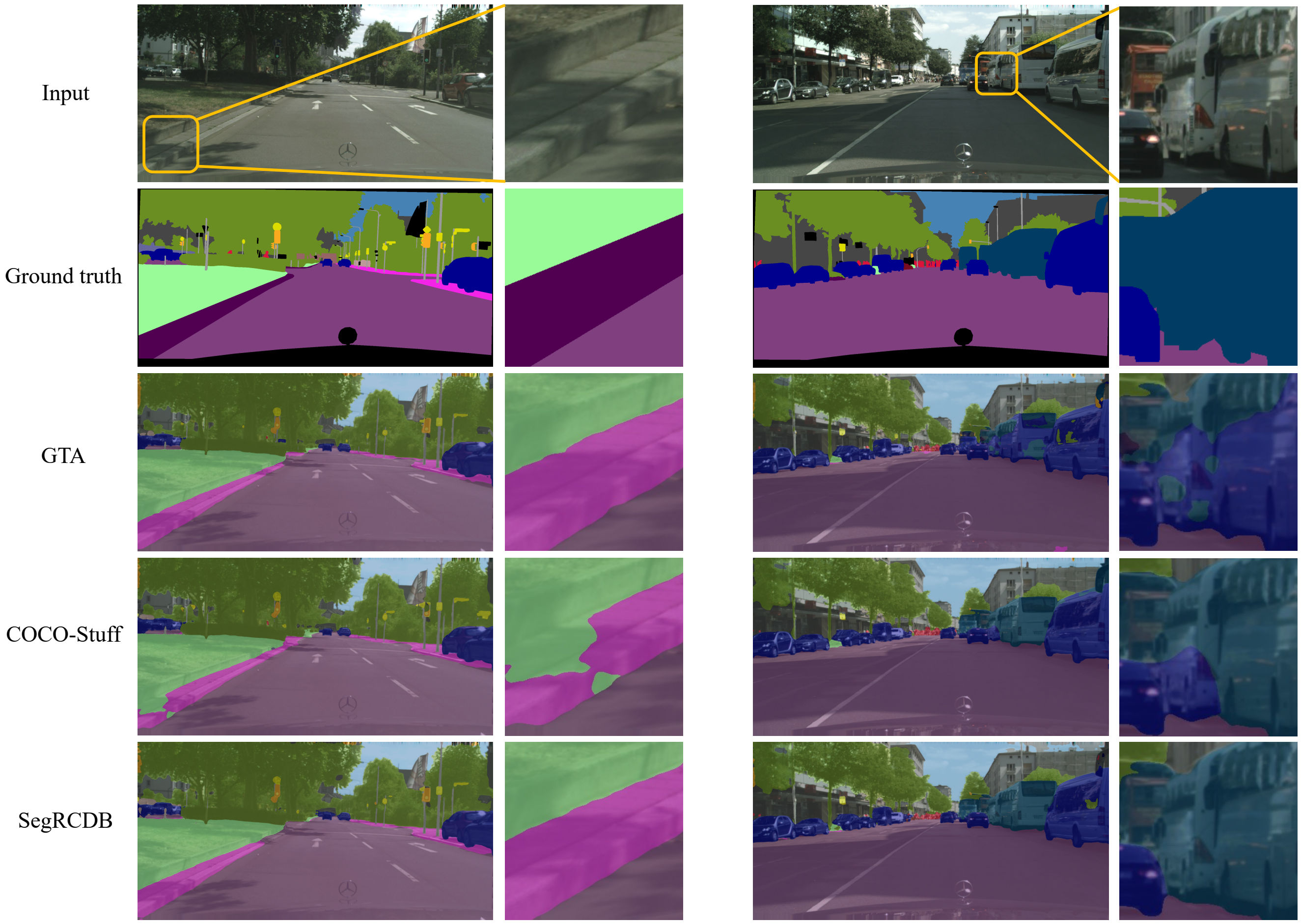

Figure 2: Visual comparison of fine-tuning result on Cityscapes. The first low represents the input images, and the second row represents the ground truth. An enlarged detail of the image is attached on the right. SegRCDB used 118k images.

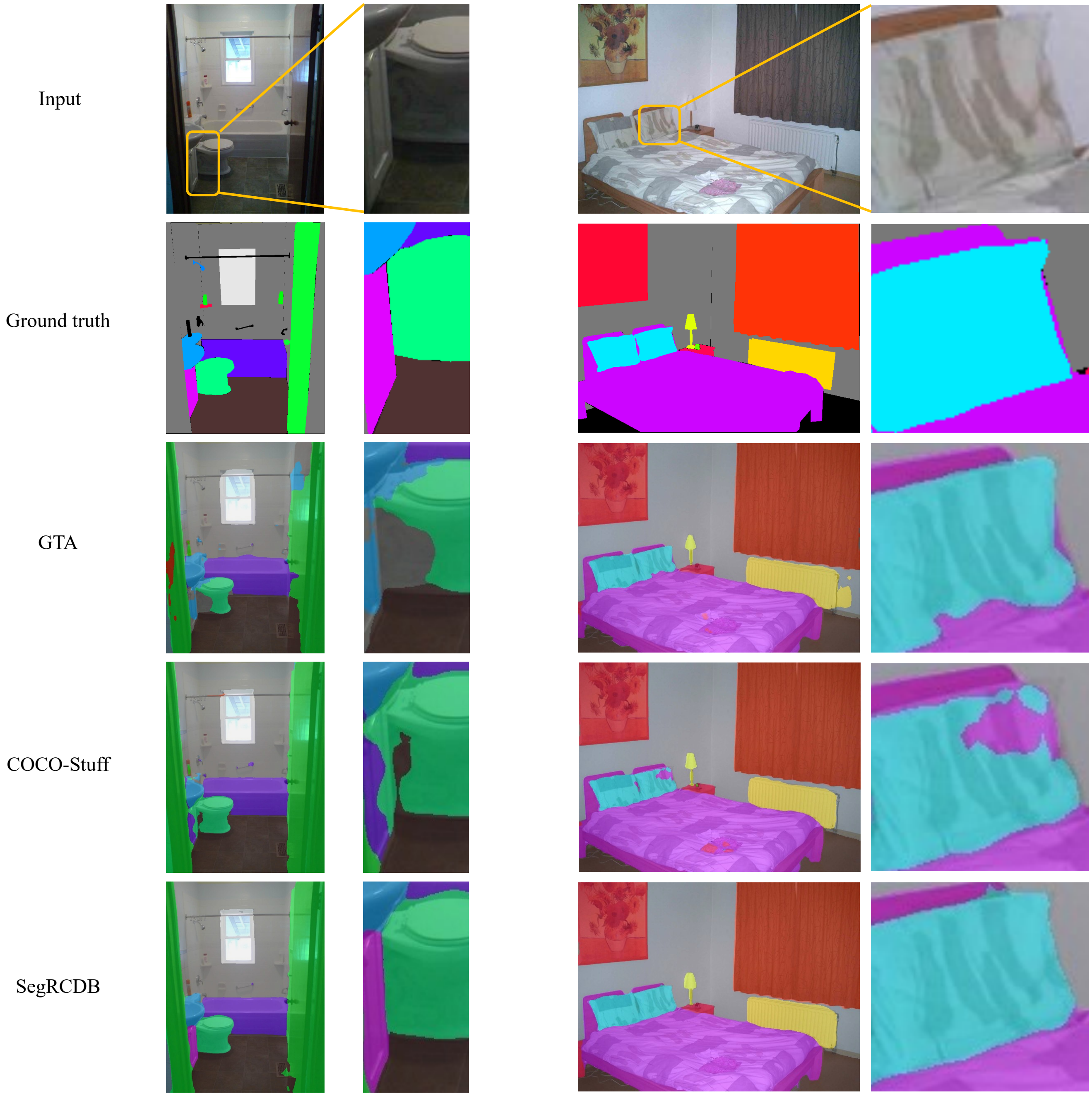

Figure 3: Visual comparison of fine-tuning result on ADE-20k. The first low represents the input images, and the second row represents the ground truth. An enlarged detail of the image is attached on the right. SegRCDB used 118k images.