Supplementary Material for "Self-Regulation for Semantic Segmentation"

We provide the experimental results of using the ensemble of feature maps and logits as the Teacher in Section S1, the experimental results of using logits for SR-F and using feature maps for SR-L in Section S2, the details and results of applying different hyperparameters in Section S3, results of using other λ_3 in Section S4, results of using other loss functions in Section S5, result comparisons between SR-F and skip connections S6, more experimental results by deploying SR-F and SR-L S7, more experimental results for weakly-supervised semantic segmentation in Section S8, and additional qualitative results in Section S9.

S1. An Ensemble of Teachers

This supplementary is for Section 4 of the main paper. In Section 4, for SR-F loss, feature maps output by the shallowest block are taken as the Shallow Teacher, and for SR-L loss, the classification logits output by the deepest block acts as the Deep Teacher. In this Section A, we provide the experimental results of using an ensemble of feature maps (or logits) as the Shallow Teacher (or the Deep Teacher). By "ensemble", we mean that the average feature maps (or logits) of all the previous (or following) blocks are used as the Shallow Teacher (or the Deep Teacher). For example for SR-F loss, the Teacher of block-3 is the average feature maps of block-1 and block-2, and the teacher of block-4 is the average feature maps of block-1, block-2, and block-3. For SR-L loss, "ensemble" is defined analogously.

We show the experimental results in Table S1. We used CONTA [9]+SPGNet [3] as the baseline model. We can observe from the middle row that the ensemble feature maps and logits (*i.e.*, $\text{SR}_{\text{ensemble}}$) can also boost the model performance of baseline, *e.g.*, by 0.3% mIoU on the *val* set of PASCAL VOC 2012 (PC) [5]. While the improvement margin is clearly smaller than that of using a single layer as the Teacher (*i.e.*, $\text{SR}_{\text{regular}}$) in our main paper. We think the reason is that ensembling multi-layer information mixes the semantic and detail representations learned by different layers and thus weakens either of them.

Methods	Settings	PC *val* %
baseline	w/o SR	67.1
baseline+SR	$\text{SR}_{\text{ensemble}}$	$67.4_{+0.3}$
baseline+SR	$\text{SR}_{\text{regular}}$	$68.5_{+1.4}$

Table S1. Experimental results on ensemble feature maps and logits on the *val* set of PASCAL VOC 2012 (PC) [5]. "w/o SR" means that there is no SR loss used. "$\text{SR}_{\text{ensemble}}$" denotes the ensemble feature maps and logits for SR loss function. "$\text{SR}_{\text{regular}}$" denotes the same implementation as the main paper. "+SR" means applying our SR loss function to train the segmentation models.

S2. SR-F on Logits and SR-L on Feature Maps

This supplementary is for Section 4 of the main paper. In Section 4, the computation SR-F is based on feature maps, while SR-L is based on logits. In this Section, we provide results of some empirical trials of using classification logits for SR-F and using feature maps for SR-L.

We show the experimental results under different settings in Table S2. We can observe from the first block of results that replacing the logits of SR-L with feature maps indeed reduces the model performance. We think the reason is that classification logits contain more semantic relationships among different object classes (than the feature maps which contain more details instead). We can observe from the second block of results that replacing the feature maps of SR-F with logits causes a great drop of the model performance. The reason is intuitive as the classification logits at the shallowest layer are the worst ones and taking them as the Teacher definitely misleads other layers.

S3. Hyperparameters

This supplementary is for Section 5 of the main paper. In this section, we show more results by using temperature scaling. In the main paper, the temperature τ in Eq.(2) is used to control the smoothness of feature maps (or logits), *i.e.*, the higher the value of τ, the greater the suppression between the maximum and minimum values in feature maps (or logits). Its practical effect is to suppress the regions with extremely high confidence in feature maps (or logits) from being overly-focused by SS models, and also to ignore the regions with low confidence, *e.g.*, small or thin objects and the rare objects on datasets. An important hy-

Methods	Settings	PC *val*
baseline	w/o SR	67.1
baseline+SR	+ MEA + SR-$L_{regular}$	68.1$_{+1.0}$
baseline+SR	+ MEA + SR-$L_{feature}$	67.3$_{+0.2}$
baseline+SR	+ MEA + SR-$L_{regular}$ + SR-$L_{feature}$	67.5$_{+0.4}$
baseline+SR	+ MEA + SR-$F_{regular}$	68.2$_{+1.1}$
baseline+SR	+ MEA + SR-F_{logits}	66.4$_{-0.7}$
baseline+SR	+ MEA + SR-$F_{regular}$ + SR-F_{logits}	68.4$_{+0.2}$

Table S2. Experimental results (mIoU, %) on the *val* set of PASCAL VOC 2012 (PC) [5] with different settings of SR-F and SR-L. "w/o SR" means that there is no SR loss used. "+" means adding the loss function term(s) to the baseline model. SR-$L_{feature}$ denotes that SR-L loss and SR-F loss are based on the feature maps and logits, respectively. "$SR_{regular}$" denotes the regular implementation as "SR" in the main paper.

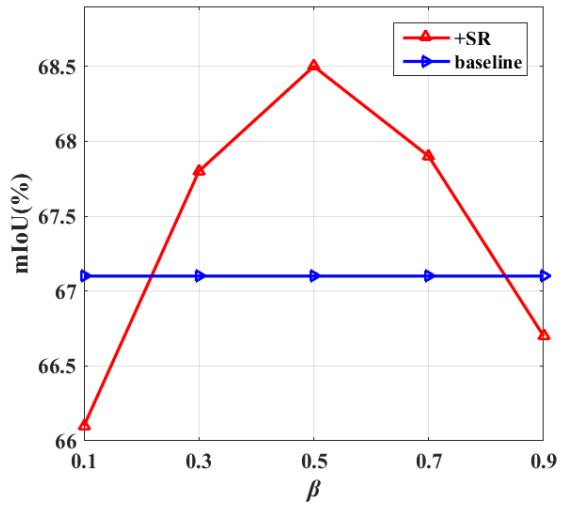

Figure S1. Experimental results on the *val* set of PASCAL VOC 2012 (PC) [5] with different upper-bound values. "+SR" means applying SR loss function to train the baseline model.

perparameter for applying τ is the difference between the maximum and minimum values in feature maps (or logits). We denote this "difference" as β. In our main paper, β is uniformly set to 0.5. Here, we use different values of β and provide the corresponding results in Figure S1. We use CONTA [9]+SPGNet [3] as baseline.

S4. Using Other λ_3

This supplementary is for Section 5 of the main paper. In this section, we provide results of using other value of λ_3 in Eq.(5). We took CONTA [9]+SPGNet [3] as the baseline model. Experiments were carried out on PASCAL VOC 2012 [5] and we show results on its *val* set for SR-F and SR-L in Figure S2, respectively. There is no SR-F or SR-L used when $\lambda_3 = 0$ (*i.e.*, only MEA is deployed on the baseline segmentation model). When $\lambda_3 = 1$, this setting corresponds to the setting of our main paper.

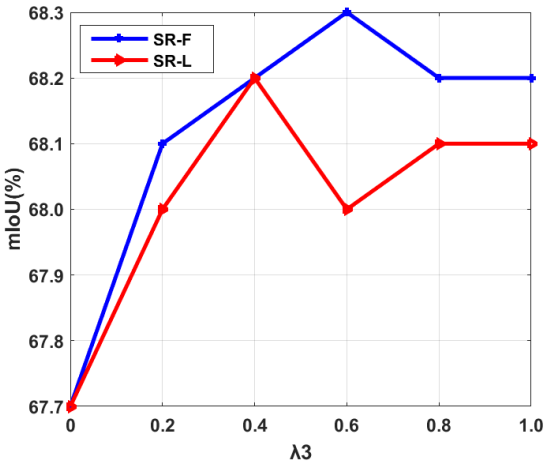

Figure S2. Experimental results on the *val* set of PASCAL VOC 2012 (PC) [5] with different λ_3.

S5. Using Other Loss Functions

This supplementary is for SR-F loss in Section 4.1 and SR-L loss in Section 4.2 of the main paper. In Section 4, the standard Cross-Entropy (CE) loss is used in both SR-F and SR-L losses. Following [10], Table S3 shows more experimental results by using other loss functions on the *val* sets of PASCAL VOC 2012 (PC) [5] and MS COCO (MC) [6]. Using L2 loss, KL divergence loss, CE loss, and the combined loss (Combined) in our SR, we can obtain 68.3%, 68.2%, 68.5%, and 53.2% mIoU on PC, and 34.1%, 34.3%, 34.5%, and 22.9% mIoU on MC, respectively. We can clearly observe that CE achieves the best performance and the combined loss has the worst performance.

Dataset	L2	KL Divergence	CE (used)	Combined
PC	68.3%	68.2%	68.5%	53.2%
MC	34.1%	34.3%	34.5%	22.9%

Table S3. Results of using different loss functions in SR.

S6. SR-F *vs* Skip Connections

This supplementary is for Section 4 of the main paper. In Section 4, we propose to use SR loss for feature regulation. In this section, we provide experimental results by using skip connections as the feature regulation manner. In Table S4, we show experimental results on the *val* sets of PASCAL VOC 2012 (PC) [5] and MS COCO (MC) [6].

We can observe that SR-F can observably achieve a better performance than skip connections on both datasets.

Dataset	SR-F	Skip Connections
PC	68.2%	52.7%
MC	34.3%	23.8%

Table S4. Result comparisons between SR-F and skip connections.

MEA_s	SR-F	SR-L	PC (mIoU%)	MC (mIoU%)
✗	✗	✗	67.1	33.6
✓	✗	✗	$67.4_{+0.3}$	$33.8_{+0.2}$
✓	✓	✗	$67.6_{+0.5}$	$33.9_{+0.3}$
✓	✗	✓	$67.7_{+0.6}$	$34.0_{+0.4}$
✓	✓	✓	$67.9_{+0.8}$	$34.2_{+0.6}$

Table S5. More experimental results by deploying SR-F and SR-L.

S7. More Experiments by Deploying SR-F/-L

This supplementary is for "ablation study" in Section 5.2 of the main paper. In our ablation study, both SR-F and SR-L are implemented on MEA. To further verify the effectiveness of SR-F and SR-L, in this section, we deploy SR-F and SR-L on MEA_s where each exit contains only the semantic segmentation loss. In Table S5, we show results on the *val* sets of PASCAL VOC 2012 (PC) [5] and MS COCO (MC) [6]. On PC dataset, we can observe that improvements of SR-F nd SR-L on MEA_s are 0.2% mIoU and 0.3% mIoU, respectively, and their sum is 0.5% mIoU which shows that SR-Fand SR-L are complementary. The same observation can be obtained on MC dataset.

S8. More Experimental Results for Weakly-Supervised Semantic Segmentation

This supplementary is for Section 5 of the main paper. In this section, we provide more experimental results for weakly-supervised semantic segmentation task. We show experimental results in Table S6. We use two pseudo-mask generation methods (*i.e.*, IRNet [1] and SEAM [7]) with two different segmentation backbones (*i.e.*, SegNet [2] and SPGNet [3]) in experiments. We can observe that using SR can consistently boost the performance of these two pseudo-mask generation methods on both SegNet and SPGNet on the *val* and *test* sets of PASCAL VOC 2012 [5].

Methods	Backbone	PC *val*	PC *test*
IRNet [1]	SegNet [2]	62.1	62.7
IRNet+SR	SegNet [2]	$63.3_{+1.3}$	$63.7_{+1.0}$
IRNet [1]	SPGNet [3]	62.5	63.0
IRNet+SR	SPGNet [3]	$63.5_{+1.0}$	$64.1_{+1.1}$
SEAM [7]	SegNet [2]	63.7	64.0
SEAM+SR	SegNet [2]	$64.7_{+1.0}$	$64.9_{+0.9}$
SEAM [7]	SPGNet [3]	62.4	62.9
SEAM+SR	SPGNet [3]	$63.4_{+1.0}$	$64.0_{+1.1}$

Table S6. Comparing to the state-of-the-arts on the *val* and *test* sets of PASCAL VOC 2012 (PC) [5] using image-level class labels. "+SR" means applying our SR loss function to train the models.

S9. More Qualitative Results

This supplementary is for Section 5 of the main paper. In this section, we provide more visualization results for both weakly-supervised and fully-supervised SS tasks as shown in Figure S3.

References

[1] Jiwoon Ahn, Sunghyun Cho, and Suha Kwak. Weakly supervised learning of instance segmentation with inter-pixel relations. In *CVPR*, 2019. 3

[2] Vijay Badrinarayanan, Alex Kendall, and Roberto Cipolla. Segnet: A deep convolutional encoder-decoder architecture for image segmentation. In *TPAMI*, 2017. 3

[3] Bowen Cheng, Liang-Chieh Chen, Yunchao Wei, Yukun Zhu, Zilong Huang, Jinjun Xiong, Thomas S Huang, Wen-Mei Hwu, and Honghui Shi. Spgnet: Semantic prediction guidance for scene parsing. In *ICCV*, 2019. 1, 2, 3, 4

[4] 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. 4

[5] Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. In *IJCV*, 2010. 1, 2, 3, 4

[6] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *ECCV*, 2014. 2, 3

[7] Yude Wang, Jie Zhang, Meina Kan, Shiguang Shan, and Xilin Chen. Self-supervised equivariant attention mechanism for weakly supervised semantic segmentation. In *CVPR*, 2020. 3

[8] Yuhui Yuan, Xilin Chen, and Jingdong Wang. Object-contextual representations for semantic segmentation. In *ECCV*, 2020. 4

[9] Dong Zhang, Hanwang Zhang, Jinhui Tang, Xiansheng Hua, and Qianru Sun. Causal intervention for weakly-supervised semantic segmentation. In *NeurIPS*, 2020. 1, 2, 4

[10] Linfeng Zhang, Jiebo Song, Anni Gao, Jingwei Chen, Chenglong Bao, and Kaisheng Ma. Be your own teacher: Improve the performance of convolutional neural networks via self distillation. In *ICCV*, 2019. 2

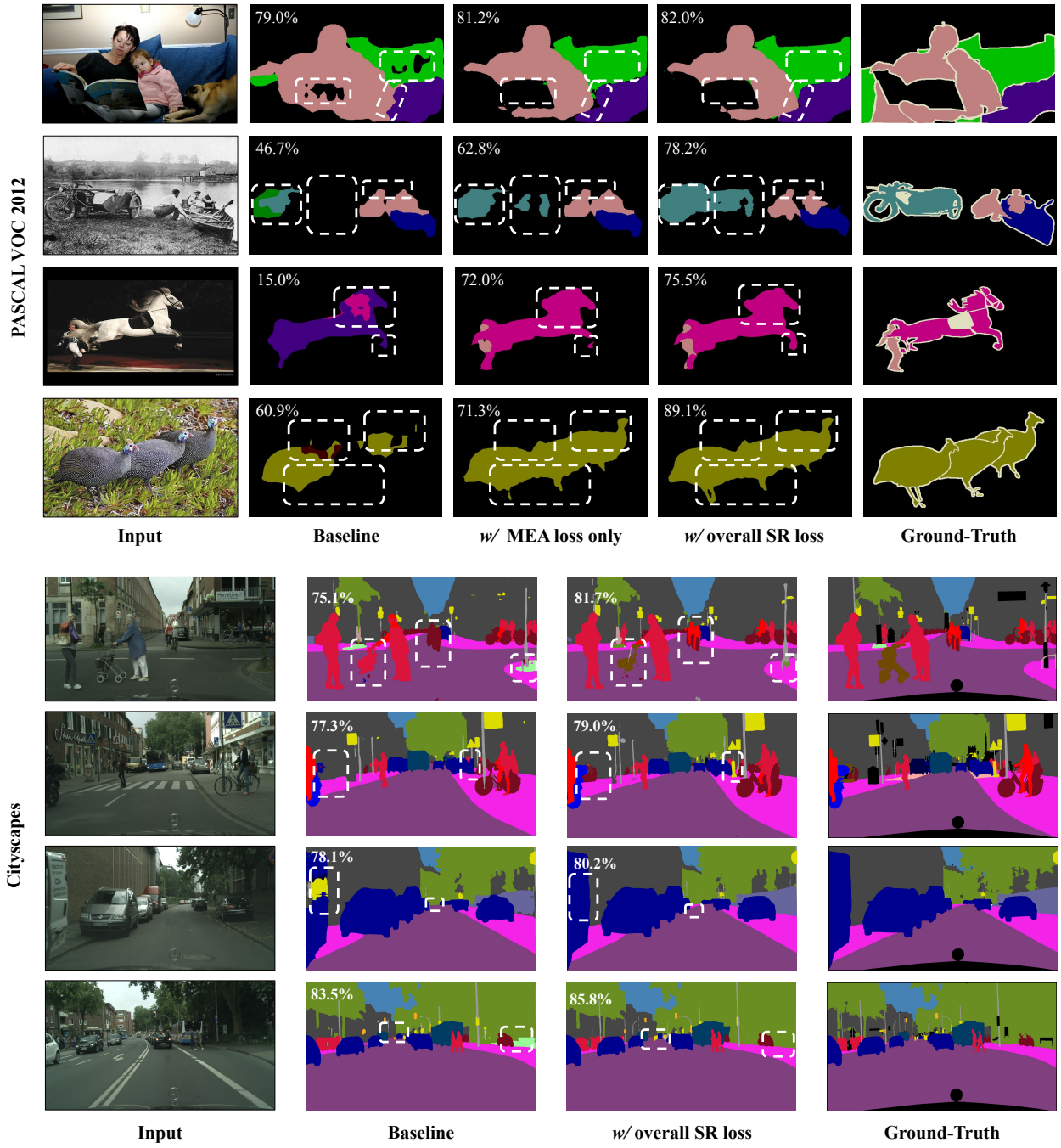

Figure S3. Visualization results for weakly-supervised semantic segmentation on the *val* set of PASCAL VOC 2012 [5] using CONTA [9]+SPGNet [3] as the baseline model, and fully-supervised semantic segmentation on the *val* set of Cityscapes [4] using OCR-Net [8] as the baseline model. "*w/* overall SR loss" means applying MEA loss, SR-F loss, and SR-L loss on baseline models. The mIoU is shown on each result image. The white dotted frames highlight the revised regions by applying our SR approach.