CFCG: Semi-Supervised Semantic Segmentation via Cross-Fusion and Contour Guidance Supervision

Shuo Li*, Yue He*, Weiming Zhang*, Wei Zhang†, Xiao Tan, Junyu Han, Errui Ding, Jingdong Wang
Baidu Inc
{lishuo16, heyue04, zhangweiming, zhangwei99, tanxiao, hanjunyu, dingerrui, wangjingdong}@baidu.com

1. More Implementation Details

1.1. Detail in Coherent-Perturbation

As mentioned in section 3.2, the strong augmentation between image-level and feature-level is coherent by the noise, blur, and erasure operations. Specially, for noise operation, we use random grayscale and color jitter in image-level, while in feature-level, we uniformly sample the noise tensor $N \sim \mathcal{U}(-0.01, 0.01)$ of the same size as the encoder's output z and get the final result $\tilde{z}$ by $\tilde{z} = (z \bigodot N) + z$ operation[5], where $\bigodot$ denotes element-wise dot product. For the blur, we use Gaussian blur in image-level and simple maxpool followed by upsampling in feature-level. For erasure, we use Cutout[3] in both image-level and feature-level.

1.2. Training Hyper-parameters

The hyperparameter λ_1 and λ_2 in Eq.1 and Eq.3 are set as 0.2 and 0.5 respectively on PASCAL VOC 2012[4], 0.2 and 6.0 respectively on Cityscapes[2].

2. More Experiments

2.1. Comparison with Supervised Baselines

We illustrate the improvements of our method compared with supervised baseline on Cityscapes under all partition protocols in Table 1. All the methods are based on DeepLabv3+ with ResNet-50 and ResNet-101.

As shown in Table 1, our CFCG consistently outperforms the supervised baseline. Specifically, the improvements of our CFCG w/ fusion inference over the baseline method are 11.24%, 8.61%, 5.17%, and 3.91% under 1/16, 1/8, 1/4, and 1/2 partition protocols separately with ResNet-50. And the gains of our CFCG w/ fusion inference over the baseline method are 10.96%, 7.41%, 5.06%, and 2.86% under 1/16, 1/8, 1/4, and 1/2 partition protocols separately with ResNet-101. The results demonstrate that our method brings more boost under the 1/16 and 1/8 partitions than under the 1/4 and 1/2 partitions, which means that the less labeled data there is, the more gain there is.

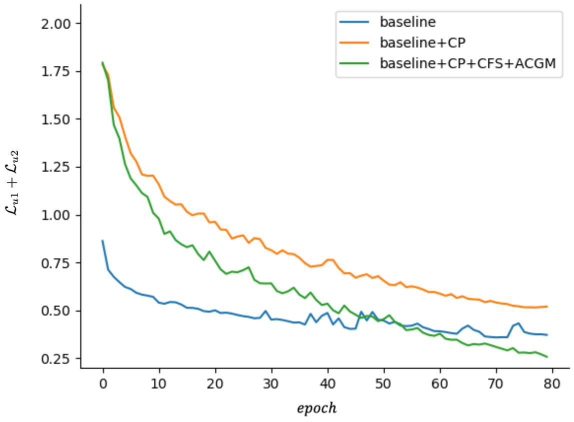

Figure 1. Visualization of $\mathcal{L}_{u1}+\mathcal{L}_{u2}$ curves on different structures, including semi-supervised baseline (blue line), semi-supervised baseline with CP (orange line), and semi-supervised baseline with CP, CFS, and ACGM (green line).

2.2. Ablation Study on Supervised Baseline

The ablation study of our strategies using the supervised baseline with DeepLabv3+ architecture on PASCAL VOC 2012 has shown in Table 2. We can see that the improvement of CFS over the baseline is 0.51%, and the improvement of the ACGM is 0.27%, which is lower than the CFCG baseline's (1.78% and 2.33%). Experiment results indicate that more significant gains generated by CFS and ACGM are achieved in a semi-supervised framework, which indicates that our CFCG is better suited to the semi-supervised framework.

*Co-first author
†Corresponding author
This work was done when Shuo Li was an intern at Baidu Inc.

Method	ResNet-50				ResNet-101			
	1/16(186)	1/8(372)	1/4(744)	1/2(1488)	1/16(186)	1/8(372)	1/4(744)	1/2(1488)
Supervised baseline	64.90	70.32	74.11	76.22	66.80	72.19	75.30	78.06
Ours(w/o fusion inference)	**76.13**	**78.49**	**78.98**	**79.76**	**77.28**	**79.09**	**80.07**	**80.59**
Ours(w/ fusion inference)	**76.14**	**78.93**	**79.28**	**80.13**	**77.76**	**79.60**	**80.36**	**80.92**

Table 1. Comparison with supervised baseline based on DeepLabv3+ on Cityscapes with ResNet-50 and ResNet-101.

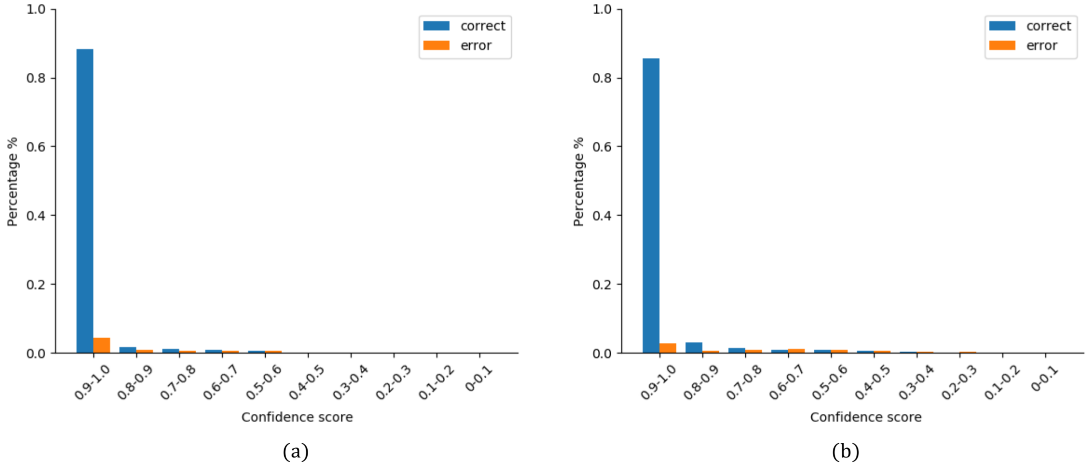

Figure 2. Illustration on the distribution of different weight maps based on the confidence-based method (a) and our ACGM (b) on the VOC dataset. The horizontal axis represents the different intervals, and the vertical axis represents the percentage of correct and error pseudo labels in the current weight interval.

Method	CFS	ACGM	mIoU
Only supervised			69.41
	√		69.92
	√	√	70.19
Semi-supervised			73.44
	√		75.22
	√	√	77.55

Table 2. Ablation study using supervised baseline and semi-supervised baseline on PASCAL VOC 2012 with DeepLabv3+ architecture.

2.3. Analysis about Cross-Fusion Supervision

From Table3 in paper, we find that the CP strategy does not bring much improvement as expected. It implies that the perturbed features may be heavy and diverse. Exist paradigms become almost unbearable, and cannot earn perturbation-diversity dividends. Thus we are driven to design a stronger paradigm under such a setting.

To prove our conjecture, we visualize the loss curves in Figure 1. We can observe that, compared with the convergence trend of semi-supervised baseline, semi-supervised baseline with CP has a higher starting point, since the difference between the prediction and the ground truth is larger under CP strategy. The large difference leads to poor convergence performance of semi-supervised baseline with CP. It suggests that heavy perturbation introduced by CP not only enriches the training data but also makes deep learning process difficult. Compared with the convergence trend of semi-supervised baseline with CP, we find that semi-supervised baseline with CP, CFS, and ACGM converges faster, which suggests that CFS and ACGM solve the heavy loss problem caused by CP and the difference between the prediction and the ground truth is narrowed in the end. In short, these visualizations further demonstrate CFCG is a simple but strong semi-supervised segmentation approach.

2.4. Analysis about Adaptive Contour Guidance Module

To compare the confirmation bias of the ACGM and the confidence-based method, we present the score distribution of weight map based on two methods. Figure 2 shows the distribution of different weight maps based on the confidence and our ACGM, we can see that, in the high score range, compared to the weight map generated by confidence-base method, the weight map gener-

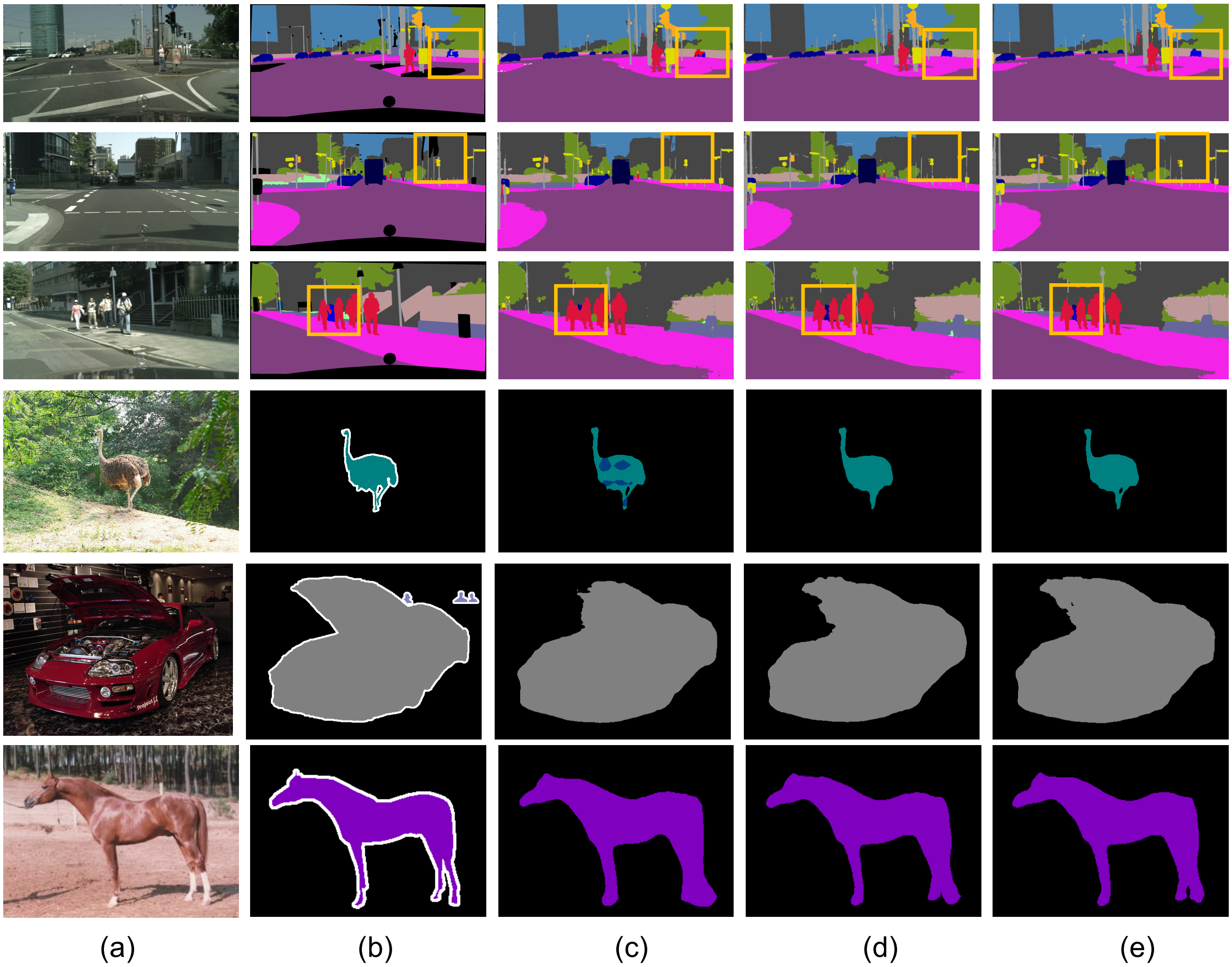

Figure 3. Qualitative results from PASCAL VOC 2012 and Cityscapes. (a) input image, (b) ground truth, (c) CPS[1] results, (d) CFCG results (w/o fusion inference) (e) CFCG results (w/ fusion inference).

ated by ACGM has fewer error/noise pseudo labels. In the low score range, the weight map generated by ACGM has more error pseudo labels. In general, The error regions are assigned lower scores in our ACGM compared to the confidence-based method, suggesting that our ACGM has a greater ability to recognize errors in pseudo labels and suppresses the errors. Moreover, We note that some of the correct regions of pseudo labels are also assigned slightly lower scores in our ACGM compared to the confidence-based method, which may be caused by the blur operation. However, compared to the improvement of error recognition ability, the effection of correct regions with slightly lower scores can be ignored since the model can easily predict correctly for the slightly lower correct regions.

2.5. Analysis on Different Partition Protocol

With the continuous appearance of techniques for SSSS, a variety of data partition protocols are proposed. As shown in Table 3, by using the partition in [6, 7], we employ the 1/16 and 1/8 partition protocols experiment on the PASCAL VOC 2012 dataset with ResNet-101 using DeepLabv3+ architecture. Specifically, the improvements of CFCG w/o fusion inference over the PCR[7] are 1.93% and 0.69% under 1/16 and 1/8 partition protocols separately. And the gains of CFCG w/ fusion inference over the PCR[7] are 2.33% and 1.44% under 1/16 and 1/8 partition protocols separately. The results imply that our CFCG scheme is efficient and robust under these proportions.

Method	1/16	1/8	1/4	1/2
U^2PL(w/CutMix)[6]	77.21	79.01	79.30	80.50
PCR[7]	78.60	80.71	80.78	80.91
Ours (w/o fusion inference)	**80.53**	**81.40**	**82.81**	**82.98**
Ours (w/ fusion inference)	**80.93**	**82.15**	**83.17**	**83.21**

Table 3. Comparison with methods which are based on U^2PL partition protocol methods on the PASCAL VOC 2012 dataset with ResNet-101 using DeepLabv3+ architecture.

3. Qualitative Results

We display some qualitative results on the Cityscapes val set (the first three rows in Figure 3) and the PASCAL VOC 2012 val set (the last three rows in Figure 3), and all the approaches are based on DeepLabv3+ with ResNet-50 network under 1/8 protocol. In the first three rows of Figure 3, for each input image, we select one point (marked as an orange box) and show their corresponding region in ground truth, CPS results[1], CFCG results w/o fusion inference, and CFCG results w/ fusion inference in columns 2,3,4 and 5 respectively. We observe that our CFCG is more noticeable for the response of small objects. For example, the motorcycle in the first row and third row, and the traffic light and pole in the second row. In the last three rows of Figure 3, We observe that our CFCG can achieve better results for object boundaries. In general, benefiting from the proposed framework with a series of components, our final results in column (d) significantly improve better performance than the other method, which demonstrates the effectiveness of our approach.

4. Limitations and Future Works

In this paper, we propose CFCG to tackle the semi-supervised semantic segmentation task. Concretely, CFS mechanism can leverage multiple learners to achieve stronger expressive power under CP. ACGM introduces the semantic contour for encouraging position relations establishment as guidance to effectively identify unreliable spatial regions in pseudo labels. Although CFCG has shown remarkable improvements, it remains challenging to further improve the quality of pseudo labels. First, CFCG w/ fusion inference achieves better performance than CFCG w/o fusion, suggesting that less underutilized knowledge still exists in small amounts. Second, there is still some room for improvement in excavating the noise of pseudo labels. The above ideas are the core problems to improve SSSS performance and generality which are still worth exploring in future research.

References

[1] Xiaokang Chen, Yuhui Yuan, Gang Zeng, and Jingdong Wang. Semi-supervised semantic segmentation with cross pseudo supervision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2613–2622, 2021. 3, 4

[2] 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 *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3213–3223, 2016. 1

[3] Terrance DeVries and Graham W Taylor. Improved regularization of convolutional neural networks with cutout. *arXiv preprint arXiv:1708.04552*, 2017. 1

[4] Mark Everingham, SM Ali Eslami, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes challenge: A retrospective. *International journal of computer vision*, 111:98–136, 2015. 1

[5] Yassine Ouali, Céline Hudelot, and Myriam Tami. Semi-supervised semantic segmentation with cross-consistency training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12674–12684, 2020. 1

[6] Yuchao Wang, Haochen Wang, Yujun Shen, Jingjing Fei, Wei Li, Guoqiang Jin, Liwei Wu, Rui Zhao, and Xinyi Le. Semi-supervised semantic segmentation using unreliable pseudo-labels. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4248–4257, 2022. 3, 4

[7] Haiming Xu, Lingqiao Liu, Qiuchen Bian, and Zhen Yang. Semi-supervised semantic segmentation with prototype-based consistency regularization. In *Advances in Neural Information Processing Systems*, 2022. 3, 4