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

Abstract

Current state-of-the-art semi-supervised semantic segmentation (SSSS) methods typically adopt pseudo labeling and consistency regularization between multiple learners with different perturbations. Although the performance is desirable, many issues remain: (1) supervisions from a single learner tend to be noisy which causes unreliable consistency regularization (2) existing pixel-wise confidence-score-based reliability measurement causes potential error accumulation as the training proceeds. In this paper, we propose a novel SSSS framework, called CFCG, which combines cross-fusion and contour guidance supervision to tackle these issues. Concretely, we adopt both image-level and feature-level perturbations to expand feature distribution thus pushing the potential limits of consistency regularization. Then, two particular modules are proposed to enable effective semi-supervised learning under heavy coherent perturbations. Firstly, Cross-Fusion Supervision (CFS) mechanism leverages multiple learners to enhance the quality of pseudo labels. Secondly, we introduce an adaptive contour guidance module (ACGM) to effectively identify unreliable spatial regions in pseudo labels. Finally, our proposed CFCG achieves gains of mIoU +1.40%, +0.89% with a single learner and +1.85%, +1.33% by fusion inference on PASCAL VOC 2012 and on Cityscapes respectively under 1/8 protocols, clearly surpassing previous methods and reaching the state-of-the-art.

1. Introduction

Semantic segmentation, an essential task of the pixel-wise classification task, has been remarkably successful with the development of deep learning. However, the training for such a problem is rather a challenge owing to costly and laborious pixel-wise manual labeling[18]. To alleviate this problem, semi-supervised semantic segmentation(SSSS) with the precious labeled data and large amounts of unlabeled data is urgently needed to liberate labor and ensure accuracy. Under such a setting, how to adequately leverage unlabeled data becomes critical.

Recently, approaches based on the combination of consistency regularization and pseudo labeling dominate SSSS research [27, 23, 6, 11]. Specially, it encourages high similarity between the predictions of perturbation for the same input but expands a confidence prediction map of an unlabeled image to a one-hot pseudo label's map. However, it still exists some problems: (1) supervisions from a single learner tend to be noisy which causes unreliable consistency regularization (2) existing pixel-wise confidence-score-based reliability measurement causes potential error accumulation as the training proceeds.

In this paper, we propose a novel SSSS framework, called CFCG, which combines cross-fusion and contour guidance supervision to tackle the above issues. First, we conduct a detailed analysis of consistency regularization. Weak-to-strong consistency regularization(WS)[27], one of the representative image-level consistency regularization methods, tries to make weakly and strongly-augmented images pair. Specifically, weak augmentation includes only flip-and-shift data augmentation while strong augmentation is heavily-distorted versions of a given image, including the noise, blur, and erasure operations. Its success comes from weak branch can produce high-quality pseudo labels while strong branch can make the training process hard by injecting noise. While in the feature-level, various types of perturbations for consistency are designed, such as VAT[25], random dropout, random noise, etc. We find that image-level strong perturbations reflected in pixel noise

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

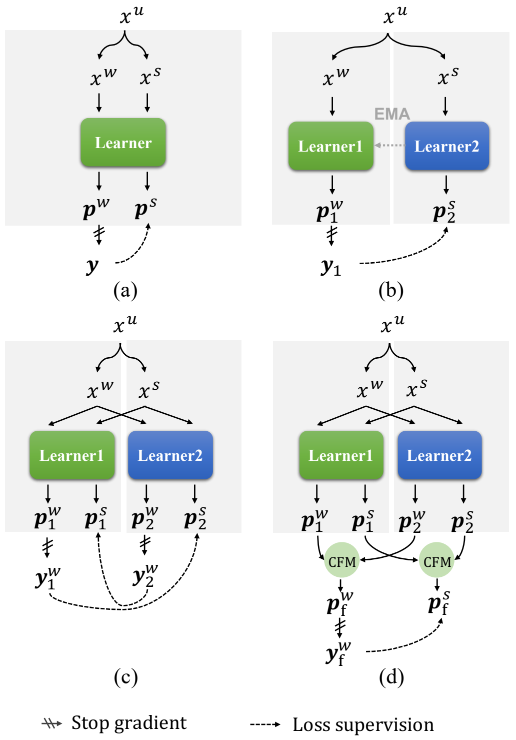

Figure 1. A comparison between typical SSSS frameworks. (a) A single learner with weak-to-strong consistency (WS) strategy. [27, 40] (b) Multiple individual learners with WS strategy. [23, 16, 39] (c) Multiple learners with symmetric WS strategy. [11] (d) Our proposed Cross-Fusion Supervision (CFS) is applied to multiple learners.

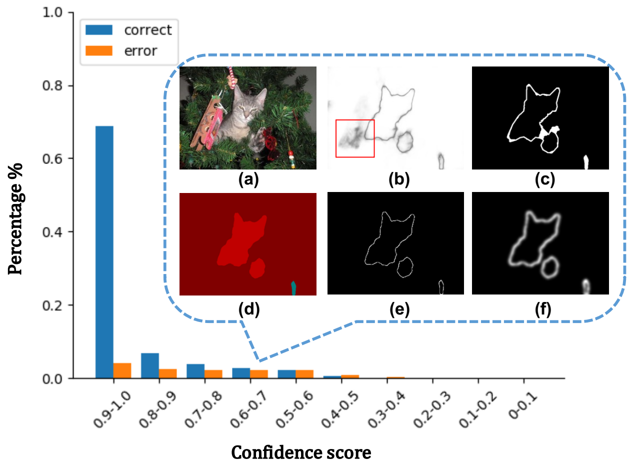

Figure 2. The confidence score distributions of presudo label on the VOC dataset. The horizontal axis represents the different confidence intervals, and the vertical axis represents the number of correct and error pseudo labels in the current confidence interval. We illustrate the problems of conventional confidence-score-based reliability measurement: (a) Input image. (b) Confidence map of the model prediction. (c) The difference between ground truth and pseudo label. (d) The predicted pseudo label. (e) The semantic contour of the pseudo label generated from (d). (f) The weight map adopted in our proposed ACGM. We can observe that: from (b) and (c), the confidence map is noisy and unreliable as highlighted by the red box; from (f) and (c) our proposed ACGM effectively identifies unreliable spatial regions.

and feature-level perturbation reflected in feature noise exhibit similar characteristics. We put both image-level strong augmentation and feature-level perturbations in the same branch to cause heavy Coherent Perturbation(CP) to push the potential limits of consistency regularization.

Fig.1 provides an extensive comparison between WS-based SSSS frameworks. Typically, as shown in Fig. 1 (a), a single learner with WS is first proposed in [40]. Then, with the advantage of mean-teacher, the SSSS framework evolves to Fig. 1 (b) in which WS strategy is applied together with multiple learners [23, 16, 39]. Fig. 1 (c) further introduces the symmetric dual-student method where both strongly-augmented images and weakly-augmented images flow into bipartite learners. It uses the pseudo label obtained from one network to supervise the other network and vice versa [11]. Different from the above works, we propose to add the Cross-Fusion Supervision (CFS) mechanism, which enables the information transfer between the multiple learners to enhance the quality of pseudo labels thus benefiting the process of the semi-supervised training process. While in the inference stage, only one model without relying on cross-fusion is able to generate high-quality results.

Next, we delve into the shortcomings of conventional confidence-score-based reliability measurement in SSSS. Fig.2 illustrates the confidence score distributions of pseudo labels on the VOC dataset. We can observe that it is almost impossible to identify unreliable pixel-wise pseudo labels via pixel-wise confidence score thresholds, which is widely adopted in semi-supervised tasks [27, 41, 23, 11, 16]. Furthermore, we can also observe from (b) and (c) of Fig.2, the confidence map is usually noisy and unreliable as highlighted by the red box. This issue leads to error accumulation as the training proceeds, thus seriously affecting the performance of semi-supervised learning in segmentation tasks. Previous pixel-wise confidence-score-based methods struggle with very limited performance gains. To address this issue, we draw inspiration from the similarity between (c) and (f) in Fig.2, where most of the unreliable spatial regions can be identified by the contour map. Therefore, we propose an adaptive contour guidance module (ACGM) that gradually applies a contour-guided weight map to re-weight training loss. In this way, whether the pixel-wise pseudo label is reliable incorporates contextual cues. Intuitively, as shown in Fig.2, our weight map (f) is acquired by softening based on the semantic contour of the pseudo label (e), and its visualization results prove that our ACGM's weight map

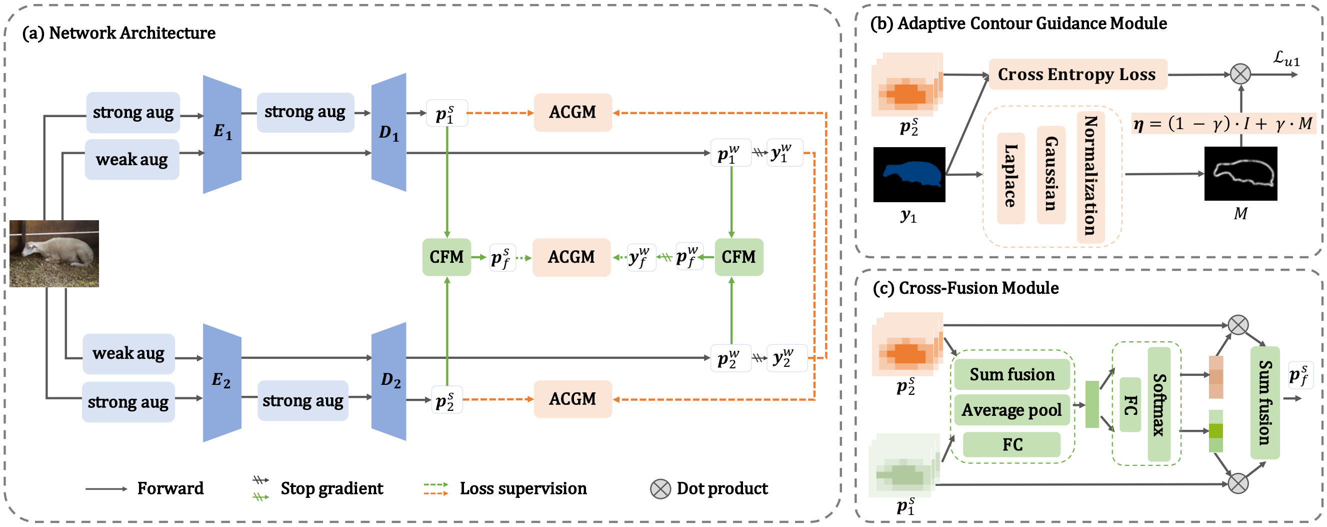

Figure 3. Illustration of our CFCG method on unlabeled data. CFM denotes the Cross-Fusion Module which is one component of Cross-Fusion supervision how to implement fusion can be seen in (c). The details of ACGM are shown in (b). CP consists of the input image augmentation and encoder feature augmentation, which further enhances the learner's learning ability. ACGM iteratively updates reliable pseudo labels to guide semi-supervised learning. CFM fuses the weak flow output and strong flow output of the two learners separately and then leverages the fusion pseudo label of weak flow to supervise the strong flow's fusion. For the labeled data, images are augmented by weak flow into the two networks and both the networks and the CFM are supervised by corresponding ground truth respectively.

can cover the error region more exactly.

To summarise, our contributions are:

- We propose a novel SSSS framework CFCG mainly including cross-fusion supervision(CFS) and the adaptive contour guidance module(ACGM). Our CFS which neatly blends CP can tap into underutilized knowledge to enhance the quality of pseudo labels during the training stage and help enhance expressive power during inference.
- The proposed ACGM introduces the semantic contour for encouraging position relations establishment as guidance to effectively identify unreliable spatial regions in pseudo labels and accurately mitigates the problem of confirmation bias.
- The experiment results present our CFCG achieves new state-of-the-art results on two commonly used benchmarks, which yield mIoU 77.10%, 78.49% with no additional calculations, and 77.55%, 78.93% by fusion inference way on PASCAL VOC 2012 and on Cityscapes separately under 1/8 protocols.

2. Related Work

2.1. Semantic Segmentation

Traditional semantic segmentation is a fundamental topic in the computer vision field. Recent years have witnessed remarkable progress in semantic segmentation: from FCN to Transfomer-like networks. Specifically, the FCN[24] is a milestone, which builds fully convolutional networks for pixel-wise prediction. Since then, extensions based on FCN[5, 2, 36] have been validated to be invaluable in resolving the capturing long-range context dependency problem. Deeplabv3[5] designs an atrous spatial pyramid pooling module to enlarge the receptive field. The SegNet[2] strengthens context cues by its encoder-decoder structure. PSPNet[36] incorporates a pyramid pooling module to embed multi-scale context features to improve FCN-base architecture. Lately. Transformer-based solutions have attracted more and more attention since Visual Transformer[9] has been designed. These methods continue to actively expand transformer skills, including extracting features from input image [31, 37], learning class embedding [28], or formulating segmentation as a simple mask classification task [7].

2.2. Semi-supervised Learning

Semi-supervised learning(SSL) has been an active research issue and related literature contains a large variety of methods[19, 4, 3, 32, 27, 34, 29]. They can be mainly categorized as self-training-based approach and consistency regularization-based approach. The former typically predicts pseudo labels to unlabeled data and then the ground truth and pseudo labels will be used together as the supervisory signal in training. The latter enforces similar predictions output by the network under different forms of perturbations. Specifically, Dual student[19] replaces the teacher of the mean-teacher framework with another student. UDA[32] and FixMatch[27] utilize confidence-based thresholding techniques to ensure the quality of pseudo labels. FlexMatch improves this strategy by considering different learning difficulties of different classes[34]. Self-adaptive confidence threshold is proposed in FreeMatch[29].

2.3. Semi-supervised Semantic Segmentation

SSSS is rather a challenge compared with SSL due to the need for dense prediction. Inspired by SSL, common SSSS mostly follows the same track where the self-training approach and consistency regularization approach are still validated to be very powerful in this field. Specifically, CCT[26] has a shared encoder and multi-decoder with a series of feature perturbations. CPS[6] achieves cross pseudo supervision for the same input via two networks with different initialization. Then, the class imbalance problem has been solved successively with the adaptive equalization learning framework[16], uncertainty guided cross-head co-training framework[11], and pixel-level contrastive learning scheme[1]. Based on the mean-teacher model, PS-MT[23] creately designs a new auxiliary teacher and a stricter confidence-weighted cross entropy loss. The U^{2}PL[30] chooses a unique way of using unreliable pseudo labels. GCT[18] aims for diverse pixel-wise tasks, which proposed a flaw detector to locate the noise of pseudo labels. Note that GCT[18], ELN[21] and other similar models try to design extra models to make them have the ability to predict flaw regions. Currently, self-supervised learning is employed for this task[38, 1, 22, 35]. For self-training, ST++[33] improves the self-training pipeline and gets better performance in an offline way. Different from them, we find the potential information and use it effectively to avoid using extra training models and training time.

3. Method

In this section, we present an overview of the proposed framework in Section 3.1. Then we describe our CP in Section 3.2. Additionally, we describe CFS's detail in Section 3.3. Finally, the ACGM is further introduced in Section 3.4.

3.1. Overview

As shown in Fig.3, our semi-supervised framework consists of two learners with the same structure, in which each learner contains a weak and a strong flow. For unlabeled images, we need to perform three steps to realize our semi-supervised learning.

Firstly, an unlabeled image is fed into the weak and strong flow of each learner, where the CP strategy further enhances the learner's ability. Secondly, we use CFS to fuse the weak flow output and strong flow output of the two learners separately and then leverage the fused pseudo label of weak flow to supervise the strong flow's fusion output. Finally, we use the pseudo label generated by weak flow to supervise the strong flow output of the other learner. During the supervision, ACGM is used to iteratively update reliable pseudo labels to guide semi-supervised learning.

For labeled images, following the previous work, images are sent to the weak flow of the two learners simultaneously and are supervised by corresponding ground truth respectively. Note that the CFS and ACGM techniques are also employed in it.

In general, given limited labeled images $D_l = \{(x_i^l, y_i^l); i \in (1, ..., N_l)\}$ and large amount of unlabeled images $D^u = \{(x_i^u); i \in (1, ..., N_u)\}$, the unsupervised loss for unlabeled images can be written as:

$$\mathcal{L}_u = \lambda_1(\mathcal{L}_{u1} + \mathcal{L}_{u2}) + \mathcal{L}_{uf}, \tag{1}$$

where $(\mathcal{L}_{u1} + \mathcal{L}_{u2})$ represents the cross pseudo supervision loss from two learners and $\mathcal{L}_{uf}$ represents the CFS loss. We use λ_1 to control the balance between cross pseudo supervision loss and CFS loss as the trade-off weight.

For the labeled images, following the previous work[11], images are augmented by weak flow into the two networks and are supervised by corresponding ground truth respectively. Note that the CFS and ACGM techniques are also employed in it. The supervised loss from labeled data is:

$$\mathcal{L}_l = \mathcal{L}_{l\_gt} + \mathcal{L}_{lf}, \tag{2}$$

where $\mathcal{L}_{l\_gt}$ represents the loss between predictions and ground truth. $\mathcal{L}_{lf}$ represents the CFS loss for labeled images, which is obtained by supervising the fused logits of CFS with ground truth. The detail about the losses is described in section 3.3. In the end, the total loss $\mathcal{L}_{total}$ is as follows:

$$\mathcal{L}_{total} = \mathcal{L}_l + \lambda_2\mathcal{L}_u. \tag{3}$$

Here, we use λ_2 to control a balance between the loss on the labeled data loss and the unlabeled data.

3.2. Weak and Strong Coherent-Perturbation

Classic weak-to-strong consistency regularization[27] can be formulated as:

$$f^w = E(\alpha(x^u)), \tag{4}$$

$$f^s = E(\mathcal{A}(\alpha(x^u))), \tag{5}$$

where x^u represents unlabeled images, E represents encoder embedding, f^w and f^s represent the encoder embedding feature generated by weak image augmentation $\alpha(\cdot)$ and strong image augmentation $\mathcal{A}(\cdot)$ respectively. To push the potential limits of consistence regularization, we propose CP Strategy, which gets into feature strong perturbation after getting the encoder output of strongly-augmented unlabeled images to cause heavily perturbed feature. Thus the f^s is rewritten as f^{s^+}:

$$f^{s^+} = \mathcal{B}(f^s), \tag{6}$$

where the $\mathcal{B}$ represents the feature perturbation. Among them, the strong flow for both image-level and feature-level includes the noise, blur, and erasure operations.

3.3. Cross-Fusion Supervision

Sub-figures (a) and (c) in Fig.3 illustrate our proposed Cross-Fusion Supervision (CFS) mechanism. In general, CFS first realizes channel-wise attention and fusion for both weakly and strongly perturbed learners, respectively. Then consistency regularization between the two learners is calculated based on the fused weak logits. Compared with previous SSSS works, in which consistency regularization is calculated between outputs of single learners, our proposed CFS leverages both learners to enhance the quality of pseudo labels thus benefiting the process of semi-supervised learning.

Specifically, our proposed CFS contains three parts: cross-learner attention, weighted fusion, and cross-supervision. In cross-learner attention, we perform an element-wise sum operation with $\mathbf{p}_1^w, \mathbf{p}_2^w \in \mathbb{R}^{C\times H\times W}$, and average pooling is used to encode the spatial representation of different semantic channels to generate $a \in \mathbb{R}^C$.

$$a = \frac{1}{H\times W}\sum_{i=1}^{H}\sum_{j=1}^{W}(\mathbf{p}_1^w(i,j) + \mathbf{p}_2^w(i,j)). \tag{7}$$

Next, $v_i = FC_i(a),\ where\ i \in \{0, 1\}$. The a is fed into different FC layers to generate two attention vectors $v_1, v_2 \in \mathbb{R}^C$ respectively.

In weighted fusion part, generated attention vectors v_1, v_2 are used to perform optionally weighted fusion for $\mathbf{p}_1^w, \mathbf{p}_2^w$. The final feature map $\mathbf{p}_f^w$ is obtained as follows:

$$\mathbf{p}_f^w = v_1 \cdot \mathbf{p}_1^w + v_2 \cdot \mathbf{p}_2^w, \tag{8}$$

where $\mathbf{p}_f^w \in \mathbb{R}^{C\times H\times W}$. In the same way, $\mathbf{p}_f^s$ is acquired by the same operation above from $\mathbf{p}_1^s, \mathbf{p}_2^s$.

Finally, argmax function is needed to choose the corresponding class $c \in \{1, ..., C\}$ with the maximal probability for $\mathbf{p}_f^w$. The argmax result is denoted as $\mathbf{y}_f^w$. In this case, the fusion loss can be written as follows:

$$\mathcal{L}_{uf} = \frac{1}{|D_u|}\sum_{x\in D_u}\frac{1}{W\times H}\boldsymbol{\eta_f}\cdot \ell_{ce}(\mathbf{p}_f^s, \mathbf{y}_f^w) \tag{9}$$

where the cross entropy loss ℓ_{ce} is used to minimize the two probability distribution terms. D_u is the batch size of unlabeled images. $\boldsymbol{\eta_f}$, a coefficient for loss, will be explained in Section 3.4.

Finally, we introduce cross-supervision which calculates consistency loss between the fused logits of weak and strong flow, respectively. As shown in Fig.3, the pseudo labels, generated from one weak CP flow, are used to supervise the predictions which are from the other strong weak CP flow. The loss $\mathcal{L}_{u1}, \mathcal{L}_{u2}$ of CFS can be formulated as:

$$\mathcal{L}_{u1} = \frac{1}{|D_u|}\sum_{x\in D_u}\frac{1}{W\times H}\boldsymbol{\eta_1}\cdot \ell_{ce}(\mathbf{p}_2^s, \mathbf{y}_1^w), \tag{10}$$

$$\mathcal{L}_{u2} = \frac{1}{|D_u|}\sum_{x\in D_u}\frac{1}{W\times H}\boldsymbol{\eta_2}\cdot \ell_{ce}(\mathbf{p}_1^s, \mathbf{y}_2^w), \tag{11}$$

where the cross entropy loss ℓ_{ce} is used to minimize the two probability distribution terms. D_u is the batch size of unlabeled images. $\boldsymbol{\eta_1}$ and $\boldsymbol{\eta_2}$ represent the loss coefficient for two learners separately.

Taking the advantage of our proposed CFS, each learner is able to receive better supervision from the other learner. During the inference stage, one single learner without fusion inference is able to generate state-of-the-art SSSS results. And the learner with fusion inference further improves the effect but subsequently with additional calculations.

3.4. Adaptive Contour Guidance Module

In most previous works, confidence-based thresholding strategy is proposed to measure the quality of pseudo labels. If the confidence is higher than the given threshold, the corresponding pseudo label will be considered reliable and retainable. However, we find that a large number of wrong pseudo labels located in the high-confidence interval are considered reliable and vice versa from Fig.2. Based on this observation, we formulate this idea as a re-weight task and propose ACGM using contour to guide the network, aiming to detect the noise of pseudo labels by relying on spatial information.

Contour-based Weight Map. First of all, the contour-based weight map $M \in \mathbb{R}^{H\times W}$ is generated by two steps: Exact and Soften. In the first step, as shown in Fig.1 (c), semantic contour maps are exacted from the pseudo labels by the Laplacian group. Specifically, we use this operator to exact contour maps to catch multi-scale information, and kernels with different strides are employed. Then upsample different scale contour maps to the original size and fuse them with 1 × 1 convolution. Next, convert it into a binary image as the semantic contour map. In the second step, image processing operations are injected into this semantic contour map to produce the final weight map. Specifically, (1) a Gaussian kernel is employed for blurring the semantic contour map, transforming it into a dense probability map. (2) normalizes all pixels of the dense probability map to range between [0, 1] to generate the final weight map.

Adaptive Loss Re-weight. As mentioned above, we use the contour-based weight map M as the coefficient of cross entropy loss to guide the model to leverage spatial information and distinguish the noise in pseudo labels. Compared with direct multiplication, we take into account that the learning ability of network in the early training stage is not enough, so we adopt a progressive strategy $\boldsymbol{\eta}$ to gradually introduce the knowledge of semantic contour:

$$\boldsymbol{\eta} = (1-\gamma)\cdot I + \gamma \cdot M, \tag{12}$$

$$\gamma = exp(a_e t) - 1\ ,\ a_e = \frac{ln(2)}{max\_epoch}. \tag{13}$$

Where $\boldsymbol{\eta} \in \mathbb{R}^{H\times W}$, and the $I \in \mathbb{R}^{H\times W}$ is the all-ones matrix. In Eq.12, we make the γ gradually learn to assign higher weight due to the fact that the network may converge in the wrong direction in the early stage of training, and in the later training stage the contour is more reliable which also means the pseudo label is more instructive. From it, the function variable t denotes the current training epoch and the constant parameter a_e is calculated to guarantee γ ranges from 0 to 1 by Eq.13.

4. Experiments

4.1. Experimental Setup

Datasets. We evaluate our framework on two different datasets: PASCAL VOC 2012 [10] and Cityscapes [8]. The PASCAL VOC 2012 dataset contains 1464 and 1449 images used for training and validation respectively. Later, it is expanded by extra relatively coarse manual annotations from the SBD dataset [13], resulting in a total of 10582 training images. We follow the previous work [14] to use the augmented set as our full training set. We further provide results for the Cityscapes dataset, which consists of 2975 training and 500 validation images with 19 individual classes.

We rigorously follow the partition protocols of Guided Collaborative Training (GCT) [18], and it divides the whole training set into two groups via randomly sub-sampling 1/2, 1/4, 1/8, and 1/16 of the whole set as the labeled set and regards the remaining images as the unlabeled set.

Evaluation. Following previous methods [12, 16, 26, 38, 6], the images are center cropped into a fixed resolution for PASCAL VOC 2012. For Cityscapes, previous methods apply slide window evaluation, and so do we. Then we adopt the mean of Intersection over Union (mIoU) as the metric to evaluate these cropped images. All results are measured on the val set on both PASCAL VOC 2012 [10] and Cityscapes [8]. Ablation studies are conducted on the blender PASCAL VOC 2012 [10] val set under 1/8 partition protocol. During the test, one network(w/o fusion inference) and two networks(w/ fusion inference) are all analyzed in our approach.

Network. We use DeepLabv3+ [5] with ResNet [15] pretrained on ImageNet [20] as our segmentation network. The decoder head is composed of separable convolution same as standard DeepLabv3+.

Implementation Details. We initialize the weights of two backbones in the two segmentation networks with the same weights pre-trained on ImageNet and initialize the weights of two segmentation heads randomly. We adopt mini-batch SGD with momentum to train our model with Sync-BN [17]. The momentum is fixed as 0.9 and the weight decay is set to 0.0005. We employ a poly learning rate policy where the initial learning rate is multiplied by $(1-\frac{iter}{max\_iter})^{0.9}$. We train PASCAL VOC 2012 for 80 epochs with a base learning rate set to 0.0025 and crop size of 512 x 512, and Cityscapes for 240 epochs with a base learning rate set to 0.02 and crop size of 800 x 800.

4.2. Comparison to State-of-the-Art Methods

We compare our method with some recent semi-supervised segmentation methods including CPS[6], U^{2}PL[30], UCC [11], PS-MT[23],[21] and ST++[33] under different partition protocols, using the same architecture and partition protocols for fairness.

PASCAL VOC 2012. Table 1 shows the comparison results on PASCAL VOC 2012. We can see that over all the partitions from 1/16 to 1/2, with both ResNet-50 and ResNet-101, our method w/o fusion inference and w/ fusion inference consistently outperforms the other methods. For example, compared to CPS[6], which can be considered as our baseline, CFCG w/o fusion inference improves by 2.00%-3.43% under all partition protocols with ResNet-50. While w/ fusion inference improves by 2.50%-3.88% under all partition protocols with ResNet-50. Additionally, our method is superior to the other state-of-art methods in various settings. To be specific, based on ResNet101, it outperforms previous state-of-art UCC [11] by 2.29% and 2.04% under the 1/8 partitions with ResNet-50 and ResNet-101 respectively. The result demonstrates our CFCG's robustness for this SSSS task.

Based on the w/o fusion inference, we find the test with w/ fusion inference brings a significant improvement in performance. Our approach with fusion inference outperforms the approach without fusion inference by 0.50% and 0.59% under 1/2 partition protocol with ResNet-50 and ResNet-101 separately.

Cityscapes. Table 2 illustrates the comparison results on the Cityscapes val set. In comparison to other state-of-the-art(SOTA) methods, our model achieves higher performance among all partition protocols with both ResNet-50 and ResNet-101 backbones. For example, our method w/o fusion inference outperforms previous state-of-art PS-MT[23] by 2.73% and 2.06% under the 1/8 and 1/4 partitions with ResNet-50 separately.

Method	ResNet-50				ResNet-101			
	1/16(662)	1/8(1323)	1/4(2646)	1/2(5291)	1/16(662)	1/8(1323)	1/4(2646)	1/2(5291)
CPS(w/CutMix)[6]	71.98	73.67	74.90	76.15	74.48	76.44	77.68	78.64
U^{2}PL(w/CutMix)[30] †	-	-	-	-	74.43	77.60	78.70	79.94
UCC[11]	74.05	74.81	76.38	76.53	76.49	77.06	79.07	79.54
PS-MT[23]	72.83	75.70	76.43	77.88	75.50	78.20	78.72	79.76
ELN[21]	-	73.20	74.63	-	-	75.10	76.58	-
ST++[33]	72.60	74.40	75.40	-	74.50	76.30	76.60	-
Ours(w/o fusion inference)	**75.00**	**77.10**	**77.72**	**78.15**	**76.82**	**79.10**	**79.96**	**80.18**
Ours(w/ fusion inference)	**75.58**	**77.55**	**78.34**	**78.65**	**77.39**	**79.40**	**80.42**	**80.77**

Table 1. Comparison with state-of-the-art methods on the PASCAL VOC 2012 dataset. Labeled images are sampled from the blended training set, which is augmented by the SBD dataset. ”w/o fusion inference” denotes directly using output which is predicted by one model without any CF operation test; ”w/ fusion inference” denotes using output which is predicted by two models with CF operation. Results of U^{2}PL with † are acquired through the open-source code repository whose partition is the same with [18].

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)
CPS(w/CutMix)[6]	74.47	76.61	77.83	78.77	74.72	77.62	79.21	80.21
U^{2}PL(w/CutMix) [30]†	-	-	-	-	70.30	74.37	76.47	79.05
U^{2}PL(w/AEL) [30]†	-	-	-	-	74.90	76.48	78.51	79.12
UCC[11]	76.02	77.60	78.28	79.54	77.17	78.71	79.59	80.57
PS-MT[23]	-	75.76	76.92	77.64	-	76.89	77.60	79.09
ELN[21]	-	-	-	-	-	70.33	73.52	75.33
ST++[33]	-	72.70	73.80	-	-	-	-	-
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 2. Comparison with state-of-the-art methods on the Cityscapes dataset. ”w/o fusion inference” denotes directly using output which is predicted by one model without any CF operation test; ”w/ fusion inference” denotes using output which is predicted by two models with CF operation. Results of U^{2}PL with † follow the paper which is different with [18]. Considering that the partition's impact is not significant on the Cityscapes dataset, we also put it in this table for reference.

CP	ACGM	CFS		mIoU
		w/o fusion inference	w/ fusion inference	
				73.08
√				73.44
√	√			75.23
√		√		75.22
√	√	√		77.10
√	√		√	77.55

Table 3. Ablation study using the 1/8 labelled ratio on PASCAL VOC 2012 under DeepLabv3+ architecture.

Overall, our architecture can adapt to various setting both on the PASCAL VOC 2012 and Cityscapes dataset and achieves stable and impressive results.

4.3. Ablation Studies

In order to deeply explore the effects of different modules, in this section, we conduct all the ablation experiments by running on PASCAL VOC 2012 under a 1/8 ratio with ResNet-50, and we use DeepLabv3+ to evaluate our results.

The effect of CP, CFS, and ACGM in our method is verified in Table 3, where we use cross pseudo supervised [6] trained with the input image augmentation as the baseline. We note that CP strategy provides a 0.36% improvement slightly, it shows that there still exists a certain amount of noise in pseudo labels. To deal with the problem, we introduced ACGM and CFS. We can see that ACGM and CFS(w/o fusion inference) increase by 1.79% and 1.78% respectively, which verify ACGM can generate better pseudo labels for SSSS task, and CFS enable to fully fuse different learners' knowledge to make the model possess strong learning ability. Further, We achieved a significant improvement of 3.66% by integrating ACGM and CFS(w/o fusion inference). Finally, compared with the baseline, we achieved a performance improvement of 4.47%, which indicate that our CFCG framework is efficient and friendly for SSSS.

Cross-fusion Supervision. To verify the effectiveness of the fusion operation, we also make inferences without CFS and simply test the model using average logits fusion of the

Fusion Strategy	mIoU
w/o CFS	75.23
Average fusion	76.01
CFS (w/o fusion inference)	77.10
CFS (w/ fusion inference)	77.55

Table 4. Comparison of different fusion modes in the reference stage. The first row represents the result of our method without CFS strategy. The second row represents the result of average model fusion. The third row represents the result of our single model with CFS for the training stage but without fusion inference. The last row represents the result of fusion inference.

two learners. As shown in Table 4, the performance of average fusion is 0.78% higher obviously, which shows that the fusion is indeed working. Moreover, comparing CFS(w/o fusion inference) testing on one learner with average fusion, the performance of the former is significantly boosted by 1.11%. It undoubtedly proves that the fusion ability of CFS is so powerful that it can learn the knowledge distribution of different learners well and test without extra computation. Finally, testing with CFS(w/ fusion inference), the fusion performance further improves by 0.45%, which tells us that the fusion strategy makes valuable improvement with additional calculations. That is, our CFS strategy is very necessary to improve the performance, which greatly integrates the learning ability of the two branch models.

Adaptive Contour Guidance Module. In section 3.4, we add parameter γ which is set to be initialized as 0 and gradually learns to assign a higher weight to 1. Several tendencies are examined in the experiment. As shown in Table 5, compared with not using the weight map by setting γ as 0, the ACGM strategy can clearly improve the performance from 75.22% to 77.10%. Conversely, constant 1 means the loss totally depends on the contour weight map. Others like log, linear, and exp tendencies keep up while the training iteration process also has been shown. In the above manner, we find the exp can improve the baseline by 1.15% to not using the weight map as the baseline. At the same time, we investigate the influence of different kernel sizes of Gaussian blur k that is used to soft binary contour maps as shown in section 3.4. From Table 5, we can see that $k = 64$ performs best on PASCAL VOC 2012. As a result for Cityscapes, we also use $k = img\_size/8$ which is 100 to generate the weight map. It is worth mentioning that the parameter k adjustment is based on the γ used in the exp manner.

Param	Value	Detail	mIoU
γ	constant 0	$\gamma = 0$	75.22
	constant 1	$\gamma = 1$	75.95
	linear	$\gamma = a_k x, 0 \leq \gamma < 1$	76.31
	exp	$\gamma = exp(a_e x) - 1, 0 \leq \gamma < 1$	**77.10**
	log	$\gamma = log(a_l x + 1), 0 \leq \gamma < 1$	75.56
k	k=32	$k = \frac{img\_size}{16}$	75.86
	k=64	$k = \frac{img\_size}{8}$	**77.10**
	k=128	$k = \frac{img\_size}{4}$	75.88

Table 5. Illustration on the performance of different trends of γ. and the performance of different kernel size k. All results are evaluated under the 1/8 partition protocol with ResNet-50 on PASCAL VOC 2012. Note that a_e is calculated by Eq.13, and a_k, a_l is in a similar way, and the results are generated by w/o fusion inference.

Contour Strategy	mIoU
w/o ACGM	75.22
Weakened	75.01
Strengthened	77.10

Table 6. Comparation on how to handle unreliable regions in the contour strategy. Note that the results are generated by w/o fusion inference.

Generally, there is a contour-weakened strategy that dynamically gives higher weight to the reliable pseudo labels while suppressing the unreliable pseudo labels. Instead of using it, we adopt the contour-strengthened strategy, which means the unreliable region acquires a higher weight, making the learner pay more attention to these regions and let the region less and less until the model is explicit. To prove this, we conduct contrast experiments as shown in Table 6, from it we can see that the contour-weakened strategy is lower than the contour-strengthened strategy by 2.09%. Further, weakening contour strategy brings no benefits even if compared with w/o ACGM. Actually, due to the unreliable pseudo labels being the difference region between the predictions of two learners, the unreliable pseudo label indeed should be emphasized. Emphasizing the different regions can promote the consistency regularization's performance on the other hand.

5. Conclusion

In this paper, we propose a novel SSSS framework called CFCG combining the CFS and ACGM. Specifically, our CFS creatively fuses the information between weak flows and strong flows respectively, through the channel-wise attention mechanism to tap into underutilized knowledge. During the test, both with and without fusion inference can achieve consistent performance gains. On the other hand, our ACGM guides the learners to adaptively and effectively identify unreliable spatial regions by relying on spatial contour information. The experiment results present our CFCG achieved new state-of-the-art results on two commonly used benchmarks, which yield mIoU 77.10%, 78.49% with no additional calculations, and 77.55%, 78.93% by fusion inference way on PASCAL VOC 2012 and on Cityscapes separately under 1/8 protocols. These comparisons have shown that our CFCG method performs favorably against state-of-the-art approaches for the semi-supervised semantic segmentation task.

References

[1] Inigo Alonso, Alberto Sabater, David Ferstl, Luis Montesano, and Ana C Murillo. Semi-supervised semantic segmentation with pixel-level contrastive learning from a class-wise memory bank. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8219–8228, 2021. 4

[2] Vijay Badrinarayanan, Alex Kendall, and Roberto Cipolla. Segnet: A deep convolutional encoder-decoder architecture for image segmentation. *IEEE transactions on pattern analysis and machine intelligence*, 39(12):2481–2495, 2017. 3

[3] David Berthelot, Nicholas Carlini, Ekin D Cubuk, Alex Kurakin, Kihyuk Sohn, Han Zhang, and Colin Raffel. Remixmatch: Semi-supervised learning with distribution alignment and augmentation anchoring. *arXiv preprint arXiv:1911.09785*, 2019. 3

[4] David Berthelot, Nicholas Carlini, Ian Goodfellow, Nicolas Papernot, Avital Oliver, and Colin A Raffel. Mixmatch: A holistic approach to semi-supervised learning. *Advances in neural information processing systems*, 32, 2019. 3

[5] Liang-Chieh Chen, George Papandreou, Florian Schroff, and Hartwig Adam. Rethinking atrous convolution for semantic image segmentation. *arXiv preprint arXiv:1706.05587*, 2017. 3, 6

[6] 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. 1, 4, 6, 7

[7] Bowen Cheng, Alex Schwing, and Alexander Kirillov. Per-pixel classification is not all you need for semantic segmentation. *Advances in Neural Information Processing Systems*, 34:17864–17875, 2021. 3

[8] 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. 6

[9] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 3

[10] 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. 6

[11] Jiashuo Fan, Bin Gao, Huan Jin, and Lihui Jiang. Ucc: Uncertainty guided cross-head co-training for semi-supervised semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9947–9956, 2022. 1, 2, 4, 6, 7

[12] Geoff French, Samuli Laine, Timo Aila, Michal Mackiewicz, and Graham Finlayson. Semi-supervised semantic segmentation needs strong, varied perturbations. *arXiv preprint arXiv:1906.01916*, 2019. 6

[13] Bharath Hariharan, Pablo Arbeláez, Lubomir Bourdev, Subhransu Maji, and Jitendra Malik. Semantic contours from inverse detectors. In *2011 international conference on computer vision*, pages 991–998. IEEE, 2011. 6

[14] Bharath Hariharan, Pablo Arbeláez, Lubomir Bourdev, Subhransu Maji, and Jitendra Malik. Semantic contours from inverse detectors. In *2011 international conference on computer vision*, pages 991–998. IEEE, 2011. 6

[15] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 6

[16] Hanzhe Hu, Fangyun Wei, Han Hu, Qiwei Ye, Jinshi Cui, and Liwei Wang. Semi-supervised semantic segmentation via adaptive equalization learning. *Advances in Neural Information Processing Systems*, 34:22106–22118, 2021. 2, 4, 6

[17] 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. pmlr, 2015. 6

[18] Zhanghan Ke, Di Qiu, Kaican Li, Qiong Yan, and Rynson WH Lau. Guided collaborative training for pixel-wise semi-supervised learning. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XIII 16*, pages 429–445. Springer, 2020. 1, 4, 6, 7

[19] Zhanghan Ke, Daoye Wang, Qiong Yan, Jimmy Ren, and Rynson WH Lau. Dual student: Breaking the limits of the teacher in semi-supervised learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6728–6736, 2019. 3, 4

[20] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolu-

tional neural networks. *Communications of the ACM*, 60(6):84–90, 2017. 6

[21] Donghyeon Kwon and Suha Kwak. Semi-supervised semantic segmentation with error localization network. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9957–9967, 2022. 4, 6, 7

[22] Xin Lai, Zhuotao Tian, Li Jiang, Shu Liu, Hengshuang Zhao, Liwei Wang, and Jiaya Jia. Semi-supervised semantic segmentation with directional context-aware consistency. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1205–1214, 2021. 4

[23] Yuyuan Liu, Yu Tian, Yuanhong Chen, Fengbei Liu, Vasileios Belagiannis, and Gustavo Carneiro. Perturbed and strict mean teachers for semi-supervised semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4258–4267, 2022. 1, 2, 4, 6, 7

[24] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3431–3440, 2015. 3

[25] Takeru Miyato, Shin-ichi Maeda, Masanori Koyama, and Shin Ishii. Virtual adversarial training: a regularization method for supervised and semi-supervised learning. *IEEE transactions on pattern analysis and machine intelligence*, 41(8):1979–1993, 2018. 1

[26] 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. 4, 6

[27] Kihyuk Sohn, David Berthelot, Nicholas Carlini, Zizhao Zhang, Han Zhang, Colin A Raffel, Ekin Dogus Cubuk, Alexey Kurakin, and Chun-Liang Li. Fixmatch: Simplifying semi-supervised learning with consistency and confidence. *Advances in neural information processing systems*, 33:596–608, 2020. 1, 2, 3, 4

[28] Robin Strudel, Ricardo Garcia, Ivan Laptev, and Cordelia Schmid. Segmenter: Transformer for semantic segmentation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 7262–7272, 2021. 3

[29] Yidong Wang, Hao Chen, Qiang Heng, Wenxin Hou, Marios Savvides, Takahiro Shinozaki, Bhiksha Raj, Zhen Wu, and Jindong Wang. Freematch: Self-adaptive thresholding for semi-supervised learning. *arXiv preprint arXiv:2205.07246*, 2022. 3, 4

[30] 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. 4, 6, 7

[31] Enze Xie, Wenhai Wang, Zhiding Yu, Anima Anandkumar, Jose M Alvarez, and Ping Luo. Segformer: Simple and efficient design for semantic segmentation with transformers. *Advances in Neural Information Processing Systems*, 34:12077–12090, 2021. 3

[32] Qizhe Xie, Zihang Dai, Eduard Hovy, Thang Luong, and Quoc Le. Unsupervised data augmentation for consistency training. *Advances in neural information processing systems*, 33:6256–6268, 2020. 3, 4

[33] Lihe Yang, Wei Zhuo, Lei Qi, Yinghuan Shi, and Yang Gao. St++: Make self-training work better for semi-supervised semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4268–4277, 2022. 4, 6, 7

[34] Bowen Zhang, Yidong Wang, Wenxin Hou, Hao Wu, Jindong Wang, Manabu Okumura, and Takahiro Shinozaki. Flexmatch: Boosting semi-supervised learning with curriculum pseudo labeling. *Advances in Neural Information Processing Systems*, 34:18408–18419, 2021. 3, 4

[35] Jianrong Zhang, Tianyi Wu, Chuanghao Ding, Hongwei Zhao, and Guodong Guo. Region-level contrastive and consistency learning for semi-supervised semantic segmentation. *arXiv preprint arXiv:2204.13314*, 2022. 4

[36] Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia. Pyramid scene parsing network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2881–2890, 2017. 3

[37] Sixiao Zheng, Jiachen Lu, Hengshuang Zhao, Xiatian Zhu, Zekun Luo, Yabiao Wang, Yanwei Fu, Jianfeng Feng, Tao Xiang, Philip HS Torr, et al. Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6881–6890, 2021. 3

[38] Yuanyi Zhong, Bodi Yuan, Hong Wu, Zhiqiang Yuan, Jian Peng, and Yu-Xiong Wang. Pixel contrastive-consistent semi-supervised semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7273–7282, 2021. 4, 6

[39] Yanning Zhou, Hang Xu, Wei Zhang, Bin Gao, and Pheng-Ann Heng. C3-semiseg: Contrastive semi-supervised segmentation via cross-set learning and dynamic class-balancing. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7036–7045, 2021. 2

[40] Yuliang Zou, Zizhao Zhang, Han Zhang, Chun-Liang Li, Xiao Bian, Jia-Bin Huang, and Tomas Pfister. Pseudoseg: Designing pseudo labels for semantic segmentation. *arXiv preprint arXiv:2010.09713*, 2020. 2

[41] Simiao Zuo, Yue Yu, Chen Liang, Haoming Jiang, Siawpeng Er, Chao Zhang, Tuo Zhao, and Hongyuan Zha. Self-training with differentiable teacher. *arXiv preprint arXiv:2109.07049*, 2021. 2