Region-Level Contrastive and Consistency Learning for Semi-Supervised Semantic Segmentation

Jianrong Zhang1†, Tianyi Wu2,3 †, Chuanghao Ding4 *, Hongwei Zhao1, Guodong Guo2,3 ‡
1College of Computer Science and Technology, Jilin University, Changchun, China
2Institute of Deep Learning, Baidu Research, Beijing, China
3National Engineering Laboratory for Deep Learning Technology and Application, Beijing, China
4College of Software, Jilin University, Changchun, China
{jrzhang20, dingch20}@mails.jlu.edu.cn, zhaohw@jlu.edu.cn,
{wutianyi01, guoguodong01}@baidu.com

Abstract
Current semi-supervised semantic segmentation methods mainly focus on designing pixel-level consistency and contrastive regularization. However, pixel-level regularization is sensitive to noise from pixels with incorrect predictions, and pixel-level contrastive regularization has a large memory and computational cost. To address the issues, we propose a novel region-level contrastive and consistency learning framework (RC2L) for semi-supervised semantic segmentation. Specifically, we first propose a Region Mask Contrastive (RMC) loss and a Region Feature Contrastive (RFC) loss to accomplish region-level contrastive property. Furthermore, Region Class Consistency (RCC) loss and Semantic Mask Consistency (SMC) loss are proposed for achieving region-level consistency. Based on the proposed region-level contrastive and consistency regularization, we develop a region-level contrastive and consistency learning framework (RC2L) for semi-supervised semantic segmentation, and evaluate our RC2L on two challenging benchmarks (PASCAL VOC 2012 and Cityscapes), outperforming the state-of-the-art.

1 Introduction
Semantic segmentation has high potential values in a variety of applications. However, training supervised semantic segmentation models requires large-scale pixel-level annotations, and such pixel-wise labeling is time-consuming and expensive. This work focuses on semi-supervised semantic segmentation, which takes advantage of a large amount of unlabeled data and limits the need for labeled examples.

Currently, many works have demonstrated that designing pixel-level consistency regularization [French et al., 2020; Ouali et al., 2020; Chen et al., 2021], and pixel-level contrastive regularization [Zhong et al., 2021; Lai et al., 2021] are beneficial for semi-supervised semantic segmentation. The first family of works utilized pixel-level label consistency under different data augmentations [French et al., 2020; Zou et al., 2021], feature perturbations [Ouali et al., 2020], network branches [Ke et al., 2020], and segmentation models [Chen et al., 2021]. These methods benefited from using the label-space consistency property on the unlabeled images. Figure 1 (a) shows the pixel-level label consistency of employing different data augmentations. Differently, the work [Lai et al., 2021] proposed the Directional Contrastive Loss to accomplish the pixel-level feature contrast, which is
shown in Figure 1 (b). Substantial progress has been made by utilizing pixel-level label consistency and pixel-level feature contrastive property. For example, PC²Seg [Zhong et al., 2021] leveraged and simultaneously enforced the consistency property in the label space and the contrastive property in the feature space, which is illustrated in Figure 1 (c). However, such pixel-level regularization is sensitive to noise from pixels with incorrect predictions. Besides, pixel-level contrastive regularization has memory and computational cost with $O(\text{pixel}_\text{num}^2)$, and usually requires designing additional negative example filtering mechanism carefully.

To overcome the above challenges, we propose to design Region-level Contrastive and Consistency Learning (RC²L) for semi-supervised semantic segmentation. As shown in Figure 1 (d), our method enforces the consistency of region classes and the contrastive property of features and masks from different regions. It was inspired by MaskFormer [Cheng et al., 2021] which formulated supervised semantic segmentation as a mask (or region) classification problem.

Specifically, we first propose a Region Mask Contrastive (RMC) loss and a Region Feature Contrastive (RFC) loss to achieve region-level contrastive property. The former pulls the masks of matched region pairs (or positive pairs) closer and pushes away the unmatched region pairs (or negative pairs), and the latter pulls the features of matched region pairs closer and pushes away the unmatched regions pairs. Furthermore, Region Class Consistency (RCC) loss and Semantic Mask Consistency (SMC) loss are proposed for encouraging the consistency of the region classes and the consistency of union regions with the same class, respectively.

Based on the proposed components, we develop a Region-level Contrastive and Consistency Learning (RC²L) framework for semi-supervised semantic segmentation. The effectiveness of our approach is demonstrated by conducting extensive ablation studies. In addition, we evaluate our RC²L on several widely-used benchmarks, e.g., PASCAL VOC 2012 and Cityscapes, and the experiment results show that our approach outperforms the state-of-the-art semi-supervised segmentation methods.

2 Related Works

Semantic segmentation. Since the emergence of Fully Convolutional Network (FCN) [Long et al., 2015], per-pixel classifications have achieved high accuracies in semantic segmentation tasks. The development of modern deep learning methods for semantic segmentation mainly focuses on how to model context [Chen et al., 2018; Wu et al., 2020b; Wu et al., 2020a]. Recently, Transformer-based methods for semantic segmentation show an excellent performance. SegFormer [Xie et al., 2021] proposed mask transformer as a decoder. More recently, MaskFormer [Cheng et al., 2021] re-formulated semantic segmentation as a mask (or region) classification task. Differently, we focus on how to make a better use of unlabeled data to perform region-level predictions for semi-supervised semantic segmentation.

Semi-supervised semantic segmentation. It is important to explore semi-supervised semantic segmentation to reduce per-pixel labeling costs. Early approach [Hung et al., 2018] used generative adversarial networks (GANs) and adversarial loss to train on unlabeled data. Consistency regularization and pseudo-labeling have also been widely explored for semi-supervised semantic segmentation, by enforcing consistency among the predictions, either from feature perturbation [Ouali et al., 2020], augmented input images [French et al., 2020], or different segmentation models [Ke et al., 2020]. Later, PseudoSeg [Zou et al., 2021] combined pixel-level labels with image-level labels to enhance the semi-supervised learning through boosting the quality of pseudo labels. CPS [Chen et al., 2021] proposed the cross pseudo supervision to apply consistency between two segmentation networks with the same architecture. Self-training has also driven the development of the state-of-the-art methods [Zoph et al., 2020; He et al., 2021]. Pixel-level contrastive based approaches have been proposed recently. CMB [Alonso et al., 2021] introduced a memory bank for positive-only contrastive learning. Directional Contrastive Loss [Lai et al., 2021] was proposed for training between pixel features and pseudo labels. PC²Seg [Zhong et al., 2021] enforced the pseudo labels to be consistent, and encouraged pixel-level contrast to pixel features. All these methods mainly focused on pixel-level regularization and conducted pixel-level predictions. Differently, our approach explores region-level regularization and performs region-level predictions.

Contrastive learning. Image-level contrastive learning has shown excellent prospects for self-supervised representation learning. SimCLR [Chen et al., 2020a] proposed to take the contrastive loss for training between images by applying different data augmentations. MoCo V2 [Chen et al., 2020b] presented a momentum encoder to reduce the requirement of large batch size. Pixel-level contrastive learning has also been proven to be beneficial for dense prediction tasks. [Wang et al., 2021] introduced pixel-level contrastive learning for supervised semantic segmentation. Different from these works, we explore how to conduct region-level contrastive learning for semi-supervised semantic segmentation.

3 Method

We first describe our framework, Region-level Contrastive and Consistency Learning (RC²L) for semi-supervised semantic segmentation. Then, we present the Region Mask Contrastive (RMC) loss and Region Feature Contrastive (RFC) loss. The former pulls the masks of matched region pairs closer and pushes away the unmatched region pairs, and the latter pulls the features of matched region pairs closer and pushes away the unmatched region pairs. Finally, we introduce the Region Class Consistency (RCC) loss and Semantic Mask Consistency (SMC) loss. RCC loss can enforce the consistency of region classes, and SMC loss can further promote the consistency of union regions with the same class.

3.1 Framework

The overall framework of our RC²L is shown in Figure 2, which consists of a student model and a teacher model. The teacher model has the same architecture as the student model but uses a different set of weights. The student model is trained with both labeled and unlabeled data.
Specifically, given an image-label pair, \( \{ x_i, z^{gt}_i \} \), \( x_i \) is the image, and \( z^{gt}_i \) is the corresponding annotations that are the set of \( N^{gt} \) ground truth segments, i.e., \( z^{gt}_i = \{ (c_i^{gt}, m_i^{gt}) | c_i^{gt} \in \{ 1, 2, \ldots, K \}, m_i^{gt} \in [0,1]^{H \times W} \}_{i=1}^{N^{gt}} \), where \( c_i \) is the ground truth class of the \( i^{th} \) segment, \( W \) and \( H \) represent the width and height of the input image. We feed the image \( x_i \) into the student model, outputting probability-mask pairs \( z^s_i = \{ (p_i^s, m_i^s) | p_i^s \in \Delta^{K+1}, m_i^s \in [0,1]^{H \times W} \}_{i=1}^{N} \), where the probability distribution \( p_i \) contains a “no object” label \( \emptyset \) and \( K \) category labels. A bipartite matching-based assignment \( \sigma \) between the set of predictions \( z^s_i \) and \( z^{gt}_i \) is conducted for computing supervised loss:

\[
L_{label}(z^s_i, z^{gt}_i) = \sum_{i=1}^{N} \left[ -\log p_{\sigma(i)}(c_i^{gt}) + \mathbb{1}_{c_i^{gt} \neq \emptyset} \mathcal{L}_{mask}(m_{\sigma(i)}^s, m_i^{gt}) \right],
\]

(1)

where \( \mathcal{L}_{mask} \) is a binary mask loss. Please refer to [Cheng et al., 2021] for more details.

For the unlabeled image \( x_u \), we feed its weak augmentation \( x_u^w \) and strong augmentation \( x_u^s \) into the teacher and student models, respectively. Here, we use short edge resize, random crop, random flip, and color augmentation as weak augmentations, and use all weak data augmentation methods and CutMix [Yun et al., 2019] as strong augmentations. The teacher is employed to generate pseudo label \( z^t_i = \{ (c_i^t, m_i^t) \}_{i=1}^{N} \), which is used to guide the training of the student. The unsupervised loss can be formulated as follow:

\[
L_{unlabel}(x_u) = \beta_1 \mathcal{L}_{RCC} + \beta_2 \mathcal{L}_{SMC} + \beta_3 \mathcal{L}_{RMC} + \beta_4 \mathcal{L}_{RFC},
\]

(2)

where \( \mathcal{L}_{RCC} \) and \( \mathcal{L}_{SMC} \) mean the Region Class Consistency loss and Semantic Mask Consistency loss (Section 3.2), respectively. \( \mathcal{L}_{RMC} \) and \( \mathcal{L}_{RFC} \) denote the Region Mask Contrastive loss and Region Feature Contrastive loss (Section 3.3), respectively. \( \beta_1, \beta_2, \beta_3 \) and \( \beta_4 \) denote the loss weight. Therefore, the total loss is defined as:

\[
\mathcal{L} = L_{label}(z^s_i, z^{gt}_i) + \alpha L_{unlabel}(x_u),
\]

(3)

where \( \alpha \) is a constant to balance between the supervised and unsupervised losses.

Following Mean Teacher [Tarvainen and Valpola, 2017], the parameters \( \theta_t \) of teacher model are an exponential moving...
average of parameters $\theta_s$ of the student model. Specifically, at every training step, the parameters $\theta_t$ of teacher network is updated as follows:

$$\theta_t = \tau \theta_t + (1 - \tau) \theta_s,$$

(4)

where $\tau \in [0, 1]$ is a decay rate.

### 3.2 Region-level Contrastive Learning

The goal of region-level contrastive learning is to increase the similarity between each region mask and feature from strong augmentation $x_w^s$ and the mask and feature of the matched region (from weak augmentation $x_w^w$), and reduce the similarity between unmatched region pairs. Specifically, we propose a Region Mask Contrastive (RMC) loss and a Region Feature Contrastive (RFC) loss to achieve region-level contrastive property. The former pulls the masks of matched region pairs (or positive pairs) closer and pushes away the unmatched region pairs (or negative pairs), and the latter pulls the features of matched region pairs closer and pushes away the unmatched regions pairs.

The weak augmentation $x_w^w$ is firstly fed into the student network, which outputs pseudo labels $z^t = \{(c_i^t, m_i^t)\}_{i=1}^{N^t}$. The strong augmentation $x_w^s$ is fed into the student network, which outputs class logits $p^t = \{p_i^t \in \Delta^{K + 1}\}_{i=1}^{N^t}$, mask embeddings $f^s = \{f_i^s \in \mathbb{R}^N\}_{i=1}^{N}$, and per-pixel features $F^s \in \mathbb{R}^{C \times H / 4 \times W / 4}$. Firstly, we obtain binary mask predictions $\{m_i^t\}_{i=1}^{N^t}$ via a dot product between the mask embeddings and per-pixel features. Then, the bipartite matching-based assignment $\sigma$ between student predictions $\{m_i^t\}_{i=1}^{N^t}$ and pseudo segment set $\{\sigma(i)\}_{i=1}^{N^t}$ is used to get the matched index set $ID = \{\sigma(i)\}_{i=1}^{N^t}$ and computing RMC Loss:

$$L_{RMC} = \sum_{i}^{N^t} -\log \frac{\exp(d(m^t_{\sigma(i)}, m^t_i)/\tau_m)}{\sum_{j \in ID, j \neq \sigma(i)} \exp(d(m^t_{\sigma(i)}, m^t_j)/\tau_m)},$$

(5)

where $\tau_m$ is a temperature hyper-parameter to control the scale of terms inside exponential, and $d(m^t_i, m^t_j) = 2|m_i \cap m_j|/(|m_i| + |m_j|)$ measures the similarity between two masks.

Next, we compute region features $r^s = \{r^s_{\sigma(i)}\}_{i=1}^{N^t}$ via combining per-pixel features $F^s$ and region mask set $\{m_i^t\}_{i=1}^{N^t}$. The process can be formulated as follow:

$$r^s_{\sigma(i)} = GAP(m^t_{\sigma(i)} \cdot F^s), \forall \sigma(i),$$

(6)

where “GAP” indicates a global average pooling operation. We compute target region features $r^t = \{r^t_i\}_{i=1}^{N^t}$ as follow:

$$r^t_i = GAP(m^t_i \cdot F^s), \forall i,$$

(7)

Here, per-pixel features $F^s$ are not used to compute the target region features, since the student features, since the student model and the teacher model have different weights and feature space. Then, Region Feature Contrastive loss is defined as:

$$L_{RFC} = \sum_{i}^{N^t} -\log \frac{\exp(\cos(r^t_i, r^t_i)/\tau_f)}{\sum_{j \in ID, j \neq \sigma(i)} \exp(\cos(r^t_i, r^t_j)/\tau_f)},$$

(8)

where $\tau_f$ is a temperature hyper-parameter to control the scale of terms inside exponential, $\cos(u, v) = \frac{u^T v}{\|u\|\|v\|}$ is the cosine similarity.

### 3.3 Region-level Consistency Learning

Different from previous works [Zhong et al., 2021; Alonso et al., 2021] which used pixel-level consistency regularization, we develop a region-level consistency learning, which consists of a Region Class Consistency (RCC) Loss and Semantic Mask Consistency (SMC) loss. The former enforces the consistency of the region classes, and the latter further promotes the consistency of the union regions with the same class.

Given the student class logits $\{p_i^t \in \Delta^{K+1}\}_{i=1}^{N}$ and the pseudo segment label $\{c_i^t, m_i^t\}_{i=1}^{N^t}$, the proposed Region Class Consistency loss is employed to enforce the class consistency of matched region pairs, which is defined as:

$$L_{RCC} = \sum_{i}^{N^t} -\log p_{\sigma(i)}^s(c_i^t),$$

(9)

Furthermore, different regions may correspond to the same class, so we design a Semantic Mask Consistency loss to promote the consistency of the union regions with the same class and the pseudo semantic mask, which can be formulated as follow:

$$L_{SMC} = \sum_{i}^{N^t} \text{Mask}(\text{Union}(\{m^t_i\}_{i=1}^{N^t}, c_i^t), m_i^t),$$

(10)

where $\text{Union}(\cdot)$ means merging regions with the same class into a single region. Following DETR [Carion et al., 2020] and MaskFormer [Cheng et al., 2021], $\text{Mask}(\cdot)$ is a linear combination of a focal loss [Lin et al., 2017] and a dice loss [Milletari et al., 2016].

### 4 Experiment

#### 4.1 Datasets

**PASCAL VOC 2012.** PASCAL VOC 2012 [Everingham et al., 2015] is a standard object-centric semantic segmentation dataset, which contains more than 13,000 images with 21 classes (20 object classes and 1 background class). In the original PASCAL VOC 2012 dataset (VOC Train), 1464 images are used for training, 1449 images for validation and 1456 images for testing. Following the common practice, we also use the augmented dataset (VOC Aug) which contains 10,582 images as the training set. For both the original and augmented datasets, 1/2, 1/4, 1/8, and 1/16 training images are used as labeled data, respectively, for conducting the semi-supervised experiments.
with power = 0.9. For PASCAL VOC 2012 dataset, we use the crop size of 512 × 512, and train our model for 120K and 160K iterations for VOC Train and VOC Aug, respectively. For Cityscapes dataset, we use the crop size of 768 × 768, and set training iterations as 120K without using any extra training data.

4.2 Comparison to the State-of-the-Arts

Results on PASCAL VOC 2012. We show the comparison results on the PASCAL VOC 2012 Val set in Table 1. All the models are trained using 4 V100 GPUs. On VOC Train, our proposed RC²L outperforms all existing pixel-level regularization methods, achieving the improvement of 1.18%, 0.53%, 1.45%, and 1.26% with partition protocols of 1/2, 1/4, 1/8, and 1/16 respectively. We also compare the results of the 1.4k/9k split, our RC outperforms all existing pixel-level regularization methods, achieving the improvement of 1.79%, 2.03%, 1.05%, and 1.08% under 1/2, 1/4, 1/8, and 1/16 partitions, respectively.

Results on Cityscapes. Table 2 shows comparison results on Cityscapes Val set. All the models are trained using 8 V100 GPUs. The performance of our method is improved by 2.53% and 2.51%, compared to the supervised baseline under partition protocols of 1/4 and 1/8, respectively. Our method also outperforms the previous state-of-the-art methods, obtaining the improvements of 1.79%, 2.03%, 1.05%, and 1.08% under 1/2, 1/4, 1/8, and 1/16 partitions, respectively.

4.3 Ablation Study

Investigating each component. We investigate the effect of each component in our methods. The experimental results are illustrated in Table 3. The baseline model only uses mask consistency between teacher and student predictions for unlabeled data. It can be seen that the baseline method achieves 69.37% and 75.26% on 1/2 VOC Train and 1/4 VOC Aug datasets, respectively. We can see that SMC loss improves the
Table 3: Ablation study of each component. The first row is the result of our baseline model.

| $\mathcal{L}_{\text{SMC}}$ | $\mathcal{L}_{\text{RCC}}$ | $\mathcal{L}_{\text{RMC}}$ | $\mathcal{L}_{\text{RFC}}$ | VOC Train (1/2) | VOC Aug (1/4) |
|----------------------|----------------------|----------------------|----------------------|----------------|----------------|
| ✓                    | ✓                    | ✓                    | ✓                    | 69.37          | 75.26          |
| ✓                    | ✓                    | ✓                    | ✓                    | 73.26          | 76.58          |
| ✓                    | ✓                    | ✓                    | ✓                    | 76.07          | 77.89          |
| ✓                    | ✓                    | ✓                    | ✓                    | 76.85          | 78.92          |
| ✓                    | ✓                    | ✓                    | ✓                    | 77.06          | 79.71          |

Table 4: Effect of different unsupervised loss weights.

| $\alpha$ | VOC Train (1/2) | VOC Aug (1/4) |
|----------|----------------|---------------|
| 1.0      | 75.65          | 79.71         |
| 1.5      | 76.60          | 79.15         |
| 2.0      | 77.06          | 79.15         |

Table 5: Study on the number of queries.

| Num. queries | VOC Train (1/2) | VOC Aug (1/4) |
|--------------|----------------|---------------|
| 100          | 74.01          | 75.83         |
| 50           | 77.06          | 79.71         |
| 20           | 76.18          | 78.37         |

Loss weight. We show the effect of loss weight $\alpha$ which is used to balance the supervised loss and unsupervised loss in Table 4. We find that $\alpha = 2$ achieves the best performance on 1/2 VOC Train set. For 1/4 VOC Aug set, $\alpha = 1$ obtains the best performance.

Number of queries. We study the number of queries on PASCAL VOC 2012 in Table 5. We can see that $N = 50$ achieves the best result on both 1/2 VOC Train and 1/4 VOC Aug, and $N = 100$ degrades the performance more than $N = 20$. This suggests that a small number of queries is sufficient to provide a good result for datasets with a few number of categories in just one image.

4.4 Visualization

Figure 3 shows the visualization results of our methods on PASCAL VOC 2012. We compare the predictions of our RC$^2$L with the ground-truth, supervised baseline, and semi-supervised consistency baseline. One can see that our RC$^2$L can correct more noisy predictions compared to the supervised baseline and the semi-supervised consistency baseline. We follow CCT and directly complete consistency learning between student outputs and pseudo labels as the semi-supervised consistency baseline. In particular, the supervised baseline mislabels some pixels in the 1st row and the 3rd row. Both the supervised baseline and the semi-supervised consistency baseline mistakenly classify some pixels in the 3rd row and the 5th row.

5 Conclusion

We have developed the Region-level Contrastive and Consistency Learning (RC$^2$L) method for semi-supervised semantic segmentation. The core contributions of our RC$^2$L are the proposed region-level contrastive and consistency regularization. The former consists of Region Mask Contrastive (RMC) loss and Region Feature Contrastive (RFC) loss, the latter contains Region Class Consistency (RCC) loss and Semantic Mask Consistency (SMC) loss. Extensive experiments on PASCAL VOC 2012 and Cityscapes have shown that our RC$^2$L outperforms the state-of-the-art semi-supervised semantic segmentation methods, demonstrating that our region-level Contrastive and Consistency regularization can achieve better results than previous pixel-level regularization.
References

[Alonso et al., 2021] 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 ICCV, 2021.

[Carion et al., 2020] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In ECCV, pages 213–229. Springer, 2020.

[Chen et al., 2018] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L. Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2018.

[Chen et al., 2020a] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. *arXiv preprint arXiv:2002.05709*, 2020.

[Chen et al., 2020b] Xinlei Chen, Haoqi Fan, Ross Girshick, and Kaiming He. Improved baselines with momentum contrastive learning. *arXiv preprint arXiv:2003.04297*, 2020.

[Chen et al., 2021] Xiaokang Chen, Yuhui Yuan, Gang Zeng, and Jingdong Wang. Semi-supervised semantic segmentation with cross pseudo supervision. In CVPR, 2021.

[Cheng et al., 2021] Bowen Cheng, Alexander G. Schwing, and Alexander Kirillov. Per-pixel classification is not all you need for semantic segmentation. In NeurIPS, 2021.

[Cords et al., 2016] Marius Cords, 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.

[Deng et al., 2009] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, pages 248–255, 2009.

[Everingham et al., 2015] Mark Everingham, SM Ali Eslami, Luc Van Gool, ChristopherKIWilliams, John Winn, and Andrew Zisserman. The pascal visual object classes challenge: A retrospective. *International Journal of Computer Vision*, 111(1):98–136, 2015.

[French et al., 2020] Geoff French, Timo Aila, Samuli Laine, Michal Mackiewicz, and Graham Finlayson. Semi-supervised semantic segmentation needs strong, high-dimensional perturbations. In *BMVC*, 2020.

[He et al., 2021] Ruifei He, Jihan Yang, and Xiaojuan Qi. Redistributing biased pseudo labels for semi-supervised semantic segmentation: A baseline investigation. In *ICCV*, pages 6930–6940, 2021.

[Hung et al., 2018] Wei-Chih Hung, Yi-Hsuan Tsai, Yan-Ting Liu, Yen-Yu Lin, and Ming-Hsuan Yang. Adversarial learning for semi-supervised semantic segmentation. In *BMVC*, 2018.

[Ke et al., 2020] Zhanghan Ke, Di Qu, Kaican Li, Qiong Yan, and Rynson WH Lau. Guided collaborative training for pixel-wise semi-supervised learning. In *ECCV*, 2020.

[Lai et al., 2021] Xin Lai, Zhuoat Tian, Li Jiang, Shu Liu, Hengshuang Zhao, Liwei Wang, and Jiaya Jia. Semi-supervised semantic segmentation with directional context-aware consistency. In *CVPR*, 2021.

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

[Lin et al., 2017] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *ICCV*, pages 2980–2988, 2017.

[Long et al., 2015] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In CVPR, 2015.

[Milletari et al., 2016] F Milletari, N Navab, SA Ahmadi, and V-net. Fully convolutional neural networks for volumetric medical image segmentation. In *JDV*, 2016.

[Miyato et al., 2018] Takeru Miyato, Shin-ichi Maeda, Shin Ishii, and Masanori Koyama. Virtual adversarial training: a regularization method for supervised and semi-supervised learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2018.

[Ouali et al., 2020] Yassine Ouali, Céline Hudelot, and Myriam Tami. Semi-supervised semantic segmentation with cross-consistency training. In CVPR, 2020.

[Tarvainen and Valpola, 2017] Antti Tarvainen and Harri Valpola. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. In *NeurIPS*, 2017.

[Wang et al., 2021] Wenguan Wang, Tianfei Zhou, Fisher Yu, Jifeng Dai, Endre Konukoglu, and Luc Van Gool. Exploring cross-image pixel contrast for semantic segmentation. In *ICCV*, pages 7303–7313, 2021.

[Wu et al., 2020a] Tianyi Wu, Yu Lu, Yu Zhu, Chuang Zhang, Ming Wu, Zhanyu Ma, and Guodong Guo. Ginet: Graph interaction network for scene parsing. In *European Conference on Computer Vision*, pages 34–51. Springer, 2020.

[Wu et al., 2020b] Tianyi Wu, Sheng Tang, Rui Zhang, Juan Cao, and Yongdong Zhang. Cgnet: A lightweight context guided network for semantic segmentation. *IEEE Transactions on Image Processing*, 30:1169–1179, 2020.

[Xie et al., 2021] Enze Xie, Wenhai Wang, Zhiding Yu, Anima Anandkumar, Jose M Alvarez, and Ping Luo. Segformer: Simple and efficient design for semantic segmentation with transformers. *arXiv preprint arXiv:2105.15203*, 2021.

[Yun et al., 2019] Sangdoo Yun, Dongyoon Han, Seong Joon Oh, Sanghyuk Chun, Junsuk Cho, and Youngjoon Yoo. Cutmix: Regularization strategy to train strong classifiers with localizable features. In *ICCV*, pages 6023–6032, 2019.

[Zhong et al., 2021] Yuanyi Zhong, Bodi Yuan, Hong Wu, Zhiquiang Yuan, Jian Peng, and Yu-Xiong Wang. Pixel contrastive-consistent semi-supervised semantic segmentation. In *ICCV*, pages 7273–7282, 2021.

[Zoph et al., 2020] Barret Zoph, Golnaz Ghiasi, Tsung-Yi Lin, Yin Cui, Hanxiao Liu, Ekin D. Cubuk, and Quoc V. Le. Rethinking pre-training and self-training. *CoRR*, abs/2006.06882, 2020.

[Zou et al., 2021] Yuliang Zou, Zizhao Zhang, Han Zhang, Chunliang Li, Xiao Bian, Jia-Bin Huang, and Tomas Pfister. Pseudoseg: Designing pseudo labels for semantic segmentation. In *ICLR*, 2021.