CARD: Semi-supervised Semantic Segmentation via Class-agnostic Relation based Denoising

Xiaoyang Wang[1,2,3], **Jimin Xiao**[1*], **Bingfeng Zhang**[1,2] and **Limin Yu**[1]
[1]Xi'an Jiaotong Liverpool University
[2]University of Liverpool
[3]Dinnar Automation Technology
wangxy@liverpool.ac.uk, {jimin.xiao, bingfeng.zhang, limin.yu}@xjtlu.edu.cn

Abstract

Recent semi-supervised semantic segmentation methods focus on mining extra supervision from unlabeled data by generating pseudo labels. However, noisy labels are inevitable in this process which prevent effective self-supervision. This paper proposes that noisy labels can be corrected based on semantic connections among features. Since a segmentation classifier produces both high and low-quality predictions, we can trace back to feature encoder to investigate how a feature in a noisy group is related to those in the confident groups. Discarding the weak predictions from the classifier, rectified predictions are assigned to the wrongly predicted features through the feature relations. The key to such an idea lies in mining reliable feature connections. With this goal, we propose a class-agnostic relation network to precisely capture semantic connections among features while ignoring their semantic categories. The feature relations enable us to perform effective noisy label corrections to boost self-training performance. Extensive experiments on PASCAL VOC and Cityscapes demonstrate the state-of-the-art performances of the proposed methods under various semi-supervised settings.

1 Introduction

The semantic segmentation model learns to predict pixel-wise categorical labels given an unseen image [Chen *et al.*, 2017]. However, its performance relies on massive pixel-level annotations, where the labeling process can be laborious and time-consuming. To tackle this limitation, semi-supervised semantic segmentation is studied to facilitate learning in a low data regime by leveraging a fraction of fully labeled images and massive unlabeled samples to achieve comparable results to the fully-supervised setting.

The key in semi-supervised semantic segmentation lies in mining extra supervision from unlabeled data. Motivated by advances in semi-supervised classification [Berthelot *et al.*, 2019; Sohn *et al.*, 2020; Zhang *et al.*, 2021], recent semi-supervised segmentation favors pseudo labeling [Chen *et al.*, 2020] and consistency regularization [Mittal *et al.*, 2019]. Their hybrid methods have achieved state-of-the-art performance [Chen *et al.*, 2021]. Recent works mainly enforce prediction consistency on differently perturbed unlabeled data [French *et al.*, 2020] and meanwhile conduct self-supervision by generating pseudo labels. However, few works have paid attention on the problem of label noise in this process. In the early stage of training, the weak classifier cannot guarantee high quality pseudo supervision and the misleading signals can hardly be corrected by the knowledge of the model itself, which leads to the confirmation bias [Arazo *et al.*, 2020] as training proceeds.

In this work, we propose an effective pseudo label denoising approach to achieve robust semi-supervised semantic segmentation. The motivation comes from the smoothness assumption [Luo *et al.*, 2018] that two near samples should have close corresponding predictions. Since the weak clas-

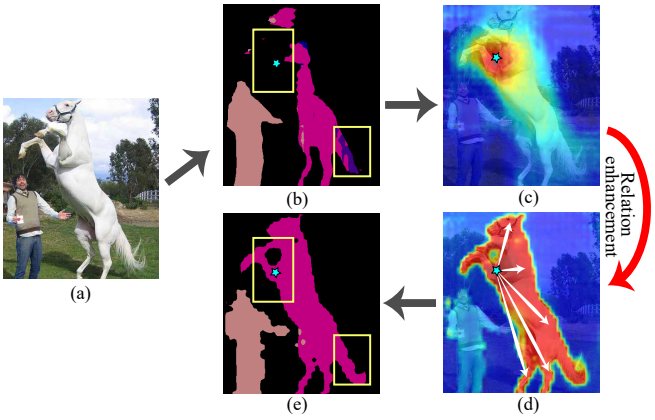

Figure 1: The noisy label correction procedure by the proposed method. (a) The input image. (b) The initial pseudo label with noisy predictions and one of them are marked as blue star. (c) Given the noisy pixel, we take features from encoder to measure the relation of its corresponding feature to all other features. The relation estimation is weak since it only focuses on its geometric neighbours. (d) The feature relation estimation is enhanced by our proposed module and the refined estimation strongly connects the pixel with its semantic neighbours. (e) The enhanced relation guides us to ensemble reliable predictions to re-calibrate the wrongly predicted pixels.

*Corresponding author

sifier trained on limited data can produce a decent amount of high-quality predictions, we assume this knowledge can be transferred to those noisy regions to re-calibrate their predictions, based on a reasonable "closeness" estimation between samples. Given a wrongly predicted feature embedding, once its semantic connections to plenty of features with correct labels are known, it is possible to conduct weighted prediction ensemble to re-estimate its prediction using its semantic neighbors. The key to such an approach relies on reliable feature relations. One straightforward way is to leverage features from the segmentation encoder to estimate their relations based on a pre-defined distance metric (*e.g.*, L2 distance). However, such direct relation estimations are inaccurate, as shown in Fig. 1(c) where the wrongly predicted pixel (blue star) only activates its geometric but not semantic neighbors, which prevents effective label correction.

To obtain reliable feature relations, we propose a class-agnostic relation network (CARNet) to enhance the relation estimation. Given features from the segmentation encoder, the proposed module maps these features into a new metric space where the feature relation is supervised in a class-agnostic way. Reliable relation estimation can be obtained through CARNet as shown in Fig. 1(d), where strong connections are built between the target feature to all its semantic neighbors. The noisy label can be corrected by transferring knowledge from plenty of reliable features with such connections. Class-agnostic supervision is the key for the CARNet, which is extracted from categorical annotations with no cost. Such supervision only focuses on the class equivalence between features, which primarily alleviates the bias caused by unbalanced classes. Therefore, it enables CARNet to provide reliable feature relation estimations. With refined feature relations, effective label denoising in self-training can be achieved. Furthermore, a progressive denoising strategy is proposed to control the pace of denoising during training in order to achieve stable and efficient label correction.

Our approach achieves state-of-the-art performances on Pascal VOC and Cityscapes under various semi-supervised settings. Note that the proposed framework is solely based on label denoising and does not contain any form of strong data augmentation, while previous works highly relied on this. Our contributions are summarized as follows:

- We propose a pseudo label denoising routine by transferring knowledge from reliable predictions to noisy labels through feature relations.
- We propose a module named CARNet to online refine feature relation estimation by enhancing their intra-class connections.
- We propose a progressive denoising strategy to conduct stable and efficient self-training. The framework achieves state-of-the-art performances.

2 Related Works

2.1 Semi-supervised Semantic Segmentation

Early works on semi-supervised semantic segmentation apply generative models and adversarial training to generate high-quality pseudo labels [Hung *et al.*, 2018; Mittal *et al.*, 2019]. Recent works pay attention to consistency regularization between predictions on different views of unlabeled data. French *et al.* [French *et al.*, 2020] introduces strong data augmentations of patch-level CutMix on unlabeled data to create randomly mixed versions of images. GCT [Ke *et al.*, 2020] introduces a multi-encoder system to enforce consistency among predictions of different encoders under feature-level perturbations. C^3-SemiSeg [Zhou *et al.*, 2021] creates region-level data augmentation to minimize the feature discrepancy between labeled and unlabeled data. CPS [Chen *et al.*, 2021] applies two differently initialized models in self-training to conduct cross pseudo supervision. Note that the noisy pseudo label problem is not well studied in previous works, which is the main focus of this paper.

2.2 Learning with Noisy Labels

Our approach is closely related to the topic of learning from noisy labels. Recent works mainly focus on mining reliable supervisions according to model disagreement. Co-teaching [Han *et al.*, 2018] trains two models simultaneously, and each model selects its small-loss samples which are used to train another model. Dual-student [Ke *et al.*, 2019] follows a similar path and adopts the idea to the field of semi-supervised classification. DMT [Feng *et al.*, 2020] leverages two differently initialized models to let each model generate offline pseudo labels for the other one. Inter-model disagreement is used to re-weight training loss to achieve noise-robust training. Unlike the above methods that try to bypass noisy supervision, our approach aims to directly correct those noisy labels based on the knowledge of feature relations.

3 Methodology

3.1 Overview

The proposed method aims to re-calibrate noisy pseudo labels by aggregating knowledge from their corresponding semantic neighbors. After conducting a warm-up on the annotated set, a student-teacher-CARNet framework is built to conduct self-training on unlabeled images. The teacher model is the exponential moving average (EMA) of the student network. The whole process of the self-training stage is shown in Fig. 2, which can be divided into the following parts: (1) Given predictions from the teacher model, a class-wise confidence bank is built and updated to achieve online label reliability estimation. It dynamically identifies both potentially noisy regions and high-quality predictions. (2) The CARNet after warm-up is continuously supervised by relation labels from reliable teacher predictions to further boost accuracy of feature relation estimation. (3) Based on the relation estimation, label denoising is conducted progressively to provide enhanced supervision for the student model.

3.2 Class-balanced Label Reliability Estimation

Since our goal is to correct noisy labels with reliable predictions, the first thing to consider is how label quality is measured. In this work, to set the criterion, we follow [Sohn *et al.*, 2020] to simply adopt prediction confidence as reliability indicator. Unlike previous work that sets a single fixed threshold for all classes, we design dynamic categorical

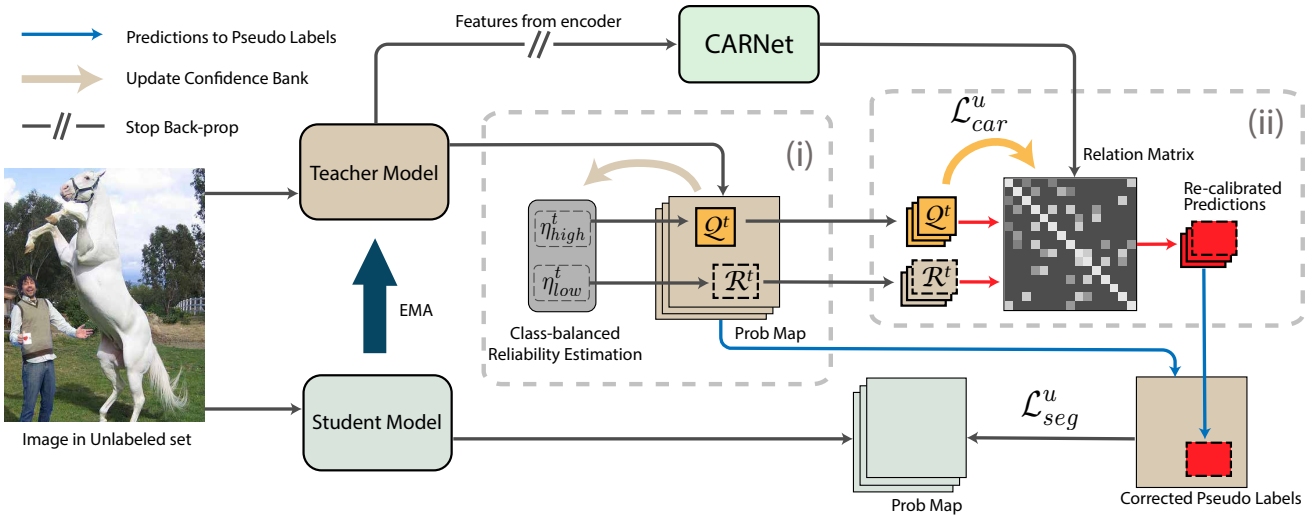

Figure 2: Overview of the proposed framework. The teacher model is the student model's exponential moving average (EMA). In (i), the confidence bank online locates potential noisy regions $\mathcal{R}^t$ and also the high confident predictions $\mathcal{Q}^t$ in a class-balanced way. In (ii), the relation matrix generated from CARNet bridges noisy regions with reliable predictions for label correction. Meanwhile, the mined reliable teacher predictions are transformed into supervision to improve CARNet. The student model is supervised with the updated pseudo labels.

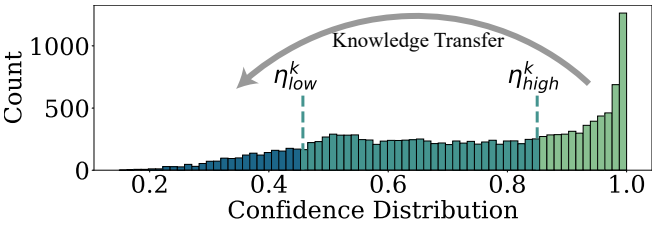

Figure 3: Confidence-based label quality measure for class k.

thresholds, considering the evolved model knowledge and the varied learning difficulties among classes. Class-wise confidence banks are built to online track the model performance on each class. For class k, a bank $\mathcal{N}^k$ collects the softmax probability from instances that are predicted as k on unlabeled data. The bank is updated with latest predictions during training as a queue and maintains a sufficient size to estimate the confidence distribution on class k. Then η^k_α, the α_{th} quantile of $\mathcal{N}^k$, can be used as the confidence threshold to determine whether a prediction is reliable. The per class threshold can guarantee class-balanced reliability estimation. By setting a low threshold η^k_{low} and a high threshold η^k_{high}, we can determine both confident and uncertain regions to conduct knowledge transfer as shown in Fig. 3.

3.3 Class Agnostic Feature Relation Mining

CARNet is proposed to measure the closeness of any two feature vectors, which serves for denoising. We exploit relation supervisions from the provided annotations and predicted pseudo labels to obtain accurate feature relation estimation. Reliable relation estimations can be obtained by performing distance calculations from its output feature maps.

Training CARNet from Annotations. In warm-up training, both segmentation model and CARNet are trained on provided annotations. To provide class-agnostic supervision, we directly transform a one-hot coded K class annotation map of size $H \times W \times K$ into a $HW \times HW$ relation matrix $\boldsymbol{G}$ as in Eq. (1).

$$\boldsymbol{G} = \boldsymbol{M}\boldsymbol{M}^T, \tag{1}$$

where each element $g_{ij} \in G$ indicates class equivalence between features of pixel i and j. Value 1 is given if they share the same class annotation and 0 otherwise. Under the binary supervision $\boldsymbol{G}$, we leverage cosine similarity between CARNet feature pair f_i and f_j to estimate feature relations:

$$a_{ij} = \cos(f_i, f_j). \tag{2}$$

Then a_{ij} is processed by a sigmoid function, outputting $\hat{a}_{ij}$, which is used in loss calculation. We gather positive feature pairs in set $\mathcal{Z}_p$ and negative pairs in $\mathcal{Z}_n$ as

$$\mathcal{Z}_p = \{(i,j)|g_{ij} = 1\}, \quad \mathcal{Z}_n = \{(i,j)|g_{ij} = 0\}. \tag{3}$$

The CARNet learns to minimize the cross-entropy loss $\mathcal{L}_{car}$ between predicted semantic relation $\hat{a}_{ij}$ and relation labels g_{ij}, as shown in Eq. (4):

$$\mathcal{L}_{car} = -\sum_{(i,j)\in\mathcal{Z}_p} \frac{\log \hat{a}_{ij}}{|\mathcal{Z}_p|} - \sum_{(i,j)\in\mathcal{Z}_n} \frac{\log(1-\hat{a}_{ij})}{|\mathcal{Z}_n|}. \tag{4}$$

Training CARNet from Mined Reliable Continuous Supervision. Supervision from annotations has brought CARNet decent performance. However, it is sub-optimal to stop training CARNet during training the segmentation model on unlabeled data. Since the CARNet is trained on features from segmentation encoder, it is beneficial to be continuously adapted to the segmentation encoder which evolves during training. More importantly, knowledge from segmentation model on unlabeled data can be transferred into relation supervision to expand training samples for CARNet. To mine

reliable relations, we adopt class-wise confidence bank from section 3.2 and set the threshold $\eta_{high}^{(t,k)}$ for class k at time t as the median value of current confidence bank $\mathcal{N}^{(t,k)}$:

$$\eta_{high}^{(t,k)} = \text{median}(\mathcal{N}^{(t,k)}). \tag{5}$$

During training, a categorical prediction on unlabeled pixel is considered reliable only if its class probability is over the current threshold $\eta_{high}^{(t,k)}$. With filtered predictions, two sets $\mathcal{Z}_p^u$ and $\mathcal{Z}_n^u$ are built the same way as in Eq. (3) to gather reliable prediction pairs on unlabeled pixels. The dynamically updated relation supervision is used to continuously supervise CARNet by minimizing $\mathcal{L}_{car}^u$:

$$\mathcal{L}_{car}^u = -\sum_{(i,j)\in\mathcal{Z}_p^u} \frac{\log \hat{a}_{ij}}{|\mathcal{Z}_p^u|} - \sum_{(i,j)\in\mathcal{Z}_n^u} \frac{\log(1-\hat{a}_{ij})}{|\mathcal{Z}_n^u|}. \tag{6}$$

3.4 Progressive Pseudo Label Online Denoising

We conduct online label corrections during training on unlabeled data. To perform stable and efficient denoising, we propose a strategy to progressively correct noisy regions in a class-balanced way while keeping high-quality labels as anchor supervision. By adopting prediction confidence as a noise indicator, the denoising starts from the least confident pseudo labels since they are most likely to be noise and mislead the self-training. The denoising is conducted in a reverse-confidence manner to firstly target on the most noisy labels and then expand target regions progressively towards those with higher confidence.

Class-balanced Denoising Curriculum. Based on the categorical confidence estimation from section 3.2, we design a class-balanced denoising curriculum for label correction. We map current training iteration t to a low threshold $\eta_{low}^{(t,k)} \leftarrow \eta_{g(t)}^{(t,k)}$ to locate noisy regions. $g(t)$ denotes the mapping, which is calculated as:

$$g(t) = \left\lceil \frac{b}{10} \times \frac{t}{T} \right\rceil \times 10, \tag{7}$$

where $\lceil . \rceil$ indicates ceil operation and T is total training iterations. b is the upper bound for quantile selection and ranges from 10 to 100. As training proceeds, the noise threshold $\eta_{low}^{(t,k)}$ starts from $\eta_{10}^{(t,k)}$ and finally reaches the up limit $\eta_b^{(t,k)}$.

Noisy Targets and Denoising Candidates. We define noisy regions in pseudo labels at step t as $\mathcal{R}^t$ in Eq. (8). Label of pixel i whose confidence is below current threshold $\eta_{low}^{(t,k)}$ needs to be rectified. Meanwhile, we mine confident predictions from current teacher model as shown in Fig. 2(i). The confident prediction set is defined as $\mathcal{Q}^t$ in Eq. (9). $\mathcal{Q}^t$ serves as references for label re-estimation of $\mathcal{R}^t$.

$$\mathcal{R}^t = \{i | p_i^{(t,k)} < \eta_{low}^{(t,k)},\ k \in \{1,...,K\}\}, \tag{8}$$

$$\mathcal{Q}^t = \{i | p_i^{(t,k)} > \eta_{high}^{(t,k)},\ k \in \{1,...,K\}\}. \tag{9}$$

Relation based Online Prediction Ensemble. As shown in Fig. 2(ii), labels of $\mathcal{R}^t$ are rectified with $\mathcal{Q}^t$ and CARNet. To correct a label at location $i \in \mathcal{R}^t$, we draw inter-feature similarities a_{ij}^t from CARNet to search for the semantic neighbors of the target. Then we turn similarities into normalized weights w_{ij}^t to determine contribution of each prediction at location j to correct pseudo labels at i:

$$w_{ij}^t = \frac{\exp(-(1-a_{ij}^t))}{\sum_{j'\in\mathcal{Q}^t}\exp(-(1-a_{ij'}^t))},\ \forall i \in \mathcal{R}^t. \tag{10}$$

Then the pseudo label prediction at location i is corrected by weighted fusion of all its updated semantic neighbours p_j^t:

$$\hat{p}_i^t = p_i^t + \sum_{j\in\mathcal{Q}^t} w_{ij}^t p_j^t,\ \forall i \in \mathcal{R}^t. \tag{11}$$

The corrected hard labels $\hat{y}_i^t$ are converted from the soft predictions as:

$$\hat{y}_i^t = \begin{cases} \underset{k}{\text{argmax}}\, \hat{p}_i^{(t,k)}, & \forall i \in \mathcal{R}^t \\ \underset{k}{\text{argmax}}\, p_i^{(t,k)}, & \text{else} \end{cases}. \tag{12}$$

The student model is then trained on the updated pseudo labels $\hat{y}_i^t$. From the view of consistency regularization, we enforce the student prediction to be consistent with both its temporal ensemble (teacher model EMA) and spatial ensemble (weighted fusion of teacher predictions).

3.5 Loss Functions

Warm up stage. Pixel-wise cross-entropy loss in Eq. (13) is applied on segmentation model, where y^a is segmentation label from annotation. $\mathcal{L}_{car}$ is used to supervise CARNet. The total loss $\mathcal{L}_w$ in warm-up stage is shown in Eq. (14):

$$\mathcal{L}_{seg} = \sum_{i=1}^{H\times W}\sum_{k=1}^{K} y_i^a \log(p_i^t), \tag{13}$$

$$\mathcal{L}_w = \mathcal{L}_{seg} + \mathcal{L}_{car}. \tag{14}$$

Self-training stage. We adopt a dynamic cross-entropy loss on outputs s_i from student model to further alleviate the effect from potential noise. The loss is formulated as

$$\mathcal{L}_{seg}^u = \sum_{i=1}^{H\times W}\sum_{k=1}^{K} w_i^t \hat{y}_i^t \log(s_i^t), \tag{15}$$

$$w_i^t = s_i^{(t,k)},\ k = \underset{k'}{\text{argmax}}\, s_i^{(t,k')}, \tag{16}$$

where w_i^t is the predicted probability to weight the sample loss. More attention is paid on confident samples while less weights are given on uncertain predictions. In this way, the model is updated in a conservative way, which leads to noise robust training. Combined with continuous supervision on CARNet, the overall loss function for unlabeled samples is

$$\mathcal{L}_u = \mathcal{L}_{seg}^u + \mathcal{L}_{car}^u. \tag{17}$$

4 Experiments

4.1 Model Implementation

Backbone Network. Since the majority of works report results on DeepLabv2 [Chen *et al.*, 2017], we adopt the same model as the base model. We follow the implementation from [Hung *et al.*, 2018] and use ResNet-101 as the backbone encoder.

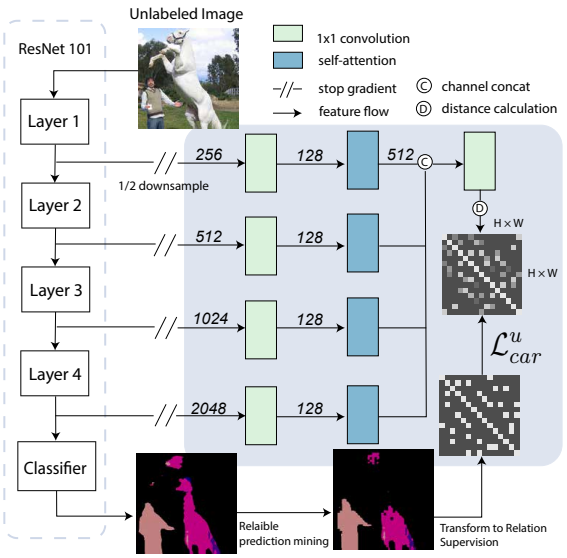

Figure 4: Architecture and training process of CARNet. The main body of the module is shown in the blue region.

CARNet Architecture. The architecture of CARNet is shown in Fig. 4. It takes multi-level feature maps from the segmentation encoder to capture comprehensive representations. The first level features are downsampled to half the original size to align with other levels. A 1×1 convolutional layer is applied, followed by a batchnorm and ReLU layer, to process level-wise features in parallel. Feature channels are all reduced to 128 before being fed into self-attention, where the feature representations are further enhanced. The enhanced multi-level feature is then concatenated on channel dimension and processed by a final 1×1 convolution layer.

4.2 Experimental Setup

Datasets. PASCAL VOC 2012 [Everingham *et al.*, 2010] with SBD annotations [Hariharan *et al.*, 2011] are commonly used benchmark for semantic segmentation. It has 20 foreground classes and a background class. The training set and validation set consist of 10,582 and 1,449 images. The images are randomly cropped to 321×321 for training. Cityscapes [Cordts *et al.*, 2016] is a dataset for urban scene understanding. It consists of 30 categories, while only 19 are used in the evaluation. The training set and validation set contain 2,975 and 500 finely annotated images. The original image resolution is 2048×1024. We resize it as 1024×512 and train the model on cropped 512×256 patches following [Hung *et al.*, 2018; French *et al.*, 2020].

Evaluation Protocols. For PASCAL VOC, 1/100, 1/50, 1/20, and 1/8 data splits are created. For Cityscapes, 1/30, 1/8, 1/4 images are randomly sampled from the dataset. The selected images are used as labeled data, while the rest are as an unlabeled set. Three folds are created for each data partition, and the final performance is the average on the three folds on the validation set. Mean intersection over union (mIoU) is adopted as the evaluation metric. We conduct single-scale evaluation without CRF post-processing.

Method	1/100	1/50	1/20	1/8	Oracle
Supervised	48.45	55.29	62.78	67.95	74.50
AdvSemiSeg	-	57.2	64.7	69.5	74.9
S4GAN+MLMT	-	63.3	67.2	71.4	75.6
DMT	63.04	67.15	69.92	72.70	74.75
ECS	-	-	-	72.95	-
ClassMix	54.18	66.15	67.77	71.00	-
Contra-SemiSeg	-	68.2	70.1	71.8	74.1
CARD (Ours)	**65.08**	**70.94**	**72.94**	**74.07**	74.50

Table 1: mIoU results (%) on PASCAL VOC 2012 *val* set. All the methods are based on DeepLabv2 with ResNet-101 backbone pretrained with COCO.

Method	1/30 (100)	1/8 (372)	1/4 (744)	Oracle
Supervised	46.85	59.05	62.02	66.82
AdvSemiSeg	-	57.1	60.5	66.2
S4GAN+MLMT	-	59.3	61.9	65.8
CutMix	51.20	60.34	63.87	67.68
ECS	-	60.26	63.77	-
DMT	54.80	63.03	-	68.16
ClassMix	54.07	61.35	63.63	-
C^3-SemiSeg	55.17	63.23	65.50	69.53
DARS	-	64.2	-	-
Contra-SemiSeg*	**58.0**	63.0	64.8	66.4
CARD (Ours)	55.63	**65.10**	**66.45**	66.82

Table 2: mIoU results (%) on Cityscapes *val* set. All the methods are based on DeepLabv2 with ResNet-101 backbone pretrained with ImageNet. * indicates model trained on higher resolution (512×512).

Implementation Details. For both PASCAL VOC and Cityscapes datasets, we use the stochastic gradient descent (SGD) optimizer with learning rate 0.001, weight decay 0.0005 and momentum 0.9. Following the common practice, we use 'poly' learning rate policy where the initial learning rate is scaled by $(1 - \text{iter}/\text{max\_iter})^{0.9}$. Among all semi-supervised protocols, the warm-up training on labeled sets lasts 10k iterations. We conduct single-stage self-training and train the model on unlabeled images for 40 epochs. The batch size is set to 8 for both training stages. The weight for updating the EMA teacher model is 0.999. For data augmentation, only random horizontal flip is applied. Upper bound b in Eq. (7) for denoising region is set as 60_{th} class-wise quantile.

4.3 Comparisons to State-of-the-art Methods

We compare the proposed method with different semi-supervised semantic segmentation methods on PASCAL VOC and Cityscapes, including AdvSemiSeg [Hung *et al.*, 2018], S4GAN+MLMT [Mittal *et al.*, 2019], DMT [Feng *et al.*, 2020], ECS [Mendel *et al.*, 2020], ClassMix [Olsson *et al.*, 2021], Contra-SemiSeg [Alonso *et al.*, 2021], CutMix [French *et al.*, 2020], C^3-SemiSeg [Zhou *et al.*, 2021] and DARS [He *et al.*, 2021]. For a fair comparison, we mainly show the results on the same backbone, *i.e.*, DeepLabv2 with ResNet-101. All results are reported by their original papers under the same settings.

Experiment	DL	RD	CAR	PRO	mIoU (%)
Supervised					56.25
ST					59.91
I	✓				63.13
II	✓	✓			57.90
III	✓	✓	✓		70.45
IV	✓	✓	✓	✓	**70.95**

Table 3: Ablation study on different components. ST: Plain self-training. DL: Dynamic self-weighted loss $\mathcal{L}_{seg}^{u}$ in Eq. (15). RD: Plain feature relation based denoising, with feature relations calculated from the segmentation encoder. CAR: Feature relation are calculated from CARNet. PRO: Progressive denoising strategy.

Results on PASCAL VOC. Table 1 presents the mIoU results on PASCAL VOC validation dataset with COCO [Lin *et al.*, 2014] pretrained weights. The proposed method improves over purely supervised baseline by 16.6%, 15.6%, 10.2%, and 6.1% under 1/100, 1/50, 1/20, and 1/8 data proportions, respectively. On 1/8 data partition, our method achieves comparable performance to fully supervised settings, with a small gap of 0.4%. Our method achieves state-of-the-art performances under all settings compared to the previous works. In particular, our method surpasses the prior best method Contra-SemiSeg by 2.7%, 2.8%, and 2.3% under 1/50, 1/20, and 1/8 protocols, respectively. With ImageNet [Deng *et al.*, 2009] pretrained weights, our method achieves similar performances and outperforms Contra-SemiSeg by 0.4%, 2.4%, and 2.2% under the same data partitions.

Results on Cityscapes. Table 2 shows the quantitative comparison of Cityscapes. Our method outperforms the supervised baseline by a large margin, with gains of 8.8%, 6.0%, and 4.4% under 1/30, 1/8, and 1/4 data proportions. In particular, our method nearly closes the performance gap between 1/4 data partition (66.45%) and oracle (66.82%).

4.4 Ablation Studies

To investigate the contribution of each component, we conduct ablation studies based on 1/50 data split in the PASCAL VOC dataset, and mIoU results are obtained on the validation set. We set two baselines to compare with: the model supervised only by labeled data (denoted as "Supervised" in Table 3) and the model trained in a plain self-training approach on initial pseudo labels (denoted as "ST" in Table 3).

Feature Relation based Denoising. We ablate the framework to manifest the effectiveness of the proposed label denoising approach. In Table 3, dynamic self-weighted loss $\mathcal{L}_{seg}^{u}$ in experiment I brings 3.2% gain over plain self-training. Upon that, in Experiment II, plain feature relation based denoising is introduced, with feature relations calculated from the segmentation encoder. The denoising approach fails, leading to a significant performance drop. The main reason is that inaccurate relation estimation fails to capture semantic connections. In Experiment III, the feature relation estimation from CARNet guides the denoising process to the right path, leading to a performance gain of 7.3% compared to Experiment I, which indicates that the CARNet plays a vital role in the relation-based denoising approach. Our progressive denoising strategy further improves the framework by 0.5% as shown in Experiment IV. During self-training, our method produces more accurate pseudo labels compared to the approach, as shown in Fig. 5.

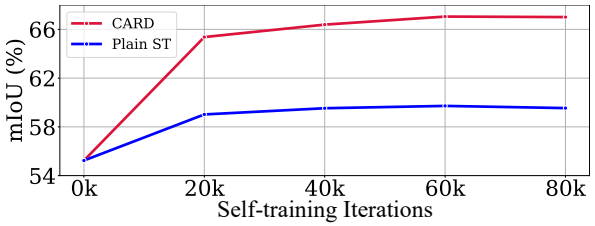

Figure 5: Quality comparison on pseudo labels generated by CARD and plain self-training.

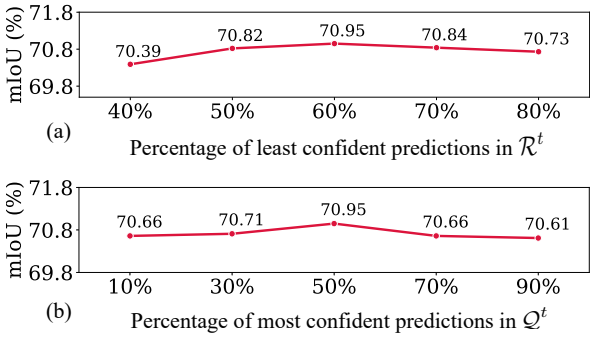

Figure 6: Ablation study on confidence thresholds. (a) Effect of denoising region $\mathcal{R}^t$. (b) Effect of $\mathcal{Q}^t$ to mine CARNet supervision.

Denoising Regions. Fig. 6(a) presents the effect of denoising region $\mathcal{R}^t$. When the denoising region increases to 60%, the performance peaks at 70.95% but sees a slight decrease when the denoising region expands further. By observation, the regions that demand correction most are the predictions below the 60% class-wise confidence threshold.

Mining Continuous CARNet Supervision. We mine relation supervision from filtered teacher outputs $\mathcal{Q}^t$ controlled by η_{high}^{t}. Fig. 6(b) shows the effect of threshold choices when generating supervision. The model performance increases when filtered teacher predictions increase from 10% to 50%. Top performance 70.95% is achieved when 50% of teacher predictions are leveraged in CARNet supervision. Overall, our method is insensitive to different confidence thresholds.

5 Conclusion

This paper has proposed a novel semi-supervised semantic segmentation framework based on effective pseudo label denoising. We present a label denoising routine by transferring confident label predictions to rectify the label for noisy regions, through reliable feature semantic connections. Class-agnostic relation network (CARNet) is designed to capture reliable feature connections, which is supervised by class equivalence. Extensive experiments demonstrate that our method outperforms the state-of-the-art methods by a significant margin, indicating the effectiveness of our method.

Acknowledgments

This work was suppoted by National Key R&D Program of China under grant SQ2021YFE020328 and National Science Foundation of China under grant 61972323.

References

[Alonso *et al.*, 2021] Iñigo 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.

[Arazo *et al.*, 2020] Eric Arazo, Diego Ortego, Paul Albert, Noel E O'Connor, and Kevin McGuinness. Pseudo-labeling and confirmation bias in deep semi-supervised learning. In *IJCNN*, 2020.

[Berthelot *et al.*, 2019] David Berthelot, Nicholas Carlini, Ian Goodfellow, Nicolas Papernot, Avital Oliver, and Colin Raffel. Mixmatch: A holistic approach to semi-supervised learning. In *NeurIPS*, 2019.

[Chen *et al.*, 2017] 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. *TPAMI*, 2017.

[Chen *et al.*, 2020] Liang-Chieh Chen, Raphael Gontijo Lopes, Bowen Cheng, Maxwell D Collins, Ekin D Cubuk, Barret Zoph, Hartwig Adam, and Jonathon Shlens. Naive-student: Leveraging semi-supervised learning in video sequences for urban scene segmentation. In *ECCV*, 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.

[Cordts *et al.*, 2016] 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.

[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*, 2009.

[Everingham *et al.*, 2010] Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *IJCV*, 2010.

[Feng *et al.*, 2020] Zhengyang Feng, Qianyu Zhou, Qiqi Gu, Xin Tan, Guangliang Cheng, Xuequan Lu, Jianping Shi, and Lizhuang Ma. Dmt: Dynamic mutual training for semi-supervised learning. *arXiv preprint arXiv:2004.08514*, 2020.

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

[Han *et al.*, 2018] Bo Han, Quanming Yao, Xingrui Yu, Gang Niu, Miao Xu, Weihua Hu, Ivor Tsang, and Masashi Sugiyama. Co-teaching: Robust training of deep neural networks with extremely noisy labels. In *NeurIPS*, 2018.

[Hariharan *et al.*, 2011] Bharath Hariharan, Pablo Arbeláez, Lubomir Bourdev, Subhransu Maji, and Jitendra Malik. Semantic contours from inverse detectors. In *ICCV*, 2011.

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

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

[Ke *et al.*, 2019] 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 *ICCV*, 2019.

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

[Lin *et al.*, 2014] 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.

[Luo *et al.*, 2018] Yucen Luo, Jun Zhu, Mengxi Li, Yong Ren, and Bo Zhang. Smooth neighbors on teacher graphs for semi-supervised learning. In *CVPR*, 2018.

[Mendel *et al.*, 2020] Robert Mendel, Luis Antonio De Souza, David Rauber, João Paulo Papa, and Christoph Palm. Semi-supervised segmentation based on error-correcting supervision. In *ECCV*, 2020.

[Mittal *et al.*, 2019] Sudhanshu Mittal, Maxim Tatarchenko, and Thomas Brox. Semi-supervised semantic segmentation with high-and low-level consistency. *TPAMI*, 2019.

[Olsson *et al.*, 2021] Viktor Olsson, Wilhelm Tranheden, Juliano Pinto, and Lennart Svensson. Classmix: Segmentation-based data augmentation for semi-supervised learning. In *WACV*, 2021.

[Sohn *et al.*, 2020] Kihyuk Sohn, David Berthelot, Chun-Liang Li, Zizhao Zhang, Nicholas Carlini, Ekin D Cubuk, Alex Kurakin, Han Zhang, and Colin Raffel. Fixmatch: Simplifying semi-supervised learning with consistency and confidence. In *NeurIPS*, 2020.

[Zhang *et al.*, 2021] Bowen Zhang, Yidong Wang, Wenxin Hou, Hao Wu, Jindong Wang, Manabu Okumura, and Takahiro Shinozaki. Flexmatch: Boosting semi-supervised learning with curriculum pseudo labeling. In *NeurIPS*, 2021.

[Zhou *et al.*, 2021] 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 *ICCV*, 2021.