RWMS: Reliable Weighted Multi-Phase for Semi-supervised Segmentation

Wensi Liu, Xiao-Yu Tang*, Chong Yang, Chunjie Yang

College of Control Science and Engineering, Zhejiang University
{ws.liu, xytang, yangchong2020, cjyang999}@zju.edu.cn

Abstract

Semantic segmentation is one of the tasks concerned in the field of computer vision. However, the cost of capturing large numbers of pixel-level annotations is expensive. Semi-supervised learning can utilize labeled and unlabeled data, providing new ideas for solving the problem of insufficient labeled data. In this work, we propose a data-reliability weighted multi-phase learning method for semi-supervised segmentation (RWMS). Under the framework of self-training, we train two different teacher models to evaluate the reliability of pseudo labels. By selecting reliable data at the image level and reweighting pseudo labels at the pixel level, multi- phase training is guided to focus on more reliable knowledge. Besides, we also inject strong data augmentations on unlabeled images while training. Through extensive experiments, we demonstrate that our method performs remarkably well compared to baseline methods and substantially outperforms them, more than 3% on VOC and Cityscapes.

Introduction

With the rapid development of deep learning, semantic segmentation plays a crucial role in computer vision. However, fully supervised semantic segmentation methods typically require a large amount of labeled data to achieve accurate segmentation, which limits their feasibility in practical applications. In comparison, semi-supervised semantic segmentation methods leverage unlabeled data to assist model training, thereby improving segmentation performance while reducing annotation costs (Peláez-Vegas, Mesejo, and Luengo 2023). Prior researches propose to apply consistency regularization (Laine and Aila 2017; Olsson et al. 2021) or entropy minimization (Xie et al. 2020; Ke et al. 2022)in semi-supervised learning. Recently, the combination of the aforementioned two methods, such as FixMatch (Sohn et al. 2020), has shown promising performance. Ke et al. propose a three-stage self-training framework that incorporates consistency regularization into the self-training process. In general, our approach is a hybrid method based on classical self-training (Lee et al. 2013) that combines strong data augmentation as a form of consistency regularization.

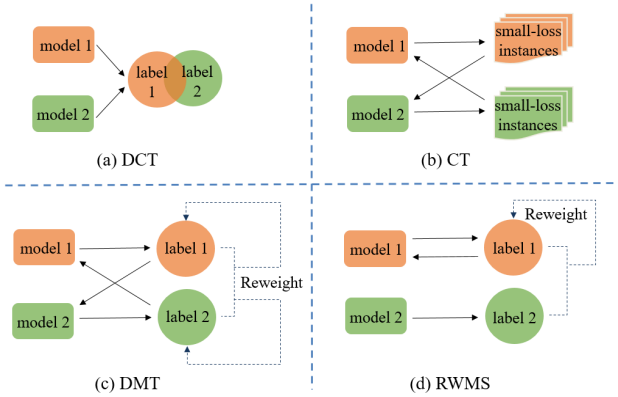


Figure 1: Frameworks for previous methods and our RWMS. We use an auxiliary model (model 2) to help make reliability judgments. Reliability is incorporated to guide the process of retraining and reweighting. More details are illustrated in the introduction section.

A difficulty faced by self-training is the noise in pseudo labels, which can contribute incorrect gradient messages to training. Some methods are used to improve the quality of pseudo labels, such as refining pseudo labels by auxiliary network (Li and Zheng 2021), screening according to the evolving stability of produced pseudo masks during the supervised training phase (Yang et al. 2022), distribution alignment (He, Yang, and Qi 2021), filtering low-confidence labels by pre-defined threshold (Zoph et al. 2020) and so on.

In fact, a single model has limited ability to correct its own errors. On the contrary, the information provided by the two different models complements each other, which can assist in better selection of high-quality pseudo labels. DCT (Qiao et al. 2018) employs diverse models and learns by maximizing their consensus on unlabeled data (Figure 1(a)). Both models in CT (Han et al. 2018) choose small-loss examples for the other model to train (Figure 1(b)). In DMT (Feng et al. 2022), the loss function is dynamically reweighted according to the disagreement between two models and each model is trained with pseudo labels generated by the other (Figure 1(c)).These mutual training methods, despite their effectiveness, necessitate the maintenance of training for two models, thereby escalating the computational resource

*Corresponding author.

requirements. Consequently, additional time and hardware resources may be necessary to complete the training process, resulting in increased costs and complexity.

Is there a way to fully integrate the knowledge of two models and become a strong competitor by only maintaining the training of the main model? In this work, we propose comparing the pseudo labels generated by two models to obtain reliability at both the image and pixel levels. We then combine this reliability information to perform multi-phase training of the main model (Figure 1(d)). Specifically, we describe a data-reliability guided multi-phase learning method for semi-supervised segmentation. From the perspective of model ensemble, there is a low probability of simultaneous prediction errors among different models. When two models produce consistent predictions for a pixel or image, it is highly likely that the prediction is correct.

First, we instantiate two teacher models. Among them, the auxiliary model is only used to judge the data reliability, and the subsequent training is carried out on the main model. At the image level, the Mean Intersection-over-Union (MIoU) of pseudo labels generated by two models on unlabeled data is used as a reliability metric. Images are prioritized based on this metric to guide multi-phase training. At the pixel level, a reweighting strategy is defined based on the confidence scores of the two pseudo labels. This strategy is applied in multi-phase training, allowing the model to pay more attention to regions with higher reliability. During multi-phase training, we also apply strong data augmentation to the unlabeled data which is a simple yet effective method.

In sum, our contributions are three-fold:

- In order to address the issue of noise in pseudo labels, we integrate the knowledge of two models to assess the quality of pseudo labels. When both models make consistent predictions, there is a high likelihood that those predictions are correct.
- We propose a data-reliability weighted multi-phase learning method (RWMS). At the image level, we guide the multi-phase learning based on reliability and focus on images of different difficulty levels in each stage. At the pixel level, we reweight the pixels based on their scores, enabling the model to pay more attention to regions with higher reliability.
- Extensive experiments and analyses have demonstrated the superiority of our method, e.g., we achieve 74.9% accuracy on Pascal VOC (Everingham et al. 2015) benchmark.

Related Work

Semantic Segmentation Model Architecture

Semantic segmentation is a vision task that classifies images at the pixel level (Mo et al. 2022). With the development of deep learning methods, the fully convolutional network (FCN) (Long, Shelhamer, and Darrell 2015) is applied to this field, which significantly outperforms traditional methods. At present, semantic segmentation methods mainly focus on multi-scale and pyramid network based models (Chen et al. 2018; Zhao et al. 2017), attention mechanisms (Li et al. 2019; Fu et al. 2019), encoder-decoder schemes (Badrinarayanan, Kendall, and Cipolla 2017; Ronneberger, Fischer, and Brox 2015) and transformer (Wang et al. 2021; Xie et al. 2021), etc. In this paper, we conduct experiments mainly using DeepLabv3+ (Chen et al. 2018), DeepLabv2 (Chen et al. 2017) and PSPNet (Zhao et al. 2017).

Semi-supervised Learning

Semi-supervised learning is a label-efficient task that aims to leverage a vast amount of unlabeled data to enhance model performance while minimizing the reliance on labeled data. It primarily consists of two paradigms: entropy minimization (Grandvalet and Bengio 2004; Xie et al. 2020; Cascante-Bonilla et al. 2021; Ke et al. 2022) and consistency regularization (Laine and Aila 2017; Tarvainen and Valpola 2017; Olsson et al. 2021). On one hand, self-training serves as a representative approach in entropy minimization, where the training data is expanded by generating pseudo labels for unlabeled data, thereby improving model's performance. How to enhance the quality of pseudo labels becomes a key challenge in this field. On the other hand, consistency regularization encourages the model to produce consistent and stable predictions when subjected to perturbations. Various types of perturbations, such as input, feature, and network perturbations, are commonly employed (Peláez-Vegas, Mesejo, and Luengo 2023).

Semi-supervised Semantic Segmentation

The primary goal of semantic segmentation is to classify every pixel in an image into specific classes or categories. However, training models for such dense prediction tasks heavily relies on extensive data and laborious manual annotations. Semi-supervised semantic segmentation is proposed to remedy this issue by leveraging the information from unlabeled data to boost segmentation performance, effectively mitigating annotation costs. Adversarial learning(Hung et al. 2018; Lei et al. 2022), consistency training(Fan et al. 2022; Yang et al. 2023) and entropy minimization (Ke et al. 2022; Kwon and Kwak 2022) are three main branches. For instance, AdvSemiSeg (Hung et al. 2018) introduces an FCN-based discriminator to distinguish ground truth mask from network predictions. DCC (Lai et al. 2021) proposes the maintenance of context-aware consistency as a means to achieve feature alignment. PS-MT (Liu et al. 2022b) extends the Mean Teacher model by introducing a novel feature perturbation method called T-VAT, significantly enhancing student model's learning efficiency. PRCL (Xie et al. 2023b) proposes to model latent representation as a multivariate Gaussian distribution, which alleviate the negative effects from inaccurate pseudo labels. Additionally, some works based on self-training have also demonstrated powerful performance in this field. Specially, ST++ (Yang et al. 2022) incorporates strong data augmentation techniques for unlabeled data and employs image-level uncertainty to guide the multi-stage self-training process. DMT (Feng et al. 2022) introduces a dynamically reweighted loss function, greatly alleviating the pseudo label noise during training. To further improve the distribution of embedding space, some recent proposed works (Liu et al. 2022a; Wang et al. 2022; Xie

et al. 2023a) also attempt to apply regularization constraints to latent features.

Method

Self-training Pipeline

Semi-supervised semantic segmentation aims to learn from both labeled dataset D^l and unlabeled dataset D^u. The optimization objective can be expressed as follows:

$$\mathcal{L} = \mathcal{L}^s + \lambda \mathcal{L}^u \tag{1}$$

where λ is a coefficient that adjusts the balance between the supervised and unsupervised losses. The commonly used loss function in this context is the cross-entropy loss, computed as:

$$CE(y, p) = - \sum_{c=1}^{class\_num} y_c \lg p_c \tag{2}$$

where $class\_num$ represents the total number of classes, and y, p denote the one-hot label and confidence label, respectively.

For labeled dataset D^l, the cross-entropy loss is computed between the predicted values and the ground truth values. When calculating the unsupervised loss for self-training, pseudo labels are employed instead of the unavailable ground truth labels. The general steps of self-training are as follows:

1. **Initialization:** The model is initialized and trained on labeled data D^l to establish a baseline.

2. **Pseudo label generation:** The trained model (teacher model) is used to predict labels for unlabeled data D^u, generating pseudo labels.

3. **Semi-supervised training:** Both labeled data D^l and pseudo labeled data are combined for retraining. The model (student model) is optimized based on the supervised loss $\mathcal{L}^s$ and the unsupervised loss $\mathcal{L}^u$.

Since self-training relies heavily on the model's predictions for unlabeled data, any inaccuracies in these pseudo labels can lead to the model overfitting to noise, thereby compromising its performance. Secondly, the coupling of predictions between the student and teacher models poses a challenge. The predictions of the student model might become excessively swayed by those of the teacher model. If the teacher model makes incorrect predictions, the student model may follow suit, leading to misguided learning directions.

To address these limitations, the rationale for applying strong data augmentation to the student model becomes evident. It should be noted that strong data augmentation is a form of consistency regularization. When applying it, we are assuming the fulfillment of both the smoothness assumption and the cluster assumption: data points with distinct labels are segregated within regions of low density, while resembling data points exhibit comparable outcomes. Therefore, if an actual perturbation is applied to an unlabeled data point, its predicted outcome should not undergo significant changes, meaning that the outputs remain consistent.

Strong data augmentation, by introducing diverse variations to the training data, provides the model with a wider array of training instances. This aids in mitigating the risk of overfitting caused by noisy pseudo labels. Additionally, strong data augmentation empowers the student model to make more independent predictions, reducing the impact of prediction coupling with the teacher model. By fortifying the student model's generalization and robustness, it becomes better equipped to handle noisy labels and uncertainties, ultimately elevating the overall effectiveness of self-training. Colorjitter, grayscale, and blur (Chen et al. 2020; Yang et al. 2022) are commonly uesd in this area.

Multi-phase Learning

The classical self-training framework, which involves the simultaneous utilization of all available unlabeled images, encounters a challenge due to the inherent variability in the reliability of individual images and their corresponding pseudo labels. To address this issue, a more logical and effective strategy emerges: one that quantifies these discrepancies and integrates them into the training process through diverse approaches. In response, we introduce a selective multi-phase training paradigm, ingeniously designed to harness data progressively from the least intricate to the more complex instances, guided by the reliability-based ranking of the unlabeled images (Figure 2). Specifically, the reliability measurement of unlabeled images is based on the image-level consistency of pseudo labels generated by two teacher models.

Under this framework, the top 50% of images, prioritized according to their reliability, advance into retraining stage 1. It is within this stage that a superior student model 1 is cultivated, showcasing heightened performance capabilities. Student model 1 can help generate higher-quality pseudo labels for the remaining less reliable images. In retraining stage 2, a comprehensive integration of all images ensues, fostering a training environment enriched by an expanded dataset and bolstered by higher-caliber pseudo labels. Consequently, this synergy leads to the emergence of a remarkably adept student model 2.

Additionally, in the multi-phase training, we introduce two additional elements. First, strong data augmentation techniques are infused, amplifying the diversity and comprehensiveness of the training data. Secondly, a pixel-reliability-based reweighting strategy is introduced, which further refines the learning process by assigning varying levels of importance to individual pixels based on their reliability. The overall process is in Algorithm 1.

Reliability Assessment

In this chapter, we explain the definition of reliability at two levels: image and pixel. It should be noted that reliability here refers to the reliability of the image/pixel and its pseudo labels, which we abbreviate as image/pixel reliability.

Image reliability. At the image level, we need a metric to distinguish the difficulty of images, allowing the model to focus on disparate levels of complexity during each stage of training. Through the utilization of two trained teacher models, each unlabeled image becomes endowed with a pair of pseudo labels. The degree of consensus between these two

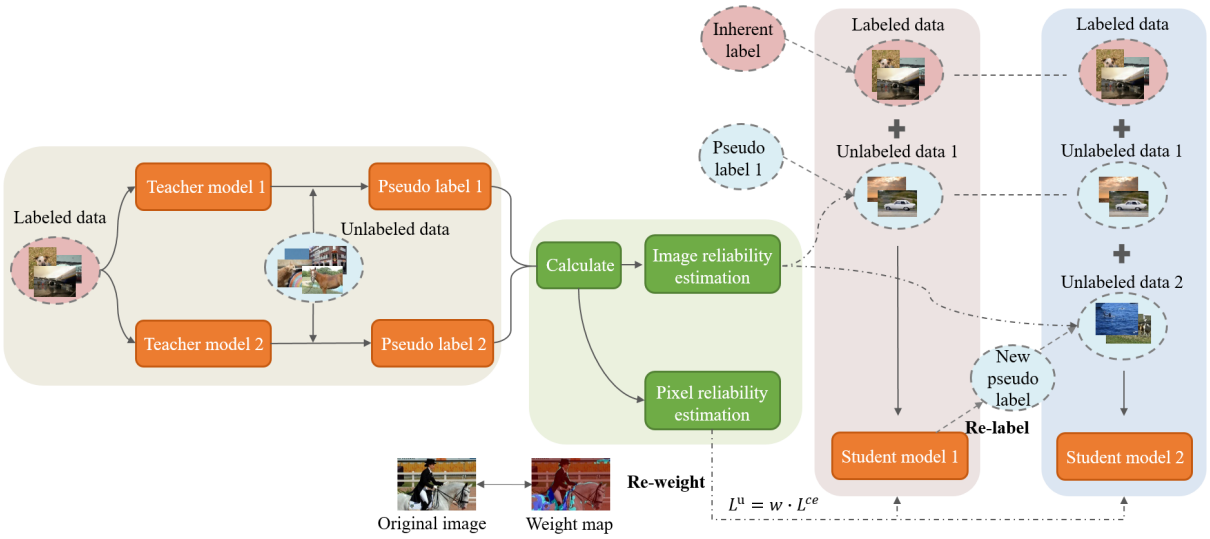


Figure 2: Overview of RWMS. Two trained teacher models generate pseudo labels on unlabeled data, and the reliability is calculated by comparing the two pseudo labels at the image level and pixel level. Image-level reliability is used to guide multi-stage training, and pixel-level reliability is used to incorporate reweighting in training.

Algorithm 1: Pseudocode

Input: Labeled training data set $D^l = \{X, Y\}$
Unlabeled training data set $D^u = \{U\}$
Teacher model T_1/T_2
Output: Trained student model S_2

1: Train T_1 and T_2 separately with D^l
2: Obtain pseudo labels $T_1(U)$ and $T_2(U)$
3: Calculate MIoU between two pseudo labels for each $ui \in U$
4: Select the top 50% reliable images (U_1) and the last 50% unreliable images (U_2) according to MIoU
5: Calculate weight graph based on two pseudo labels for each $ui \in U$
6: Let $D^{u1} = \{U_1, T_1(U_1)\}$
7: Retraining 1: train S_1 on $D^l \cup D^{u1}$ with weight graphs
8: Let $D^{u2} = \{U_2, S_1(U_2)\}$
9: Retraining 2: train S_2 on $D^l \cup D^{u1} \cup D^{u2}$ with weight graphs

Return S_2

pseudo labels serves as an indicator of label reliability. To quantify this consistency, we opt for the MIoU as evaluation metric. A higher MIoU value signifies an increased likelihood of overall correct predictions for pseudo label 1 or pseudo label 2, which implies higher reliability of the pseudo labels and suggests that the image is easier or more reliable. During the early stage of retraining, the strategic selection of more reliable images for training can progressively enhance the model's performance and alleviate the issue of overfitting caused by label noise. The unreliable images are reattributed with pseudo labels by student model 1, and these pseudo labels exhibit a heightened level of reliability. In the later stage of retraining, images of varying levels of difficulty are all engaged in the training process. Guided by the influence of higher-quality pseudo labels, the training of the student model is enhanced, leading to an improved student model 2.

Pixel reliability. Through the utilization of the training framework mentioned above, we possess the capability to partially alleviate the problem of noise interference by filtering out difficult images during the initial training stage. However, it is imperative to acknowledge that within even the most uncomplicated images, the existence of pixels or regions lacking in reliability presents a formidable challenge, one that has the potential to disrupt the model's training trajectory. Thus, solely conducting an evaluation of reliability at the level of the entire image is an inadequate approach. In light of this limitation, we put forth a strategy grounded in the concept of pixel reliability for reweighting. To elaborate, this framework is predicated on the assignment of an individual reliability score to each constituent pixel. By virtue of this reweighting scheme, the influence of pixels marked by their unreliability is effectively curtailed, concurrently amplifying the significance attributed to zones with heightened reliability.

The process of getting predicted results for unlabeled images through two teacher models can be represented as follows:

$$P_k = T_k(U) \rightarrow Y_k \quad k = 1, 2 \tag{3}$$

where $P_1(P_2)$ represents the segmentation confidence map, which is obtained by applying softmax normalization to the network output. $Y_1(Y_2)$ is the predicted one-hot label map, referred to as pseudo label 1 (pseudo label 2) in the previous context. For pixel point i, $p_{1i}(p_{2i})$ represents the confidence vector, while $y_{1i}(y_{2i})$ represents the corresponding one-hot vector. The dimensionality of the vector is equal to

the number of classes in the dataset being processed (denoted as $class\_num$).

The channel index of the non-zero value in y_{1i} is denoted as $class_{1i}$, which represents the class predicted by teacher model 1 for pixel i. Assuming the value at position $class_{1i}$ in p_{2i} is ranked as the x^{th} position in the descending order of all channel values. Then, at pixel point i, the weight for pseudo label 1 is determined as follows:

$$w_i = ae^{-b(x-1)} \tag{4}$$

where a and b are predefined parameters, and x takes values from 1 to $class\_num$. At the pixel level, when the class predicted by teacher model 1 with the maximum channel value (i.e., $class_{1i}$), ranks higher in the predictions of teacher model 2, the predictions of the two models indicate greater consistency. This implies a higher likelihood of y_{1i} being correct at that pixel position. Consequently, the pixel point becomes more predictable and should be assigned a higher weight.

By using the aforementioned method, a weight value is computed for each pixel of unlabeled images, so that each unlabeled image can get a corresponding weight graph $w^{H\times W}$ (H for height and W for width). Incorporating these weight maps into the training of the student models enables the models to focus on more certain knowledge.

The loss function for unlabeled images ($\mathcal{L}^u$) can be formalized as follows:

$$\mathcal{L}^u = \begin{cases} \frac{1}{N}\sum w_i CE(T_1(i), S(i)), & i \in U1 \\ \\ \frac{1}{N}\sum w_i CE(S_1(i), S(i)), & i \in U2 \end{cases} \tag{5}$$

where for the top 50% reliable images (U_1), the pseudo labels are generated by teacher model 1, and are no longer updated. For the last 50% unreliable images (U_2), the pseudo labels are generated by the student model 1 (S_1). During the retraining phase, the predicted outputs of the training student model (S) are used to calculate the cross-entropy loss with the pseudo labels, and they are assigned different weights. This, along with the supervised loss $\mathcal{L}^s$, guides the model's training.

Experiments

Dataset. Pascal VOC 2012 (Everingham et al. 2015) is an important standard semantic segmentation dataset, initially consisting of 1464 images for training and 1449 images for validation. In order to increase the number of training samples, previous researchers adopt a method of introducing relatively lower-quality annotations from the SBD dataset (Hariharan et al. 2011), forming an augmented training set with a total of 10582 images. Cityscapes (Cordts et al. 2016) is a genuine dataset of urban environments. The training and validation subsets consist of 2975 and 500 images, respectively.

Network architectures. We evaluate two network structures for the main teacher model: DeepLabv3+ with ResNet50/101. The auxiliary teacher model remains PSPNet with ResNet50. The student model maintains the same framework as the main model.

Implementation details. We initialize the backbone network by using pre-trained weights on ImageNet (Russakovsky et al. 2015). Training is performed using the SGD optimizer with the momentum of 0.9 and the weight decay of 0.0001. We employ a polynomial learning rate decay strategy, as used in previous works: $(1-\frac{iter}{max\_iter})^{0.9}$ (Chen et al. 2021; Yang et al. 2022; Liu et al. 2022b). The training is conducted for 80 epochs on both datasets.

Images are cropped to 321x321 on Pascal and 721x721 on Cityscapes. For labeled images, we apply weak data augmentation, including random scaling and random flipping. For unlabeled images, we additionally apply strong data augmentation, such as colorjitter, grayscale, and blur. When performing positional augmentation on the images, the corresponding weight maps are also adjusted accordingly. Since we divide the retraining process into two stages, we consider the top 50% of images with the higher reliability scores as reliable images and treat the rest as unreliable images.

Comparison with Existing Methods

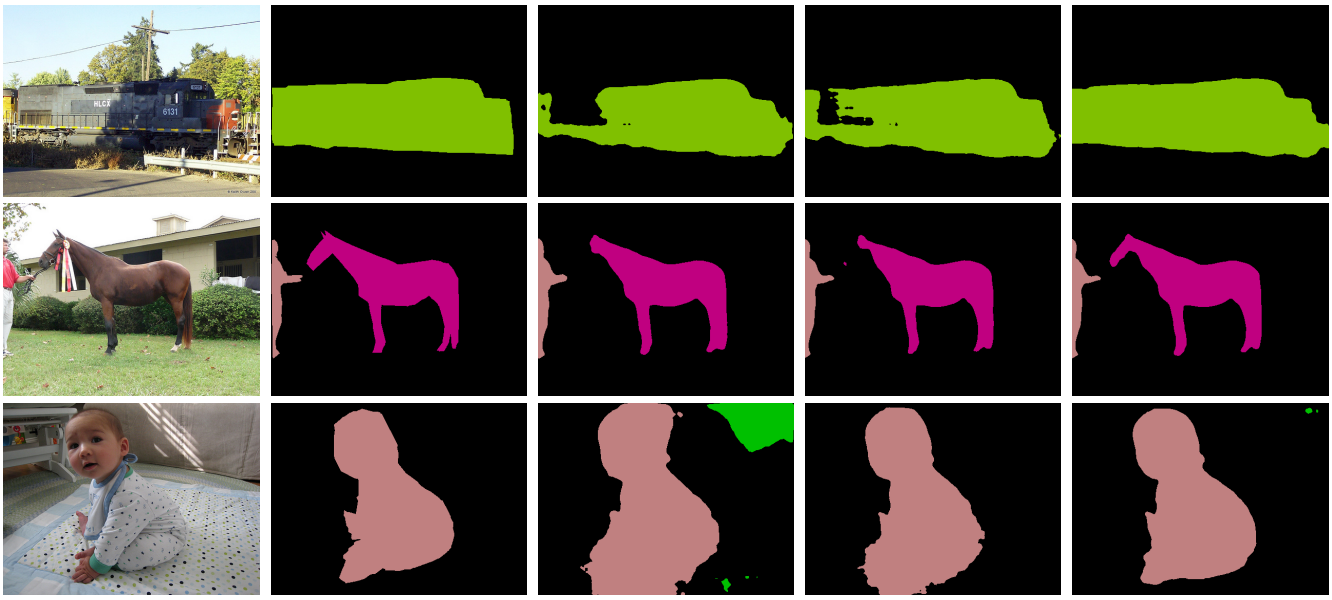
We show the improvements of our method compared with others under various partition protocols on the Pascal VOC 2012 in Table 1. We first discuss the case of fully supervised learning on labeled data only (i.e., SupOnly). The lower MIoU in this case also justifies the need for semi-supervised learning. Secondly, we choose traditional self-training as the baseline model, characterized by single-stage retraining and operations without strong data augmentation. Specifically, when using ResNet50, our method improves by 3.4%, 3.0% and 2.7% compared to the baseline method under the 1/16, 1/8 and 1/4 partition protocols, respectively. In the case of less labeled data, the performance improvement is more obvious. Visual results can be seen in Figure 3. Besides that, we compare our method with some recent semi-supervised segmentation methods, including: ECS (Mendel et al. 2020), GCT (Ke et al. 2020), DCC (Lai et al. 2021), ELN (Kwon and Kwak 2022) and SOSA (Huang and Zhang 2023). As shown in Table 1, our proposed strategy significantly outperforms existing methods in different network frameworks. Moreover, in Table 2, the experimental results on Cityscapes further demonstrate the generalization ability of our method, compared with DMT (Feng et al. 2022), CutMix (French et al. 2020), ClassMix (Olsson et al. 2021), etc.

Ablation Studies

Ablation studies are performed using DeepLabv3+ with ResNet50 on Pascal VOC 2012. We opt for PSPNet with ResNet50 as the supplementary teacher model architecture and 1323 (1/8) labeled images for supervised training, unless changing the model architecture or partition ratio as an experimental parameter.

Effectiveness of Multi-phase Retraining Strategy

In Table 3, we investigate the impact of multi-phase retraining. As a point of comparison, we randomly divide all un-



(a) Input Images (b) Ground Truth (c) SupOnly (d) Baseline (e) RWMS

Figure 3: Qualitative comparison from Pascal VOC 2012.

Main Network	Method	1/16	1/8	1/4
DeepLabv3+ ResNet50	SupOnly	64.9	69.2	70.2
	Baseline	68.8	71.9	72.4
	ECS	-	70.2	72.6
	DCC	70.1	72.4	74.0
	ELN	-	73.2	74.6
	Ours	**72.2**	**74.9**	**75.1**
DeepLabv3+ ResNet101	SupOnly	67.8	71.5	73.4
	Baseline	71.9	74.2	75.4
	GCT	67.2	72.5	75.1
	DCC	72.4	74.6	76.3
	ELN	-	75.1	76.6
	SOSA	-	74.5	75.7
	Ours	**74.0**	**75.7**	**76.8**

Table 1: Comparison on Pascal VOC 2012. Best results are in bold.

Main Network	Method	1/30	1/8	1/4
DeepLabv3+ ResNet101	DMT	54.8	63.0	-
	CutMix	55.7	65.8	68.3
	ClassMix	-	61.4	63.6
DeepLabv3+ ResNet50	SupOnly	55.1	65.8	68.4
	Baseline	58.1	68.7	69.3
	ECS	-	67.4	70.7
	DCC	-	69.7	72.7
	Ours	**60.8**	**71.9**	**73.2**

Table 2: Comparison on Cityscapes.

Method	1/16	1/8	1/4
One-phase retraining	69.9	72.1	73.0
Random multi-phase retraining	70.0	72.3	72.8
Selective multi-phase retraining (Ours)	**72.2**	**74.9**	**75.1**

Table 3: Effectiveness of selective multi-phase retraining. Random multi-phase retraining cannot improve performance steadily. Our approach benefits from a carefully designed selection strategy.

labeled images into two groups and sequentially incorporate them in the multi-phase training process outlined by our method. The results gathered in Table 3 reflect the following observations: (1) Increasing the number of training stages arbitrarily does not inherently lead to improved results; in certain scenarios, it may even yield outcomes worse than single-phase retraining. (2) Our multi-phase training based on image reliability consistently outperforms the corresponding single-phase retraining. (3) The improvement in the effectiveness of our method comes from the correct classification of reliable and unreliable images, rather than simply increasing the number of retraining stages.

We further undertake a quantitative assessment to rigorously quantify the differentials between the sets of reliable and unreliable images. As depicted in Figure 4, the MIoU of the two pseudo labels on the reliable image set is significantly higher than that on the unreliable image set. This conspicuous discrepancy serves as compelling evidence that our method is proficient in effectively discerning images with differential complexities, and the division based on reliability is rational.

Analysis of Teacher Model 2 Selection

In our proposed methodology, teacher model 1 assumes the role of the primary model. In the context of our experimental setup, it remains consistent as Deeplabv3+ with ResNet50.

Teacher model 2	PSPNet with ResNet50	DeepLabv2 with ResNet101	DeepLabv3+ with ResNet101
Difference	0.66	0.69	0.71
MIoU	**74.9**	74.3	74.1

Table 4: Analysis of the difference between teacher models. The representation of the difference is the MIoU of the pseudo labels produced by the two teacher models.

Method	1/16	1/8	1/4
W/o reweighting	71.6	73.8	74.7
Reweighting (Ours)	**72.2**	**74.9**	**75.1**

Table 5: Effectiveness of reweighting strategy compared with ignoring pixel reliability.

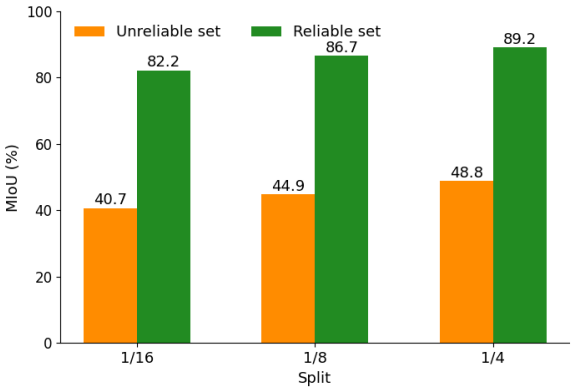


Figure 4: The reliability quality comparison of the reliable and unreliable images.

As for the selection of teacher model 2, it exhibits variability. In Table 4, we discuss the impact of differences between teacher model 1 and 2. For each unlabeled image, we first calculate the MIoU of its two pseudo labels and then compute the overall average for all images. Lower values indicate more pronounced discrepancies between the teacher models. We find that the semi-supervised learning benefits from dissimilarities between the models. On one hand, the diverse knowledge acquired from two markedly distinct models introduces a wealth of learning information. On the other hand, the likelihood of disparate models confidently committing identical errors remains low, thereby diminishing the influence of label noise on learning.

Effectiveness of Reweighting Strategy

We further investigate the impact of reweighting. We compare our method with the case where no reweighting is performed. Both scenarios adopt multi-phase training and the results are listed in Table 5. We can observe that relying solely on image reliability handling is insufficient. The weight graphs in Figure 5 measure the consistency of pseudo labels assigned by two teacher models to the same image. For regions where the decisions are inconsistent, we find it challenging to ensure the reliability of the pseudo labels, especially when pseudo label 1 is supposed to provide supervision signals for retraining. Since we do not have enough confidence in guaranteeing the reliability of pseudo label 1, reducing attention to these regions is a relatively prudent strategy. Through this approach, our model can focus more on specific areas and reduce attention to potential label noise.

Figure 5: Examples of reweighting. The left column is the original images, and the right column is the heat maps after reweighting. Blue regions have greater unreliability, and these regions are assigned with lower weights.

As shown in Figure 5, the disagreements between the teacher models are mainly concentrated at the boundaries of objects. In fact, these border regions often present a more intricate challenge in terms of classification when compared to the interior of the objects. This observation aligns coherently with our intuitive cognitive expectations.

Conclusion

In this work, we propose a data reliability weighted multi-phase semi-supervised segmentation method, named RWMS. A structure of dual-teacher models is proposed to generate reliable evaluation metrics. At the image level, we use image reliability ranking to guide the multi-phase training, allowing each stage to focus on more reliable images. At the pixel level, we reweight the pixels based on their reliability, enabling the model to concentrate on more confident knowledge. Benefiting from the refined grading of reliability, we have new criteria for selecting the quality of pseudo labels, allowing the model to focus its attention on knowledge with lower noise. A series of experiments have indicated the superiority of our approach.

Acknowledgments

This work was supported by the National Nature Science Foundation of China (Grant No. 62103365), Zhejiang Provincial Natural Science Foundation, China (No. LQ22F030002) and the Fundamental Research Funds for the Central Universities (Zhejiang University NGICS Platform).

References

Badrinarayanan, V.; Kendall, A.; and Cipolla, R. 2017. Segnet: A deep convolutional encoder-decoder architecture for image segmentation. *IEEE transactions on pattern analysis and machine intelligence*, 39(12): 2481–2495.

Cascante-Bonilla, P.; Tan, F.; Qi, Y.; and Ordonez, V. 2021. Curriculum labeling: Revisiting pseudo-labeling for semi-supervised learning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, 6912–6920.

Chen, L.-C.; Papandreou, G.; Kokkinos, I.; Murphy, K.; and Yuille, A. L. 2017. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE transactions on pattern analysis and machine intelligence*, 40(4): 834–848.

Chen, L.-C.; Zhu, Y.; Papandreou, G.; Schroff, F.; and Adam, H. 2018. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *Proceedings of the European conference on computer vision (ECCV)*, 801–818.

Chen, T.; Kornblith, S.; Norouzi, M.; and Hinton, G. 2020. A Simple Framework for Contrastive Learning of Visual Representations. In III, H. D.; and Singh, A., eds., *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, 1597–1607. PMLR.

Chen, X.; Yuan, Y.; Zeng, G.; and Wang, J. 2021. Semi-supervised semantic segmentation with cross pseudo supervision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2613–2622.

Cordts, M.; Omran, M.; Ramos, S.; Rehfeld, T.; Enzweiler, M.; Benenson, R.; Franke, U.; Roth, S.; and Schiele, B. 2016. The cityscapes dataset for semantic urban scene understanding. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 3213–3223.

Everingham, M.; Eslami, S. A.; Van Gool, L.; Williams, C. K.; Winn, J.; and Zisserman, A. 2015. The pascal visual object classes challenge: A retrospective. *International journal of computer vision*, 111: 98–136.

Fan, J.; Gao, B.; Jin, H.; and Jiang, L. 2022. 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*, 9947–9956.

Feng, Z.; Zhou, Q.; Gu, Q.; Tan, X.; Cheng, G.; Lu, X.; Shi, J.; and Ma, L. 2022. DMT: Dynamic mutual training for semi-supervised learning. *Pattern Recognition*, 130: 108777.

French, G.; Laine, S.; Aila, T.; Mackiewicz, M.; and Finlayson, G. 2020. Semi-supervised semantic segmentation needs strong, varied perturbations. arXiv:1906.01916.

Fu, J.; Liu, J.; Tian, H.; Li, Y.; Bao, Y.; Fang, Z.; and Lu, H. 2019. Dual attention network for scene segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 3146–3154.

Grandvalet, Y.; and Bengio, Y. 2004. Semi-supervised learning by entropy minimization. *Advances in neural information processing systems*, 17.

Han, B.; Yao, Q.; Yu, X.; Niu, G.; Xu, M.; Hu, W.; Tsang, I.; and Sugiyama, M. 2018. Co-teaching: Robust training of deep neural networks with extremely noisy labels. *Advances in neural information processing systems*, 31.

Hariharan, B.; Arbeláez, P.; Bourdev, L.; Maji, S.; and Malik, J. 2011. Semantic contours from inverse detectors. In *2011 international conference on computer vision*, 991–998. IEEE.

He, R.; Yang, J.; and Qi, X. 2021. Re-distributing biased pseudo labels for semi-supervised semantic segmentation: A baseline investigation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 6930–6940.

Huang, W.; and Zhang, F. 2023. Semi-Supervised Semantic Segmentation with Structured Output Space Adaption. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 1–5. IEEE.

Hung, W.-C.; Tsai, Y.-H.; Liou, Y.-T.; Lin, Y.-Y.; and Yang, M.-H. 2018. Adversarial Learning for Semi-supervised Semantic Segmentation. In *British Machine Vision Conference*.

Ke, R.; Aviles-Rivero, A. I.; Pandey, S.; Reddy, S.; and Schönlieb, C.-B. 2022. A three-stage self-training framework for semi-supervised semantic segmentation. *IEEE Transactions on Image Processing*, 31: 1805–1815.

Ke, Z.; Qiu, D.; Li, K.; Yan, Q.; and Lau, R. W. 2020. 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*, 429–445. Springer.

Kwon, D.; and Kwak, S. 2022. Semi-supervised semantic segmentation with error localization network. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 9957–9967.

Lai, X.; Tian, Z.; Jiang, L.; Liu, S.; Zhao, H.; Wang, L.; and Jia, J. 2021. Semi-supervised semantic segmentation with directional context-aware consistency. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 1205–1214.

Laine, S.; and Aila, T. 2017. Temporal Ensembling for Semi-Supervised Learning. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net.

Lee, D.-H.; et al. 2013. Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks. In *Workshop on challenges in representation learning, ICML*, volume 3, 896. Atlanta.

Lei, T.; Zhang, D.; Du, X.; Wang, X.; Wan, Y.; and Nandi, A. K. 2022. Semi-supervised medical image segmentation using adversarial consistency learning and dynamic convolution network. *IEEE Transactions on Medical Imaging*.

Li, H.; and Zheng, H. 2021. A Residual Correction Approach for Semi-supervised Semantic Segmentation. In *Pattern Recognition and Computer Vision: 4th Chinese Conference, PRCV 2021, Beijing, China, October 29–November 1, 2021, Proceedings, Part IV 4*, 90–102. Springer.

Li, X.; Zhong, Z.; Wu, J.; Yang, Y.; Lin, Z.; and Liu, H. 2019. Expectation-maximization attention networks for semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 9167–9176.

Liu, S.; Zhi, S.; Johns, E.; and Davison, A. 2022a. Bootstrapping Semantic Segmentation with Regional Contrast. In *International Conference on Learning Representations*.

Liu, Y.; Tian, Y.; Chen, Y.; Liu, F.; Belagiannis, V.; and Carneiro, G. 2022b. Perturbed and strict mean teachers for semi-supervised semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 4258–4267.

Long, J.; Shelhamer, E.; and Darrell, T. 2015. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 3431–3440.

Mendel, R.; De Souza, L. A.; Rauber, D.; Papa, J. P.; and Palm, C. 2020. Semi-supervised segmentation based on error-correcting supervision. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXIX 16*, 141–157. Springer.

Mo, Y.; Wu, Y.; Yang, X.; Liu, F.; and Liao, Y. 2022. Review the state-of-the-art technologies of semantic segmentation based on deep learning. *Neurocomputing*, 493: 626–646.

Olsson, V.; Tranheden, W.; Pinto, J.; and Svensson, L. 2021. Classmix: Segmentation-based data augmentation for semi-supervised learning. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 1369–1378.

Peláez-Vegas, A.; Mesejo, P.; and Luengo, J. 2023. A Survey on Semi-Supervised Semantic Segmentation. *arXiv preprint arXiv:2302.09899*.

Qiao, S.; Shen, W.; Zhang, Z.; Wang, B.; and Yuille, A. 2018. Deep co-training for semi-supervised image recognition. In *Proceedings of the european conference on computer vision (ECCV)*, 135–152.

Ronneberger, O.; Fischer, P.; and Brox, T. 2015. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18*, 234–241. Springer.

Russakovsky, O.; Deng, J.; Su, H.; Krause, J.; Satheesh, S.; Ma, S.; Huang, Z.; Karpathy, A.; Khosla, A.; Bernstein, M.; et al. 2015. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115: 211–252.

Sohn, K.; Berthelot, D.; Carlini, N.; Zhang, Z.; Zhang, H.; Raffel, C. A.; Cubuk, E. D.; Kurakin, A.; and Li, C.-L. 2020. Fixmatch: Simplifying semi-supervised learning with consistency and confidence. *Advances in neural information processing systems*, 33: 596–608.

Tarvainen, A.; and Valpola, H. 2017. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. *Advances in neural information processing systems*, 30.

Wang, W.; Xie, E.; Li, X.; Fan, D.-P.; Song, K.; Liang, D.; Lu, T.; Luo, P.; and Shao, L. 2021. Pyramid vision transformer: A versatile backbone for dense prediction without convolutions. In *Proceedings of the IEEE/CVF international conference on computer vision*, 568–578.

Wang, Y.; Wang, H.; Shen, Y.; Fei, J.; Li, W.; Jin, G.; Wu, L.; Zhao, R.; and Le, X. 2022. Semi-supervised semantic segmentation using unreliable pseudo-labels. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 4248–4257.

Xie, B.; Li, S.; Li, M.; Liu, C. H.; Huang, G.; and Wang, G. 2023a. Sepico: Semantic-guided pixel contrast for domain adaptive semantic segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Xie, E.; Wang, W.; Yu, Z.; Anandkumar, A.; Alvarez, J. M.; and Luo, P. 2021. SegFormer: Simple and efficient design for semantic segmentation with transformers. *Advances in Neural Information Processing Systems*, 34: 12077–12090.

Xie, H.; Wang, C.; Zheng, M.; Dong, M.; You, S.; Fu, C.; and Xu, C. 2023b. Boosting semi-supervised semantic segmentation with probabilistic representations. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, 2938–2946.

Xie, Q.; Luong, M.-T.; Hovy, E.; and Le, Q. V. 2020. Self-training with noisy student improves imagenet classification. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 10687–10698.

Yang, L.; Qi, L.; Feng, L.; Zhang, W.; and Shi, Y. 2023. Revisiting weak-to-strong consistency in semi-supervised semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 7236–7246.

Yang, L.; Zhuo, W.; Qi, L.; Shi, Y.; and Gao, Y. 2022. St++: Make self-training work better for semi-supervised semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 4268–4277.

Zhao, H.; Shi, J.; Qi, X.; Wang, X.; and Jia, J. 2017. Pyramid scene parsing network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2881–2890.

Zoph, B.; Ghiasi, G.; Lin, T.-Y.; Cui, Y.; Liu, H.; Cubuk, E. D.; and Le, Q. 2020. Rethinking pre-training and self-training. *Advances in neural information processing systems*, 33: 3833–3845.