Pruning-Guided Curriculum Learning for Semi-Supervised Semantic Segmentation

Heejo Kong[1] Gun-Hee Lee[2] Suneung Kim[3] Seong-Whan Lee[3†]
[1]Department of Brain and Cognitive Engineering, Korea University, Seoul, South Korea
[2]Department of Computer Science and Engineering, Korea University, Seoul, South Korea
[3]Department of Artificial Intelligence, Korea University, Seoul, South Korea
{hj_kong, gunhlee, se_kim, sw.lee}@korea.ac.kr

Abstract

This study focuses on improving the quality of pseudo-labeling in the context of semi-supervised semantic segmentation. Previous studies have adopted confidence thresholding to reduce erroneous predictions in pseudo-labeled data and to enhance their qualities. However, numerous pseudo-labels with high confidence scores exist in the early training stages even though their predictions are incorrect, and this ambiguity limits confidence thresholding substantially. In this paper, we present a novel method to resolve the ambiguity of confidence scores with the guidance of network pruning. A recent finding showed that network pruning severely impairs the network generalization ability on samples that are not yet well learned or represented. Inspired by this finding, we refine the confidence scores by reflecting the extent to which the predictions are affected by pruning. Furthermore, we adopted a curriculum learning strategy for the confidence score, which enables the network to learn gradually from easy to hard samples. This approach resolves the ambiguity by suppressing the learning of noisy pseudo-labels, the confidence scores of which are difficult to trust owing to insufficient training in the early stages. Extensive experiments on various benchmarks demonstrate the superiority of our framework over state-of-the-art alternatives.

1. Introduction

Thanks to the growth of deep supervised learning, we have witnessed the remarkable advances of semantic segmentation [34, 4, 5, 52] over the last decade. However, this success is highly dependent on large-scale training datasets [7, 9], the construction of which is labor intensive and time consuming owing to the high cost of pixel-level labeling. To address this problem, semi-supervised learning (SSL) [27, 44, 42, 46, 47] has attracted attention for semantic segmentation, in which it is assumed that only a fraction of the entire dataset is labeled.

Figure 1. Heatmap visualizations of pixel-wise confidence scores for foreground objects. The confidence scores are estimated by the trained network for 10 epochs (out of a total of 80 epochs). The predicted scores of the red region are higher than those of the blue ones. This figure indicates that many confident pixels exist despite their predictions being incorrect in the early training stage.

The key challenge in semi-supervised semantic segmentation is the effective training of the network by leveraging unlabeled data. Pseudo-labeling [28] is a typical solution that assigns the most probable class that is predicted on the unlabeled sample as the pseudo ground truth. Due to the simple yet effective approach, recent studies [26, 1, 18, 54] have commonly adopted pseudo-labeling. Unfortunately, these methods suffer from the confirmation bias [2] that is caused by incorrect pseudo-labels, which directly degrades the performance of the training network. Previous studies tackle this bias by exploiting confidence thresholding [1, 54, 36, 38]. They have attempted to prevent the learning of incorrect predictions by only reflecting highly confident samples with confidence scores that exceed a predefined threshold.

However, in practice, the effectiveness of confidence thresholding is limited due to the ambiguity of confidence scores, and we observed that this ambiguity is naturally further intensified in the early training stages, as illustrated in Fig. 1. This figure indicates that many highly confident pixels exist, although their predictions are erroneous, which

cannot be filtered out by confidence thresholding. These confident but erroneous predictions are assigned as pseudo-labels and learned as noise for subsequent training epochs. This noise accumulation directly causes a confirmation bias.

In the early training stages, the network lacks generalization ability for specific samples owing to insufficient training. These samples directly cause noisy pseudo-labels with high confidence scores, even though the predictions are incorrect. Hence, we argue that if the extent to which the samples are trained from the network in the current stage can be estimated, it will be possible to reduce the noise by refining their confidence scores effectively. To this end, this study borrows the concept from recent empirical finding [19] for image classification. They showed that network pruning severely impairs deep neural network (DNN) generalization and memorization on insufficiently trained samples due to the sparsity of the training distribution. Inspired by this finding, our study applied network pruning as a practical tool to identify samples that are not yet well trained in the network.

In this study, we propose pruning-guided curriculum learning (PGCL), which is a novel method for resolving the ambiguity of confidence scores in the early training stages. Specifically, we first measure the similarity of the features that are extracted from the original and pruned networks for the same pixel. The measured similarity is considered to be the extent to which the samples have been trained in the current stage. Subsequently, we refine the confidence score by leveraging the similarity based on the curriculum policy so that the network learns gradually from easy to hard samples. This approach enhances the pseudo-labeling quality by preventing the network from learning the noisy pseudo-labels, the confidence scores of which are difficult to trust owing to insufficient training. The proposed PGCL is beneficial for being easily applied to the existing methods, and it effectively improves the segmentation performance.

Following the well-known benchmarks, we conduct extensive experiments on the PASCAL VOC [9] and Cityscapes [7] datasets, and the results demonstrate the effectiveness of our proposed method. In summary, our main contributions are three-fold:

- To enhance the quality of pseudo-labeling, we propose a novel method to refine the ambiguity of confidence score with the guidance of network pruning.
- We design PGCL framework that gradually trains the segmentation network from easy to hard samples based on the proposed refinement scheme. It is simple yet effective and can be easily incorporated into the existing SSL methods.
- Extensive experiments on PASCAL VOC 2012 and Cityscapes demonstrate that our proposed method surpasses current state-of-the-art alternatives.

2. Related Works

Semantic segmentation. Semantic segmentation [9, 7, 40, 29] is a fundamental task in computer vision that assigns semantic labels to each pixel in an image. The introduction of the FCN [34] achieved significant advances in the task, and recent studies have exploited this method, in which the three aspects of resolution, context, and edge have been studied. Studies on resolution to obtain accurate high-resolution outputs have attempted to leverage the encoder-decoder structure [5, 41] or dilated convolutional layer [4, 49]. Studies on context have aimed to obtain more diverse spatial contexts, for example, PSPNet [52] and ASPP [4]. Several studies have attempted to enhance the segmentation quality of the edge area, including PointRend [24], and SegFix [50]. However, their performance relies heavily on large-scale datasets, which require expensive label consumption.

Semi-supervised semantic segmentation. Early methods [22, 30, 43] in semi-supervised semantic segmentation tend to use generative adversarial networks [13], which train the unlabeled data as an adversarial loss. In recent years, the dominant streams have included consistency regularization [27, 44, 38, 12] and self-training [46, 47, 18, 48]. Consistency regularization enforces the consistency of the predictions with different perturbations on the same input, and hence, allows the learned decision boundary to be located in the low-density region. Self-training assigns the predictions on unlabeled data as pseudo-labels through the pre-trained network using a labeled dataset and retrains the network with both the labeled and pseudo-labeled data. Recent attempts have exploited a holistic method [42, 54, 6] that combines consistency regularization and self-training. Moreover, several studies [26, 1, 32, 53] have employed pixel-wise contrastive learning and the performance of semi-supervised semantic segmentation has been improved considerably.

Curriculum learning. Curriculum learning [3] is a training strategy that gradually incorporates easier to harder samples during training, thereby imitating the meaningful learning order in human curricula. Previous studies have revealed that curriculum learning offers the advantage of improving the network generalization capacity and convergence speed in several scenarios [39, 14, 20].

Network pruning beyond compression. Network pruning [15, 31, 33] is a primary way of removing the redundant weights of DNNs, and it effectively prevents the wastage of both computation and memory while preserving network performance. By contrast, recent studies [19, 10, 11] have attempted to explore network pruning beyond simply an ad-hoc compression tool from the perspective of its deeper connections with DNN memorization and generalization. The study most relevant to ours is that of Hooker et al. [19], who examined the impact of network pruning on the generalization properties in image classification. The authors empir-

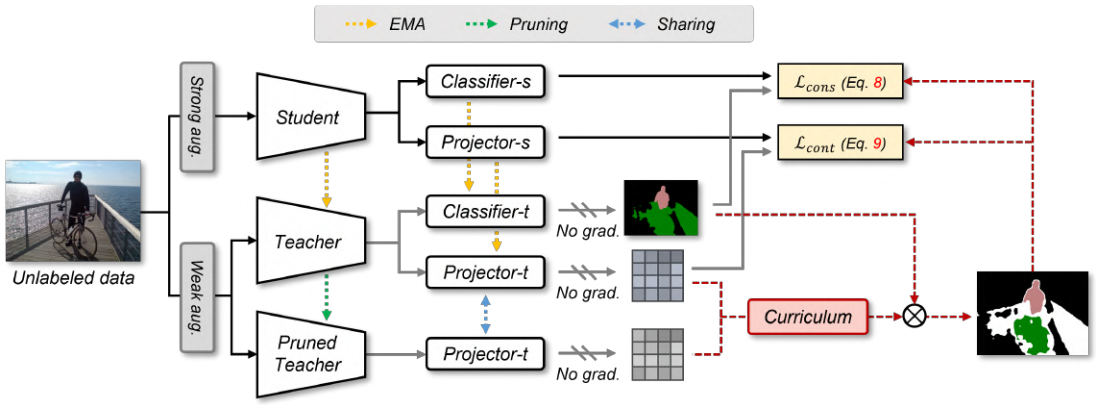


Figure 2. Overview of our SSL framework incorporating PGCL, which consists of three branches: the student, teacher, and pruned teacher networks. The student network is trained with pseudo-labels in two different manners, namely consistency loss (Eq. (8)) and pixel-wise contrastive loss (Eq. (9)). The teacher network generates pseudo-labels and it is updated by an EMA of the weights the student network. The pruned teacher network refines the generated pseudo-labels by leveraging the similarity between the output features extracted by the teacher and pruned teacher networks for the same input, and it is updated by applying network pruning to the teacher network. PGCL enables training on unlabeled data to be robust against noises on pseudo-labels by identifying and disregarding pixels with predictions that are difficult to trust owing to insufficient training.

ically demonstrated that pruning the trained classifier had a greater impact on certain examples or classes, such as the most difficult and long-tailed images, owing to the introduction of sparsity. This provided novel insight into which network pruning exposes the potential weakness of the trained network, and several studies [21, 45] have adopted this insight in their methods.

3. Proposed Method

3.1. Overview

Following the setting of semi-supervised semantic segmentation [1, 37, 25], we train the network by leveraging both a small set of labeled data D_l and a large set of unlabeled data D_u. The overall loss function is designed to minimize the sum of the supervised loss $\mathcal{L}_{\text{sup}}$ for the labeled dataset and unsupervised loss $\mathcal{L}_{\text{unsup}}$ for the unlabeled dataset:

$$\mathcal{L} = \mathcal{L}_{\text{sup}} + \mathcal{L}_{\text{unsup}}. \tag{1}$$

The supervised loss for the labeled dataset applies a standard pixel-wise cross-entropy loss between the predicted logits and given ground truth labels:

$$\mathcal{L}_{sup} = -\frac{1}{N} \sum_{x_i \in B_l}^{N} y_i^T \log \left(g \circ f \left(\theta; x_i \right) \right), \tag{2}$$

where x_i represents the i-th pixel-wise input of the labeled dataset, y_i represents the one-hot vector label of the pixel. B_l denotes the labeled data in each batch, and $g \circ f$ denotes the composition function of encoder f with the learnable weights θ and classifier head g.

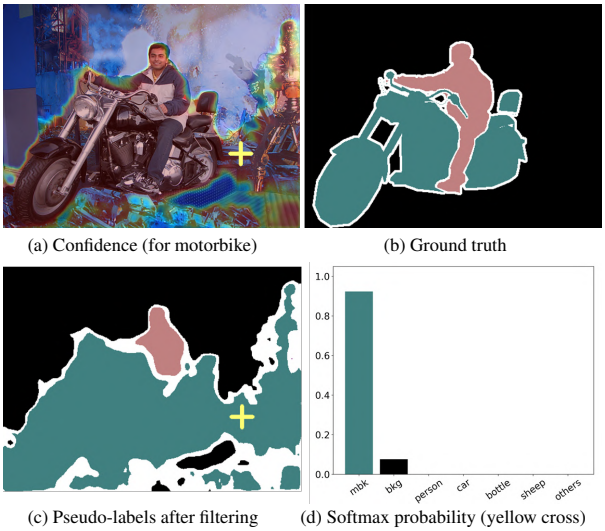
In this study, we propose a novel framework for the robust learning of unlabeled data. An overview of the proposed framework is presented in Fig. 2. Specifically, the framework consists of three branches with the same architecture but different update rules, which are student, teacher, and pruned teacher networks. The student network directly learns the unlabeled data as the main branch and the weights θ of the network are updated using gradient descent to optimize the unsupervised loss function. The teacher network generates pseudo-labels for the supervision of the student network. Following previous works [1, 12, 32], we adopt the mean teacher framework [44], which enables the teacher network to provide more stable pseudo-labels. The weights $\hat{\theta}$ of the teacher network are updated by the exponential moving average (EMA) of the weights θ with an update ratio α:

$$\hat{\theta}_t = \alpha \hat{\theta}_{t-1} + (1 - \alpha)\theta_t, \tag{3}$$

The pruned teacher network is used to refine the pseudo-labels generated by the teacher network. The weights $\hat{\theta}^p$ of the pruned teacher network are updated by applying network pruning [15] to the teacher network with weights $\hat{\theta}$. Both networks process the same input to obtain an output pair, and the similarity of the pair is used to enable adaptive training on the pseudo-label of each sample. In the following sections, we introduce the details of the proposed method (Sec. 3.2), and then describe how the learning of semi-supervised semantic segmentation is enhanced by leveraging this method (Sec. 3.3).

3.2. Pruning-Guided Curriculum Learning

Confidence thresholding [1, 54, 36, 38] is a typical solution for reducing noisy samples in the generated pseudo-labels and enhancing their quality. However, this criterion suffers from the ambiguity of confidence scores for insufficiently trained samples in the early training stages. For



(a) Confidence (for motorbike) (b) Ground truth

(c) Pseudo-labels after filtering (d) Softmax probability (yellow cross)

Figure 3. Illustration of ambiguity of confidence scores. (a) Heatmap of pixel-wise confidence scores for the motorbike, estimated by the trained network for 10 epochs (out of a total of 80 epochs). (b) Pixel-wise ground truth image. (c) Pixel-wise pseudo-labels filtered by confidence thresholding (0.9), where the pixels within the white region are not assigned as pseudo-labels. (d) Category-wise softmax probability of yellow cross pixel, which has a high confidence score of over 0.9 despite its prediction being incorrect.

example, the yellow cross pixel in Fig. 3 has a high confidence score of over 0.9, although it is incorrectly predicted as a motorbike. Fig. 3 (c) depicts the pseudo-label image following filtering with a threshold of 0.9 applied to all predictions, and it indicates that several erroneous predictions are still trained as supervision.

We design the novel PGCL method to resolve this ambiguity and to improve the pseudo-labeling quality. The key approach of our PGCL method is twofold: (1) refining the confidence score by leveraging the similarity of the extracted features from the teacher and pruned teacher networks for the same pixel, and (2) applying the curriculum policy to the refinement process, which enables the network to train gradually from easy to hard samples. Through this approach, we aim to prevent the network from learning on noisy pseudo-labels, the predictions of which are difficult to trust owing to insufficient training in the early stages.

Pruned teacher construction. To resolve the ambiguity of the confidence score, we first construct a pruned teacher network by leveraging network pruning. The crux of the construction is to remove as many weights as possible while maintaining the performance on the labeled dataset. Through this approach, we aim to impair the generalization ability of the network only on under-trained samples. To stabilize the pruned network, network pruning is applied only to the teacher encoder, whereas their projector heads share the weights.

Specifically, we obtain a pruning mask M by applying the simplest magnitude-based pruning [15] to the teacher encoder $f(\hat{\theta};)$. To save on computational overheads, the pruning mask is lazy-updated [21] at the beginning of every epoch; that is, all iterations in the same epoch adopt the same mask. The obtained pruning mask is applied to the teacher encoder at every iteration to construct a pruned teacher encoder, $f(\hat{\theta}^p;) = f(M \circ \hat{\theta};)$.

Confidence score refinement. Network pruning significantly impairs the network generalization or memorization ability on poorly learned or poorly represented samples [19]. Therefore, our method is designed to refine the confidence score by reflecting the extent to which the individual samples are affected by network pruning. Through this approach, our method prevents the learning of noisy pseudo-labels, which may lead to incorrect predictions even though their confidence scores are high, due to insufficient learning.

We first measure the similarity between the pixel embedding pair of the teacher and pruned teacher networks for the same pixel input i. Each pixel embedding is extracted by the composition function of the encoder and projector head, and the cosine similarity is applied as a similarity metric as follows:

$$d(\tilde{z}_i, \tilde{z}_i^p) = (1 + \tilde{z}_i \cdot \tilde{z}_i^p)/2, \tag{4}$$

where $\tilde{z}_i$ and $\tilde{z}_i^p$ represent the normalized pixel embeddings extracted by the teacher and pruned teacher networks, and $d(\cdot,\cdot) \in [0,1]$ denotes the normalized similarity. The similarity is closer to zero when the i-th sample is more strongly affected by network pruning.

Subsequently, the measured similarity is embedded into the confidence score estimated by the softmax probability with hyperparameter β to control the influence of the pruning:

$$s_i = \tilde{p}_i \cdot d\left(\tilde{z}_i, \tilde{z}_i^p\right)^\beta, \tag{5}$$

where $\tilde{p}_i$ denotes the softmax probability estimated by the teacher network for pixel i. Through this approach, our proposed method resolves the ambiguity by reducing the confidence for samples with a prediction that is difficult to trust because of insufficient training in the current stage.

Using the refined score in Eq. (5), we employ confidence thresholding, similar to previous methods [1, 54, 36, 38], as follows:

$$\omega_i = \begin{cases} 1, s_i \geq \gamma \\ 0, \text{ otherwise} \end{cases}. \tag{6}$$

Curriculum policy. Although the proposed refinement method can effectively reduce noisy pseudo-labels, a trade-off exists in that it also suppresses the learning of hard samples. As such samples are more likely to be close to object boundaries or belong to long-tail classes, they usually contribute significantly to improving the class discriminative ability. Therefore, we introduce a curriculum learning strat-

egy that also considers the learning of hard samples to train the unlabeled data effectively.

Specifically, we design a self-paced function to control the influence of network pruning by adjusting β as a learning pace parameter in Eq. (5). By gradually decreasing the pace parameter as learning progresses, the influence of pruning is reduced on the confidence score:

$$\beta_t = \beta_{\max} - (\beta_{\max} + \beta_{\min}) \left(\frac{t}{t_{\max}} \right)^{\sigma}, \tag{7}$$

where β_{max} and β_{min} represent the maximum and minimum values of β, respectively. t denotes the current training epoch, and t_{max} denotes the epoch at the end of training. The hyperparameter σ controls how rapidly the pace parameter decreases during the training process. Through this approach, the proportion of hard samples is gradually increased during the entire training period while suppressing the learning of noisy pseudo-labels in the early learning phase.

3.3. Semi-Supervised Learning with PGCL

Our proposed PGCL can be easily applied to existing methods, making their learning on unlabeled data more robust. Therefore, by applying the PGCL to previous SSL approaches, we aim to demonstrate that it can effectively improve the segmentation performance. To this end, this study leverages both pixel-wise contrastive learning [26, 32, 25], which has recently exhibited remarkable performance, and consistency [54, 38, 6], which is the most common approach for the task.

Pixel-wise cross-entropy loss is applied to the consistency regularization loss $\mathcal{L}_{\text{cons}}$, similar to Eq. (2). To generate pseudo-labels and their indicator functions in Eq. (6), we use weakly augmented unlabeled data for the teacher and pruned teacher networks. For the student network, the strongly augmented unlabeled data is processed to improve the generalization ability as follows:

$$\mathcal{L}_{\text{cons}} = -\frac{1}{N} \sum_{x_i \in B_u}^{N} \omega_i \cdot \hat{y}_i^T \log \left(f\left(\theta; \hat{A} \circ x_i\right)\right), \tag{8}$$

where $\hat{y}_i$ denotes the one-hot vector of the pseudo label for the pixel i, and $\hat{A}$ represents the strong augmentation operator. Let B_u denote the unlabeled data in each batch. Using the proposed indicator function ω_i, the segmentation network can only be trained on valid pixels.

We adopt the pixel-wise InfoNCE loss for the contrastive learning loss $\mathcal{L}_{\text{cont}}$. Let z_{i^+} and z_{i^-} denote the positive and negative keys for the anchor embedding z_i, respectively. The positive key z_{i^+} is the mean representation of all pixel embeddings with a predicted class same as the class of pixel i, and the negative key z_{i^-} is sampled from the rest pixel embeddings in the same training batch. $s\left(z_i, z_{i^+}\right) = \exp\left(z_i \cdot z_{i^+}/\tau\right)$ indicates the similarity metric between the embedding pair, and τ is the temperature hyper-parameter. Formally, $\mathcal{L}_{\text{cont}}$ defined as:

$$\mathcal{L}_{\text{cont}} = -\frac{1}{N} \sum_{x_i \in B_u}^{N} \omega_i \cdot \log \frac{s\left(z_i, \tilde{z}_{i^+}\right)}{s\left(z_i, \tilde{z}_{i^+}\right) + \sum_{i^- \in \mathcal{N}_i} s\left(z_i, \tilde{z}_{i^-}\right)}, \tag{9}$$

where $\mathcal{N}_i$ denotes the set of negative keys for the anchor pixel i, and ω_i represents the indicator function for pixel i, as in Eq. (8). z_i and $\tilde{z}_i$ are normalized pixel embeddings from the student and teacher networks, respectively.

Subsequently, the overall function of the unsupervised loss is the weighted sum of the consistency regularization loss $\mathcal{L}_{\text{cons}}$ and pixel-wise contrastive loss $\mathcal{L}_{\text{cont}}$ as follows:

$$\mathcal{L}_{\text{unsup}} = \lambda_{\text{cons}} \mathcal{L}_{\text{cons}} + \lambda_{\text{cont}} \mathcal{L}_{\text{cont}}, \tag{10}$$

where λ_{cons} and λ_{cont} are the hyperparameters used to control the intensities of the two losses.

4. Experiments

4.1. Experimental Setting

Datasets. Our experiments are conducted on PASCAL VOC 2012 [9] and Cityscapes [7] datasets. PASCAL VOC 2012 is a standard segmentation benchmark consisting of 20 semantic classes and 1 background class. The dataset has three separate subsets which are training, validation, and testing consist of 1464, 1449, 1456 images, respectively. Following the common practice, we use Segmentation Boundary Dataset (SBD) [16] as the augmented set with 9,118 additional training images. Cityscapes is a real urban scene dataset with 19 semantic classes for foreground objects and background stuffs. The training and validation splits contain 2975 and 500 images, respectively. We compare our method on several portions of labeled data. For PASCAL VOC 2012, we use the three partition protocols, 1/20, 1/8, and 1/4, while 1/8, 1/4, 1/2 are used for Cityscapes.

Data augmentation. We employ the same data augmentation strategy for training in PASCAL VOC 2012 and Cityscapes. All training images are first randomly resized by a ratio between 0.5 and 2, and then random cropping is applied to them (320×320 for PASCAL VOC 2012 and 720×720 for Cityscapes). Random horizontal flip is applied to the cropped images with the probability of 0.5. Random grayscale, random Gaussian blur, and color jittering are adopted to strong augmentation in Eq. (8) and (9) with the probability of 0.2, 0.5, and 0.8, respectively. Additionally, we employ the CutMix [51] for strong augmentation following the previous studies [12, 6, 32].

Implementation details. We use ResNet-50, 101 [17] pretrained on ImageNet [8] as the backbone and DeepLab v3+

Methods	SegNet	Backbone	1/20 (530)	1/8 (1323)	1/4 (2645)	Full (10582)
GCT [23]	DL2	R101	-	72.14	73.62	75.73
ClassMix [37]	DL2	R101	67.77	71.00	72.45	-
Alonso et al. [1]	DL2	R101	70.00	71.60	-	74.10
ECS [35]	DL3+	R50	-	70.22	72.60	76.29
CAC [26]	DL3+	R50	-	72.40	74.00	76.50
CPS [6]	DL3+	R50	-	73.67	74.90	-
ELN [25]	DL3+	R50	70.52	73.20	74.63	-
Baseline	DL3+	R50	62.10	68.20	70.40	77.00
Ours	DL3+	R50	**70.90**	**75.20**	**76.00**	-
CAC [26]	DL3+	R101	-	74.60	76.30	78.20
CPS [6]	DL3+	R101	-	76.44	77.68	-
ELN [25]	DL3+	R101	72.52	75.10	76.58	-
Baseline	DL3+	R101	67.30	71.50	74.00	78.80
Ours	DL3+	R101	**73.60**	**76.80**	**77.90**	-

Table 1. Performance (mIoU) on the PASCAL VOC 2012 validation set under different partition protocols. "Baseline" stands for the results of supervised training on the labeled dataset only.

Methods	SegNet	Backbone	1/8 (372)	1/4 (744)	1/2 (1487)	Full (2975)
CutMix [12]	DL2	R101	60.34	63.87	-	67.68
ClassMix [37]	DL2	R101	61.35	63.63	66.29	-
Alonso et al. [1]	DL2	R101	64.40	65.90	-	67.30
ECS [35]	DL3+	R50	67.38	70.70	72.89	74.76
CAC [26]	DL3+	R50	69.70	72.70	-	77.50
Alonso et al. [1]	DL3+	R50	70.00	71.60	-	74.20
ELN [25]	DL3+	R50	70.33	73.52	75.33	-
Baseline	DL3+	R50	61.20	66.20	72.00	78.90
Ours	DL3+	R50	**71.20**	**73.90**	**76.80**	-

Table 2. Performance (mIoU) on the Cityscapes validation set under different partition protocols. "Baseline" stands for the results of supervised training on the labeled dataset only.

[5] as the decoder. The projection head consists of two "1x1 Conv-BN -ReLU" blocks with a hidden and output dimensions of 128 and 256, respectively. For $\mathcal{L}_{\text{cont}}$ in Eq. (9), the τ is set to 0.5, and the number of negative samples is set to 19200 for PASCAL VOC 2012 and 14400 for Cityscapes. We set both λ_{cons} and λ_{cont} to 1.0, and a fixed threshold γ in Eq. (6) is set to 0.7.

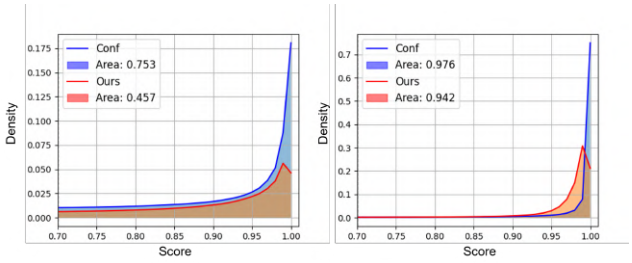
We adopt mini-batch Stochastic Gradient Descent (SGD) optimizer with momentum which is fixed as 0.9, and the weight decay is set to 0.0001. The poly scheduling is used to decay the learning rate during the training process: $lr = lr_{\text{base}} \cdot (1 - iter/total\_iter)^{0.9}$. For the training on PASCAL VOC 2012, we set the base learning rates to 0.001 and 0.01 for backbone and the rest parameters respectively, and the entire training epochs to 80 with batch sizes of 16. For the training on Cityscapes, we use the base learning rate with 0.01 and 0.1 for backbone and the rest parameters respectively, and the entire training epochs as 200 with batch sizes of 8. To stabilize training, the network is trained only with supervised learning in the first 3 epochs for PASCAL (while 5 epochs for Cityscapes).

4.2. Results

Comparison to state-of-the-art methods. To demonstrate the superiority of our proposed method PGCL, we compare our method to the current state-of-the-art methods and baseline (training on labeled data only). We adopt the mean Intersection-over-Union (mIoU) metric to evaluate the segmentation performance. All results are reported on the validation set for both PASCAL VOC and Cityscapes datasets. We abbreviate DeepLab v2 to DL2, DeepLab v3+ to DL3+, and ResNet-50, 101 to R50, R101. Table 1 shows the comparison results on PASCAL VOC. PGCL achieves the highest performance for all partition protocols (1/20, 1/8, and 1/4) as well as for both ResNet-50 and ResNet-101 backbone networks. In particular, our PGCL surpasses ECS [35] and ELN [25] by a large margin, and it is considered that our approach from the perspective of DNN memorization/generalization is more effective than error localization and error correction schemes leveraging the auxiliary networks to be trained. Further, to demonstrate the generalization ability of our method, we conduct experiments on Cityscapes with three partition protocols (1/8, 1/4, and 1/2), as shown in Table 2. This table shows that our method still outperforms other state-of-the-art methods.

	Acc.	Precision	Recall	F1-score
Conf.	92.73	77.50	84.12	79.94
Ours	**95.24**	**83.98**	**85.66**	**84.16**

Table 3. Pixel-level accuracy, Precision, Recall, and F1 score on the confidence thresholding (Conf.) and our proposed methods. Reported scores are averages of all the results of each class. The experiment is conducted on validation set of PASCAL VOC 2012.



(a) False (b) True

Figure 4. Probability density functions (PDFs) comparison between the confidence (Conf.) and our proposed scores. 'False' and 'True' represent the cases in which the predicted pseudo-labels are wrong and right, respectively.

Analysis on the quality of pseudo-labeling. To demonstrate the effectiveness of our PGCL for improving the quality of pseudo-labeling, we compare our method to confidence thresholding [1, 54, 36, 38]. We conduct experiments on the validation set of PASCAL VOC with ResNet-50 and DeepLab v3+. The trained network for 10 epochs only (out of a total of 80 epochs) is used, and hyper-parameters β and γ are set to 1.0 and 0.7, respectively. 'Conf.' represents the filtering method using the confidence score estimated by softmax probability same as previous works, and 'Ours' is the proposed method leveraging the refined confidence score that is introduced in Eq. (5) and Eq. (6). As shown in

	$\mathcal{L}_{\text{unsup}}$	Ref.	Cur.	CutMix	1/8	1/4
Sup. only					68.2	70.4
	✓				73.1	74.1
	✓	✓			72.8	74.2
	✓	✓	✓		74.2	75.1
	✓			✓	74.0	75.1
	✓	✓	✓	✓	**75.2**	**76.0**

Table 4. Ablation study on the effectiveness of each component in PGCL. The experiment is conducted on the unlabeled data of the given ratio of 1/8 and 1/4 to PASCAL VOC 2012. **Ref:** Pruning-guided confidence score refinement in Eq. (5). **Cur.:** Curriculum policy in Eq. (7). **CutMix:** CutMix augmentation for strongly augmented images.

Table 3, our proposed method outperforms the confidence thresholding by +2.51%, +6.48%, +1.54%, and +4.22% in pixel accuracy, precision, recall, and F1-score, respectively. These results demonstrate that our method effectively disregards the noisy pseudo-labels in the early training stages. Fig. 4 shows probability density functions (PDFs) of the confidence (Conf.) and our proposed scores exceeding the pre-defined threshold γ (0.7). 'False' and 'True' represent the cases in which the predicted pseudo-labels are incorrect or correct respectively, and 'Area' indicates the ratio of pseudo-labels reflected in training. As shown in this figure, the area of false pseudo-labels is greatly suppressed from 0.753 to 0.457 (-0.296), whereas the area of true pseudo-labels is almost preserved from 0.976 to 0.942 (-0.034).

4.3. Ablation Studies

We report the ablation studies and experiments for the setting of hyper-parameters in this section. All experiments are conducted with the ResNet-50 and DeepLab v3+ for the segmentation network.

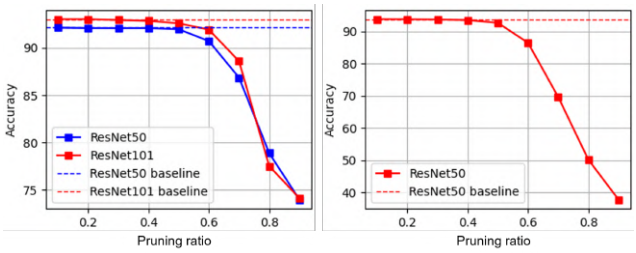
Effectiveness of each component. We conduct ablation studies to investigate the contributions of each component in our proposed method. All the ablations are under 1/8 and 1/4 partition protocols on PASCAL VOC validation set, and Table 4 shows the results. We use the model trained with only the supervised loss as our baseline, achieving mIoU of 68.2% and 70.4% under 1/8 and 1/4 proportions of labeled data, respectively. Leveraging unsupervised loss without PGCL improves the baseline by +4.9% under the 1/8 split (+3.7% under the 1/4 split). Simply applying the proposed refinement scheme even worsens the performance from 73.1% to 72.8% under the 1/8 split. It is considered a limitation caused by the lack of training on hard samples, as mentioned in Sec 3.2. After applying the curriculum policy to overcome the limitation, the proposed PGCL surpasses the case of simply applying the unsupervised loss as well as the baseline by a large margin. In addition, the table shows that the proposed method effectively improves performance with CutMix augmentation as well.

	$\mathcal{L}_{\text{cons}}$	$\mathcal{L}_{\text{cont}}$	$\mathcal{L}_{\text{cons}} + \mathcal{L}_{\text{cont}}$
Conf.	73.6	73.1	74.0
Ours	**73.9**	**74.2**	**75.2**

Table 5. Ablation study on the effectiveness of our PGCL in different loss components. $\mathcal{L}_{\text{cons}}$ and $\mathcal{L}_{\text{cont}}$ represents consistency regularization loss and pixel-wise contrastive loss in Eq. (8) and (9), respectively.

	σ				
β_{max}	0.1	0.3	0.5	1.0	1.5
0.5	74.0	73.8	74.0	74.1	74.4
1.0	74.3	**75.2**	74.8	74.1	73.8
1.5	74.1	74.3	74.6	74.2	73.6

Table 6. Performance (mIoU) on PASCAL VOC validation set under different β_{max} and σ for self-pacing function in Eq. (7).



(a) PASCAL VOC (b) Cityscapes

Figure 5. Pixel-level accuracy of training sets under different pruning ratios. "Baseline" stands for the results without applying network pruning, i.e., the pruning ratio of "baseline" equals to zero.

Ablation study on the different loss functions. Table 5 shows the mIoU performance over different loss components, consistency regularization loss $\mathcal{L}_{\text{cons}}$ and pixel-wise contrastive loss $\mathcal{L}_{\text{cont}}$. The experiments are conducted on PASCAL VOC with a ratio of 1/8. 'Conf.' and 'Ours' denote the different filtering methods using the confidence score estimated by softmax probability and our proposed score introduced in Eq. (5) and (6), respectively. In the case of $\mathcal{L}_{\text{cons}}$, applying our method brings an improvement by +0.3%, while in the case of $\mathcal{L}_{\text{cont}}$, the performance improves by +1.1%. Since the proposed method considers similarity in feature space, it demonstrates that our proposed method is more effective in contrastive learning, which directly learns the similarity between embedding features.

Ablation study on hyper-parameters. Table 6 shows the results of a grid search varying two hyper-parameters β_{max} and σ (β_{max} is fixed to 0) for self-pacing function in Eq. (7). The experiments are conducted on PASCAL VOC validation set using a 1/8 split. As can be seen, we found that $\beta_{\text{max}} = 1.0$ and $\sigma = 0.3$ achieve the best result, therefore adopting these values in all other experiments. Moreover, to set up the appropriate pruning ratio, we study the pixel-level accuracy under different pruning ratios for training sets of PASCAL VOC and Cityscapes using 1/8 and 1/4 partition protocols, respectively. We set the pruning ratio to 0.6 for

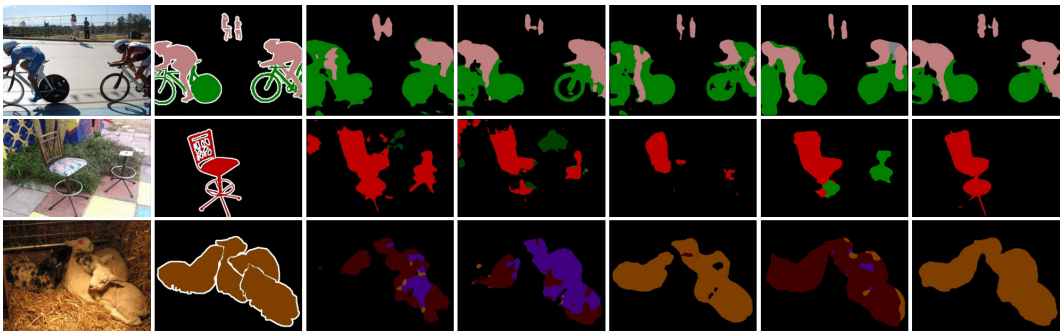


Figure 6. Qualitative results for comparison PGCL to previous state-of-the-art methods on PASCAL VOC 2012 validation set. "Baseline" stands for the results of supervised training on the labeled dataset only. For a fair comparison, all models are trained with a 1/8 split.

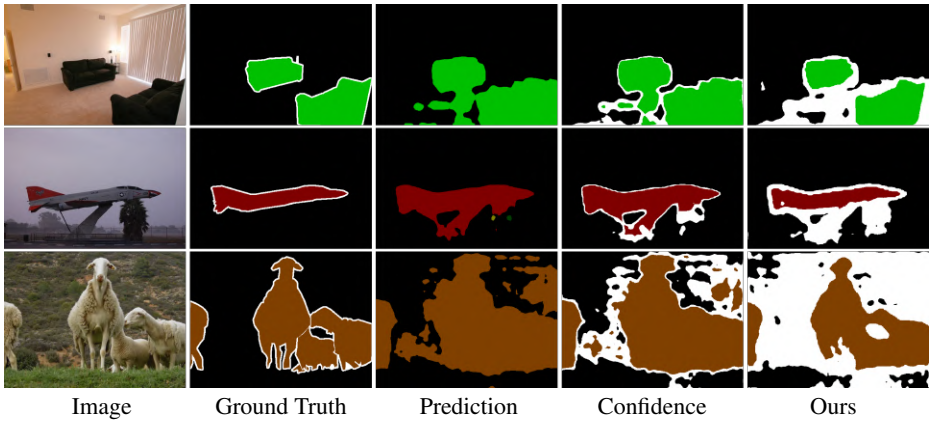


Figure 7. Qualitative results for comparison the pseudo-labeling quality between confidence thresholding and our PGCL on PASCAL VOC 2012 validation set. All predictions are estimated by the trained network for 10 epochs (out of a total of 80 epochs) with a 1/8 split. The white region indicates pixels that are not assigned as pseudo-labels, i.e., their scores are lower than the predefined threshold (0.7).

PASCAL VOC and 0.5 for Cityscapes to reduce as many weights as possible while preserving the performance of the training set.

Qualitative results. Fig. 6 presents the qualitative results for comparison of our proposed method PGCL to current state-of-the-art methods [26, 6, 25] and baseline (i.e., trained with supervised training on labeled dataset only) on PASCAL VOC validation set. For a fair comparison, all models are composed of the DeepLab v3+ decoder and ResNet-50 backbone, and they are trained with a 1/8 split. As can be seen, we observe that the results of our PGCL are generally superior to others. To analyze pseudo-labeling quality, we further display some qualitative pseudo-labeling results of confidence thresholding [1, 54, 36, 38] and our PGCL on PASCAL VOC validation set. This figure shows that our method effectively suppresses the learning of noisy pseudo-labels compared to confidence thresholding.

5. Conclusion

In this paper, we proposed a novel pruning-guided curriculum learning for semi-supervised semantic segmentation. In order to resolve the ambiguity of the confidence score in the early training stages, our method refines the score by reflecting the similarity of the features that are extracted from the original and pruned networks for the same pixel. Through this approach, our method effectively enhances the quality of pseudo-labeling by preventing the learning of noisy pseudo-labels that are difficult to trust due to insufficient training. To the best of our knowledge, this study is the first work utilizing network pruning for robust learning on unlabeled data in semi-supervised semantic segmentation. Extensive experiments show that our approach outperforms the previous state-of-the-art methods.

6. Acknowledgement

This work was supported by Institute of Information & communications Technology Planning Evaluation (IITP) grant funded by the Korea government(MSIT) (No. 2019-0-00079, Artificial Intelligence Graduate School Program(Korea University), No.2022-0-00984,Development of Artificial Intelligence Technology for Personalized Plug-and-Play Explanation and Verification of Explanation)

References

[1] Inigo Alonso, Alberto Sabater, David Ferstl, Luis Montesano, and Ana C Murillo. Semi-supervised semantic segmentation with pixel-level contrastive learning from a class-wise memory bank. In *Proc. IEEE International Conference on Computer Vision (ICCV)*, 2021.

[2] Eric Arazo, Diego Ortego, Paul Albert, Noel E O'Connor, and Kevin McGuinness. Pseudo-labeling and confirmation bias in deep semi-supervised learning. In *Proc. International Joint Conference on Neural Networks (IJCNN)*, 2020.

[3] Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston. Curriculum learning. In *Proc. International Conference on Machine Learning (ICML)*, 2009.

[4] 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 (TPAMI)*, 2017.

[5] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *Proc. European Conference on Computer Vision (ECCV)*, 2018.

[6] Xiaokang Chen, Yuhui Yuan, Gang Zeng, and Jingdong Wang. Semi-supervised semantic segmentation with cross pseudo supervision. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.

[7] 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 *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[8] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2009.

[9] Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *International Journal of Computer Vision (IJCV)*, 2010.

[10] Jonathan Frankle and Michael Carbin. The lottery ticket hypothesis: Finding sparse, trainable neural networks. In *Proc. International Conference on Learning Representations (ICLR)*, 2018.

[11] Jonathan Frankle, Gintare Karolina Dziugaite, Daniel M Roy, and Michael Carbin. Stabilizing the lottery ticket hypothesis. *arXiv preprint arXiv:1903.01611*, 2019.

[12] Geoffrey French, Samuli Laine, Timo Aila, Michal Mackiewicz, and Graham D. Finlayson. Semi-supervised semantic segmentation needs strong, varied perturbations. In *Proc. British Machine Vision Conference (BMVC)*, 2020.

[13] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Proc. Advances in Neural Information Processing Systems (NeurIPS)*, 2014.

[14] Sheng Guo, Weilin Huang, Haozhi Zhang, Chenfan Zhuang, Dengke Dong, Matthew R Scott, and Dinglong Huang. Curriculumnet: Weakly supervised learning from large-scale web images. In *Proc. European Conference on Computer Vision (ECCV)*, 2018.

[15] Song Han, Jeff Pool, John Tran, and William Dally. Learning both weights and connections for efficient neural network. In *Proc. Advances in Neural Information Processing Systems (NeurIPS)*, 2015.

[16] Bharath Hariharan, Pablo Arbeláez, Lubomir Bourdev, Subhransu Maji, and Jitendra Malik. Semantic contours from inverse detectors. In *Proc. IEEE International Conference on Computer Vision (ICCV)*, 2011.

[17] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[18] Ruifei He, Jihan Yang, and Xiaojuan Qi. Re-distributing biased pseudo labels for semi-supervised semantic segmentation: A baseline investigation. In *Proc. IEEE International Conference on Computer Vision (ICCV)*, 2021.

[19] Sara Hooker, Aaron Courville, Gregory Clark, Yann Dauphin, and Andrea Frome. What do compressed deep neural networks forget? *arXiv preprint arXiv:1911.05248*, 2019.

[20] Yuge Huang, Yuhan Wang, Ying Tai, Xiaoming Liu, Pengcheng Shen, Shaoxin Li, Jilin Li, and Feiyue Huang. Curricularface: adaptive curriculum learning loss for deep face recognition. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

[21] Ziyu Jiang, Tianlong Chen, Bobak J Mortazavi, and Zhangyang Wang. Self-damaging contrastive learning. In *Proc. International Conference on Machine Learning (ICML)*, 2021.

[22] Tarun Kalluri, Girish Varma, Manmohan Chandraker, and CV Jawahar. Universal semi-supervised semantic segmentation. In *Proc. IEEE International Conference on Computer Vision (ICCV)*, 2019.

[23] Zhanghan Ke, Di Qiu, Kaican Li, Qiong Yan, and Rynson WH Lau. Guided collaborative training for pixel-wise semi-supervised learning. In *Proc. European Conference on Computer Vision (ECCV)*, 2020.

[24] Alexander Kirillov, Yuxin Wu, Kaiming He, and Ross Girshick. Pointrend: Image segmentation as rendering. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

[25] Donghyeon Kwon and Suha Kwak. Semi-supervised semantic segmentation with error localization network. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

[26] Xin Lai, Zhuotao Tian, Li Jiang, Shu Liu, Hengshuang Zhao, Liwei Wang, and Jiaya Jia. Semi-supervised semantic segmentation with directional context-aware consistency. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.

[27] Samuli Laine and Timo Aila. Temporal ensembling for semi-supervised learning. In *Proc. International Conference on Learning Representations (ICLR)*, 2017.

[28] Dong-Hyun Lee et al. Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks. In *ICML Workshop on challenges in representation learning*, 2013.

[29] Seong-Whan Lee and Sang-Yup Kim. Integrated segmentation and recognition of handwritten numerals with cascade neural network. *IEEE Transactions on Systems, Man, and Cybernetics (TSMC)*, 1999.

[30] Daiqing Li, Junlin Yang, Karsten Kreis, Antonio Torralba, and Sanja Fidler. Semantic segmentation with generative models: Semi-supervised learning and strong out-of-domain generalization. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.

[31] Hao Li, Asim Kadav, Igor Durdanovic, Hanan Samet, and Hans Peter Graf. Pruning filters for efficient convnets. In *Proc. International Conference on Learning Representations (ICLR)*, 2017.

[32] Shikun Liu, Shuaifeng Zhi, Edward Johns, and Andrew Davison. Bootstrapping semantic segmentation with regional contrast. In *Proc. International Conference on Learning Representations (ICLR)*, 2021.

[33] Zhuang Liu, Jianguo Li, Zhiqiang Shen, Gao Huang, Shoumeng Yan, and Changshui Zhang. Learning efficient convolutional networks through network slimming. In *Proc. IEEE International Conference on Computer Vision (ICCV)*, 2017.

[34] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.

[35] Robert Mendel, Luis Antonio de Souza, David Rauber, Joao Paulo Papa, and Christoph Palm. Semi-supervised segmentation based on error-correcting supervision. In *Proc. European Conference on Computer Vision (ECCV)*, 2020.

[36] Sudhanshu Mittal, Maxim Tatarchenko, and Thomas Brox. Semi-supervised semantic segmentation with high-and low-level consistency. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 2019.

[37] Viktor Olsson, Wilhelm Tranheden, Juliano Pinto, and Lennart Svensson. Classmix: Segmentation-based data augmentation for semi-supervised learning. In *Proc. Winter Conference on Applications of Computer Vision (WACV)*, 2021.

[38] Yassine Ouali, Céline Hudelot, and Myriam Tami. Semi-supervised semantic segmentation with cross-consistency training. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

[39] Anastasia Pentina, Viktoriia Sharmanska, and Christoph H Lampert. Curriculum learning of multiple tasks. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.

[40] Myung-Cheol Roh, Tae-Yong Kim, Jihun Park, and Seong-Whan Lee. Accurate object contour tracking based on boundary edge selection. *Pattern Recogntion*, 2007.

[41] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Proc. International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, 2015.

[42] Kihyuk Sohn, David Berthelot, Nicholas Carlini, Zizhao Zhang, Han Zhang, Colin A Raffel, Ekin Dogus Cubuk, Alexey Kurakin, and Chun-Liang Li. Fixmatch: Simplifying semi-supervised learning with consistency and confidence. In *Proc. Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

[43] Nasim Souly, Concetto Spampinato, and Mubarak Shah. Semi supervised semantic segmentation using generative adversarial network. In *Proc. IEEE International Conference on Computer Vision (ICCV)*, 2017.

[44] Antti Tarvainen and Harri Valpola. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. In *Proc. Advances in Neural Information Processing Systems (NeurIPS)*, 2017.

[45] Zhihua Wang, Haotao Wang, Tianlong Chen, Zhangyang Wang, and Kede Ma. Troubleshooting blind image quality models in the wild. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.

[46] Qizhe Xie, Zihang Dai, Eduard Hovy, Thang Luong, and Quoc Le. Unsupervised data augmentation for consistency training. In *Proc. Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

[47] Qizhe Xie, Minh-Thang Luong, Eduard Hovy, and Quoc V Le. Self-training with noisy student improves imagenet classification. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

[48] Lihe Yang, Wei Zhuo, Lei Qi, Yinghuan Shi, and Yang Gao. St++: Make self-training work better for semi-supervised semantic segmentation. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

[49] Fisher Yu and Vladlen Koltun. Multi-scale context aggregation by dilated convolutions. In *Proc. International Conference on Learning Representations (ICLR)*, 2016.

[50] Yuhui Yuan, Jingyi Xie, Xilin Chen, and Jingdong Wang. Segfix: Model-agnostic boundary refinement for segmentation. In *Proc. European Conference on Computer Vision (ECCV)*, 2020.

[51] Sangdoo Yun, Dongyoon Han, Seong Joon Oh, Sanghyuk Chun, Junsuk Choe, and Youngjoon Yoo. Cutmix: Regularization strategy to train strong classifiers with localizable features. In *Proc. IEEE International Conference on Computer Vision (ICCV)*, 2019.

[52] Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia. Pyramid scene parsing network. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[53] Yuanyi Zhong, Bodi Yuan, Hong Wu, Zhiqiang Yuan, Jian Peng, and Yu-Xiong Wang. Pixel contrastive-consistent semi-supervised semantic segmentation. In *Proc. IEEE International Conference on Computer Vision (ICCV)*, 2021.

[54] Yuliang Zou, Zizhao Zhang, Han Zhang, Chun-Liang Li, Xiao Bian, Jia-Bin Huang, and Tomas Pfister. Pseudoseg: Designing pseudo labels for semantic segmentation. In *Proc. International Conference on Learning Representations (ICLR)*, 2020.