Semi-supervised Semantic Segmentation with Error Localization Network

Donghyeon Kwon[1]   Suha Kwak[1,2]
Dept. of CSE, POSTECH[1]   Graduate School of AI, POSTECH[2]
http://cvlab.postech.ac.kr/research/ELN/

Abstract

This paper studies semi-supervised learning of semantic segmentation, which assumes that only a small portion of training images are labeled and the others remain unlabeled. The unlabeled images are usually assigned pseudo labels to be used in training, which however often causes the risk of performance degradation due to the confirmation bias towards errors on the pseudo labels. We present a novel method that resolves this chronic issue of pseudo labeling. At the heart of our method lies error localization network (ELN), an auxiliary module that takes an image and its segmentation prediction as input and identifies pixels whose pseudo labels are likely to be wrong. ELN enables semi-supervised learning to be robust against inaccurate pseudo labels by disregarding label noises during training and can be naturally integrated with self-training and contrastive learning. Moreover, we introduce a new learning strategy for ELN that simulates plausible and diverse segmentation errors during training of ELN to enhance its generalization. Our method is evaluated on PASCAL VOC 2012 and Cityscapes, where it outperforms all existing methods in every evaluation setting.

1. Introduction

Recent advances in semantic segmentation have been attributed to supervised learning of deep neural networks [6, 7, 33, 37, 50] on large-scale datasets [10, 12, 15, 32]. However, collecting training data for semantic segmentation is labour-intensive and time-consuming due to the prohibitive cost of pixel-wise class labeling, which often leads to a dataset limited in terms of the number of annotated data and class diversity. To address this issue, label efficient learning, such as semi-supervised learning [3, 5, 19, 24, 26, 27, 29, 30, 35, 36], unsupervised learning [9, 46], weakly supervised learning [1, 2, 4, 11, 23, 28, 42, 48], and synthetic-to-real domain adaptation [20, 25, 31, 44, 45, 47, 53], has been proposed for semantic segmentation.

This paper studies semi-supervised learning of semantic segmentation, which assumes that only a subset of training images are assigned segmentation labels while the others remain unlabeled. Undoubtedly, the key to the success of this task is to utilize the unlabeled images effectively. Self-training [5, 27, 35, 49] and contrastive learning [3, 29, 51, 52] are techniques commonly used for the purpose in literature. Self-training generates pseudo labels of unlabeled images using a model trained on labeled ones, and uses them for supervised learning. Meanwhile, contrastive learning forces feature vectors corresponding to the same pseudo label to be close to each other. Although these techniques have improved the performance of semi-supervised semantic segmentation substantially, they share a common drawback: Since predictions for unlabeled images are usually corrupted by errors, learning using such predictions as supervision causes *confirmation bias* towards the errors and returns corrupted models consequently. Most existing methods alleviate this issue simply by not using uncertain predictions as supervision [3, 19, 29, 36], but their performance depends heavily on hand-tuned thresholds.

A recent approach deals with errors on pseudo labels by learning and exploiting an auxiliary network that corrects the errors [27, 35]; this model, called error correction network (ECN), learns from the difference between predictions of the main segmentation network and their ground truth labels on the labeled subset of training images. Ideally, ECN can significantly improve the quality of pseudo labels, but in practice, its advantage is often limited due to the challenges in its training. Since the segmentation network is quickly overfitted to a small number of labeled images, its outputs used as input to ECN do not cover a wide variety of prediction errors that ECN faces in testing, which results in limited generalization capability of ECN.

We present a novel method that is also dedicated to handling errors on pseudo labels yet better generalizes to those of arbitrary unlabeled images. The core of our method is the error localization network (ELN), which identifies pixels with erroneous pseudo labels in the form of binary segmentation. As will be demonstrated empirically, simply disregarding invalid pseudo labels, instead of correcting them, is sufficient to alleviate the confirmation bias and to learn accurate segmentation models. More importantly, since er-
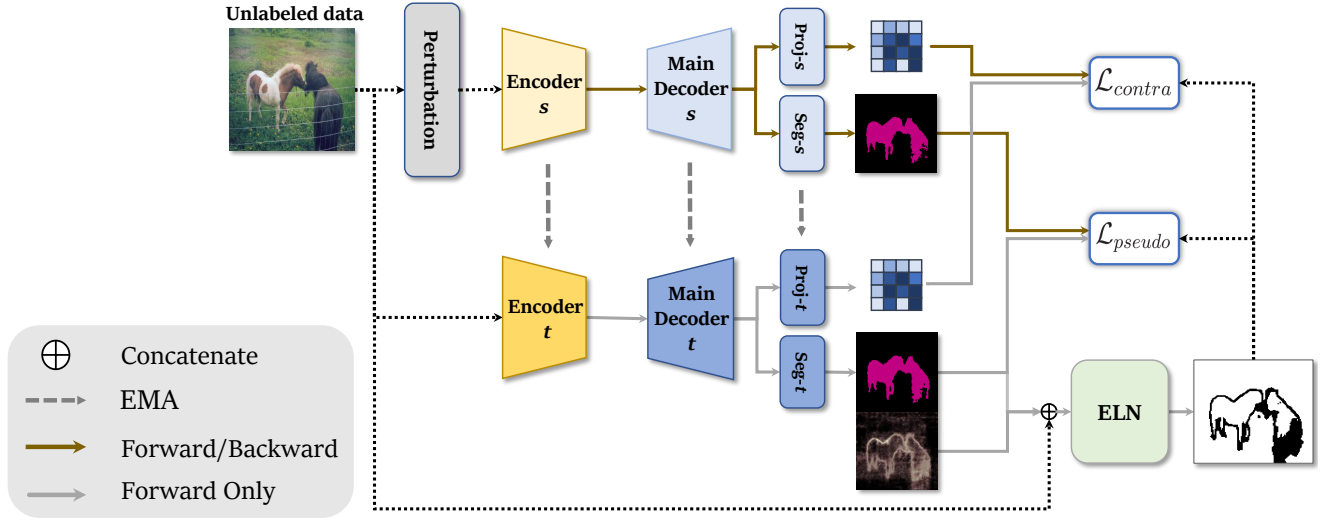
Figure 1. Our semi-supervised learning framework incorporating ELN. It employs two segmentation networks, the student (s), which will be our final model, and the teacher (t) used for generating pseudo labels. The student is trained using the pseudo labels of the teacher in two different ways, self-training and contrastive learning. To be specific, the decoder has two heads, one for segmentation (*Seg*) and the other for feature embedding (*Proj*); self-training and contrastive learning are applied to outputs of the *Seg* and *Proj* heads, respectively. Then the teacher is updated by an exponential moving average (EMA) of the student. ELN allows both self-training and contrastive learning to be robust against noises on pseudo labels by identifying and disregarding pixels whose pseudo labels are likely to be noisy.

ror localization is a class-agnostic subproblem of error correction and accordingly easier to solve, it is more straightforward to train an accurate and well-generalizable network for the target task.

Moreover, we design a novel training strategy for ELN to further improve its generalization. Specifically, we attach multiple auxiliary decoders to the main segmentation network and train them to achieve different accuracy levels so that they simulate the segmentation network at different training stages. ELN is then trained to localize errors on the predictions given by the auxiliary decoders as well as the main segmentation network. This strategy improves generalization of ELN since such predictions used as input to ELN potentially exhibit error patterns that the segmentation network causes during self-training with unlabeled images.

The trained ELN is then used for semi-supervised learning of semantic segmentation; the overall pipeline incorporating ELN is illustrated in Fig. 1. Our framework exploits unlabeled images in two ways: self-training and contrastive learning, both relying on pseudo labels. To this end, we adopt two segmentation networks: A student network, which will be our final model, and a teacher network generating pseudo labels and updated by an exponential moving average of the student. Self-training is done by learning the student using pseudo labels produced by the teacher. Meanwhile, contrastive learning encourages embedding vectors of the student and teacher to be similar if their pseudo labels are identical. ELN helps improve the effect of both

self-training and contrastive learning by filtering out potentially erroneous pseudo labels.

Following the convention, the proposed method is evaluated on the PASCAL VOC 2012 [12] and Cityscapes [10] datasets while varying the number of labeled training images, and it demonstrates superior performance to previous work on both of the datasets. In brief, our main contribution is three-fold.

- We propose *error localization*, a new approach to dealing with errors on pseudo labels. It is simple yet effective and can be naturally incorporated with self-training and contrastive learning. Moreover, we empirically demonstrate the superiority of error localization to error correction.

- We develop a new strategy for generating diverse and plausible prediction errors intentionally during the training of ELN. This improves the generalization of ELN even using a small number of labeled data for training.

- Segmentation networks trained by our method achieves the state of the art on two benchmark datasets, PASCAL VOC 2012 and Cityscapes, in every setting.

2. Related Work

Semantic segmentation. The goal of semantic segmentation is to generate dense pixel-wise classification. Starting with FCN [21], which replaced the classifier's last fully-connected layer with a fully convolutional layer for the
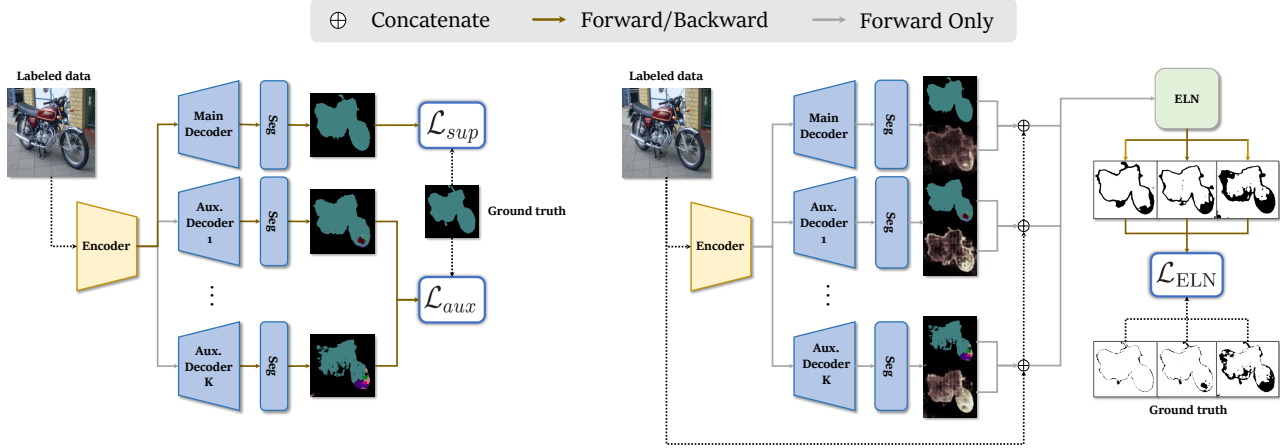
Figure 2. Training ELN along with the main segmentation network and the auxiliary decoders. (*left*) The main segmentation network is trained with the ordinary cross-entropy loss $\mathcal{L}_{sup}$, but the auxiliary decoders are trained with constrained cross-entropy losses $\mathcal{L}_{aux}$ so that they are inferior to the main segmentation network, and their predictions contain plausible and diverse errors intentionally. (*right*) All predictions from the decoders are used as training input to ELN, which learns to localize errors on the predictions. Note that ELN and other components are trained simultaneously, although their training processes are drawn separately in this figure for brevity.

first time, various approaches have been studied early. An encoder-decoder structure has been proposed to obtain an accurate high-resolution output [37, 40], and structures such as ASPP [6] and PSPNet [50] have been exploited to obtain more diverse spatial contexts. An attention mechanism has been studied to obtain a global relation [13, 22]. However, the success of these models requires a large amount of data, which costs expensive labour.

Semi-supervised semantic segmentation. Attempts to reduce the cost by applying a semi-supervised learning scheme have been studied intensely. Several methods [24, 30, 36] based on GAN and adversarial learning have been studied to reduce the gap between prediction on unlabeled and labeled data. One of the techniques frequently used in semi-supervised learning [14, 26, 39, 41] is consistency regularization. It allows the decision boundary to be located in a low-density region by using constraints to make the outputs of various perturbed inputs consistent with each other. Another approach [5, 27, 35, 49], self-training, is a method of generating pseudo labels with unlabeled data by pre-trained model and training the model with both labeled and pseudo labeled data. Recently, various methods [3, 29, 51, 52] have applied contrastive learning [16] to semantic segmentation in a semi-supervised manner, showing significant performance improvement.

Self-correction networks for semi-supervised semantic segmentation. The idea of correcting pseudo labels by an auxiliary network has been studied in [27, 35]. They presented networks that correct errors of pseudo labels by learning the difference between predicted and ground truth segmentation labels. However, it is challenging to train such networks effectively in the semi-supervised learning setting

since the segmentation network is quickly overfitted to the labeled data, leading to a poor generalization of correction networks. To address this generalization issue of the previous work, we introduce a new auxiliary task called error localization, and present ELN and its training strategy.

3. Proposed Method

Our framework consists of two-stage, learning ELN using labeled data and semi-supervised learning with ELN. The major issue in the first stage is the lack of diversity in predictions of the main network, which leads to a poor generalization of ELN. To address this issue, in addition to the main segmentation network (encoder $\mathcal{E}$, decoder $\mathcal{D}$), we employ auxiliary decoders ($\mathcal{D}_{aux}^1, ..., \mathcal{D}_{aux}^K$) that are learned to be inferior to the main segmentation network intentionally; predictions of the auxiliary decoders will depict plausible and diverse errors. ELN is learned along with the segmentation network and auxiliary decoders to identify errors on their predictions. The overall procedure of ELN training is illustrated in Fig. 2.

In the second stage, the trained ELN is then used for semi-supervised learning of semantic segmentation, where unlabeled images are exploited in two ways, self-training, and contrastive learning. The role of ELN is to identify pixels whose pseudo labels are likely to be erroneous so that we disregard such pixels in the process of self-training and contrastive learning for stable and effective training.

The remainder of this section presents details of the two stages of our method.

3.1. Learning ELN Using Labeled Data

At first, the main segmentation network is pre-trained with the standard pixel-wise cross-entropy loss $\mathcal{L}_{sup}$ on the set of labeled images D_L. Let $\mathcal{L}_{ce}(P, Y)$ denote the standard pixel-wise cross-entropy between segmentation prediction P and its ground truth label Y:

$$\mathcal{L}_{ce}(P, Y) = -\sum_i Y_i^\top \log(P_i), \tag{1}$$

where i is the index indicating each pixel of the input and Y_i is the one-hot vector of the ground truth for pixel i. Let $P = \mathcal{D}(\mathcal{E}(X))$ denote segmentation prediction of the main network for an image X. $\mathcal{L}_{sup}$ is then given by

$$\mathcal{L}_{sup} = \frac{1}{|D_L|} \sum_{X \in D_L} \mathcal{L}_{ce}(P, Y), \tag{2}$$

where Y is the ground truth of the input image X.

When the pre-training is completed, each auxiliary decoder is trained similarly to the main network but with a constrained cross-entropy loss, which is minimized only up to a certain multiple of $\mathcal{L}_{ce}(P, Y)$ and its gradient is not propagated beyond the auxiliary decoder. Let K be the number of all auxiliary decoders and k be their index. Then the total loss for the K auxiliary decoders, denoted by $\mathcal{L}_{aux}$, is given by

$$\mathcal{L}_{aux} = \frac{1}{|D_L|} \sum_{X \in D_L} \sum_{k=1}^{K} \mathbb{1}\{\mathcal{L}_{ce}(P^k, Y) > \alpha^k \cdot \mathcal{L}_{ce}(P, Y)\} \cdot \mathcal{L}_{ce}(P^k, Y), \tag{3}$$

where $P^k = \mathcal{D}^k(\mathcal{E}(X))$ denotes segmentation prediction of the kth decoder, and α^k indicates a scale hyper-parameter for constraining the loss applied to the kth auxiliary decoder. Training the auxiliary decoders in this way enables them to produce plausibly erroneous predictions, which are used as training input to ELN.

Given an image and its segmentation prediction as input, ELN is trained to localize errors on the prediction through supervised learning, where true locations of the errors are revealed by comparing the prediction with its ground truth counterpart. Let E^k be the pixel-wise entropy map of P^k and $B^k = \text{ELN}(X \oplus P^k \oplus E^k)$ denote the prediction of ELN in the form of binary segmentation map, where $\oplus$ represents channel-wise matrix concatenation. Then the binary cross-entropy loss for ELN, $\mathcal{L}_{\text{ELN}}$, is given by

$$\mathcal{L}_{\text{ELN}} = \frac{1}{|D_L| \cdot (K+1)} \sum_{X \in D_L} \sum_{k=0}^{K} \mathcal{L}_{ce}(B^k, M^k), \tag{4}$$

where M^k denotes the ground truth mask of B^k; M_i^k is 1 if the prediction of pixel i is correct and 0 otherwise. Note that $k = 0$ denotes the main decoder.

Despite using the auxiliary decoders, the population of pixel-level binary labels in M^k is typically biased to 1 (correct), which impairs the error identification ability of ELN. To alleviate this, a re-weighting factor is applied to pixels with incorrect predictions in $\mathcal{L}_{\text{ELN}}$ for balanced training. Let $\mathcal{L}_{wce}$ denote a weight re-adjusted pixel-wise cross-entropy between segmentation prediction P and its binary ground truth label Y:

$$\mathcal{L}_{wce}(P, Y) =$$
$$-\sum_i \left(\mathbb{1}\{Y_i = 0\} \frac{\sum_j \mathbb{1}\{Y_j = 1\}}{\sum_j \mathbb{1}\{Y_j = 0\}} Y_i^\top \log(P_i) \right.$$
$$\left. + \mathbb{1}\{Y_i = 1\} Y_i^\top \log(P_i) \right). \tag{5}$$

Then the loss in Eq. (4) is revised as

$$\mathcal{L}_{\text{ELN}} = \frac{1}{|D_L| \cdot (K+1)} \sum_{X \in D_L} \sum_{k=0}^{K} \mathcal{L}_{wce}(B^k, M^k). \tag{6}$$

The total loss minimized in the first stage for labeled data is as follows:

$$\mathcal{L}_{labeled} = \mathcal{L}_{sup} + \mathcal{L}_{aux} + \mathcal{L}_{\text{ELN}}. \tag{7}$$

Note that losses in $\mathcal{L}_{labeled}$ are jointly optimized in the first stage, although in pre-training $\mathcal{L}_{sup}$ is solely minimized.

3.2. Semi-supervised Learning with ELN

After learning ELN, the main segmentation network is trained on the set of unlabeled images D_U with two losses, a self-training loss and a pixel-wise contrastive loss. We adopt the mean teacher framework [43], which allows the teacher network to provide more stable pseudo supervision to the student network. Weights $\tilde{\theta}$ of the teacher $(\tilde{\mathcal{E}}, \tilde{\mathcal{D}})$ are updated by the exponential moving average of weights θ of the student $(\mathcal{E}, \mathcal{D})$ with an update ratio β:

$$\tilde{\theta}_t = \beta \tilde{\theta}_{t-1} + (1 - \beta) \theta_t. \tag{8}$$

The proposed self-training loss $\mathcal{L}_{pseudo}$ is the pixel-wise cross-entropy loss like $\mathcal{L}_{sup}$, but is applied only to valid pixels identified by ELN. Let $\tilde{P} = \tilde{\mathcal{D}}(\tilde{\mathcal{E}}(X))$ denote segmentation prediction of the teacher network and $P^a = \mathcal{D}(\mathcal{E}(A \cdot X))$ denote that of the student network, where A is the perturbation operator applied to the input image X. Also, let $\tilde{B} = \text{ELN}(X \oplus \tilde{P} \oplus \tilde{E})$ be the binary segmentation output of ELN. Then $\mathcal{L}_{pseudo}$ is given by

$$\mathcal{L}_{pseudo} = -\frac{1}{|D_U|} \sum_{X \in D_U} \sum_i \lfloor \tilde{B} \rceil_i \cdot \hat{Y}_i^\top \log(P_i^a), \tag{9}$$

where $\lfloor \rceil$ is a function rounding to the nearest integer and $\hat{Y}_i$ denotes the one-hot vector of the pseudo label for pixel

i. Through the rounded binary mask, the main segmentation network can be trained on valid pixels only.

In order to further improve the quality of learned features, we adopt a pixel-wise contrastive loss $\mathcal{L}_{contra}$. Specifically, in this loss, features whose pseudo labels are the same attract each other while those from different categories are pushed away in the feature space. Instead of applying the loss on a single image, we expand its range to the whole input batch for considering various feature relations, leading to a significant performance improvement. For a given input, let Ω_p^i denote a set of pixels belonging to the class of pixel i and Ω_n^i denote a set of pixels that do not belong to the class of pixel i. Also, let $\mathbf{d}$ represent a distance function, $\mathbf{d}(f_1, f_2) = \exp(\cos(f_1, f_2)/\tau)$, where $\cos$ means the cosine similarity and τ is a temperature hyperparameter. The pixel-wise contrastive loss $\mathcal{L}_{contra}$ is then given by

$$\mathcal{L}_{contra} = \\ -\frac{1}{|V|} \sum_{i \in V} \sum_{j \in \Omega_p^i} \log \frac{\mathbf{d}(f_i, \tilde{f}_j)}{\mathbf{d}(f_i, \tilde{f}_j) + \sum_{k \in \Omega_n^i} \mathbf{d}(f_i, \tilde{f}_k)}, \quad (10)$$

where V denotes the set of valid pixels on D_U, f_i and $\tilde{f}_i$ are feature embeddings of pixel i from the student and teacher networks, respectively.

The total loss for unlabeled data is as follows:

$$\mathcal{L}_{unlabeled} = \mathcal{L}_{pseudo} + \mathcal{L}_{contra}. \quad (11)$$

Note that labeled data are also involved in training through $\mathcal{L}_{labeled}$. When training is completed, only the student network is used at inference since the others, including ELN, are all auxiliary modules that support semi-supervised learning of the student.

4. Experiments

4.1. Network Architecture

We use DeepLab v3+ [8] with ResNet [18] backbone as our segmentation network since it has been adopted in recent papers [3, 29, 35] and shares a similar structure with Deeplab v2 [6], that has been widely used in literature [14, 27, 36, 38].

The proposed model mainly consists of two types of networks, the main segmentation network and the ELN. Each network is formed with an encoder and decoder. The encoder includes a ResNet [18] backbone, and the decoder (including the auxiliary decoders) contains sub-modules such as an atrous spatial pyramid pooling layer [6], a pixel-wise classifier for segmentation (*Seg* in Fig. 1), and a projector for feature embedding (*Proj* in Fig. 1). The last two modules are implemented by two 1×1 convolutional layers and one intermediate ReLU activation layer.

Method	SegNet	Backbone	1/20	1/8	1/4	Full
CutMix [14]	DL2	R101	66.48	67.60	-	72.54
S4GAN+MLMT [36]	DL2	R101	62.9	67.3	-	73.2
GCT [27]	DL2	R101	-	72.14	73.62	75.73
Alonso et al. [3]	DL2	R101	67.8	69.9	-	72.6
Baseline	DL3+	R50	59.88	67.63	70.56	76.6
ECS [35]	DL3+	R50	-	70.22	72.60	76.29
Xin et al. [29]	DL3+	R50	-	72.4	74.0	76.5
Alonso et al. [3]	DL3+	R50	69.1	71.8	-	75.9
Ours	DL3+	R50	70.52	73.20	74.63	-
Baseline	DL3+	R101	64.47	69.52	72.95	78.24
CutMix [14]	DL3+	R101	69.57	72.45	-	76.73
Xin et al. [29]	DL3+	R101	-	74.6	76.3	78.2
Ours	DL3+	R101	**72.52**	**75.10**	**76.58**	-

Table 1. mIoU value in the PASCAL VOC 2012 *val* set with different labeled-unlabeled ratios. All results of our experiments are averaged from three different subsets of the same ratio.

Method	SegNet	Backbone	1/8	1/4	1/2	Full
CutMix [14]	DL2	R101	60.34	63.87	-	67.68
S4GAN [36]	DL2	R101	59.3	61.9	-	65.8
Alonso et al. [3]	DL2	R101	63.0	64.8	-	66.4
Baseline	DL3+	R50	59.88	61.86	67.63	77.70
ECS [35]	DL3+	R50	67.38	70.70	72.89	74.76
Xin et al. [29]	DL3+	R50	69.7	72.7	-	77.5
Alonso et al. [3]	DL3+	R50	70.0	71.6	-	74.2
Ours	DL3+	R50	**70.33**	**73.52**	**75.33**	-

Table 2. mIoU value in the Cityscapes *val* set with different labeled-unlabeled ratios. All results of our experiments are averaged from three different subsets of the same ratio.

We adopt ResNet-50 or ResNet-101 as the backbone of the main network, and ResNet-34 for ELN. The backbones are pre-trained on ImageNet, but since the input to ELN is the concatenation of an image and tensors, its first convolutional layer is accordingly re-designed and fine-tuned.

4.2. Implementation Details

Datasets. We conduct experiments on two different datasets, PASCAL VOC 2012 [12] and Cityscapes [10]. PASCAL VOC 2012 is a standard semantic segmentation dataset consisting of 21 classes including the background class. The dataset has three separate subsets for training, validation, and testing; the subsets consist of 1464, 1449, 1456 images, respectively. Following the common practice, we use additional 9118 training images from the Segmentation Boundary (SBD) Dataset [17]. During training on PASCAL VOC 2012, we resize images to 512×512 pixels. Cityscapes [10] is a dataset of urban driving scenes with 19 classes for objects and background stuffs. It consists of training, validation, and testing splits with 2975, 500, and 1525 images, respectively. Images of the dataset is randomly cropped to 512×1024 pixels.

Data augmentation. Random horizontal flip is applied to both training datasets with the probability of 0.5. As the perturbation operator for the semi-supervised learning, we adopt color jittering and random grayscale with the probability of 0.2.

Optimizer. AdamW [34] is adopted with learning rate 1e-4 and weight decay 1e-5.

Hyper-parameters. For both labeled and unlabeled data, the size of a mini-batch is 6 on PASCAL VOC 2012 and 4 on Cityscapes. We assign 20 and 50 to the first and second auxiliary decoders, respectively. The temperature value τ of $\mathcal{L}_{contra}$ is set to 0.5. The update ratio β is set to 0.995.

Evaluation Metrics. We adopt the mean Intersection-over-Union (mIoU) as an evaluation metric. During evaluation, image of PASCAL VOC 2012 are resized to 512×512 and those of Cityscapes are used as-is. We conduct experiments on several proportions of labeled data to unlabeled data for validating our method under different conditions. For PASCAL VOC 2012, we use three ratios, 1/20, 1/8, and 1/4, while 1/8, 1/4, and 1/2 are used for Cityscapes.

4.3. Results

Performance analysis on semantic segmentation. To demonstrate the superior performance of our method, we compare the method with recent state-of-the-art models and training on labeled data only (Baseline). The results of our method on PASCAL VOC 2012 are listed in Table 1. We abbreviate Deeplab v2 to DL2, Deeplab v3+ to DL3+ and ResNet-50 to R50, ResNet-101 to R101. To test the performance of our method under various conditions, we conduct experiments on three ratios (1/20, 1/8, 1/4) with ResNet-50 and ResNet-101 as a backbone network, respectively. As we can see from the table, our method achieves superior performance over all other works for both backbone networks. It is considered that our error localization concept is much more effective than error correction in the semi-supervised scheme from the comparison of results between ECS [35] and Ours. We achieve higher performance than ECS with less labeled data; note that the performance of ECS is 70.22 in the 1/8 ratio, while Ours is 70.52 in the 1/20. Moreover, we conduct experiments on Cityscapes on three ratios (1/8, 1/4, 1/2) to show the generalization capability of our method. The results are displayed in Table 2, showing that our method still outperforms other methods. Fig. 3 and 4 show qualitative results of our method under various ratio conditions.

Performance analysis on error localization network. We further conduct additional experiments to compare our method with two similar approaches to demonstrate the effectiveness of ELN. We conduct experiments on the ratio of 1/20 to PASCAL VOC 2012 with ResNet-50 as a backbone network. As the first thing to compare, we consider a simple error correction network (s-ECN) which has a sim-

Method	ELN	s-ECN	Threshold
mIoU	**70.52**	67.14	67.77

Table 3. mIoU value of ELN, s-ECN and confidence score threshold method. The experiment is conducted in a *val* set of PASCAL VOC 2012.

Method	ELN	s-ECN	Threshold
Precision	0.6961	0.7060	0.7054
Recall	0.9673	0.8294	0.8783
F1 score	**0.7881**	0.7424	0.7627

Table 4. Precision, Recall, and F1 score of the ELN, s-ECN, and confidence score threshold method. Reported scores are averages of all the results of each image. To compare s-ECN, we only consider its error localization ability, not a correction. The experiment is conducted on the unlabeled data of the given ratio of 1/20 to PASCAL VOC 2012.

# of aux. decoders	0	1	2	3
α	-	20	20,50	20,50,100
mIoU	69.89	70.20	70.52	**71.13**

Table 5. An ablation study for the model performance according to the number of decoders and loss constrain parameter α.

ilar learning strategy as ELN; s-ECN is trained with pixel-wise cross-entropy loss and yields a corrected segmentation prediction as an output, not a binary mask. We choose another method, performing confidence score threshold on the output of segmentation prediction after softmax layer, without an additional network. As we can see from Table 3, ELN achieves the highest mIoU value over the other two approaches. We also conduct another experiment to understand how well each method performs error localization to unseen data. In Table 4, ELN shows the highest F1 score among methods. Note that results of s-ECN are worse than Threshold; it emphasizes the limitations of the error correction scheme, implying that it does not work as intended due to its harsh training condition. In Fig. 5 and 6, we display our qualitative results of segmentation prediction and its binary mask.

4.4. Ablation Studies

We conduct ablation studies to investigate the impacts of each component of the proposed method. The experiment is based on PASCAL VOC 2012 with the ratio of 1/20, and results are averaged over three times. We use ResNet-50 as the backbone of the main segmentation network.

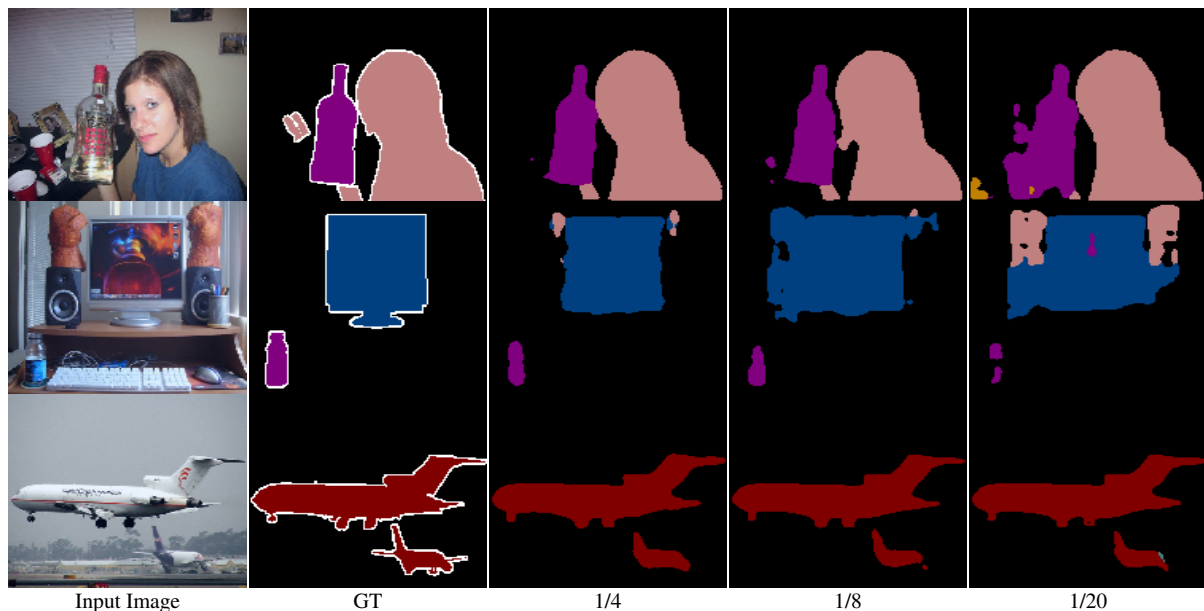**Different number of auxiliary decoders.** The auxiliary

Figure 3. Qualitative results on a *val* set of PASCAL VOC 2012 in various proportions of labeled data to unlabeled data.
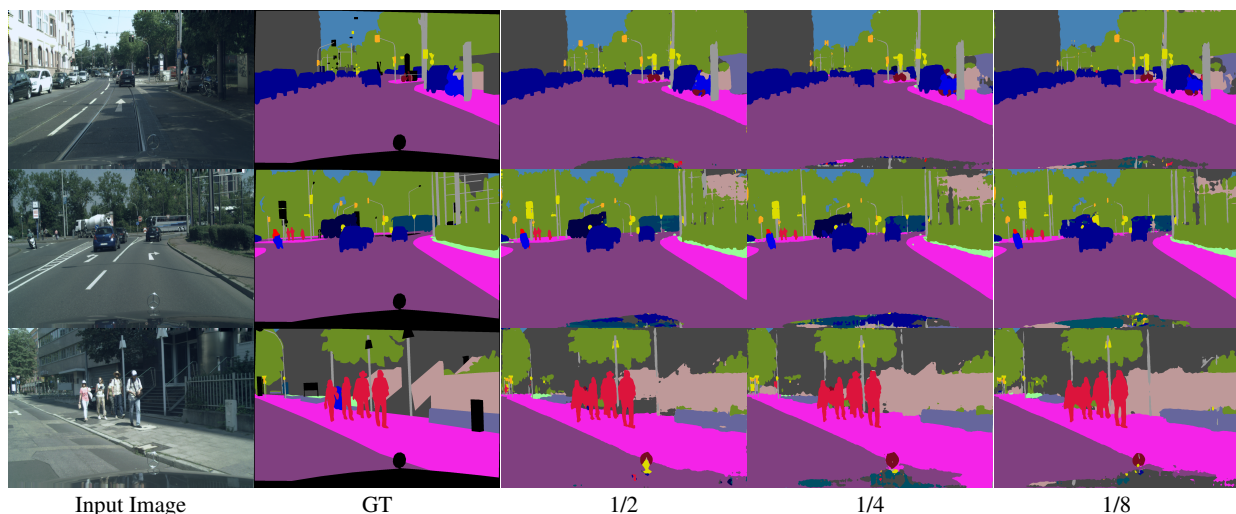

Figure 4. Qualitative results on a *val* set of Cityscapes in various proportions of labeled data to unlabeled data.

decoder plays a critical role in ELN learning. We experiment with how the auxiliary decoders affect the overall performance by changing the number of auxiliary decoders and constraints parameters. The result is listed in Table 5. As reported from the experiments, the performance improves as the number of decoders increases and a high loss constrain value is applied. It shows that various quality of segmentation predictions helps the effective learning of the ELN. Note that we could achieve sufficient performance improvement with only two auxiliary decoders.

Different loss combination in Eq. (11). In the semi-supervised learning stage, $\mathcal{L}_{pseudo}$ and $\mathcal{L}_{contra}$ are jointly

Threshold	$\mathcal{L}_{pseudo}$	$\mathcal{L}_{contra}$	$\mathcal{L}_{pseudo} + \mathcal{L}_{contra}$
67.77	69.14	69.30	**70.52**

Table 6. An ablation study on different loss combinations in mIoU. "Threshold" in the table is the method using none of the two losses associated with ELN but applying a confidence score threshold.

optimized. The proposed pixel-wise contrastive loss, $\mathcal{L}_{contra}$, enables the training of feature embeddings in more diverse contexts by learning the relation between images in a batch with a standard pixel-wise cross-entropy loss
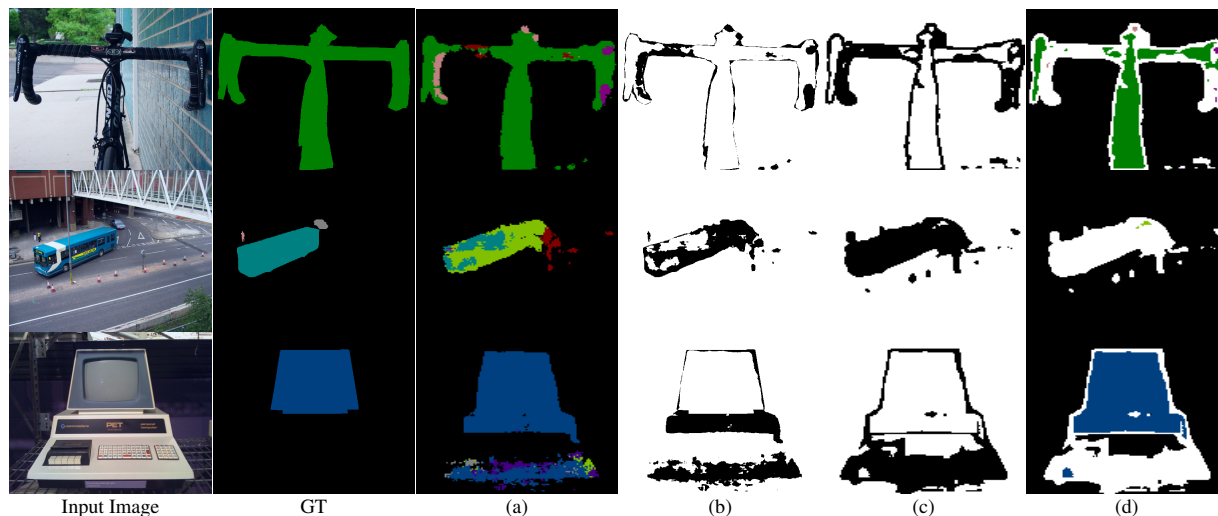
Figure 5. Qualitative results on unlabeled data of training set on PASCAL VOC 2012 in the labeled ratio of 1/20. (a) Segmentation prediction from the main segmentation network. (b) Ground truth binary mask. (c) Binary mask predicted by ELN. (d) Filtered segmentation prediction by the predicted binary mask. Erroneous predictions colored in white in (d) are not used as pseudo labels.
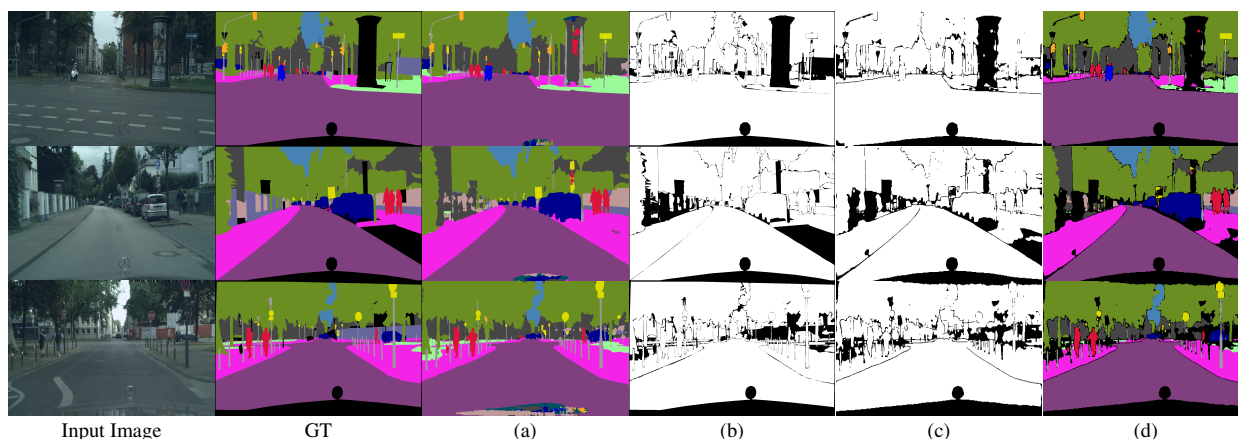

Figure 6. Qualitative results on unlabeled data of training set on Cityscapes in the labeled ratio of 1/8. (a) Segmentation prediction from the main segmentation network. (b) Ground truth binary mask. (c) Binary mask predicted by ELN. (d) Filtered segmentation prediction by the predicted binary mask. Erroneous predictions colored in white in (d) are not used as pseudo labels.

$\mathcal{L}_{pseudo}$. We make a comparison to investigate effect of each loss term in Eq. (11). As shown in Table 6, each term contributes to the performance, and using both of them improves the most.

5. Conclusion

We have presented a novel training framework suitable for semi-supervised semantic segmentation tasks. To mitigate the performance degradation caused by confirmation bias due to invalid pseudo labels, we have proposed error localization network (ELN) and its training scheme. Our experiments validated that ELN effectively removes error of pseudo labels for unseen data, which demonstrate that our

learning strategy using erroneous predictions simulated by auxiliary decoders is helpful. Our method achieved the state of the art on both of the PASCAL VOC 2012 and Cityscapes datasets with high generalization capability.

limitations. Due to the additional auxiliary networks, our method needs a relatively larger amount of GPU memory during training, and as the number of auxiliary decoders increases, larger memory footprint is required. ELN sometimes failed to indicate erroneous predictions that the main segmentation network has strong confidence (*i.e.*, low-entropy).

References

[1] Jiwoon Ahn, Sunghyun Cho, and Suha Kwak. Weakly supervised learning of instance segmentation with inter-pixel relations. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 1

[2] Jiwoon Ahn and Suha Kwak. Learning pixel-level semantic affinity with image-level supervision for weakly supervised semantic segmentation. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 1

[3] Iñigo Alonso, Alberto Sabater, David Ferstl, Luis Montesano, and Ana C Murillo. Semi-supervised semantic segmentation with pixel-level contrastive learning from a class-wise memory bank. In *Proc. IEEE International Conference on Computer Vision (ICCV)*, 2021. 1, 3, 5

[4] Liyi Chen, Weiwei Wu, Chenchen Fu, Xiao Han, and Yuntao Zhang. Weakly supervised semantic segmentation with boundary exploration. In *Proc. European Conference on Computer Vision (ECCV)*, 2020. 1

[5] Liang-Chieh Chen, Raphael Gontijo Lopes, Bowen Cheng, Maxwell D. Collins, Ekin D. Cubuk, Barret Zoph, Hartwig Adam, and Jonathon Shlens. Naive-student: Leveraging semi-supervised learning in video sequences for urban scene segmentation. In *Proc. European Conference on Computer Vision (ECCV)*, 2020. 1, 3

[6] 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. 1, 3, 5

[7] 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. 1

[8] 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. 5

[9] Jang Hyun Cho, Utkarsh Mall, Kavita Bala, and Bharath Hariharan. Picie: Unsupervised semantic segmentation using invariance and equivariance in clustering. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 1

[10] 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 under-standing. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 1, 2, 5

[11] Zhang Dong, Zhang Hanwang, Tang Jinhui, Hua Xiansheng, and Sun Qianru. Causal intervention for weakly supervised semantic segmentation. In *Proc. Neural Information Processing Systems (NeurIPS)*, 2020. 1

[12] 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. 1, 2, 5

[13] Haijie Tian Yong Li Yongjun Bao Zhiwei Fang and Jing Liu Hanqing Lu Jun Fu. Dual attention network for scene segmentation. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 3

[14] Geoff French, Samuli Laine, Timo Aila, Michal Mackiewicz, and Graham Finlayson. Semi-supervised semantic segmentation needs strong, varied perturbations. In *Proc. British Machine Vision Conference (BMVC)*, 2020. 3, 5

[15] Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012. 1

[16] Raia Hadsell, Sumit Chopra, and Yann LeCun. Dimensionality reduction by learning an invariant mapping. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2006. 3

[17] 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. 5

[18] 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)*, June 2016. 5

[19] Ruifei He, Jihan Yang, and Xiaojuan Qi. Redistributing biased pseudo labels for semi-supervised semantic segmentation: A baseline investigation. In *Proc. IEEE International Conference on Computer Vision (ICCV)*, 2021. 1

[20] Judy Hoffman, Eric Tzeng, Taesung Park, Jun-Yan Zhu, Phillip Isola, Kate Saenko, Alexei Efros, and Trevor Darrell. Cycada: Cycle-consistent adversarial domain adaptation. In *Proc. International Conference on Machine Learning (ICML)*, 2018. 1

[21] Judy Hoffman, Dequan Wang, Fisher Yu, and Trevor Darrell. Fcns in the wild: Pixel-level adversarial and constraint-based adaptation. *arXiv preprint arXiv:1612.02649*, 2016. 2

[22] Zilong Huang, Xinggang Wang, Lichao Huang, Chang Huang, Yunchao Wei, and Wenyu Liu. Cc-net: Criss-cross attention for semantic segmentation. In *Proc. IEEE International Conference on Computer Vision (ICCV)*, 2019. 3

[23] Zilong Huang, Xinggang Wang, Jiasi Wang, Wenyu Liu, and Jingdong Wang. Weakly-supervised semantic segmentation network with deep seeded region growing. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 1

[24] Tarun Kalluri, Girish Varma, Manmohan Chandraker, and C V Jawahar. Universal semi-supervised semantic segmentation. In *Proc. IEEE International Conference on Computer Vision (ICCV)*, 2019. 1, 3

[25] Guoliang Kang, Yunchao Wei, Yi Yang, Yueting Zhuang, and Alexander Hauptmann. Pixel-level cycle association: A new perspective for domain adaptive semantic segmentation. In *Proc. Neural Information Processing Systems (NeurIPS)*, 2020. 1

[26] Rihuan Ke, Angelica Aviles-Rivero, Saurabh Pandey, Saikumar Reddy, and Carola-Bibiane Schönlieb. A three-stage self-training framework for semi-supervised semantic segmentation. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 1, 3

[27] Zhanghan Ke, Di Qiu, Kaican Li, Qiong Yan, and Rynson W.H. Lau. Guided collaborative training for pixel-wise semi-supervised learning. In *Proc. European Conference on Computer Vision (ECCV)*, 2020. 1, 3, 5

[28] Suha Kwak, Seunghoon Hong, and Bohyung Han. Weakly supervised semantic segmentation using superpixel pooling network. In *Proc. AAAI Conference on Artificial Intelligence (AAAI)*, 2017. 1

[29] 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. 1, 3, 5

[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. 1, 3

[31] Yunsheng Li, Lu Yuan, and Nuno Vasconcelos. Bidirectional learning for domain adaptation of semantic segmentation. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 1

[32] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft COCO: common objects in context. In *Proc. European Conference on Computer Vision (ECCV)*, 2014. 1

[33] 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. 1

[34] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *Proc. International Conference on Learning Representations (ICLR)*, 2019. 6

[35] Robert Mendel, Luis Antonio de Souza Jr, David Rauber, João Paulo Papa, and Christoph Palm. Semi-supervised segmentation based on error-correcting supervision. In *Proc. European Conference on Computer Vision (ECCV)*, 2020. 1, 3, 5, 6

[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. 1, 3, 5

[37] Hyeonwoo Noh, Seunghoon Hong, and Bohyung Han. Learning deconvolution network for semantic segmentation. In *Proc. IEEE International Conference on Computer Vision (ICCV)*, 2015. 1, 3

[38] Viktor Olsson, Wilhelm Tranheden, Juliano Pinto, and Lennart Svensson. Classmix: Segmentation-based data augmentation for semi-supervised learning. In *Proc. IEEE Winter Conf. on Applications of Computer Vision (WACV)*, 2021. 5

[39] Yassine Ouali, Celine Hudelot, and Myriam Tami. Semi-supervised semantic segmentation with cross-consistency training. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 3

[40] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Proc. Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, 2015. 3

[41] 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. Neural Information Processing Systems (NeurIPS)*, 2020. 3

[42] Guolei Sun, Wenguan Wang, Jifeng Dai, and Luc Van Gool. Mining cross-image semantics for weakly supervised semantic segmentation. In *Proc. European Conference on Computer Vision (ECCV)*, 2020. 1

[43] Antti Tarvainen and Harri Valpola. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. In *Proc. Neural Information Processing Systems (NeurIPS)*, 2017. 4

[44] Yi-Hsuan Tsai, Wei-Chih Hung, Samuel Schulter, Kihyuk Sohn, Ming-Hsuan Yang, and Manmohan Chandraker. Learning to adapt structured output space for semantic segmentation. In *Proc. IEEE Conference*

on Computer Vision and Pattern Recognition (CVPR), 2018. 1

[45] Yi-Hsuan Tsai, Kihyuk Sohn, Samuel Schulter, and Manmohan Chandraker. Domain adaptation for structured output via discriminative patch representations. In *Proc. IEEE International Conference on Computer Vision (ICCV)*, 2019. 1

[46] Wouter Van Gansbeke, Simon Vandenhende, Stamatios Georgoulis, and Luc Van Gool. Unsupervised semantic segmentation by contrasting object mask proposals. In *Proc. IEEE International Conference on Computer Vision (ICCV)*, 2021. 1

[47] Tuan-Hung Vu, Himalaya Jain, Maxime Bucher, Matthieu Cord, and Patrick Pérez. Advent: Adversarial entropy minimization for domain adaptation in semantic segmentation. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 1

[48] Xiang Wang, Shaodi You, Xi Li, and Huimin Ma. Weakly-supervised semantic segmentation by iteratively mining common object features. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 1

[49] Qizhe Xie, Minh-Thang Luong, Eduard Hovy, and Quoc V. Le. Self-training with noisy student improves imagenet classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020. 1, 3

[50] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia. Pyramid scene parsing network. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6230–6239, July 2017. 1, 3

[51] Yuanyi Zhong, Bodi Yuan, Hong Wu, Zhiqiang Yuan2, Jian Peng, and Yu-Xiong Wang. Pixel contrastive-consistent semi-supervised semantic segmentation. In *Proc. IEEE International Conference on Computer Vision (ICCV)*, 2021. 1, 3

[52] Yanning Zhou, Hang Xu, and Wei Zhang. c^3-semiseg: Contrastive semi-supervised segmentation via crossset learning and dynamic class-balancing. In *Proc. IEEE International Conference on Computer Vision (ICCV)*, 2021. 1, 3

[53] Yang Zou, Zhiding Yu, BVK Kumar, and Jinsong Wang. Unsupervised domain adaptation for semantic segmentation via class-balanced self-training. In *Proc. European Conference on Computer Vision (ECCV)*, 2018. 1