ST++: Make Self-training Work Better for Semi-supervised Semantic Segmentation

Lihe Yang[1] Wei Zhuo[3] Lei Qi[4,1] Yinghuan Shi[1,2]* Yang Gao[1]
[1]State Key Laboratory for Novel Software Technology, Nanjing University
[2]National Institute of Healthcare Data Science, Nanjing University
[3]Tencent [4]Southeast University
lihe.yang.cs@gmail.com weizhuo@tencent.com qilei@seu.edu.cn {syh, gaoy}@nju.edu.cn

Abstract

Self-training via pseudo labeling is a conventional, simple, and popular pipeline to leverage unlabeled data. In this work, we first construct a strong baseline of self-training (namely ST) for semi-supervised semantic segmentation via injecting strong data augmentations (SDA) on unlabeled images to alleviate overfitting noisy labels as well as decouple similar predictions between the teacher and student. With this simple mechanism, our ST outperforms all existing methods without any bells and whistles, e.g., iterative re-training. Inspired by the impressive results, we thoroughly investigate the SDA and provide some empirical analysis. Nevertheless, incorrect pseudo labels are still prone to accumulate and degrade the performance. To this end, we further propose an advanced self-training framework (namely ST++), that performs selective re-training via prioritizing reliable unlabeled images based on holistic prediction-level stability. Concretely, several model checkpoints are saved in the first stage supervised training, and the discrepancy of their predictions on the unlabeled image serves as a measurement for reliability. Our image-level selection offers holistic contextual information for learning. We demonstrate that it is more suitable for segmentation than common pixel-wise selection. As a result, ST++ further boosts the performance of our ST. Code is available at https://github.com/LiheYoung/ST-PlusPlus.

1. Introduction

Fully-supervised semantic segmentation learns to assign pixel-wise semantic labels via generalizing from numerous densely annotated images. Despite the rapid progress [9, 60], the pixel-wise manual labeling is costly, laborious, and even infeasible, precluding its deployment in some scenes such as medical image analysis. To avert the labor-intensive procedure, semi-supervised semantic segmentation has been proposed to learn a model from a handful of labeled images along with abundant unlabeled images.

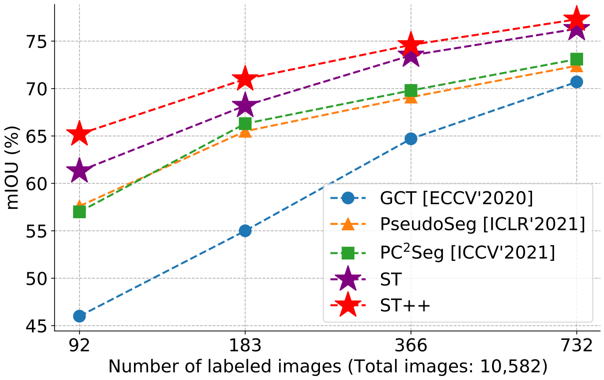

Figure 1. Performance comparisons between our ST/ST++ and the state-of-the-art methods on the Pascal. It is worth noting that the proposed ST and ST++ surpass previous best results significantly, especially in the extremely scarce-label regime, *e.g*., 92 labels.

The core challenge in semi-supervised setting lies in how to effectively utilize the unlabeled images. Prior works in semi-supervised learning (SSL) propose to apply entropy minimization [33, 53] or consistency regularization [21, 52] on unlabeled images. With increasingly sophisticated mechanisms introduced to this field, FixMatch [47] breaks the trend and achieves inspiring results via integrating both strategies into a hybrid framework with few hyper-parameters. Motivated by the tremendous progress in SSL, recent works in semi-supervised semantic segmentation have evolved from GANs-based methods [20, 38] to delving into consistency regularization from the segmentation perspective, such as enforcing consistent predictions of the same unlabeled image under strong-weak perturbations [66], of the same local patch from different contextual crops [31], and of the same unlabeled image between dual differently initialized models [12, 17].

[*] Corresponding author. Work supported by National Key Research and Development Program of China (2019YFC0118300), NSFC Major Program (62192783), CAAI-Huawei MindSpore Project (CAAIXSJLJJ-2021-042A), China Postdoctoral Science Foundation Project (2021M690609), and Jiangsu Natural Science Foundation Project (BK20210224).

Nevertheless, are the delicate mechanisms indispensable for semi-supervised semantic segmentation? More importantly, is the straightforward self-training scheme [33] proposed around a decade ago already out-of-date for this task? In this work, we intuitively and empirically present two simple and effective techniques to bring back the classical self-training method as a strong competitor again.

The self-training [33] is commonly regarded as a form of entropy minimization in SSL, since the re-trained student is supervised with hard labels produced by the teacher which is trained on labeled data. However, they suffer severe coupling issue, *i.e*., making similar predictions on the same input. We notice that, however, injecting strong data augmentations (SDA) on unlabeled images is extremely beneficial to decouple their predictions as well as alleviate overfitting on noisy pseudo labels. Despite the simplicity, self-training with SDA significantly surpasses existing methods without any bells and whistles, *e.g*., without the need of iterative re-training [55], manually choosing a threshold for filtering incorrect labels [66], or repetitively producing pseudo labels for each training minibatch [12, 25, 66]. Inspired by the impressive results, we thoroughly investigate the SDA and find that it is fully compatible with the off-the-shelf augmentation strategies in contrastive learning [10], *e.g*., colorjitter, grayscale, and blur, which deteriorate clean data distribution but perform surprisingly well for unlabeled data. Besides, we examine individual effectiveness of each data augmentation, and observe that the simple colorjitter plays most effectively and different augmentations are complementary to each other. Formally, this basic self-training framework with SDA is named as ST in this work, serving as a strong baseline for our full method.

Another longstanding but underestimated issue is that the classical self-training framework utilizes all unlabeled images at the same time. Nevertheless, different unlabeled images cannot be equally easy [34, 44, 50] and the corresponding pseudo labels cannot be equally reliable, leading to severe confirmation bias [2] and potential performance degradation when iteratively optimizing the model with those ill-posed pseudo labels. To this end, we further propose an advanced ST++ framework based on our ST, that automatically selects and prioritizes more reliable images in the re-training phase to produce higher-quality artificial labels on the remaining less reliable images. The measurement for the reliability or uncertainty of an unlabeled image is to compute the holistic stability of the evolving pseudo masks in different iterations during the entire training course. Note that, different from the common practice of manually setting a fixed confidence threshold to filter low-confidence pixels [66], we demonstrate that our image-level selection based on the stability of evolving predictions can provide holistic contextual regions for model training, which is more appropriate to the segmentation task.

It is worth noting that the classical self-training pipeline [33] is attracting increasing attention [43, 53] in the semi-supervised setting. Our work differs from them in that we empirically and systematically study the effectiveness of strong data augmentations on unlabeled data and further propose an advanced self-training framework with selective re-training property. More concrete differences are discussed in detail in the related work. Our main findings and contributions are summarized as follows:

- We construct a strong baseline (ST) of self-training in semi-supervised semantic segmentation via injecting strong data augmentations on unlabeled images during re-training. Motivated by the promising performance, we provide intuitive explanations and systematically investigate the role of SDA and each augmentation.
- Built on our ST, to alleviate the potential performance degradation incurred by incorrect pseudo labels, we further propose an advanced self-training framework ST++, that performs selective re-training via prioritizing reliable images based on holistic prediction-level stability in the entire training course. We demonstrate that the image-level selection is more suitable for segmentation task compared with pixel-wise selection.
- The ST and ST++ both outperform previous methods across extensive settings and architectures on the Pascal and Cityscapes dataset, with few hyper-parameters.

2. Related Work

Semi-supervised learning. Two main branches of methods are proposed in recent years, namely consistency regularization [4, 21, 32, 39, 45, 51, 52] and entropy minimization [7, 33, 42, 53]. Consistency regularization enforces the current optimized model to yield stable and consistent predictions under various perturbations [45,52], *e.g*., shape and color, on the same unlabeled data. Earlier works also save several checkpoints [32] or maintain a teacher whose parameters are the exponential moving average of the updating student [51] to produce more reliable artificial labels for student model. On the other hand, entropy minimization, popularized by the self-training pipeline [21, 33], leverages unlabeled data in an explicit bootstrapping manner, where unlabeled data is assigned with pseudo labels to be jointly trained with manually labeled data. Different from prior works, MixMatch [6] harvests advantages of both methodologies and proposes a hybrid framework to exploit the unlabeled data from the two perspectives. FixMatch [47] inherits the spirit from MixMatch but simplifies unnecessary mechanisms. As a further extension, the most recent work FlexMatch [58] utilizes the inherent learning status to filter low-confidence labels with class-wise thresholds.

Semi-supervised semantic segmentation. Slightly different from the trend in SSL, preliminary works [26, 38, 49] in

Algorithm 1: ST Pseudocode

```
Input: Labeled training set $\mathcal{D}^l = \{(x_i, y_i)\}_{i=1}^M$,
       Unlabeled training set $\mathcal{D}^u = \{u_i\}_{i=1}^N$,
       Weak/strong augmentations $\mathcal{A}^w/\mathcal{A}^s$,
       Teacher/student model $T/S$
Output: Fully trained student model $S$
Train $T$ on $\mathcal{D}^l$ with cross-entropy loss $\mathcal{L}_{ce}$
Obtain pseudo labeled $\hat{\mathcal{D}}^u = \{(u_i, T(u_i))\}_{i=1}^N$
Over-sample $\mathcal{D}^l$ to around the size of $\hat{\mathcal{D}}^u$
for minibatch $\{(x_k, y_k)\}_{k=1}^B \subset (\mathcal{D}^l \cup \hat{\mathcal{D}}^u)$ do
    for $k \in \{1, \ldots, B\}$ do
        if $x_k \in \mathcal{D}^u$ then
            $x_k, y_k \leftarrow \mathcal{A}^s(\mathcal{A}^w((x_k, y_k))$
        else
            $x_k, y_k \leftarrow \mathcal{A}^w(x_k, y_k)$
        $\hat{y}_k = S(x_k)$
    Update $S$ to minimize $\mathcal{L}_{ce}$ of $\{(\hat{y}_k, y_k)\}_{k=1}^B$
return $S$
```

Algorithm 2: ST++ Pseudocode

```
Input: Same as Algorithm 1
Output: Same as Algorithm 1
Train $T$ on $\mathcal{D}^l$ and save $K$ checkpoints $\{T_j\}_{j=1}^K$
for $u_i \in \mathcal{D}^u$ do
    for $T_j \in \{T_j\}_{j=1}^K$ do
        Pseudo mask $M_{ij} = T_j(u_i)$
    Compute $s_i$ with Equation 4 and $\{M_{ij}\}_{j=1}^K$
Select $R$ highest scored images to compose $\mathcal{D}^{u_1}$
$\mathcal{D}^{u_2} = \mathcal{D}^u - \mathcal{D}^{u_1}$
$\hat{\mathcal{D}}^{u_1} = \{(u_k, T(u_k))\}_{u_k \in \mathcal{D}^{u_1}}$
Train $S$ on $(\mathcal{D}^l \cup \hat{\mathcal{D}}^{u_1})$ with ST re-training
$\hat{\mathcal{D}}^{u_2} = \{(u_k, S(u_k))\}_{u_k \in \mathcal{D}^{u_2}}$
Re-initialize $S$
Train $S$ on $(\mathcal{D}^l \cup \hat{\mathcal{D}}^{u_1} \cup \hat{\mathcal{D}}^{u_2})$ with ST re-training
return $S$
```

semi-supervised semantic segmentation tend to utilize the Generative Adversarial Networks (GANs) [20] as an auxiliary supervision signal for the unlabeled data. However, GANs are not easy to optimize and may suffer the problem of mode collapse [46]. Therefore, also inspired by the success in SSL, subsequent methods [1, 12, 18, 24, 25, 31, 41, 55, 59, 61, 62, 66] manage to tackle this task with simpler mechanisms, such as enforcing similar predictions under multiple perturbed embeddings [41], under two different contextual crops [31], and between dual differently initialized models [12]. As an extension of FixMatch [47], PseudoSeg [66] adapts the weak-to-strong consistency to segmentation scenario and further applies a calibration module to refine the pseudo masks. Despite fancy mechanisms proposed and rapid progress made, nevertheless, we hope to raise a new observation to this field that, the plainest self-training framework coupled with strong data augmentations (SDA) is indeed effective enough to obtain state-of-the-art performance without any bells and whistles, *e.g.*, iterative re-training or setting a threshold to filter unreliable pixels.

Self-training. The self-training via pseudo labeling is an explicit and classical method originating from around a decade ago [33]. Recently, it is increasingly attracting attention from multiple fields, such as fully-supervised image recognition [43, 53, 54, 63], semi-supervised learning [7, 17, 48, 55], and domain adaptation [30, 64, 65]. In the semi-supervised setting, particularly, it has been revisited in several tasks, including image classification [7], object detection [48], and semantic segmentation [17, 55]. Among them, the most related ones to us are [48, 55]. Nevertheless, our work is fundamentally different from [55] in that we demonstrate appropriate SDA on unlabeled data are extremely beneficial to the semi-supervised learner, while [55] designs their method based on the assumption that excessive data augmentations are destructive to clean data distribution. Another work [48] addresses object detection task via manually designing task-relevant augmentations, whereas our SDA is common in image recognition but previously neglected in semi-supervised segmentation. Moreover, both aforementioned works adopt the plain training pipeline, whereas we further propose the ST++ to safely exploit unlabeled images in a curriculum learning manner [5].

Uncertainty estimation. Previous method [29] estimates model uncertainty with a Bayesian analysis. However, limited by the computational burden of Bayesian inference, some other methods adopt Dropout [19, 35] and data augmentations [3] to measure the uncertainty. In the semi-supervised setting, FixMatch [47] simply sets a confidence threshold to filter uncertain samples, and DMT [17] maintains two differently initialized networks to highlight disagreed regions. Compared with them, our method estimates image-level uncertainty via measuring the holistic prediction stability of evolving masks without the need of training extra networks or manually choosing the threshold, making it universal to other scenes. Also, the model learns holistic contextual regions in high-confidence images, which is more stable and appropriate to the segmentation task.

3. Method

3.1. Problem Definition

Semi-supervised semantic segmentation aims to generalize from a combination set of pixel-wise labeled images $\mathcal{D}^l = \{(x_i, y_i)\}_{i=1}^M$ and unlabeled images $\mathcal{D}^u = \{u_i\}_{i=1}^N$, where in most cases $N \gg M$. In most works, the overall optimization target is formalized as:

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

where λ acts as a tradeoff between labeled and unlabeled data. It can be a fixed value [47, 48] or be scheduled during training [27]. The unsupervised loss $\mathcal{L}^u$ is the key point to distinguish different semi-supervised methods, while the supervised loss $\mathcal{L}^s$ is typically the cross-entropy loss between predictions and manually annotated masks.

3.2. Plainest Self-training Scheme

We simplify the plainest form of self-training from [33]. It includes three steps without the need of iterative training:

1. [Supervised Learning] Train a teacher model T on $\mathcal{D}^l$ with cross-entropy loss.
2. [Pseudo Labeling] Predict one-hot hard pseudo labels on $\mathcal{D}^u$ with T to obtain $\hat{\mathcal{D}}^u = \{(u_i, T(u_i))\}_{i=1}^N$.
3. [Re-training] Re-train a student model S on the union set $\mathcal{D}^l \cup \hat{\mathcal{D}}^u$ for final evaluation.

Here, the unsupervised loss $\mathcal{L}^u$ can be formulated as:

$$\mathcal{L}^u_{plain} = \mathrm{H}\big(T(x), S(\mathcal{A}^w(x))\big), \tag{2}$$

where T and S map the image x to the output space. $\mathcal{A}^w$ applies random weak data augmentations to the raw image. H denotes entropy minimization between student and teacher.

Discussion. Since self-training has largely lagged behind consistency based methods in SSL [47], we provide intuitive explanations for the promising performance it might achieve in semantic segmentation. Self-training is deemed to heavily rely on the initial model trained with labeled data, which cannot be well satisfied in the scarce-label regime of SSL. However, the situation is different in our task, since all labeled images are densely annotated and supervised, which means that even only tens of labeled images are available, millions of pixel-level samples can be utilized for training, yielding a well-performed model for pseudo labeling.

3.3. ST: Inject SDA on Unlabeled Images

The above self-training scheme has long been criticized for its ill-posed property that errors in pseudo labels will accumulate and considerably degrade student performance when iteratively overfitting the incorrect supervision. Moreover, in such a bootstrapping process, inadequate information is introduced during re-training, leading to severe coupling issue between the teacher and re-trained student. Concretely, the re-trained S is enforced to learn the pseudo labels from T in a supervised manner. However, considering the same network structure and similar initialization of T and S, they are prone to make similar true or false predictions on the unlabeled images, hence the student S fails to learn extra information except entropy minimization.

In order to break out of the aforementioned two dilemmas, *i.e.*, overfitting noisy labels and prediction coupling between the student and teacher, we propose to inject strong data augmentations (SDA) on unlabeled images during the re-training phase to pose a more challenging optimization target for the student model. The SDA here is named opposite to the weak or basic augmentations adopted in regular fully-supervised semantic segmentation, including random resizing, random cropping and random flipping. As for specific choices of SDA, we manage to maintain a universal strategy across different datasets or settings, rather than searching for the most appropriate ones in each dataset. To simplify the choice, we adopt the off-the-shelf SDA in [10, 11], which includes colorjitter, grayscale, and blur. Apart from these color transformations, we introduce another spatial transformation Cutout [15] to compose our full SDA. In our ST, the unsupervised objective is formalized as:

$$\mathcal{L}^u_{ST} = \mathrm{H}\big(T(x), S\big(\mathcal{A}^s(\mathcal{A}^w(x))\big)\big), \tag{3}$$

where $\mathcal{A}^s$ applies strong data augmentations to the input.

Previous works adjust the impact of labeled and unlabeled data through non-uniform sampling within a minibatch [47] or selecting a hyper-parameter to re-weight the supervised and unsupervised loss [17]. In our ST, we simplify this choice via directly over-sampling $\mathcal{D}^l$ to around the same scale as $\mathcal{D}^u$ and then sampling uniformly from the combined dataset. With this modification, no extra hyper-parameters are introduced and the semi-supervised learner is optimized in a totally fully-supervised fashion. The pseudocode of our ST framework is present in **Algorithm 1** and it works as a strong baseline for our full method.

Discussion. Despite the simplicity of our ST, it surpasses existing state-of-the-art methods even without iterative re-training [55]. Compared with other online methods [12, 25, 66] that repetitively assign pseudo labels for each coming minibatch, our ST annotates unlabeled images only once and the training is conducted in a fully-supervised fashion. Besides, we do not manually fine-tune the choice of SDA. The SDA, harmful to the clean data distribution, but is vital to unlabeled images, please refer to Table 5 for detail.

3.4. ST++: Select and Prioritize Reliable Images

Despite the impressive results obtained by the straightforward ST framework, however, it treats each unlabeled sample equally and leverages them in the same way without considering the inherent reliability and difficulty of individual sample. The incorrect predictions in some hard examples may incur negative impact of the training process. Therefore, in current advanced ST++ framework, we further propose a selective re-training scheme via prioritizing reliable unlabeled samples to safely exploit the whole unlabeled set in an easy-to-hard curriculum learning manner [5].

Previous works estimate the reliability or uncertainty of an image or pixel from different perspectives, such as taking the final softmax output as the confidence distribution

Network	Method	1/16	1/8	1/4	Network	Method	1/16	1/8	1/4
		(662)	(1323)	(2645)			(662)	(1323)	(2645)
PSPNet ResNet-50	SupOnly	63.8	67.2	69.6	DeepLabv3+ ResNet-50	SupOnly	64.8	68.3	70.5
	CCT [41]	62.2	68.8	71.2		ECS [37]	-	70.2	72.6
	DCC [31]	67.1	71.3	72.5		DCC [31]	70.1	72.4	74.0
	ST	***69.1***	***73.0***	***73.2***		**ST**	***71.6***	***73.3***	***75.0***
	ST++	**69.9**	**73.2**	**73.4**		**ST++**	**72.6**	**74.4**	**75.4**
DeepLabv2 ResNet-101	SupOnly	64.3	67.6	69.5	DeepLabv3+ ResNet-101	SupOnly	66.3	70.6	73.1
	AdvSSL [26]	62.6	68.4	69.9		S4GAN [38]	69.1	72.4	74.5
	S4L [57]	61.8	67.2	68.4		GCT [28]	67.2	72.5	75.1
	GCT [28]	65.2	70.6	71.5		DCC [31]	72.4	74.6	76.3
	ST	***68.6***	***71.6***	***72.5***		**ST**	***72.9***	***75.7***	***76.4***
	ST++	**69.3**	**72.0**	**72.8**		**ST++**	**74.5**	**76.3**	**76.6**

Table 1. Results on Pascal VOC. Labeled images are selected from **augmented** training set. The fraction (*e.g.*, 1/16) and number (*e.g.*, 662) denote the proportion and number of labeled images. SupOnly (supervised baseline): no unlabeled data are leveraged, thus the model is only trained with labeled data. The best results are marked in **bold**, while the second best ones are in ***italic bold***.

and filtering low-confidence pixels by pre-defined threshold [47, 63], as well as training two differently initialized models to predict the same unlabeled sample and re-weighting the uncertainty-aware loss with their disagreements [17]. In our ST++, we hope to measure the reliability with a single training model without manually choosing the confidence threshold. And for a more stable evaluation of reliability, we filter out unreliable samples based on image-level information rather than widely adopted pixel-level information. The image-level selection also enables the model to learn more holistic contextual patterns during training.

Specifically, we observe that there is a positive correlation between the segmentation performance and the evolving stability of produced pseudo masks during the supervised training phase. Therefore, the more reliable and better predicted unlabeled images can be selected based on their evolving stability during training. More formally, considering an unlabeled image $u_i \in \mathcal{D}^u$, for K checkpoints $\{T_j\}_{j=1}^{K}$ saved during training, we predict the pseudo masks of u_i with them to obtain $\{M_{ij}\}_{j=1}^{K}$. Since training model tends to converge and achieve the best performance in the late training stage, we evaluate the meanIOU between each earlier pseudo mask and the final mask. The meanIOU can serve as a measurement for stability and further reflect the reliability of the unlabeled image along with pseudo mask:

$$s_i = \sum_{j=1}^{K-1} \texttt{meanIOU}\left(M_{ij}, M_{iK}\right), \tag{4}$$

where s_i is the stability score, reflecting the reliability of u_i.

Obtaining the stability score of all unlabeled images, we sort the whole unlabeled set based on these scores, and select the top R images with the highest scores for the first re-training phase. With a better optimized student model, the remaining unreliable images are re-labeled and the second re-training phase is conducted on the full combination of manually labeled and pseudo labeled data. The pseudocode of our ST++ method is illustrated in **Algorithm 2**.

4. Experiments

4.1. Setup

Dataset. The Pascal VOC 2012 [16] is composed of 1464 images for training and 1449 images for validation originally. And the training set can be augmented via introducing relatively lower-quality annotations from the SBD dataset [22], resulting in 10582 training images. The Cityscapes [13] contains 2975 images with fine-grained masks for training and 500 images for validation.

Network structure. In the past few years, different models and backbones are utilized. In order to conduct a comprehensive comparison, we evaluate four network structures, namely PSPNet [60] with ResNet-50 [23], DeepLabv3+ [9] with ResNet-50/101, and DeepLabv2 [8] with ResNet-101. The DeepLabv2 model is initialized with the MS COCO [36] pre-trained parameters following [28], while the other backbones are pre-trained on ImageNet [14].

Implementation details. We maintain the same hyper-parameters between supervised pre-training of the teacher and semi-supervised re-training of the student. Specifically, the batch size is set as 16 for both Pascal and Cityscapes with two and four NVIDIA V100 GPUs respectively (actually, two 2080Ti GPUs are already enough for all experiments on Pascal). We use the SGD optimizer for training, where the initial base learning rate of the backbones is set as 0.001 on Pascal and 0.004 on Cityscapes. The learning rate of the randomly initialized segmentation head is 10 times larger than that of backbones. We use the poly scheduling to decay the learning rate during the training process: $lr = baselr \times (1 - \frac{iter}{total\_iter})^{0.9}$. The model is trained for 80 epochs on Pascal and 240 epochs on Cityscapes. For

Method	ResNet-50 / ResNet-101		
	1/16 (662)	1/8 (1323)	1/4 (2645)
SupOnly	64.0 / 68.4	69.0 / 73.3	71.7 / 74.7
CCT [41]	65.2 / 67.9	70.9 / 73.0	73.4 / 76.2
CutMix-Seg [18]	68.9 / 72.6	70.7 / 72.7	72.5 / 74.3
GCT [28]	64.1 / 69.8	70.5 / 73.3	73.5 / 75.3
CPS [12]	68.2 / 72.2	73.2 / 75.8	74.2 / 77.6
CPS† [12]	72.0 / ***74.5***	73.7 / 76.4	74.9 / ***77.7***
ST	***72.2*** / 74.0	***74.8*** / ***76.9***	***75.5*** / 77.6
ST++	**73.2 / 74.7**	**75.5 / 77.9**	**76.0 / 77.9**

Table 2. Results on Pascal VOC using a modified ResNet with the deep stem block, following CPS [12]. †: CPS also adopts CutMix [56] to further boost the performance.

weak data augmentations, the training images are randomly flipped and resized between 0.5 and 2.0. They are further cropped to 321x321 on Pascal and 721x721 on Cityscapes. For the strong data augmentations on unlabeled images, we use colorjitter with the same intensity as [66], grayscale, blur same as [11], and Cutout with random values filled. The Cutout regions are ignored in loss computation. In the pseudo labeling phase, all unlabeled images are predicted with test-time augmentation, which contains five scales and horizontal flipping. The testing images are evaluated on their original resolution and no post-processing techniques are adopted. The reliable images are measured with three checkpoints that are evenly saved at 1/3, 2/3, 3/3 total iterations during training. We simply treat the top 50% highest scored images as reliable ones and the remaining ones as unreliable. It is worth noting that, for fair comparison with most existing works, we do not incorporate any advanced optimization strategies, such as OHEM in [25], auxiliary supervision in [12, 25], nor SyncBN into our method. We also choose a relatively smaller cropping size during training to save memory, compared with CPS [12] (321 *vs.* 512 on Pascal and 721 *vs.* 800 on Cityscapes).

4.2. Comparison with State-of-the-Art Methods

Two frameworks based on classical self-training scheme are proposed in this work, namely ST and ST++. In this section, we extensively compare both of our frameworks with previous methods across a variety of datasets and settings.

Pascal VOC 2012. Most previous works uniformly sample labeled images from the augmented training set, which contains 10,582 images in total. As shown in Table 1, our proposed ST already outperforms existing methods remarkably with the four network architectures across extensive settings. Moreover, with the advanced ST++, the performance is further boosted consistently. Besides, the significant margin between our ST/ST++ and the supervised only (SupOnly) results proves our successful and effective exploitation on the abundant unlabeled images. Among most recent methods, CPS [12] adopts a stronger ResNet with the deep stem block, therefore we also modify our backbone to conduct a fair comparison in Table 2.

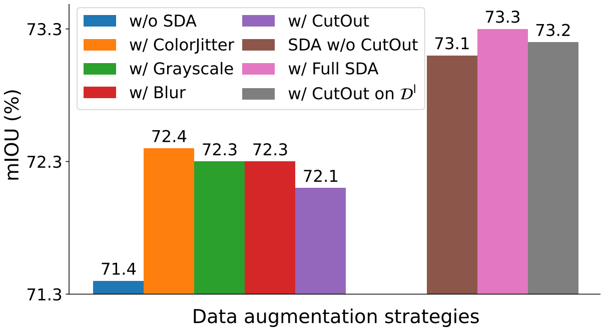

Figure 2. Effectiveness of each data augmentation. All strong data augmentations contributes to the final success. The inferior result of further applying Cutout on labeled images (73.2 *vs.* 73.3) demonstrates that Cutout in SDA works as a strong regularization for unlabeled data rather than a common trick for training.

Method	# Labeled images (Total: 10582)				
	92	183	366	732	1464
SupOnly	50.7	59.1	65.0	70.6	74.1
GCT [28]	46.0	55.0	64.7	70.7	-
CutMix-Seg [18]	55.6	63.2	68.4	69.8	-
PseudoSeg [66]	57.6	65.5	69.1	72.4	73.2
CPS [12]	***64.1***	67.4	71.7	75.9	-
PC^{2}Seg [61]	57.0	66.3	69.8	73.1	74.2
ST	61.3	***68.2***	***73.5***	***76.3***	**78.9**
ST++	**65.2**	**71.0**	**74.6**	**77.3**	**79.1**
Fully-supervised setting (10582 images): 78.2					

Table 3. Results on Pascal VOC using DeepLabv3+ with ResNet-101. Labeled images are selected from **original** training set.

Some recent works sample labeled images from the high-quality original training set. We evaluate our method under this setting in Table 3. Without iterative training adopted as [55], our ST and ST++ framework surpass the previous state-of-the-art methods impressively, even outperform the fully-supervised setting with only 1464 labeled images.

Cityscapes. As shown in Table 4, across a wide range of the number of labeled images, *e.g.*, from 744 to mere 100, both of our methods obtain the state-of-the-art results under a fair comparison with previous methods. It is worth noting that, our method with ResNet-50 backbone even surpasses other works with ResNet-101 backbone by a large margin.

Discussion with previous works. [18] explores the strong, varied perturbations in the semi-supervised semantic segmentation and finds that Cutout and CutMix play an important role in the consistency regularization. Besides, [66] inherits the spirit from FixMatch [47] and applies strong-weak perturbations for consistency regularization. Differ-

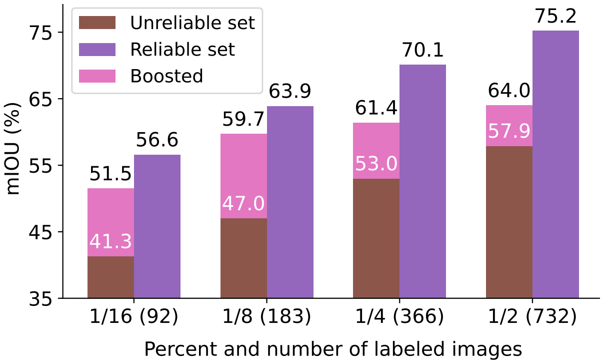

Figure 3. Pseudo mask quality of the reliable and unreliable images selected by ST++. The *Boosted* means the improved mIOU when re-labeling the unreliable images with the model trained on reliable images compared with only trained with labeled images.

Method	1/30 (100)	1/8 (372)	1/4 (744)
	DeepLabv3+, ResNet-101		
DMT [17]	54.8	63.0	-
CutMix-Seg [18]	55.7	65.8	68.3
ClassMix [40]	-	61.4	63.6
PseudoSeg [66]	61.0	69.8	72.4
	DeepLabv3+, ResNet-50		
SupOnly	55.1	65.8	68.4
DCC [31]	-	69.7	72.7
ST	***60.9***	***71.6***	***73.4***
ST++	**61.4**	**72.7**	**73.8**

Table 4. Results on Cityscapes. It is worth noting that our method with ResNet-50 already surpasses others with ResNet-101.

ent from these one-stage works, we find that a simple offline self-training scheme produces more stable and consistent artificial masks. Moreover, coupled with strong data augmentations which are uncommon in supervised scenarios, *e.g.*, colorjitter, grayscale, and blur on unlabeled data, the offline two-stage pipeline can enforce the consistency across various strong perturbations in a broader stage-wise scope, without the limitation from current minibatch. And empirically, according to extensive validations, our proposed ST surpasses all the prior methods impressively.

4.3. Ablation Studies

The main findings and contributions of this work lie in 1) strong data augmentations (SDA) on unlabeled data and 2) selective re-training. In this section, we examine the actual effectiveness of the two components in detail. We conduct our ablation studies with DeepLabv3+ and ResNet-50 on the Pascal VOC. Unless otherwise specified, 1323 (1/8) labeled images are sampled from the augmented training set.

Effectiveness of the SDA in ST. As aforementioned, SDA is composed of colorjitter, grayscale, blur, and Cutout. Here we validate their effectiveness from three perspectives. (1) Firstly, we show the results when no SDA is adopted in Table 5. It can be seen that the model performance degrades across all settings, proving that SDA is vital to the success of self-training. (2) Following this, we further examine the effect of SDA on labeled images. According to the results in Table 5, the labeled images are negatively affected by SDA, indicating that it may deteriorate the clean distribution of labeled images. (3) Finally, in order to gain a better intuition of the individual benefit of the four data augmentations, we add each of them to unlabeled images in Figure 2.

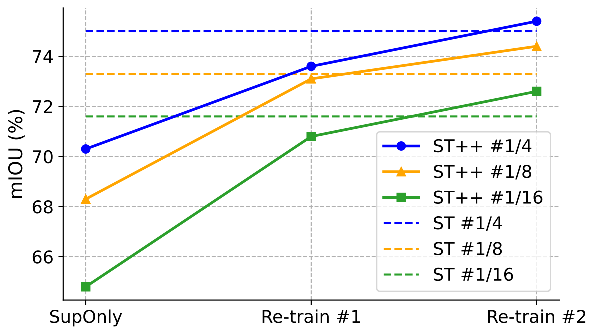

Figure 4. Training process of our proposed ST++. ST++ first trains the model on reliable images (Re-train #1) and then on all re-labeled images (Re-train #2). It can be seen the first stage re-training already obtains competitive results as our ST, further proving that the image-level selection is appropriate and necessary.

Apply SDA on		1/16	1/8	1/4
labeled data	unlabeled data	(662)	(1323)	(2645)
		70.9	71.4	73.5
✓	✓	71.0	73.0	74.3
	✓	**71.6**	**73.3**	**75.0**

Table 5. Effectiveness of full SDA. The first line without applying SDA is the plainest self-training [33]. The best results of only applying SDA on unlabeled data indicates that a more challenging optimization target for unlabeled data is vital to the success. And SDA on labeled data may destroy the clean data distribution.

Besides these, since Cutout is proposed as a common trick in image recognition, we attempt to also apply it to labeled images (gray column in Figure 2). The inferior results prove that our proposed ST benefits from the strong perturbations on unlabeled images rather than common tricks.

Effectiveness of selective re-training in ST++. We first measure the quality of the pseudo masks from reliable and unreliable set respectively. As shown in Figure 3, the mean-IOU gap between pseudo masks from reliable set and unreliable set is all larger than 15%, indicating that it is reasonable to select and prioritize the reliable set. The obtained better student can produce more accurate pseudo masks on the remaining unreliable images. The improvement of un-

Method	1/16 (662)	1/8 (1323)	1/4 (2645)
One-stage re-training (our ST)	71.6	73.3	75.0
Random two-stage re-training	71.3	73.9	74.7
Selective re-training (our ST++)	**72.6**	**74.4**	**75.4**

Table 6. Effectiveness of the selective re-training in ST++. ST++ does not benefit from random two-stage re-training process, but the progressive reliable-to-unreliable selective re-training pipeline.

Proportion of reliable images	25%	50% (default)	75%
mIOU (%)	74.0	74.4	74.5

Table 7. Ablation study on the proportion of reliable images. The default hyper-parameter 50% is effective enough.

reliable image masks is remarkable as the *Boosted* column.

We further check whether our pipeline benefits from the correct selection of reliable unlabeled images or merely the two-stage training pipeline. We randomly divide the whole unlabeled training set into two parts, training with one part first and then re-labeling the remaining ones. The model is finally jointly optimized on the full combination of manually labeled and pseudo labeled images. As shown in Table 6, the semi-supervised model does not benefit from the random two-stage training, in some cases even inferior to its one-stage counterpart. As a comparison, our selective re-training based on image-level stability and reliability consistently outperforms the one-stage re-training pipeline.

Since the re-training phase is conducted in a two-stage manner in our ST++, we examine the improvement after each stage in Figure 4. It can be easily seen that after the first stage, where only half the unlabeled images are exploited, the ST++ already obtains competitive results as ST, revealing the high quality of the selected reliable images.

We also conduct ablation studies on the proportion of selected reliable images. As shown in Table 7, the default hyper-parameter 50% is effective enough. Our ST++ is also robust to other values and the 75% is even slightly better.

Comparison between image-level and pixel-level selective re-training. The contribution in our ST++ is image-level selective re-training, where we decompose the re-training phase into two sub-phases and prioritize reliable images for the first phase re-training. We argue that image-level selection is more stable and provides complete context information for training. Therefore, we compare it with pixel-level selection, where high-confidence pixels are selected for the first phase re-training, and the remaining pixels are re-labeled with a better student for the second phase re-training. Following [63, 66], we set the confident threshold as 0.5. As demonstrated in Table 8, the image-level selection brings consistent improvements over our ST framework and is superior to the pixel-level counterpart.

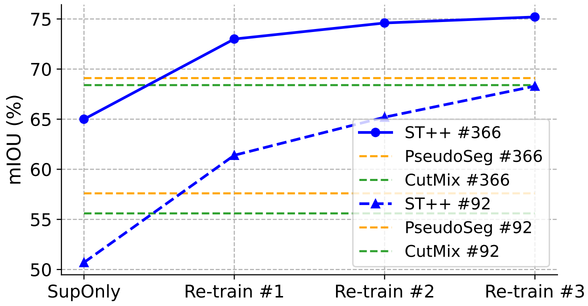

Figure 5. Effectiveness of iterative training. We also show results of PseudoSeg [66] and CutMix-Seg [18] for a clear comparison.

Method	1/16 (662)	1/8 (1323)	1/4 (2645)
our ST (w/o iterative re-train)	71.6	73.3	75.0
Image-level re-train (our ST++)	**72.6**	**74.4**	**75.4**
Pixel-level re-train phase #1	71.3	73.5	74.9
Pixel-level re-train phase #2	71.3	73.8	74.7

Table 8. Comparison between image-level (our ST++) and common pixel-level selective re-training (setting a threshold).

Effectiveness of iterative training. As aforementioned, for simplicity and efficiency, we choose not to conduct iterative training. However, it is possible to further boost the final performance via switching the teacher-student role and re-labeling unlabeled images for several times. Therefore, we examine the effectiveness of iterative training in our ST++ in Figure 5. An extra stage (Re-train #3 in the figure) is conducted, where all the unlabeled images are re-labeled with the best learned model in the second re-training stage and the student is re-trained. In the extremely scarce-label regime of only 92 and 366 labeled images, with the extra re-training stage, the performance can be further boosted from 65.2% to 68.3% and 74.6% to 75.2% respectively.

5. Conclusion

In this work, we firstly construct a strong self-training baseline for semi-supervised semantic segmentation via introducing strong data augmentations to unlabeled images, in hope of alleviating overfitting noisy labels as well as decoupling similar predictions between the teacher and student. Moreover, an advanced framework is proposed to progressively leverage the unlabeled images. With extensive experiments conducted across a variety of benchmarks and settings, both of our ST and ST++ framework outperform previous methods by a large margin. Based on the inspiring results, we further examine the effectiveness of each component in detail and provide some empirical analysis. We hope this simple yet effective framework can serve as a strong baseline or competitor for future works in this field.

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 *ICCV*, 2021. 3

[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 *IJCNN*, 2020. 2

[3] Murat Seckin Ayhan and Philipp Berens. Test-time data augmentation for estimation of heteroscedastic aleatoric uncertainty in deep neural networks. In *MIDL*, 2018. 3

[4] Philip Bachman, Ouais Alsharif, and Doina Precup. Learning with pseudo-ensembles. In *NeurIPS*, 2014. 2

[5] Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston. Curriculum learning. In *ICML*, 2009. 3, 4

[6] David Berthelot, Nicholas Carlini, Ian Goodfellow, Nicolas Papernot, Avital Oliver, and Colin Raffel. Mixmatch: A holistic approach to semi-supervised learning. In *NeurIPS*, 2019. 2

[7] Paola Cascante-Bonilla, Fuwen Tan, Yanjun Qi, and Vicente Ordonez. Curriculum labeling: Revisiting pseudo-labeling for semi-supervised learning. In *AAAI*, 2021. 2, 3

[8] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *TPAMI*, 2017. 5

[9] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *ECCV*, 2018. 1, 5

[10] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *ICML*, 2020. 2, 4

[11] Xinlei Chen, Haoqi Fan, Ross Girshick, and Kaiming He. Improved baselines with momentum contrastive learning. *arXiv:2003.04297*, 2020. 4, 6

[12] Xiaokang Chen, Yuhui Yuan, Gang Zeng, and Jingdong Wang. Semi-supervised semantic segmentation with cross pseudo supervision. In *CVPR*, 2021. 1, 2, 3, 4, 6

[13] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *CVPR*, 2016. 5

[14] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, 2009. 5

[15] Terrance DeVries and Graham W Taylor. Improved regularization of convolutional neural networks with cutout. *arXiv:1708.04552*, 2017. 4

[16] Mark Everingham, SM Ali Eslami, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes challenge: A retrospective. *IJCV*, 2015. 5

[17] Zhengyang Feng, Qianyu Zhou, Qiqi Gu, Xin Tan, Guangliang Cheng, Xuequan Lu, Jianping Shi, and Lizhuang Ma. Dmt: Dynamic mutual training for semi-supervised learning. *arXiv:2004.08514*, 2020. 1, 3, 4, 5, 7

[18] Geoff French, Timo Aila, Samuli Laine, Michal Mackiewicz, and Graham Finlayson. Semi-supervised semantic segmentation needs strong, high-dimensional perturbations. In *BMVC*, 2020. 3, 6, 7, 8

[19] Yarin Gal and Zoubin Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *ICML*, 2016. 3

[20] Ian J Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. In *NeurIPS*, 2014. 1, 3

[21] Yves Grandvalet and Yoshua Bengio. Semi-supervised learning by entropy minimization. In *NeurIPS*, 2005. 1, 2

[22] Bharath Hariharan, Pablo Arbeláez, Lubomir Bourdev, Subhransu Maji, and Jitendra Malik. Semantic contours from inverse detectors. In *ICCV*, 2011. 5

[23] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 5

[24] Ruifei He, Jihan Yang, and Xiaojuan Qi. Re-distributing biased pseudo labels for semi-supervised semantic segmentation: A baseline investigation. In *ICCV*, 2021. 3

[25] Hanzhe Hu, Fangyun Wei, Han Hu, Qiwei Ye, Jinshi Cui, and Liwei Wang. Semi-supervised semantic segmentation via adaptive equalization learning. In *NeurIPS*, 2021. 2, 3, 4, 6

[26] Wei-Chih Hung, Yi-Hsuan Tsai, Yan-Ting Liou, Yen-Yu Lin, and Ming-Hsuan Yang. Adversarial learning for semi-supervised semantic segmentation. In *BMVC*, 2018. 2, 5

[27] Jisoo Jeong, Seungeui Lee, Jeesoo Kim, and Nojun Kwak. Consistency-based semi-supervised learning for object detection. In *NeurIPS*, 2019. 4

[28] Zhanghan Ke, Kaican Li Di Qiu, Qiong Yan, and Rynson WH Lau. Guided collaborative training for pixel-wise semi-supervised learning. In *ECCV*, 2020. 5, 6

[29] Alex Kendall and Yarin Gal. What uncertainties do we need in bayesian deep learning for computer vision? In *NeurIPS*, 2017. 3

[30] Ananya Kumar, Tengyu Ma, and Percy Liang. Understanding self-training for gradual domain adaptation. In *ICML*, 2020. 3

[31] 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 *CVPR*, 2021. 1, 3, 5, 7

[32] Samuli Laine and Timo Aila. Temporal ensembling for semi-supervised learning. In *ICLR*, 2017. 2

[33] Dong-Hyun Lee et al. Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks. In *ICML Workshop*, 2013. 1, 2, 3, 4, 7

[34] Xiaoxiao Li, Ziwei Liu, Ping Luo, Chen Change Loy, and Xiaoou Tang. Not all pixels are equal: Difficulty-aware semantic segmentation via deep layer cascade. In *CVPR*, 2017. 2

[35] Xiaobo Liang, Lijun Wu, Juntao Li, Yue Wang, Qi Meng, Tao Qin, Wei Chen, Min Zhang, and Tie-Yan Liu. R-drop: Regularized dropout for neural networks. *arXiv:2106.14448*, 2021. 3

[36] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *ECCV*, 2014. 5

[37] Robert Mendel, Luis Antonio de Souza, David Rauber, João Paulo Papa, and Christoph Palm. Semi-supervised segmentation based on error-correcting supervision. In *ECCV*, 2020. 5

[38] Sudhanshu Mittal, Maxim Tatarchenko, and Thomas Brox. Semi-supervised semantic segmentation with high-and low-level consistency. *TPAMI*, 2019. 1, 2, 5

[39] Takeru Miyato, Shin-ichi Maeda, Masanori Koyama, and Shin Ishii. Virtual adversarial training: a regularization method for supervised and semi-supervised learning. *TPAMI*, 2018. 2

[40] Viktor Olsson, Wilhelm Tranheden, Juliano Pinto, and Lennart Svensson. Classmix: Segmentation-based data augmentation for semi-supervised learning. In *WACV*, 2021. 7

[41] Yassine Ouali, Céline Hudelot, and Myriam Tami. Semi-supervised semantic segmentation with cross-consistency training. In *CVPR*, 2020. 3, 5, 6

[42] Hieu Pham, Zihang Dai, Qizhe Xie, and Quoc V Le. Meta pseudo labels. In *CVPR*, 2021. 2

[43] Ilija Radosavovic, Piotr Dollár, Ross Girshick, Georgia Gkioxari, and Kaiming He. Data distillation: Towards omni-supervised learning. In *CVPR*, 2018. 2, 3

[44] Zhongzheng Ren, Raymond Yeh, and Alexander Schwing. Not all unlabeled data are equal: learning to weight data in semi-supervised learning. In *NeurIPS*, 2020. 2

[45] Mehdi Sajjadi, Mehran Javanmardi, and Tolga Tasdizen. Regularization with stochastic transformations and perturbations for deep semi-supervised learning. In *NeurIPS*, 2016. 2

[46] Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training gans. In *NeurIPS*, 2016. 3

[47] Kihyuk Sohn, David Berthelot, Chun-Liang Li, Zizhao Zhang, Nicholas Carlini, Ekin D Cubuk, Alex Kurakin, Han Zhang, and Colin Raffel. Fixmatch: Simplifying semi-supervised learning with consistency and confidence. In *NeurIPS*, 2020. 1, 2, 3, 4, 5, 6

[48] Kihyuk Sohn, Zizhao Zhang, Chun-Liang Li, Han Zhang, Chen-Yu Lee, and Tomas Pfister. A simple semi-supervised learning framework for object detection. *arXiv:2005.04757*, 2020. 3, 4

[49] Nasim Souly, Concetto Spampinato, and Mubarak Shah. Semi supervised semantic segmentation using generative adversarial network. In *ICCV*, 2017. 2

[50] Ruoqi Sun, Xinge Zhu, Chongruo Wu, Chen Huang, Jianping Shi, and Lizhuang Ma. Not all areas are equal: Transfer learning for semantic segmentation via hierarchical region selection. In *CVPR*, 2019. 2

[51] Antti Tarvainen and Harri Valpola. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. In *NeurIPS*, 2017. 2

[52] Qizhe Xie, Zihang Dai, Eduard Hovy, Minh-Thang Luong, and Quoc V Le. Unsupervised data augmentation for consistency training. In *NeurIPS*, 2020. 1, 2

[53] Qizhe Xie, Minh-Thang Luong, Eduard Hovy, and Quoc V Le. Self-training with noisy student improves imagenet classification. In *CVPR*, 2020. 1, 2, 3

[54] I Zeki Yalniz, Hervé Jégou, Kan Chen, Manohar Paluri, and Dhruv Mahajan. Billion-scale semi-supervised learning for image classification. *arXiv:1905.00546*, 2019. 3

[55] Jianlong Yuan, Yifan Liu, Chunhua Shen, Zhibin Wang, and Hao Li. A simple baseline for semi-supervised semantic segmentation with strong data augmentation. In *ICCV*, 2021. 2, 3, 4, 6

[56] 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 *ICCV*, 2019. 6

[57] Xiaohua Zhai, Avital Oliver, Alexander Kolesnikov, and Lucas Beyer. S4l: Self-supervised semi-supervised learning. In *ICCV*, 2019. 5

[58] Bowen Zhang, Yidong Wang, Wenxin Hou, Hao Wu, Jindong Wang, Manabu Okumura, and Takahiro Shinozaki. Flexmatch: Boosting semi-supervised learning with curriculum pseudo labeling. In *NeurIPS*, 2021. 2

[59] Pan Zhang, Bo Zhang, Ting Zhang, Dong Chen, and Fang Wen. Robust mutual learning for semi-supervised semantic segmentation. *arXiv:2106.00609*, 2021. 3

[60] Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia. Pyramid scene parsing network. In *CVPR*, 2017. 1, 5

[61] Yuanyi Zhong, Bodi Yuan, Hong Wu, Zhiqiang Yuan, Jian Peng, and Yu-Xiong Wang. Pixel contrastive-consistent semi-supervised semantic segmentation. In *ICCV*, 2021. 3, 6

[62] Yanning Zhou, Hang Xu, Wei Zhang, Bin Gao, and Pheng-Ann Heng. C3-semiseg: Contrastive semi-supervised segmentation via cross-set learning and dynamic class-balancing. In *ICCV*, 2021. 3

[63] Barret Zoph, Golnaz Ghiasi, Tsung-Yi Lin, Yin Cui, Hanxiao Liu, Ekin D Cubuk, and Quoc V Le. Rethinking pre-training and self-training. In *NeurIPS*, 2020. 3, 5, 8

[64] Yang Zou, Zhiding Yu, BVK Kumar, and Jinsong Wang. Unsupervised domain adaptation for semantic segmentation via class-balanced self-training. In *ECCV*, 2018. 3

[65] Yang Zou, Zhiding Yu, Xiaofeng Liu, BVK Kumar, and Jinsong Wang. Confidence regularized self-training. In *CVPR*, 2019. 3

[66] 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 *ICLR*, 2021. 1, 2, 3, 4, 6, 7, 8