HagSeg: Hardness-adaptive Guidance for Semi-supervised Semantic Segmentation

Anonymous authors
Paper under double-blind review

Abstract

Recently, semi-supervised semantic segmentation has achieved promising performance with a small fraction of labeled data. However, most existing studies treat all unlabeled data equally and barely consider the differences and training difficulties among unlabeled instances. Differentiating unlabeled instances can promote instance-specific supervision to adapt to the model's evolution dynamically. In this paper, we emphasize the cruciality of instance differences and propose an instance-specific and hardness-adaptive guidance for semi-supervised semantic segmentation, named as HagSeg. Relying on the model's performance, HagSeg employs the class-weighted symmetric intersection-over-union to evaluate the hardness of each unlabeled instance and then supervises the training on unlabeled data in a hardness-adaptive manner. Specifically, HagSeg learns from unlabeled instances progressively by weighing their corresponding consistency losses based on the evaluated hardness. Meanwhile, HagSeg dynamically adjusts the augmentation for each instance such that the distortion degree of augmented instances is adapted to the model's generalization capability across the training course. Not integrating additional losses and training procedures, HagSeg can obtain remarkable performance gains against current state-of-the-art approaches on segmentation benchmarks under different semi-supervised partition protocols.

1 Introduction

Though semantic segmentation studies (Long et al., 2015; Chen et al., 2018) have achieved significant progress, their enormous success relies on large datasets with high-quality pixel-level annotations. Semi-supervised semantic segmentation (SSS) (Hung et al., 2018; Mittal et al., 2019) has been proposed as a powerful solution to mitigate the requirement for labeled data. Recent research on SSS has two main branches, including the self-training (ST) (Lee et al., 2013) and consistency regularization (CR) (Tarvainen & Valpola, 2017) based approaches. Yang et al. (2022) follows a self-training paradigm and performs a selective re-training scheme to train on labeled and unlabeled data alternatively. Differently, CR-based works (Ouali et al., 2020b; Liu et al., 2022) tend to apply data or model perturbations and enforce the prediction consistency between two differently-perturbed views for unlabeled data. In both branches, recent research (French et al., 2020; Yuan et al., 2021; Hu et al., 2021) demonstrates that strong data perturbations like CutMix can significantly benefit the SSS training. To further improve the SSS performance, current state-of-the-art approaches (Alonso et al., 2021; Wang et al., 2022) integrate the advanced contrastive learning techniques into the CR-based approach to exploit the unlabeled data more efficiently. Works in (Ibrahim et al., 2020; Kwon & Kwak, 2022) aim to rectify the pseudo-labels through training an additional correcting network.

Despite their promising performance, SSS studies along this line come at the **cost** of introducing extra network components or additional training procedures. What's worse, majorities of them treat unlabeled data equally and completely ignore the differences and learning difficulties among unlabeled samples. For instance, **randomly and indiscriminately** perturbing unlabeled data can inevitably over-perturb some hard-to-train instances. Such over-perturbations exceed the generalization capability of the model and hinder effective learning from unlabeled data. As discussed in (Yuan et al., 2021), it may also hurt the data distribution. In addition, in most SSS studies, final consistency losses on different unlabeled instances are minimized in an **average** manner. However, blindly averaging can implicitly emphasize some hard-to-train instances and result in model overfitting to noisy supervision.

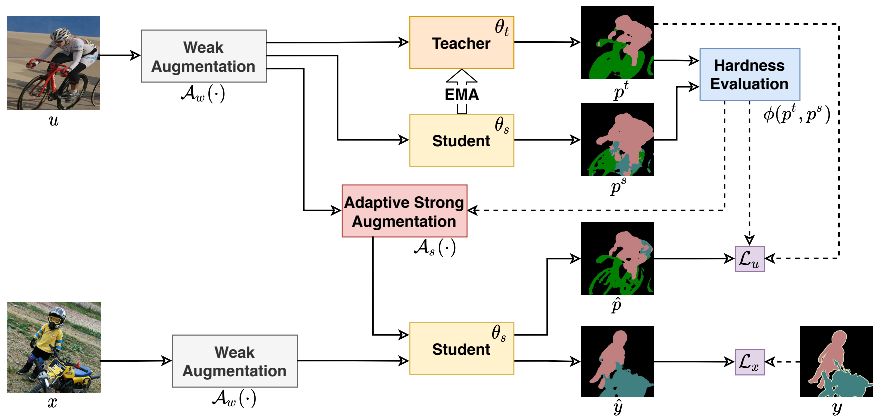


Figure 1: Diagram of our proposed HagSeg. In a teacher-student framework, labeled data (x, y) is used to train the student model, parameterized by θ_s, by minimizing the supervised loss $\mathcal{L}_x$. Unlabeled data u, weakly augmented by $\mathcal{A}_w(\cdot)$, is first fed into both the student and teacher models to obtain predictions p^s and p^t, respectively. Then we evaluate the hardness of each unlabeled instance by strategy $\phi(p^t, p^s)$. Such hardness information can be subsequently utilized: 1) to apply an adaptive augmentation, denoted by $\mathcal{A}_s(\cdot)$, on unlabeled data to obtain the student model's prediction $\hat{p}$; 2) to weigh the unsupervised loss $\mathcal{L}_u$ in a hardness-dependent manner. The teacher model's weight, θ_t, is updated by the exponential moving average of θ_s across the training course.

In this paper, we emphasize the cruciality of instance differences and differentiate unlabeled data in terms of instance hardness. First, the hardness, as a measure of the training difficulty, can vary widely between different instances and between different training epochs of the same instance. Its evaluation is closely related to the training status of the model, *e.g.,* a hard-to-train sample can become easier with the evolution of the model. Second, injecting such information into the SSS procedure is advantageous for processing unlabeled data in a more reasonable and discriminative way. Since the hardness is assessed based on the model's performance, we can leverage such information to adjust the two critical operations in SSS, *i.e.,* data perturbations and loss evaluations, to adapt to the training state of the model dynamically.

Motivated by these observations, we propose an **instance-specific and hardness-adaptive guidance** for semi-supervised semantic segmentation (**HagSeg**) to boost the SSS performance. As shown in Figure 1, following a standard consistency regularization framework, HagSeg jointly trains the student and teacher models in a mutually-beneficial manner. The teacher model is an ensemble of historical student models and generates stable pseudo-labels for unlabeled data. Intuitively, hard-to-train instances undergo considerable disagreement between predictions of the teacher and student models. Thus in HagSeg, we first evaluate the instance hardness of each unlabeled sample by calculating the class-weighted symmetric intersection-over-union (IoU) between the segmentation predictions of the student and teacher models. Then based on the evaluation, we perform hardness-adaptive data perturbations on each unlabeled instance and minimize an instance-specific weighted consistency loss to train models in a curriculum-like manner. In this way, different unlabeled instances are perturbed and weighted in a dynamic fashion, which can better adapt to the model's generalization capability throughout the training processes.

Benefiting from the instance-specific and hardness-adaptive design, HagSeg obtains state-of-the-art (SOTA) performance on Pascal VOC 2012 and Cityscapes datasets under different partition protocols. For example, our method obtains a high mIOU of 75.3% with only 183 labeled data on VOC 2012, which is 17.8% higher than the supervised baseline and 4.3% higher than previous SOTA performance. Our main contributions are summarized as follows,

- Hagseg can boost the SSS performance by highlighting the instance differences, without introducing extra network components or training losses.
- We design a hardness-evaluating strategy for unlabeled instances in segmentation tasks, based on the class-weighted teacher-student symmetric IoU.
- We propose an instance-specific and hardness-adaptive SSS framework that injects instance hardness into data perturbation and loss evaluation to dynamically adapt to the model's evolution.

2 HagSeg

The goal of semi-supervised semantic segmentation is to generalize a segmentation model by effectively leveraging a labeled training set $D_x = \{(x_i, y_i)\}_{i=1}^{|D_x|}$ and a large unlabeled training set $D_u = \{u_i\}_{i=1}^{|D_u|}$, with typically $|D_x| \ll |D_u|$. In our method, following the consistency regularization (CR) based semi-supervised classification approaches (Sohn et al., 2020; Xie et al., 2020), we aim to train the segmentation encoder and decoder on both labeled and unlabeled data simultaneously. In each iteration, given a batch of labeled samples $\mathcal{B}_x = \{(x_i, y_i)\}_{i=1}^{|\mathcal{B}_x|}$ and unlabeled samples $\mathcal{B}_u = \{u_i\}_{i=1}^{|\mathcal{B}_u|}$, the overall training loss can be formulated as,

$$\mathcal{L} = \mathcal{L}_x + \lambda_u \mathcal{L}_u, \tag{1}$$

where λ_u is a scalar hyper-parameter to adjust the relative importance between the supervised loss $\mathcal{L}_x$ on $\mathcal{B}_x$ and the unsupervised loss $\mathcal{L}_u$ on $\mathcal{B}_u$. Without introducing extra losses or network components, HagSeg can effectively evaluate the instance hardness and then supervise the training on unlabeled data in a hardness-adaptive fashion across the training course. In this section, we first introduce our proposed HagSeg at a high level in Sec. 2.1 and then present the detailed designs in terms of the hardness evaluation in Sec. 2.2 and the hardness-adaptive guidance in Sec. 2.3.

2.1 Overview

Built on top of the CR-based semi-supervised framework, HagSeg jointly trains a student model with learnable weights θ_s and a teacher model with learnable weights θ_t in a mutually-beneficial manner. The complete algorithm is shown in algorithm 1. On the one hand, the teacher model is updated by the exponential moving averaging of the student weights, *i.e.,*

$$\theta_t \leftarrow \alpha\theta_t + (1-\alpha)\theta_s, \tag{2}$$

where α is a common momentum parameter, set as 0.996 by default. On the other hand, the student model relies on the pseudo-labels generated by the teacher model to be trained on the unlabeled data. Specifically, the student model is trained via minimizing the total loss $\mathcal{L}$ in Equation 1, which consists of two cross-entropy loss terms, $\mathcal{L}_u$ and $\mathcal{L}_x$, applied on labeled and unlabeled data, respectively. Let $\mathrm{H}(z_1, z_2)$ denote the cross-entropy loss between prediction distributions z_1 and z_2. The supervised loss $\mathcal{L}_x$ is calculated as,

$$\mathcal{L}_x = \frac{1}{|\mathcal{B}_x|} \sum_{i=1}^{|\mathcal{B}_x|} \frac{1}{H \times W} \sum_{j=1}^{H \times W} \mathrm{H}(\hat{y}_i(j), y_i(j)), \tag{3}$$

where $\hat{y}_i = f_{\theta_s}(\mathcal{A}_w(x_i))$, represents the segmentation result of the student model on the i-th weakly-augmented labeled instance. j represents the j-th pixel on the image or the corresponding segmentation mask with a resolution of $H \times W$. The weak augmentation $\mathcal{A}_w$ includes standard resizing, cropping, and flipping operations. Importantly, the way to leverage the unlabeled data is the key to semi-supervised learning and also the crucial part differentiating our method from others. In most CR-based studies, the standard (std) unsupervised loss $\mathcal{L}_u^{std}$ is simply,

$$\mathcal{L}_u^{std} = \frac{1}{|\mathcal{B}_u|} \sum_{i=1}^{|\mathcal{B}_u|} \frac{1}{H \times W} \sum_{j=1}^{H \times W} \mathbb{1}(\max(p_i^t(j)) \geq \tau) \mathrm{H}(\hat{p}_i(j), p_i^t(j)), \tag{4}$$

where $\hat{p}_i = f_{\theta_s}(\mathcal{A}_s^{std}(u_i))$ represents the segmentation output of the student model on the i-th unlabeled instance augmented by $\mathcal{A}_s^{std}$, while $p_i^t = f_{\theta_t}(\mathcal{A}_w(u_i))$ represents the segmentation outputs of the teacher model on the i-th weakly-augmented unlabeled instance. τ is a predefined confidence threshold to select high-confidence predictions. $\mathcal{A}_s^{std}$ represents standard **instance-agnostic** strong augmentations, including intensity-based data augmentations (Cubuk et al., 2020) and CutMix (Yun et al., 2019) as shown in Table 5 of the appendix. However, such operations are limited in ignoring the differences and learning difficulties among unlabeled samples.

Differently, in our proposed HagSeg, we treat each instance discriminatively and provide instance-specific supervision on the training of unlabeled data. As shown in Figure 1, we first evaluate the hardness of each weakly-augmented unlabeled instance via strategy ϕ, and then employ the **instance-specific and hardness-adaptive** guidance on the strong augmentations $\mathcal{A}_s$ as well as the calculations of unsupervised loss $\mathcal{L}_u$, which are elaborated in following sections.

Algorithm 1 HagSeg algorithm in a mini-batch.

Input: Labeled batch $\mathcal{B}_x = \{(x_i, y_i)\}_{i=1}^{|\mathcal{B}_x|}$, unlabeled batch $\mathcal{B}_u = \{u_i\}_{i=1}^{|\mathcal{B}_u|}$ ($|\mathcal{B}_x| = |\mathcal{B}_u|$), hardness evaluation strategy ϕ, weak augmentation $\mathcal{A}_w(\cdot)$, adaptive strong augmentation $\mathcal{A}_s(\cdot)$
Parameter: confidence threshold τ, unsupervised loss weight λ_u

```
1: L_x = 1/|B_x| Σ_{i=1}^{|B_x|} 1/(H×W) Σ_{j=1}^{H×W} H(ŷ_i(j), y_i(j))    // calculate the supervised loss.
2: for u_i ∈ B_u do
3:     p_i^s = f_{θ_s}(A_w(u_i))    // obtain segmentation predictions on weakly-augmented instances.
4:     p_i^t = f_{θ_t}(A_w(u_i))    // obtain pseudo-labels from the teacher model.
5:     γ_i = φ(p_i^t, p_i^s)    // evaluate the hardness of each instance.
6: end for
7: L_u = 1/|B_u| Σ_{i=1}^{|B_u|} γ_i/(2H×W) Σ_{j=1}^{H×W} [1(max(p_i^t(j)) ≥ τ)H(f_{θ_s}(A_s^I(u_i)), p_i^t(j))+
   1(max(p_i^{t'}(j)) ≥ τ)H(f_{θ_s}(A_s^C(u_i)), p_i^{t'}(j))] //calculate hardness-adaptive consistency loss
8: return L = L_x + λ_u L_u
```

2.2 Hardness evaluation of unlabeled instances

In HagSeg, we evaluate the instance hardness to differentiate different unlabeled data. In most hardness-related studies, the instantaneous or historical training losses Zhou et al. (2020); Smith et al. (2014) to the ground truth are used to assess the instance hardness. However, in semi-supervised segmentation, evaluating the hardness of unlabeled data is challenging at 1) lacking accurate ground-truth labels and 2) dynamic changes closely related to the model performance. A "hard" sample can become easier with the evolution of the model, but such dynamics cannot be easily identified without accurate label information. Since it is more difficult for the teacher and student models to achieve consensus on a hard instance, we design a symmetric class-weighted IoU between the segmentation results of the student and teacher models to evaluate the instantaneous hardness. Such evaluation, denoted by ϕ, can be regarded as a function of the model performance and dynamically estimate the training difficulties of unlabeled crops throughout the training process.

Specifically, as shown in Figure 1, we first obtain the segmentation predictions p_i^s and p_i^t on the i-th weakly-augmented unlabeled instance, from the student and teacher models, respectively,

$$p_i^s = f_{\theta_s}(\mathcal{A}_w(u_i)), \quad \rho_i^s = \frac{1}{H \times W} \sum_{j=1}^{H\times W} \mathbb{1}(\max(p_i^s(j)) \geq \tau) \tag{5}$$

$$p_i^t = f_{\theta_t}(\mathcal{A}_w(u_i)), \quad \rho_i^t = \frac{1}{H \times W} \sum_{j=1}^{H\times W} \mathbb{1}(\max(p_i^t(j)) \geq \tau), \tag{6}$$

where ρ_i^s and ρ_i^t represent the high-confidence ratios on p_i^s and p_i^t, respectively. Let $\text{wIOU}(z_1, z_2)$ denote the class-weighted IoU between segmentation predictions z_1 and z_2. Note that, this evaluation is not commutative, *i.e.,* $\text{wIOU}(z_1, z_2) \neq \text{wIOU}(z_2, z_1)$. To make wIoU valid for hardness evaluation at each iteration, the symmetric hardness γ_i for i-th unlabeled instance is calculated as,

$$\gamma_i = \phi(p_i^t, p_i^s) = 1 - [\frac{\rho_i^s}{2}\text{wIOU}(p_i^s, p_i^t) + \frac{\rho_i^t}{2}\text{wIOU}(p_i^t, p_i^s)] \tag{7}$$

where $1/2$ ensures the hardness is in $[0, 1]$. In this way, the hard instance that requires better generalization ability will hold a large value of γ while the easy one will be identified by a small γ.

2.3 Hardness-adaptive guidance

With the estimated hardness for each unlabeled instance, we carefully inject such information into the training process by instance-specific hardness-adaptive strong perturbations and loss modifications. Specifically, we first leverage the instance hardness for adaptive augmentations both individually and mutually. By "individually", we adjust the intensity-based augmentation applied on each instance according to its absolute hardness value; by "mutually", we replace random pairs of unlabeled data in CutMix with specific **hard-easy pairs** assigned by sorting the corresponding hardness. Moreover, instead of indiscriminately averaging the losses, we **weigh** the losses of different unlabeled instances by multiplying their corresponding hardness. We present these details below.

2.3.1 HARDNESS-ADAPTIVE STRONG AUGMENTATIONS

The popular strong augmentations in recent semi-supervised segmentation studies mainly consist of two different types: intensity-based augmentation and CutMix, as shown in Table 5 of the appendix. In HagSeg, we apply instance-specific adjustments to both types of augmentations.

Intensity-based augmentations. Standard intensity-based data augmentations randomly select two kinds of image operations from an augmentation pool and apply them to the weakly-augmented instances. However, as discussed by Yuan et al. (2021), strong augmentations may hurt the data distribution and degrade the segmentation performance, especially during the early training phase. Unlike distribution-specific designs (Yuan et al., 2021), we simply adjust the augmentation degree for an unlabeled instance by mixing its strongly-augmented and weakly-augmented outputs. Formally, the ultimate augmented output of the i-th unlabeled instance, $\mathcal{A}_s^I(u_i)$, can be obtained by,

$$\mathcal{A}_s^I(u_i) \leftarrow \gamma_i \mathcal{A}_s^I(u_i) + (1-\gamma_i)\mathcal{A}_w(u_i) \tag{8}$$

where the distortion caused by the intensity-based strong augmentation is proportionally weakened by the corresponding weakly-augmented output. In this way, hard instances with large hardness are not perturbed significantly so that the model will not be challenged on potentially out-of-distribution cases. On the other hand, easier instances with lower values of γ, which have been well fitted by the model, can be further learned from their strongly-augmented variants. Such hardness-adaptive augmentations can better adjust to the model's generalization ability.

CutMix-based augmentations. CutMix (Yun et al., 2019) is a widely adopted technique to boost semi-supervised semantic segmentation. It is applied between unlabeled instances with a predefined probability. It can randomly copy a region from one instance to another, and so do their corresponding segmentation results. The augmentation pairs are generated randomly. Differently, in HagSeg, we improve the standard CutMix by a hardness-adaptive design, which is distinct in two ways: **1)** the mean hardness determines the trigger probability of CutMix augmentation over the mini-batch instead of using a predefined hyper-parameter; **2)** the copy-and-paste pairs are assigned specifically between the hard and easy samples. According to the instance hardness, we obtain two sequences by sorting unlabeled samples of a mini-batch in the ascending and descending orders, respectively. We then aggregate two sequences element-by-element to generate the hard-easy pairs. Formally, given a specific hard-easy pair, (u_m, u_n), the hardness-adaptive CutMix can be expressed as,

$$\left.\begin{aligned} \mathcal{A}_s^C(u_m) &\leftarrow M_m \odot u_n + (\mathbf{1}-M_m)\odot u_m \\ p_m^{t'} &\leftarrow M_m \odot p_n^t + (\mathbf{1}-M_m)\odot p^t, \\ \mathcal{A}_s^C(u_n) &\leftarrow M_n \odot u_m + (\mathbf{1}-M_n)\odot u_n \\ p_n^{t'} &\leftarrow M_n \odot p_m^t + (\mathbf{1}-M_n)\odot p_n^t \end{aligned}\right\}\text{, by a trigger probability of } \overline{\gamma} = \frac{1}{|\mathcal{B}_u|}\sum_{n=1}^{|\mathcal{B}_u|}\gamma_n \tag{9}$$

where M_m and M_n denote the randomly generated region masks for u_m and u_n, respectively. Besides, the pseudo-labels need to be revised accordingly after applying CutMix data augmentations, obtaining $p_m^{t'}$ and $p_n^{t'}$. This mutual augmentation is applied following a Bernoulli process, *i.e.,* triggered only when a uniformly random probability is higher than the average hardness $\overline{\gamma}$.

2.3.2 HARDNESS-ADAPTIVE UNSUPERVISED LOSS

Considering the learning difficulty of each instance, we design a hardness-adaptive unsupervised loss to learn from unlabeled data differentially. Inspired by curriculum learning (Bengio et al., 2009), we prioritize the training on easy samples over hard ones. Precisely, we weigh the unsupervised losses for each instance by multiplying their corresponding easiness, evaluated by one minus hardness. Combined with hardness-adaptive augmentations, we can calculate the unsupervised loss by,

$$\begin{aligned} \mathcal{L}_u = \frac{1}{|\mathcal{B}_u|}\sum_{i=1}^{|\mathcal{B}_u|}\frac{1-\gamma_i}{2H\times W}\sum_{j=1}^{H\times W} & [\mathbb{1}(\max(p_i^t(j)) \geq \tau)\mathrm{H}(f_{\theta_s}(\mathcal{A}_s^I(u_i)), p_i^t(j)) + \\ & \mathbb{1}(\max(p_i^{t'}(j)) \geq \tau)\mathrm{H}(f_{\theta_s}(\mathcal{A}_s^C(u_i)), p_i^{t'}(j))]. \end{aligned} \tag{10}$$

Since the hardness is evaluated upon each (weakly augmented) image instance, under its guide, the two strong augmentations are performed separately rather than in a cascading manner. In this way, the model will not be trained on over-distorted variants, and our hardness-adaptive designs can be effectively utilized.

Table 1: Comparison with SOTA methods on **PASCAL VOC 2012** `val` set under different partition protocols. Labeled images are sampled from the *blended* training set (augmented by SBD dataset), including $10,583$ samples in total. ‡ means the results are obtained by setting the output_stride as 8 in DeepLabV3+ (16 for others). † means running more epochs in PSMT. * denotes our reproduced results. Best results are highlighted in **bold**.

Method	ResNet-50			ResNet-101		
	1/16 (662)	1/8 (1323)	1/4 (2646)	1/16 (662)	1/8 (1323)	1/4 (2646)
Supervised*	63.8	69.0	72.5	67.4	72.1	74.7
MT (Tarvainen & Valpola, 2017)	66.8	70.8	73.2	70.6	73.2	76.6
CCT (Ouali et al., 2020b)	65.2	70.9	73.4	68.0	73.0	76.2
CutMix-Seg (French et al., 2020)	68.9	70.7	72.5	72.6	72.7	74.3
GCT (Ke et al., 2020)	64.1	70.5	73.5	69.8	73.3	75.3
CAC (Lai et al., 2021)	70.1	72.4	74.0	72.4	74.6	76.3
CPS (Chen et al., 2021)	72.0	73.7	74.9	74.5	76.4	77.7
PSMT† (Liu et al., 2022)	72.8	75.7	76.4	75.5	78.2	78.7
ELN (Kwon & Kwak, 2022)	70.5	73.2	74.6	72.5	75.1	76.6
ST++ (Yang et al., 2022)	72.6	74.4	75.4	74.5	76.3	76.6
HagSeg (ours)	74.8	76.5	77.0	76.5	77.9	78.1
U^{2}PL‡ (Wang et al., 2022)	72.0	75.2	76.2	74.4	77.6	78.7
HagSeg (ours)‡	**75.9**	**76.7**	**77.1**	**77.2**	**78.4**	**79.3**

3 Experiments

In this section, we examine the efficacy of our method on standard semi-supervised semantic segmentation benchmarks and conduct extensive ablation studies to further verify the superiority.

Dataset and backbone. Following recent SOTAs (Chen et al., 2021; Yang et al., 2022) in semi-supervised segmentation, we adopt DeepLabv3+ (Chen et al., 2018) based on Resnet (He et al., 2016) as our segmentation backbone and investigate the test performance on Pascal VOC2012 (Everingham et al., 2015) and Cityscapes (Cordts et al., 2016), in terms of the mean intersection-over-union (mIOU). The classical VOC2012 consists of 21 classes with 1464 training and 1449 validation images. As a common practice, the blended training set is also involved, including additional 9118 training images from the Segmentation Boundary (SBD) dataset (Hariharan et al., 2011). Cityscapes is a large dataset on urban street scenes with 19 segmentation classes. It consists of 2975 training and 500 validation images with fine annotations.

Implementation details. For both the student and the teacher models, we load the ResNet weights pre-trained on ImageNet (Deng et al., 2009) for the encoder and randomly initialize the decoder. An SGD optimizer with a momentum of 0.9 and a polynomial learning-rate decay with an initial value of 0.01 are adopted to train the student model. The total training epoch is 80 for VOC2012 and 240 for Cityscapes. Following (Wang et al., 2022), training images are randomly cropped into 513×513 and 769×769 for Pascal VOC2012 and Cityscapes, respectively. On Cityscapes, we also use the sliding evaluation to examine the performance on validation images with a resolution of 1024×2048. We set $\mathcal{B}_u = \mathcal{B}_x = 16$ and adopt the sync-BN for all runs.

3.1 Comparison with State-of-the-Art Methods

In this section, we demonstrate the superior performance of our HagSeg on both classic and blended VOC 2012 and Cityscapes under different semi-supervised partition protocols. It is noteworthy that, on blended VOC, U^{2}PL (Wang et al., 2022) prioritizes selecting high-quality labels from classic VOCs. Instead, we randomly sample labels from the entire dataset and adopt the same partitions as specified in (Chen et al., 2021; Liu et al., 2022). Therefore, we reproduce corresponding results on U^{2}PL and evaluate HagSeg with different output_strides, 8 and 16, respectively, for fair comparisons.

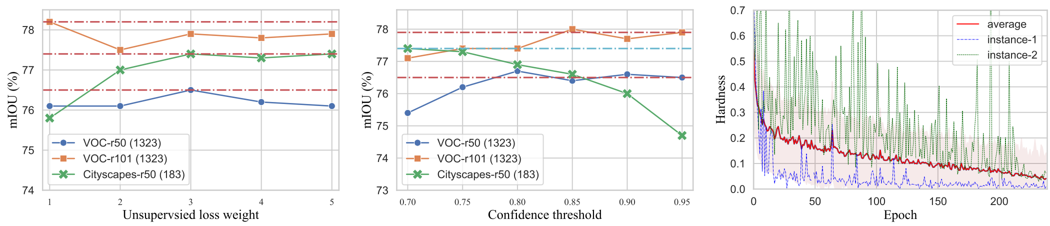
Table 2: Comparison with SOTA methods on *classic* **PASCAL VOC 2012** `val` set under different partition protocols. Labeled images are sampled from the official VOC `train` set, including $1,464$ samples in total. Results are reported using Resnet-101. All notations are the same as in Table 1.

Method	1/16 (92)	1/8 (183)	1/4 (366)	1/2 (732)	Full (1464)
Supervised *	45.5	57.5	66.6	70.4	72.9
CutMix-Seg (French et al., 2020)	52.2	63.5	69.5	73.7	76.5
PseudoSeg (Zou et al., 2021)	57.6	65.5	69.1	72.4	73.2
PC^{2}Seg (Zhong et al., 2021)	57.0	66.3	69.8	73.1	74.2
CPS (Chen et al., 2021)	64.1	67.4	71.7	75.9	-
PSMT (Liu et al., 2022)	65.8	69.6	76.6	78.4	80.0
ST++ (Yang et al., 2022)	65.2	71.0	74.6	77.3	79.1
HagSeg (ours)	68.8	74.4	78.5	79.5	81.2
U^{2}PL‡ (Wang et al., 2022)	68.0	69.2	73.7	76.2	79.5
HagSeg‡(ours)	**70.0**	**75.3**	**79.1**	**80.2**	**82.0**

Table 3: Comparison with SOTA methods on **Cityscapes** `val` set under different partition protocols. Labeled images are sampled from the Cityscapes `train` set, including $2,975$ samples in total. Results are reported using Resnet-50. * and † represent reproduced results in HagSeg and U^{2}PL, respectively. Results with ‡ are obtained by setting the output_stride as 8 in DeepLabV3+.

Method	1/16 (186)	1/8 (372)	1/4 (744)	1/2 (1488)
Supervised *	64.0	69.2	73.0	76.4
MT Tarvainen & Valpola (2017)	66.1	72.0	74.5	77.4
CCT (Ouali et al., 2020b)	66.4	72.5	75.7	76.8
GCT (Ke et al., 2020)	65.8	71.3	75.3	77.1
CPS (Chen et al., 2021)	74.4	76.6	77.8	78.8
CPS† (Wang et al., 2022)	69.8	74.3	74.6	76.8
PSMT (Liu et al., 2022)	-	75.8	76.9	77.6
ELN (Kwon & Kwak, 2022)	-	70.3	73.5	75.3
ST++ (Yang et al., 2022)	-	72.7	73.8	-
U^{2}PL * (Wang et al., 2022)	67.8	72.5	74.8	77.1
HagSeg (ours)	74.3	77.4	78.1	79.3
U^{2}PL‡* (Wang et al., 2022)	69.0	73.0	76.3	78.6
HagSeg (ours)‡	**75.2**	**78.0**	**78.2**	**80.2**

PASCAL VOC 2012. In Tables 1 and 2, we compare our HagSeg with recent SOTA methods on blended and classic VOC, respectively. We can clearly see from Table 1 that HagSeg can consistently outperform others regardless of using ResNet-50 or ResNet-101 as the segmentation encoder. The performance gain becomes more noticeable and clear as fewer labels are available. *e.g.*, in the 1/16 partition, HagSeg can improve the supervised baseline by 11% and 9.1% when using ResNet-50 and ResNet-101 as the encoders, respectively, and improve the ST++ (Yang et al., 2022) by 2.2% and 2.0%, accordingly. Checking the results among different partitions, we can also observe that HagSeg can even obtain better performance while using fewer labels compared to other SOTAs. For example, HagSeg can obtain a high mIOU of 75.9% using only 662 labels, while U^{2}PL requires 1323 labels to obtain a comparable performance of 75.2% mIOU on blended VOC. It suggests our method is more label efficient and potentially a good solution for label-scarce scenarios. In classic VOC with high-quality labels, our methods can outperform SOTA methods by a notable margin, as shown in Table 2. We attribute this improvement to the hardness-adaptive guidance that treats each unlabeled instance differently and effectively leverages them by instance-specific strategies in HegSeg. Generally, in both classic and blended cases, reserving a large feature map (*i.e.,* set output_stride=8) can slightly improve the test performance.



(a) λ_u (b) τ (c) hardness

Figure 2: We examine the effect of the loss weight and confidence threshold on VOC and Cityscapes under the 1/8 protocol in Figure (a) and (b), respectively. (c) shows how the mean instance hardness varies across the training course on Cityscapes under the 1/4 partition. Best viewed on screen.

Table 4: Ablation studies on our HagSeg. We examine the effectiveness of the hardness-adaptive guidance on the unsupervised loss, intensity-based and CutMix augmentations, respectively. Results are reported on **PASCAL VOC 2012** under the 1/8 (1323) partition using Resnet-101 as the backbone. Improvements over the supervised baseline are marked in blue.

Unsupervised Loss		**Intensity Augs**		**CutMix Augs**		**mIOU**
standard	+hardness	standard	+hardness	standard	+hardness	
-	-	-	-	-	-	72.1 (supervised)
✓	-	-	-	-	-	74.7 (2.6↑)
✓	✓	-	-	-	-	**75.5** (**3.4**↑)
✓	-	✓	-	-	-	75.4 (3.3↑)
✓	-	✓	✓	-	-	**76.5** (**4.4**↑)
✓	-	-	-	✓	-	76.1 (4.0↑)
✓	-	-	-	✓	✓	**76.9** (**4.8**↑)
✓	-	✓	-	✓	-	76.7 (4.6↑)
✓	✓	✓	✓	✓	✓	**77.9** (**5.8**↑)

Cityscapes. In table 3, we evaluate our method on more challenging Cityscapes with ResNet-50 as the segmentation encoder. HagSeg with output_stride= 8 can achieve high mIOUs of 75.2%, 78.1%, 78.2%, 80.2%, in four different splits (1/16, 1/8, 1/4, 1/2), respectively. When output_stride= 16, given only 186 labeled images, HagSeg can obtain a notable performance gain of 10.3% against the supervised baseline and 6.5% against the previous best, U^{2}PL. Not relying on any pseudo-rectifying networks (Kwon & Kwak, 2022) or extra self-supervised supervisions (Wang et al., 2022), HagSeg achieves substantially better performance than the previous SOTAs, especially with fewer labels. Despite the simplicity of Hagseg, the impressive performance further demonstrates the effectiveness and importance of our instance-specific and hardness-adaptive guidance. Surely, regardless of different semi-supervised approaches, we can see from Tables 3 that providing more labeled samples can easily improve the semi-supervised performance.

3.2 Ablations Studies

We conduct ablation studies in the 1/8 partitions of blended VOC and Cityscapes, and examine the impact of the hardness-adaptive guidance and approach-related hyper-parameters.

Effectiveness of hardness-adaptive guidance. The key of HagSeg lies in the instance-specific and hardness-adaptive guidance. In Table 4, we conduct a series of experiments on VOC2012 dataset to demonstrate its effectiveness on three components, the unsupervised loss, intensity-based and CutMix augmentations, respectively. We can observe that, performing hardness-adaptive guidance can consistently improve the standard operations, yielding around 1% improvements on all standard counterparts. The powerfulness of strong augmentations can also be witnessed, as discussed in Yang et al. (2022). As a whole, Hagseg can bring an improvement of 5.8% against the supervised baseline.

Impact of hyper-parameters. In Figure 2, we investigate the influence of different λ_u and τ on both datasets. It can be seen from Figure 2(a) that Hagseg is not very sensitive to the loss weight on VOC while a large λ_u is beneficial for Cityscapes. By default, we set $\lambda_u = 3$ for all runs. According to Figure 2(b), we set $\tau = 0.95$ for VOC and $\tau = 0.7$ for Cityscapes as default settings. This is simply because Cityscapes is a more challenging dataset requiring better discriminating ability and using a high-threshold will prevent models effectively learning from unlabeled samples. We can see from Figure 2(c) that both the mean and standard deviation of hardness evaluations on unlabeled data decrease as training processes and the model performance improves. Specifically, easy instances (like Instance-1) can hold a low hardness from the very beginning, while the hardness of hard instances (like Instance-1) fluctuates but eventually decreases.

4 Related Work

Recent studies on CR-based semi-supervised learning have achieved impressive improvements in classification tasks (Ouali et al., 2020a). Based on clustering assumptions, these methods enforce prediction consistency on the unlabeled sample with different perturbations. Early works like Mean-Teacher (Tarvainen & Valpola, 2017) aimed to generate a more robust and accurate pseudo-label using ensemble techniques. VAT (Miyato et al., 2018), UDA (Xie et al., 2020), and MixMatch (Berthelot et al., 2019) then improved the performance by using more advanced augmentations, like adversarial perturbations (Goodfellow et al., 2015), randomAug (Cubuk et al., 2020), and Mixup (Zhang et al., 2017). More recent research intended to introduce additional training and supervision, like using contrastive learning (Zhao et al., 2022), distribution alignment (Berthelot et al., 2020), and Sinkhorn-Knopp clustering (Tai et al., 2021), to further enhance the performance.

Motivated by the progress in semi-supervised classification, some studies aim to achieve dense segmentation performance with only a fraction of labels. Generally, recent jobs can be categorized into two main groups. 1) rectifying the pseudo-labels by training extra correcting networks (Ibrahim et al., 2020; Mendel et al., 2020; Kwon & Kwak, 2022), re-balancing the classes (He et al., 2021), or using multiple predictions (Liu et al., 2022); 2) exploring more supervisions by using extra losses (Chen et al., 2021), utilizing stronger augmentations (Yang et al., 2022; Yuan et al., 2021), or applying the advanced contrastive learning (Wang et al., 2022; Zhong et al., 2021; Alonso et al., 2021; Zhou et al., 2021). These studies show promising results at the cost of integrating extra network components or additional training processes. To the best of our knowledge, all the existing studies indiscriminately perturb unlabeled samples and minimize an average consistency loss over all unlabeled samples. Differently, we differentiate different samples in terms of the learning difficulty, evaluated as instance hardness. We utilize the hardness to guide the training process and achieve new SOTA performance on several semi-supervised semantic segmentation benchmarks.

Instance hardness (Smith et al., 2014; Prudêncio et al., 2015; Smith & Martinez, 2016; Chang et al., 2017) has been widely studied in hard example mining (Yuan et al., 2017) and curriculum learning (Zhou et al., 2020). Their evaluation mainly depends on the instantaneous or historical training losses with respect to ground truths. Lacking accurate label information makes hardness measurements of unlabeled instances much more challenging. Some works (Yuan et al., 2017; Jin et al., 2018) perform qualitative hardness analysis by using a threshold to select hard samples after ranking instances' losses. However, quantitative hardness analysis, especially on segmentation tasks, is still under-explored. In HagSeg, we propose a class-weighted symmetric metric to evaluate the hardness of unlabeled instances in segmentation tasks effectively.

5 Conclusion

In this paper, we highlight the instance uniqueness and propose an instance-specific and hardness-adaptive guidance (HagSeg) for semi-supervised semantic segmentation. Relying on our class-weighted symmetric hardness-evaluating strategies, our method can treat each unlabeled instance discriminatively and employ hardness-adaptive augmentation and loss weighting strategies in a instance-specific manner. Without introducing extra network components or additional training losses, HagSeg can remarkably improve the SSS performance. We hope our work can inspire future semi-supervised studies to explore more model-related dynamic strategies and leverage unlabeled data more efficiently and effectively.

REFERENCES

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.

Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston. Curriculum learning. In *ICML*, 2009.

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

David Berthelot, Nicholas Carlini, Ekin D Cubuk, Alex Kurakin, Kihyuk Sohn, Han Zhang, and Colin Raffel. Remixmatch: Semi-supervised learning with distribution alignment and augmentation anchoring. In *8th International Conference on Learning Representations (ICLR)*, 2020.

Haw-Shiuan Chang, Erik Learned-Miller, and Andrew McCallum. Active bias: Training more accurate neural networks by emphasizing high variance samples. *Advances in Neural Information Processing Systems*, 30, 2017.

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.

Xiaokang Chen, Yuhui Yuan, Gang Zeng, and Jingdong Wang. Semi-supervised semantic segmentation with cross pseudo supervision. In *CVPR*, 2021.

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.

Ekin D Cubuk, Barret Zoph, Jonathon Shlens, and Quoc V Le. Randaugment: Practical automated data augmentation with a reduced search space. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, pp. 702–703, 2020.

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.

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.

Geoff French, Timo Aila, Samuli Laine, Michal Mackiewicz, and Graham Finlayson. Semi-supervised semantic segmentation needs strong, high-dimensional perturbations. In *BMVC*, 2020.

Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. In *ICLR*, 2015.

Bharath Hariharan, Pablo Arbeláez, Lubomir Bourdev, Subhransu Maji, and Jitendra Malik. Semantic contours from inverse detectors. In *2011 international conference on computer vision*, pp. 991–998. IEEE, 2011.

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

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

Hanzhe Hu, Fangyun Wei, Han Hu, Qiwei Ye, Jinshi Cui, and Liwei Wang. Semi-supervised semantic segmentation via adaptive equalization learning. In *NeurIPS*, 2021.

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.

Mostafa S Ibrahim, Arash Vahdat, Mani Ranjbar, and William G Macready. Semi-supervised semantic image segmentation with self-correcting networks. In *CVPR*, 2020.

SouYoung Jin, Aruni RoyChowdhury, Huaizu Jiang, Ashish Singh, Aditya Prasad, Deep Chakraborty, and Erik Learned-Miller. Unsupervised hard example mining from videos for improved object detection. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 307–324, 2018.

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

Donghyeon Kwon and Suha Kwak. Semi-supervised semantic segmentation with error localization network. In *CVPR*, 2022.

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.

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

Yuyuan Liu, Yu Tian, Yuanhong Chen, Fengbei Liu, Vasileios Belagiannis, and Gustavo Carneiro. Perturbed and strict mean teachers for semi-supervised semantic segmentation. In *CVPR*, 2022.

Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *CVPR*, 2015.

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.

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

Takeru Miyato, Shin-ichi Maeda, Masanori Koyama, and Shin Ishii. Virtual adversarial training: a regularization method for supervised and semi-supervised learning. *IEEE transactions on pattern analysis and machine intelligence*, 41(8):1979–1993, 2018.

Yassine Ouali, Céline Hudelot, and Myriam Tami. An overview of deep semi-supervised learning. *arXiv preprint arXiv:2006.05278*, 2020a.

Yassine Ouali, Céline Hudelot, and Myriam Tami. Semi-supervised semantic segmentation with cross-consistency training. In *CVPR*, 2020b.

RB Prudêncio, José Hernández-Orallo, and Adolfo Martınez-Usó. Analysis of instance hardness in machine learning using item response theory. In *Second International Workshop on Learning over Multiple Contexts in ECML*, 2015.

Michael R Smith and Tony Martinez. A comparative evaluation of curriculum learning with filtering and boosting in supervised classification problems. *Computational Intelligence*, 32(2):167–195, 2016.

Michael R Smith, Tony Martinez, and Christophe Giraud-Carrier. An instance level analysis of data complexity. *Machine learning*, 95(2):225–256, 2014.

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.

Kai Sheng Tai, Peter Bailis, and Gregory Valiant. Sinkhorn label allocation: Semi-supervised classification via annealed self-training. In *ICML*, 2021.

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

Yuchao Wang, Haochen Wang, Yujun Shen, Jingjing Fei, Wei Li, Guoqiang Jin, Liwei Wu, Rui Zhao, and Xinyi Le. Semi-supervised semantic segmentation using unreliable pseudo-labels. In *CVPR*, 2022.

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

Lihe Yang, Wei Zhuo, Lei Qi, Yinghuan Shi, and Yang Gao. St++: Make self-training work better for semi-supervised semantic segmentation. In *CVPR*, 2022.

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.

Yuhui Yuan, Kuiyuan Yang, and Chao Zhang. Hard-aware deeply cascaded embedding. In *ICCV*, pp. 814–823, 2017.

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.

Hongyi Zhang, Moustapha Cisse, Yann N Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. *arXiv preprint arXiv:1710.09412*, 2017.

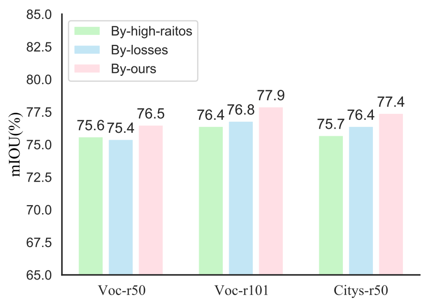
Zhen Zhao, Luping Zhou, Lei Wang, Yinghuan Shi, and Yang Gao. Lassl: Label-guided self-training for semi-supervised learning. In *AAAI*, 2022.

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

Tianyi Zhou, Shengjie Wang, and Jeffrey Bilmes. Curriculum learning by dynamic instance hardness. *Advances in Neural Information Processing Systems*, 33:8602–8613, 2020.

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.

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.



(a)

(b)

Figure 3: (a) using different hardness-evaluating strategies (b) qualitative results on Pascal VOC2012 using 183 fine labels. Columns from left to right denote the original images, the ground-truth, the supervised segmentation results, and the HagSeg segmentation results, respectively.

A Qualitative Results

We present some segmentation results on Pascal VOC 2012 in Figure 3(b) under the 183 partition protocol, using the Resnet-101 as the encoder. We can see that many mis-classified pixels and ignored segmentation details like arms in the supervised-only results are corrected in HagSeg.

B Data Augmentations

Table 5: List of various image transformations in HagSeg.

Weak Augmentations	
Random scale	Randomly resizes the image by $[0.5, 2.0]$.
Random flip	Horizontally flip the image with a probability of 0.5.
Random crop	Randomly crops an region from the image (513×513, 769×769).
Strong1: intensity-based Augmentations	
Identity	Returns the original image.
Invert	Inverts the pixels of the image.
Autocontrast	Maximizes (normalize) the image contrast.
Equalize	Equalize the image histogram.
Gaussian blur	Blurs the image with a Gaussian kernel.
Contrast	Adjusts the contrast of the image by [0.05, 0.95].
Sharpness	Adjusts the sharpness of the image by [0.05, 0.95].
Color	Enhances the color balance of the image by [0.05, 0.95]
Brightness	Adjusts the brightness of the image by [0.05, 0.95]
Hue	Jitters the hue of the image by [0.0, 0.5]
Posterize	Reduces each pixel to [4,8] bits.
Solarize	Inverts all pixels of the image above a threshold value from [1,256).
Strong2: CutMix augmentation	
CutMix	Copy and paste random size regions among different unlabeled images.

C More on hardness-evaluations

In Figure 3(a), we explore more hardness-evaluating strategies and compare corresponding performance with our class-weighted symmetric IOU evaluations in HagSeg. The "high-ratios" means using the mean high-confident ratio of the student's and teacher's segmentation results. The "losses" is evaluated by calculating the cross-entropy losses, using teacher's outputs as target labels. In terms of segmentation tasks, our proposed evaluation strategy is more appropriate and superior.