# Semi-supervised Semantic Segmentation with Mutual Knowledge Distillation

# Jianlong Yuan

Alibaba Group gongyuan.yjl@alibaba-inc.com

#### Oi Oian

Alibaba Group qi.qian@alibaba-inc.com

#### Fan Wang

Alibaba Group fan.w@alibaba-inc.com

#### Jinchao Ge

The University of Adelaide jinchao.ge@adelaide.edu.au

# **Zhibin Wang**

Alibaba Group zhibin.waz@alibaba-inc.com

#### Yifan Liu

The University of Adelaide yifan.liu04@adelaide.edu.au

#### **Abstract**

Consistency regularization has been widely studied in recent semi-supervised semantic segmentation methods. Remarkable performance has been achieved, benefiting from image, feature, and network perturbations. To make full use of these perturbations, in this work, we propose a new consistency regularization framework called mutual knowledge distillation (MKD). We innovatively introduce two auxiliary mean-teacher models based on the consistency regularization method. More specifically, we use the pseudo label generated by one mean teacher to supervise the other student network to achieve a mutual knowledge distillation between two branches. In addition to using image-level strong and weak augmentation, we also employ feature augmentation considering implicit semantic distributions to add further perturbations to the students. The proposed framework significantly increases the diversity of the training samples. Extensive experiments on public benchmarks show that our framework outperforms previous state-of-the-art(SOTA) methods under various semi-supervised settings. Code is available at: https://github.com/jianlong-yuan/semi-mmseg

# 1 Introduction

Semantic segmentation is one of the fundamental tasks in visual understanding, classifying each pixel in the image into a predefined set of categories. Recent work in semantic segmentation [1–6] has made tremendous progress in supervised learning with the help of large-scale datasets [7–10]. However, labeling such datasets is labor-intensive and time-consuming in dense prediction problems, owing to 60 times more effort than image-level labeling [11]. To address this limitation, semi-supervised learning [12–15] tries to learn a model with a set of limited labeled images and a large set of unlabeled images. Obviously, the key to the success of this task is to augment more training samples to avoid overfitting the noise in the training signals.

Current state-of-the-art semi-supervised semantic segmentation methods explore the potential of consistency regularization. The similarity between the outputs from different perturbations is enhanced during training. These perturbations are generated with augmentations on images [14, 16, 15], features [17] and various initialized networks [18, 19]. A typical network perturbation method, CPS [18],

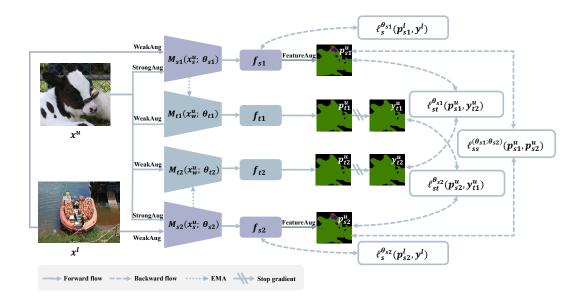


Figure 1: For each image  $x^u$ , we apply weak augmentation (WeakAug) on the teacher network and strong augmentation (StrongAug) on the student network. We apply feature augmentation on the outputs of the feature f from the students  $M(\cdot)$ .  $\mathbf{p}$  denotes the logits,  $\theta$  presents parameters of the model and  $\mathbf{y}^u$  is the one hot label generated from  $\mathbf{p}^u$ . We train the model by minimizing the consistency loss  $\ell_{st}$  and  $\ell_{ss}$  on the unlabeled set, and the cross-entropy loss  $\ell_s$  on the labeled set.

feeds the same image into two different initialized networks and uses the pseudo labels generated from one branch to supervise the other branch. Although the network perturbation achieves promising results, image-level and feature-level augmentations are neglected. Moreover, two branches are optimized with back-propagation without moving average during training. Thus, the model 'forgets' important historical information along with the training steps as stated in previous research [20–22].

To further improve the performance of the semi-supervised semantic segmentation models, we propose a novel mutual knowledge distillation (MKD) framework. Building upon two branches of co-training [18] with different initialized parameters, we further employ two auxiliary mean teacher models to record the information during the training process and provide extra supervision. The pseudo label generated from one teacher network supervises the other student and vice versa. Weak augmentation is applied to teachers' input images to increase confidence in the prediction. Also, students' input images are strongly augmented to diversify samples. Inspired by the implicit semantic data augmentation [23, 24] verified in the classification problem, we further augment the students' features. The pseudo labels from the teacher network will be more reliable, while the student network can be trained on more diverse and challenging samples. Extensive experiments verify that the proposed MKD framework benefits from the designed perturbations.

Our approach achieves state-of-the-art performance on the PASCAL VOC 2012 [7] and Cityscapes [8] under various splits of semi-supervised settings. We improve the previous SOTA method [18] by 5.81% in terms of mIoU with the most challenging 1/16 partition protocol. Our main contributions are summarized as follows.

- We propose a new consistency regularization framework, namely the mutual knowledge distillation (MKD) framework. It consists of two different initialized student networks and two corresponding mean teacher networks. The knowledge from one teacher network is used to supervise the other student branch, and vice versa.
- Image-level and feature-level augmentations are employed to increase the diversity of the training samples. Strong augmentation and weak augmentation are applied to the student and teacher networks, respectively. A semantic feature augmentation is further applied to the student network.

 We empirically demonstrate the effectiveness of our approach. Our MKD framework achieves state-of-the-art performance on PASCAL VOC 2012 and Cityscapes datasets under various semi-supervised settings. A detailed ablation study verifies the effectiveness of different components in the proposed framework.

# 2 Related Work

**Semantic segmentation.** Various methods [25, 2–5, 1, 6] have been proposed starting with FCN [25], which trains a pixel-level classifier with a fully convolutional network. It is worth mentioning that our work is based on DeepLabV3Plus [2] which applies a spatial pyramid pooling structure and an Encoder-decoder structure to refine the object boundaries. However, these models' success benefits from carefully annotated labels which are hard and time-consuming to obtain.

Semi-supervised semantic segmentation. Semi-supervised learning(SSL) proposes a new perspective to tackle semantic segmentation with the requirements of a large amount of annotated data. SSL can be grouped into consistency regularization based methods [20, 26, 13, 18, 27-31] and self-training based methods [14, 32, 16, 33, 34]. Consistency-based methods enforce the model to generate the same prediction from augmented images and original ones. Temporal ensembling [35] implements the idea of ensemble multiple checkpoints on the epoch of students. Especially, mean teacher [20, 26, 3] employs the idea of the exponential moving average of the model parameters to update the weights of the teacher model. Moreover, the student model is supervised under the pseudo label generated by the teacher model. Co-training for consistency [19, 18] feeds the same image into two different initialized networks and uses the pseudo labels generated from one branch to supervise the other branch. U<sup>2</sup>PL [29] provided a method to select reliable annotations from unreliable candidate pixels. Self-training [14, 16, 34, 32] based methods aim at generating pseudo labels to enlarge the training set. It uses a teacher model to generate the pseudo-labels based on suitable data augmentation and threshold. Unlike these methods, our approach has two student networks and two auxiliary mean teacher networks. Furthermore, we consider the image, feature, and network perturbations in the same framework.

**Data augmentation.** We describe augmentation in SSL from three aspects: image-level augmentation[36], feature-level augmentation, and network-level augmentation[37, 38]. For example, FixMatch[13] treats weakly-augmented samples as a more reliable anchor and constrains its output to be the same as strongly-augmented data. While UDA[39] draws a similar conclusion that uses weakly-augmented data and complex-augmented data to generate similar output. CutMix[15] is a widely adopted technique that generates pseudo labels and implicitly implements the idea of entropy minimization by making the decision boundary pass through a low-density region of distribution. CCT[17] or GCT[19] employ a similar idea to realize feature argumentation with cross-confidence consistency. As for feature augmentation, we believe that consistency-based SSL methods with auxiliary networks could be considered network-level augmentation. In this paper, we propose to apply complex augmentation (image-level augmentation, feature-level augmentation, and network augmentation) to the students and weak augmentation to the teachers.

# 3 Method

We first present the settings for a typical semi-supervised semantic segmentation task. Labeled datasets and unlabeled datasets are denoted as  $D^l = \{(\boldsymbol{x}_i^l, \boldsymbol{y}_i^l)\}_{i=1}^{|D^l|}$  and  $D^u = \{(\boldsymbol{x}_i^u)\}_{i=1}^{|D^u|}$  with  $|D^l| << |D^u|$ , where  $\boldsymbol{x}_i \in \mathcal{X}^{H \times W \times 3}$  is the input RGB image with the size of  $H \times W$  and  $\boldsymbol{y}_i \in \mathcal{Y}^{H \times H \times C}$  represents the pixel-level one-hot label map for C classes. We aim at training an end-to-end segmentation model with a massive amount of unlabeled and few labeled data in a semi-supervised learning manner. We propose a novel consistency regularization framework based on mutual knowledge distillation, as described in Sec. 3.1. The augmentation methods used during training are discussed in Sec. 3.2. Finally, the training procedure is introduced in Sec. 3.3.

# 3.1 Mutual Knowledge Distillation Framework

**Overview.** The proposed MKD framework is illustrated in Figure 1. We have four branches, including two *baseline student networks* and two *auxiliary mean teacher networks*. The labeled images are fed

into the students and optimized with the normal cross-entropy loss  $\mathcal{L}_s$  between ground truth labels. The unlabeled images with strong (weak) augmentation are fed into the student (teacher) networks. Furthermore, feature augmentation is also applied to the output features of the student networks. Each student network is trained under the supervision of the pseudo labels generated by the other student network ( $\mathcal{L}_{ss}$ ) and by the other teacher network ( $\mathcal{L}_{st}$ ). The knowledge between the two branches is transferred with the proposed MKD framework. Details for each part are described as follows.

Baseline student networks. The Co-training approach is to construct two networks with the same structure and different initialization and constrain the outputs between networks to be consistent. Our baseline students build upon the previous SOTA co-training method, CPS [18]. Student networks are defined as  $M_{s1}$  and  $M_{s2}$ , respectively. Network structures for  $M_{s1}$  and  $M_{s2}$  are the same, but the parameters are initialized differently with  $\theta_{s1}$  and  $\theta_{s2}$ . For example, with an input image x, the output features from baseline student networks are defined as  $f_{s1}$  and  $f_{s2}$ , respectively. Following the typical co-training baseline [18], each student network is supervised by the pseudo labels generated by the other student network, which is denoted as  $\mathcal{L}_{ss}$ .

Auxiliary mean teacher networks. It is verified in previous works [21, 22, 20] that the mean teacher can record historical information to improve the performance of the model. It does not need to be optimized and adds relatively little computation. Inspired by previous works, we add two auxiliary mean teacher networks to build our MKD framework, denoted as  $M_{t1}$  and  $M_{t2}$ . The network structure of the teacher is the same as the network structure of the student. However, the mean teacher does not require a back-propagation during training. The parameters are updated by the corresponding student model according to exponential moving average(EMA) 1, where  $\gamma$  controls the speed of updates and  $i \in [1,2]$  represents the index of the branch.

$$\theta_{ti} = \gamma \theta_{ti} + (1 - \gamma)\theta_{si} \tag{1}$$

Mutual knowledge distillation. As illustrated in Figure 1, labeled samples are used to train student models, and losses are calculated using supervised loss. Unlabelled samples, after strong augmentation, are fed into the two student models with feature augmentation to obtain different outputs( $p_{s1}^u, p_{s2}^u$ ). Similarly, samples after weak augmentation are fed into the two teacher models to obtain different outputs( $p_{t1}^u, p_{t2}^u$ ). There are two main objectives for the proposed MKD framework. First, let the teacher network update smoothly and produce high-confidence predictions on easy samples. Second, let the student network learn more challenging samples. Thus, we apply all three levels of perturbations to the samples fed into the student network. To apply the network perturbation, the output pseudo label  $\mathbf{y}_{1t}^u$  from the  $M_{t1}$  is used to supervise the logits map  $\mathbf{p}_{2s}^u$  from  $M_{s2}$ , and vice versa. The Equation 2 is the consistency loss between teacher models and student models.

$$\mathcal{L}_{st}^{u} = \ell_{st}^{\theta_{1}} \left( \mathbf{p}_{1s}^{u}, \mathbf{y}_{t2}^{u} \right) + \ell_{st}^{\theta_{2}} \left( \mathbf{p}_{2s}^{u}, \mathbf{y}_{t1}^{u} \right) \tag{2}$$

where  $\mathbf{y}_{t1}^u$  and  $\mathbf{y}_{t2}^u$  denotes one hot labels of the teachers' outputs.

# 3.2 Augmentation

To increase the diversity of the samples for the student network in our MKD framework, we apply image-level and feature level augmentations to generate perturbations.

**Image augmentation.** Image augmentation is based on weak (WDA) and strong (SDA) augmentation pairs. Weak augmentation (e.g., image flipping, cropping, resize) is applied to the images passed to the teacher models. In addition, strong augmentation (e.g., image flipping, cropping, resize, cutMix, random select an operator from color jitter, blur, gray-scale, equalize and solarize) is applied to the same ones fed to the student model to improve overall generalization. Motivated by the distribution of batch normalization [14], we do not take into account many strong color augmentation operations. Particularly, the CutMix [15] augmentation is achieved by applying a binary mask m that combines two images using the function  $x = (1-m) \odot x_i + m \odot x_j$ . We apply CutMix by combining two input images in the batch for student models and apply the same binary mask on the feature of teachers' logits with  $p = (1-m) \odot p_i + m \odot p_j$ . Then we apply p to supervising students.

**Feature augmentation.** To further increase the diversity of training samples for the student network, we introduce feature augmentation, which is implemented by manipulating infinite meaningful

semantic transformation directions to make target semantics change by moving features in space. Thus, we can change the image at a semantic level without adding an auxiliary network. Inspired by the previous work [23, 24], we found that feature augmentation is efficient in semi-supervised semantic segmentation.

First, we briefly review the semantic augmentation for supervised learning [24]. To obtain appropriate translation directions in the feature space, ISDA [24] explicitly augment each  $\boldsymbol{f}_i$  at M times to form an augmented feature set  $\left\{\left(\boldsymbol{f}_i^1,y_i\right),\ldots,\left(\boldsymbol{f}_i^M,y_i\right)\right\}_{i=1}^N$  of size MN, where  $\boldsymbol{f}_i^m$  is the  $m^{th}$  sample of augmented features for sample  $x_i$ . Based on the cross-entropy loss, the network could be optimized by minimizing  $\mathcal{L}_M(\theta) = \frac{1}{N}\sum_{i=1}^N \frac{1}{M}\sum_{m=1}^M -\log\left(\frac{e^{\boldsymbol{w}_{y_i}^T\boldsymbol{f}_i^m+b_{y_i}}}{\sum_{j=1}^C e^{\boldsymbol{w}_j^T\boldsymbol{f}_i^m+b_j}}\right)$ . Then, when  $M\to\infty$ , the CE loss under all possible augmented features can be obtained, and eventually, the upper limit of the loss is:

$$\mathcal{L}_{\infty}(\theta \mid \mathbf{\Sigma}) \leq \frac{1}{N} \sum_{i=1}^{N} \log(\sum_{j=1}^{C} e^{\mathbf{v}_{jy_i}^{\mathrm{T}} \mathbf{f}_i + (b_j - b_{y_i}) + \frac{\lambda}{2} \mathbf{v}_{jy_i}^{\mathrm{T}} \Sigma_{y_i} \mathbf{v}_{jy_i}}). \tag{3}$$

where  $\boldsymbol{v}_{jy_i}^{\mathrm{T}} = \boldsymbol{w}_{j}^{\mathrm{T}} - \boldsymbol{w}_{y_i}^{\mathrm{T}}$ ,  $\Sigma_{y_i}$  denotes co-variance of  $y_i$  class. And  $\Sigma$  is obtained by statistics on the data set in [23, 24].

To incorporate the semantic augmentation for semi-supervised semantic segmentation, we first analyze the structure of ISDA loss 3. Compared to the standard cross entropy loss, Equation 3 ends up with just one more term( $\frac{\lambda}{2} \boldsymbol{v}_{jy_i}^{\mathrm{T}} \Sigma_{y_i} \boldsymbol{v}_{jy_i}$ ), so we can treat  $\boldsymbol{v}_{jy_i}^{\mathrm{T}} \boldsymbol{f}_i + (b_j - b_{y_i}) + \frac{\lambda}{2} \boldsymbol{v}_{jy_i}^{\mathrm{T}} \Sigma_{y_i} \boldsymbol{v}_{jy_i}$  as an augmented feature. Each pixel in students' features could be calculated by  $\mathbf{p}_s = \boldsymbol{v}^{\mathrm{T}} \boldsymbol{f}_s + b + \frac{\lambda}{2} \boldsymbol{v}^{\mathrm{T}} \Sigma \boldsymbol{v}$ . As can be seen, feature augmentation requires category information, but for unlabeled data, this does not apply. Thus, we use pseudo-labels instead.

#### 3.3 Optimization of the Framework

The full training loss for the whole framework is described in Equation 4, where  $\alpha$  and  $\beta$  are the loss weights.

$$\mathcal{L} = \mathcal{L}_{s}^{l} + \alpha \mathcal{L}_{st}^{u} + \beta \mathcal{L}_{ss}^{u} \tag{4}$$

The first loss in Equation 4 is the supervised segmentation loss for student models, defined as Equation 5, where  $\ell_{ce}$  is the cross-entropy loss function, y presents the ground truth, and  $\theta_1$  and  $\theta_2$  are model parameters of different students.

$$\mathcal{L}_{s}^{l} = \ell_{ce}^{\theta_{1}} \left( \mathbf{p}_{s1}^{l}, \mathbf{y}^{l} \right) + \ell_{ce}^{\theta_{2}} \left( \mathbf{p}_{s2}^{l}, \mathbf{y}^{l} \right) \tag{5}$$

The second term is described in Equation 2, which is the consistency loss based on cross-entropy. The last term in Equation 4 is the consistency loss between students same as CPS [18].

We show our MKD framework in Algorithm 1. We initialize student models with different random initialization parameters and set the same parameters for its teacher network. After obtaining perturbed data from the input images with SDA and WDA, we first use EMA to update the teachers' parameters. In addition, we follow 3.2 to obtain a segmentation map after CutMix. Moreover, the teachers' network features are followed in the same CutMix way to get the corresponding results. We calculate meaningful semantic augmented features by equation 3.2 as FD. And finally, as described in Fig. 1, we follow the equation 4 to train the model.

# 4 Experiments

In this section, we first introduce the implementation details for semi-supervised semantic segmentation. Then we report our results compared with other state-of-the-art methods. Finally, we perform a series of ablation experiments and analyze the results in detail.

# Algorithm 1: Semi-supervised Learning Framework

```
1 Labeled images: images x_i^l and corresponding labels y_i^l from D^l
 2 Unlabeled images: images without labels \{(x_1^u, x_2^u, ..., x_m^u)\} from D^u
 3 Initialization Student Randomly initialize two student models
 4 Initialization Teacher Apply same initialization of each student to corresponding teacher.
 5 for step = 1, ..., n_{steps} do
         update M_{t1}, M_{t2}
         x^l, y^l = \text{Sample}(D^l)
         x^u=Sample(D^u)
         x_w^l, x_w^u, x_s^u = WDA(x^l), WDA(x^u), SDA(x^u)
10
         p_{t1}^u, p_{t2}^u = M_{t1}(x_w^u; \theta_{t1}), M_{t2}(x_w^u; \theta_{t2})
         p_{s1}^l, p_{s2}^l = M_{s1}(x_w^l; \theta_{s1}), M_{s2}(x_w^l; \theta_{s2}),
11
         p_{s1}^u, p_{s2}^u = FD(M_{s1}(x_s^u; \theta_{s1})), FD(M_{s2}(x_s^u; \theta_{s2}))
12
13
         loss_{sup} = \ell(p_{s1}^{l}, y^{l}) + \ell(p_{s2}^{l}, y^{l})
         loss_{st} = \ell(p_{s1}^u, max(p_{t2}^u)) + \ell(p_{s2}^u, max(p_{t1}^u))
14
         loss_{ss} = \ell(p_{s1}^u, max(p_{s2}^u)) + \ell(p_{s2}^u, max(p_{s1}^u))
15
         loss = loss_{sup} + \alpha loss_{st} + \beta loss_{ss}
16
17 end
```

#### 4.1 Implementation Details

**Datasets.** Following previous methods [18, 14, 29], experiments are performed on two widely used image segmentation datasets, PASCAL VOC 2012 [7] and Cityscapes [8]. PASCAL VOC 2012 [7] is a standard semantic segmentation benchmark with 21 classes, including the background. The standard PASCAL VOC 2012 datasets (VOC) have 1464 images for training, 1449 images for validation, and 1456 images for testing. Following the previous works [2], we combine the Pascal VOC with 9118 training images from the Segmentation Boundary Dataset (SBD) [40] as VOCAug. During training on VOCAug and VOC, we employ a crop size of  $512 \times 512$ . Cityscapes [8] consists of 2975/500/1525 finely annotated urban scene images with resolution  $2048 \times 1024$  for training/validation/testing respectively. The segmentation performance is evaluated over 19 challenging categories. We use a training crop size of  $1024 \times 512$ .

**Training.** Our method is implemented on MMSegmentation [41]. Following DeepLabV3Plus [2], we use the "poly" learning rate policy where the initial learning rate is multiplied by  $(1-iter/iter_{max})^{0.9}$ . The initial learning rate is 0.0025 for PASCAL VOC 2012 and 0.01 for Cityscapes. Specifically, the batch size is set to 16 for all datasets, and all training was performed on the four NVIDIA A100. We train the network with mini-bath stochastic gradient descent(SGD). The momentum is fixed as 0.9, and the weight decay is set to 0.0005.

**Network architecture.** We use DeepLabv3plus [2] with ResNet [42] pre-trained on ImageNet [43] as our segmentation network. The decoder head is composed of separable convolution same as standard DeepLabv3plus.

**Evaluation metrics** Following [2], we adopt the mean Intersection over Union(mIoU) as the evaluation metrics. All results are estimated on the val set. Ablation studies are performed on the VOC val set under the 1/16 partition protocol. Particularly, we report results via only single-scale testing.

# 4.2 Comparison with State-of-the-art Methods

We conduct the comparison experiments with state-of-the-art algorithms in Table 1 and Table 2.

Results on PASCAL VOC 2012 Dataset. Table 1 shows comparison results on PASCAL VOC 2012 dataset. Assuming that we utilize 1464 images for training, the proposed framework accomplishes 60.60%, 66.74%, 71.01%, 72.73% with ResNet50 and 65.35%, 70.18%, 74.44%, 75.90 with ResNet-101, individually, which outperform the previous method under most of the settings. Note that we achieve similar performance with CPS [18] under 1/16 partitions, which is trained with only 92 label images as the number of labeled images is too small to generate reliable labels for the teacher network. When employing VOC as labeled images and SBD as unlabeled data, as shown in lines 11 and 12 in Table 1, our method improved by 3.14% based on ResNet-50 and 4.96% based on ResNet-101 compared with previous SOTA [14]. When more unlabeled data are introduced, the

same splits as [18, 29, 44] are used, i.e., 1/n as labeled data and 10582-1/n as unlabeled data. When we change the unlabeled partition for the VOC dataset, under 1/16, 1/8, 1/4, 1/2 and full partition protocols, our method achieves 64.92%, 69.48%, 74.25%, 75.89%, 77.59% based on ResNet-50 and 69.10%, 74.63%, 76.76%, 78.66% and 80.02% based on ResNet-101. It shows that more unlabeled images could bootstrap performance. In particular, compared with the previous best method, our method improves by 1.12% and 3.63% under 1/8 and 1/4 partition protocols. It is also demonstrated that our method is more effective on fewer data. In addition, we set a confidence of 0.95 for selecting regions above this threshold to calculate the loss. Our method can significantly improve the results.

Table 2 compares our method with the other state-of-the-art methods on VOCAug. To make a fair comparison, we train our MKD framework under different split lists following previous work. Our method outperforms the U²PL by 1.23%, 0.73%, 0.25% and 0.10% with the same split and based on ResNet-101. Using the same split as CPS, the proposed method achieves 74.02%, 75.24%, 76.09%, 76.92% with ResNet50 and 75.90%, 76.59%, 77.62%, 78.94 with ResNet-101, respectively, which performs favorably against the previous state-of-the-art methods. As the amount of data increases, the performance gap between the various methods becomes smaller, and it also proves that the segmentation task does not require a lot of labeled data.

**Results on Cityscapes Dataset.** As shown in Table 2, our method achieves 75.47%, 78.07%, 79.95, 80.52 based ResNet-50 and 77.19%, 79.20%, 80.80% and 81.04% based ResNet-101 both under 1/16, 1/8, 1/4 and 1/2 partition protocols with same split as CPS [18]. In addition, we improve by 5.01%, 1.61%, 1.81%, and 1.69% with the same split as U<sup>2</sup>PL [29]. Our method outperforms the existing state-of-the-art method by a notable margin. Specifically, we report results with single-scale testing. We attribute this significant improvement to the fact that the Cityscapes dataset is relatively redundant in itself, so the teacher model can provide more accurate pseudo-labeling.

Table 1: Compared with state-of-the-art methods on the Pascal VOC 2012 val set under different partition protocols. Here '1/n' means that we use '1/n' labeled dataset and the remaining images in the training set are used as unlabeled dataset. † means we introduce more unlabeled dataset with a total of 10582 images. \* denotes we set 0.95 as a threshold for confidence to calculate the loss.

Method	1/16 (92)	1/8 (183)	1/4 366	1/2 (732)	Full (1464)
					1'uii (1404)
MT-R101 [20]	48.70	55.81	63.01	69.16	-
AdvSemSeg-R101 [31]	39.69	47.58	59.97	65.27	68.40
VAT-R101 [45]	36.92	49.35	56.88	63.34	-
CCT-R50 [17]	33.10	47.60	58.80	62.10	69.40
CutMix-Seg-R101 [15]	55.58	63.20	68.36	69.84	-
$PC^2Seg-R101$ [46]	57.00	66.28	69.78	73.05	74.15
GCT-R101 [19]	46.04	54.98	64.71	70.67	-
PseudoSeg-R101 [33]	57.60	65.50	69.14	72.41	73.23
CPS-R101 [18]	64.07	67.42	71.71	75.88	-
SimpleBaseline-R101 [14]	-	-	-	-	75.00
Ours-R50	60.60	66.74	71.01	72.73	78.14
Ours-R101	65.35	70.18	74.44	75.90	79.96
PS-MT-R101 [44] <sup>†</sup>	65.80	69.58	76.57	78.42	80.01
ST++-R101 [16] <sup>†</sup>	65.20	71.00	74.60	77.30	79.10
$U^{2}PL-R101[29]^{\dagger}$	67.98	69.15	73.66	76.16	79.49
Ours-R50 <sup>†</sup>	64.92	69.48	74.25	75.89	77.59
Ours-R101 <sup>†</sup>	69.10	74.63	76.76	78.66	80.02
Ours-R101 <sup>†*</sup>	74.40	75.92	77.74	78.68	80.80

#### 4.3 Ablation Study

In this subsection, we conduct experiments to explore the effectiveness of each proposed module under different semi-supervised settings. All the ablation studies are conducted on Pascal VOC 2012 under 1/16 partition protocols.

**Impact of Each Module.** As illustrated in Table 3, we conduct a series of experiments to identify each module's performance. We take co-training as our baseline, just the same as CPS. We first try to add

Table 2: Comparison with state-of-the-art on the PASCAL VOCAug and Cityscapes val set under different partition protocols. The VOCAug trainset consists of 10,582 labeled samples in total.  $\dagger$  means the same split as U<sup>2</sup>PL. Other methods use the same split as CPS. \* presents the approach reproduced by [29].

Method		ResN	et-50			ResNo	et-101	
	1/16(662)	1/8(1323)	1/4(2646)	1/2(5291)	1/16(662)	1/8(1323)	1/4(2646)	1/2(5291)
			Pascal	VOC 2012				
MT [20]	66.77	70.78	73.22	75.41	70.59	73.20	76.62	77.61
CCT [17]	65.22	70.87	73.43	74.75	67.94	73.00	76.17	77.50
CutMix-Seg [15]	68.90	70.70	72.46	74.49	72.56	72.69	74.25	75.89
GCT [19]	64.05	70.47	73.45	75.20	69.77	73.30	75.25	77.1
AEL [27]	-	-	-	-	77.20	77.57	78.06	80.29
CPS [18]	68.21	73.20	74.24	75.91	72.18	75.83	77.55	78.6
CPS w/ CutMix [18]	71.98	73.67	74.90	76.15	74.48	76.44	77.68	78.6
PS-MT [44]	72.83	75.70	76.43	77.88	75.50	78.20	78.72	79.7
ST++ [16]	73.20	75.50	76.00	-	74.70	77.90	77.90	
U <sup>2</sup> PL <sup>†</sup> [29]	-	-	-	-	77.21	79.01	79.30	80.50
Ours	74.02	75.24	76.09	76.92	75.90	76.59	77.62	78.9
Ours <sup>†</sup>	76.03	78.03	77.56	77.37	78.44	79.74	79.55	80.6
		ResNet-50		ResNet-101				
	1/16(186)	1/8(372)	1/4(744)	1/2(1488)	1/16(186)	1/8(372)	1/4(744)	1/2(1488)
			City	vscapes				
MT [20]	66.14	72.03	74.47	77.43	68.08	73.71	76.53	78.59
CCT [17]	66.35	72.46	75.68	76.78	69.64	74.48	76.35	78.29
CutMix-Seg [28]	-	-	-	-	72.13	75.83	77.24	78.93
COTT F103							76.45	78.5
GCT [19]	65.81	71.33	75.30	77.09	66.90	72.96	76.45	10.5
AEL [27]	65.81	71.33	75.30	77.09	66.90 75.83	72.96 77.90	76.45	
	65.81 - 69.79	71.33 - 74.39	75.30 - 76.85	77.09 - 78.64				80.28
AEL [27]	-	-	-	-	75.83	77.90	79.01	80.28 80.08
AEL [27] CPS [18] CPS w/ CutMix [18]	69.79	74.39	76.85	78.64	75.83 70.50	77.90 75.71	79.01 77.41	80.2 80.0 80.2
AEL [27] CPS [18] CPS w/ CutMix [18]	69.79 74.47	74.39	76.85	78.64 78.77	75.83 70.50 74.72	77.90 75.71 77.62	79.01 77.41 79.21	80.2 80.0 80.2 76.8
AEL [27] CPS [18] CPS w/ CutMix [18] CPS* <sup>†</sup> [18]	69.79 74.47	74.39	76.85	78.64 78.77	75.83 70.50 74.72 69.78	77.90 75.71 77.62 74.31	79.01 77.41 79.21 74.58	80.2 80.0 80.2 76.8
AEL [27] CPS [18] CPS w/ CutMix [18] CPS*† [18] SimpleBaseline[14]	69.79 74.47	74.39 76.61	76.85 77.83	78.64 78.77 -	75.83 70.50 74.72 69.78	77.90 75.71 77.62 74.31	79.01 77.41 79.21 74.58	80.28 80.08 80.2 76.8 78.70
AEL [27] CPS [18] CPS w/ CutMix [18] CPS*† [18] SimpleBaseline[14] PS-MT [44]	69.79 74.47	74.39 76.61	76.85 77.83	78.64 78.77 -	75.83 70.50 74.72 69.78	77.90 75.71 77.62 74.31 74.10	79.01 77.41 79.21 74.58 77.80	78.36 80.28 80.08 80.21 76.81 78.70

naive mean teacher(MT) and find that the results do not improve and even lead to training instability. Analyzing the reason, it may be because the teacher and student models are too similar, leading to collapse. So, we tried to add the mutual mean teacher and found that the collapse disappeared, which resulted in a 1.37% performance improvement. In order to introduce more perturbations, we also added the strategies of strong and weak augmentation(SDA) and feature augmentation(FD), which accompanied 3.7% and 1.94% performance improvement, respectively. Combining the two methods, SDA and Mutual MT, resulted in a 4.84% gain. The final combination of all methods yields a performance improvement of 5.2%.

**Impact of**  $\gamma$  **in Equation 1.** We investigate the influence of  $\gamma$  that is used to update the mean teacher model as shown in Equation 1. From Table 4, we can see that  $\gamma = 0.4$  performs best on PASCAL VOC 2012 under 1/16 partition protocols. We use  $\gamma = 0.4$  in our approach for all the experiments.

**Impact of**  $\beta$  **in Equation 4.** Table 5studies the loss of weight. We find that  $\beta = 1.0$  and  $\beta = 1.5$  achieves the best performance. Compared with no more branches, our method improves by 1.14% and 1.36%. So more branches do get more diversity.

Table 4: Ablation study for  $\gamma$  in Equation 1, Table 5: Impact of Weight of different loss which controls the speed of the parameter upbetween mean teacher and cross supervision. dating in the teacher model. Here we set  $\alpha=1.5$  as default.

g in the teacher moder.			11010	$\alpha = 1.0 \text{ as actaunt.}$				
		mIoU			В	mIoU		
$\gamma$	ResNet50	ResNet101		ρ -	ResNet50	ResNet101		
•	0.0	72.41	74.53		0.0	72.52	74.44	
	0.2	72.68	74.60		0.5	72.80	75.81	
	0.4	72.82	75.34		1.0	73.66	75.44	
	0.6	72.74	75.17		1.5	73.43	75.80	
	0.8	72.78	75.07		2.0	73.49	75.64	

Table 3: Ablation study on the proposed semi-supervised learning framework. The model here is Deeplabv3Plus with ResNet50 backbone. Co-training denotes the baseline the same as CPS. 'MT' means standard mean teacher. 'Mutual MT' presents mutual knowledge distillation. 'SDA' denotes strong data augmentation. 'FDA' is feature augmentation. All experiments are under the 1/16 partition protocol(662 images).

Co-training	MT	Mutual MT	SDA	FDA	mIoU
$\overline{}$					68.82
$\checkmark$	$\checkmark$				68.85
$\checkmark$		✓			70.19
$\checkmark$			$\checkmark$		72.52
$\checkmark$				$\checkmark$	70.76
$\checkmark$		✓	$\checkmark$		73.66
$\checkmark$		✓	$\checkmark$	$\checkmark$	74.02

# 4.3.1 Qualitative Results

Figure 2 shows the quantitative results of different methods on the PASCAL VOC 2012 datasets. We can see that Co-training can not reasonably separate the objects (especially large-sized objects such as cows, boat, sheep, motor bike, and bottle) completely while ours corrects these errors. Compared to Co-training, our method also performs well on these hard examples, such as potted plant and chair.

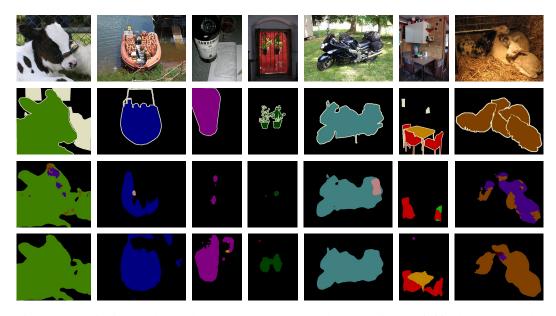


Figure 2: Qualitative results on the PASCAL VOC 2012 datasets using 1/16(662) labeled samples and ResNet50. The second line shows all the ground truth. The third line presents all baseline of co-training. The fourth line denotes our method. The proposed semi-supervised approach produces improved results compared to co-training.

# 5 Conclusion

This paper proposes a new consistency learning scheme, called mutual knowledge distillation (MKD), to finish semantic segmentation with two auxiliary mean-teacher models and the combination of strong-weak augmentation and feature augmentation. Experiments show that results outperform the state-of-the-art method on PASCAL VOC 2012 and Cityscapes. Notably, our framework's performance improved significantly when the label data is few. For example, our method improves 5.81% with ResNet-50 on the 1/16 split of PASCAL VOC 2012 and 6.69% with ResNet-101 on the 1/16 split of Cityscapes compared with the co-training baseline.

# References

- [1] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, "Pyramid scene parsing network," in *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, pp. 2881–2890, 2017.
- [2] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder-decoder with atrous separable convolution for semantic image segmentation," in *Proc. Eur. Conf. Comp. Vis.*, pp. 801–818, 2018.
- [3] J. Wang, K. Sun, T. Cheng, B. Jiang, C. Deng, Y. Zhao, D. Liu, Y. Mu, M. Tan, X. Wang, *et al.*, "Deep high-resolution representation learning for visual recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 10, pp. 3349–3364, 2020.
- [4] J. Yuan, Z. Deng, S. Wang, and Z. Luo, "Multi receptive field network for semantic segmentation," in *Proc. Winter Conf. on Appl. of Comp. Vis.*, pp. 1883–1892, IEEE, 2020.
- [5] E. Xie, W. Wang, Z. Yu, A. Anandkumar, J. M. Alvarez, and P. Luo, "Segformer: Simple and efficient design for semantic segmentation with transformers," *Proc. Advances in Neural Inf. Process. Syst.*, vol. 34, 2021.
- [6] W. Zhang, J. Pang, K. Chen, and C. C. Loy, "K-Net: Towards unified image segmentation," in *Proc. Advances in Neural Inf. Process. Syst.*, 2021.
- [7] M. Everingham, S. Eslami, L. Van Gool, C. K. Williams, J. Winn, and A. Zisserman, "The pascal visual object classes challenge: A retrospective," *Int. J. Comput. Vision*, vol. 111, no. 1, pp. 98–136, 2015.
- [8] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele, "The cityscapes dataset for semantic urban scene understanding," in *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, pp. 3213–3223, 2016.
- [9] R. Mottaghi, X. Chen, X. Liu, N.-G. Cho, S.-W. Lee, S. Fidler, R. Urtasun, and A. Yuille, "The role of context for object detection and semantic segmentation in the wild," in *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, 2014.
- [10] B. Zhou, H. Zhao, X. Puig, S. Fidler, A. Barriuso, and A. Torralba, "Scene parsing through ade20k dataset," in *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, pp. 633–641, 2017.
- [11] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft coco: Common objects in context," in *Proc. Eur. Conf. Comp. Vis.*, pp. 740–755, Springer, 2014.
- [12] D. Berthelot, N. Carlini, I. Goodfellow, N. Papernot, A. Oliver, and C. A. Raffel, "Mixmatch: A holistic approach to semi-supervised learning," *Proc. Advances in Neural Inf. Process. Syst.*, vol. 32, 2019.
- [13] K. Sohn, D. Berthelot, N. Carlini, Z. Zhang, H. Zhang, C. A. Raffel, E. D. Cubuk, A. Kurakin, and C.-L. Li, "Fixmatch: Simplifying semi-supervised learning with consistency and confidence," *Proc. Advances in Neural Inf. Process. Syst.*, vol. 33, pp. 596–608, 2020.
- [14] J. Yuan, Y. Liu, C. Shen, Z. Wang, and H. Li, "A simple baseline for semi-supervised semantic segmentation with strong data augmentation," in *Proc. IEEE Int. Conf. Comp. Vis.*, pp. 8229–8238, 2021.
- [15] S. Yun, D. Han, S. J. Oh, S. Chun, J. Choe, and Y. Yoo, "Cutmix: Regularization strategy to train strong classifiers with localizable features," in *Proc. IEEE Int. Conf. Comp. Vis.*, pp. 6023–6032, 2019.
- [16] L. Yang, W. Zhuo, L. Qi, Y. Shi, and Y. Gao, "St++: Make self-training work better for semi-supervised semantic segmentation," arXiv preprint arXiv:2106.05095, 2021.
- [17] Y. Ouali, C. Hudelot, and M. Tami, "Semi-supervised semantic segmentation with cross-consistency training," in *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, pp. 12674–12684, 2020.
- [18] X. Chen, Y. Yuan, G. Zeng, and J. Wang, "Semi-supervised semantic segmentation with cross pseudo supervision," in *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, pp. 2613–2622, 2021.
- [19] Z. Ke, D. Qiu, K. Li, Q. Yan, and R. W. Lau, "Guided collaborative training for pixel-wise semi-supervised learning," in *Proc. Eur. Conf. Comp. Vis.*, pp. 429–445, Springer, 2020.
- [20] A. Tarvainen and H. Valpola, "Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results," Proc. Advances in Neural Inf. Process. Syst., vol. 30, 2017.
- [21] K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick, "Momentum contrast for unsupervised visual representation learning," in *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, pp. 9729–9738, 2020.

- [22] J.-B. Grill, F. Strub, F. Altché, C. Tallec, P. Richemond, E. Buchatskaya, C. Doersch, B. Avila Pires, Z. Guo, M. Gheshlaghi Azar, et al., "Bootstrap your own latent-a new approach to self-supervised learning," Proc. Advances in Neural Inf. Process. Syst., vol. 33, pp. 21271–21284, 2020.
- [23] Y. Wang, X. Pan, S. Song, H. Zhang, G. Huang, and C. Wu, "Implicit semantic data augmentation for deep networks," in *Proc. Advances in Neural Inf. Process. Syst.*, pp. 12635–12644, 2019.
- [24] Y. Wang, G. Huang, S. Song, X. Pan, Y. Xia, and C. Wu, "Regularizing deep networks with semantic data augmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, 2021.
- [25] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, pp. 3431–3440, 2015.
- [26] D. Berthelot, N. Carlini, E. D. Cubuk, A. Kurakin, K. Sohn, H. Zhang, and C. Raffel, "Remixmatch: Semi-supervised learning with distribution alignment and augmentation anchoring," *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, 2019.
- [27] H. Hu, F. Wei, H. Hu, Q. Ye, J. Cui, and L. Wang, "Semi-supervised semantic segmentation via adaptive equalization learning," *Proc. Advances in Neural Inf. Process. Syst.*, vol. 34, 2021.
- [28] G. French, T. Aila, S. Laine, M. Mackiewicz, and G. Finlayson, "Semi-supervised semantic segmentation needs strong, high-dimensional perturbations," *Proc. British Machine Vis. Conf.*, 2019.
- [29] Y. Wang, H. Wang, Y. Shen, J. Fei, W. Li, G. Jin, L. Wu, R. Zhao, and X. Le, "Semi-supervised semantic segmentation using unreliable pseudo-labels," *arXiv preprint arXiv:2203.03884*, 2022.
- [30] N. Souly, C. Spampinato, and M. Shah, "Semi supervised semantic segmentation using generative adversarial network," in *Proc. IEEE Int. Conf. Comp. Vis.*, pp. 5688–5696, 2017.
- [31] W.-C. Hung, Y.-H. Tsai, Y.-T. Liou, Y.-Y. Lin, and M.-H. Yang, "Adversarial learning for semi-supervised semantic segmentation," in *Proc. British Machine Vis. Conf.*, 2018.
- [32] R. He, J. Yang, and X. Qi, "Re-distributing biased pseudo labels for semi-supervised semantic segmentation: A baseline investigation," in *Proc. IEEE Int. Conf. Comp. Vis.*, pp. 6930–6940, 2021.
- [33] Y. Zou, Z. Zhang, H. Zhang, C.-L. Li, X. Bian, J.-B. Huang, and T. Pfister, "Pseudoseg: Designing pseudo labels for semantic segmentation," *Proc. Int. Conf. Learn. Representations*, 2021.
- [34] Q. Xie, M.-T. Luong, E. Hovy, and Q. V. Le, "Self-training with noisy student improves imagenet classification," in *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, pp. 10687–10698, 2020.
- [35] S. Laine and T. Aila, "Temporal ensembling for semi-supervised learning," Proc. Int. Conf. Learn. Representations, 2016.
- [36] P. Zhang, B. Zhang, T. Zhang, D. Chen, and F. Wen, "Robust mutual learning for semi-supervised semantic segmentation," *arXiv preprint arXiv:2106.00609*, 2021.
- [37] Y. Ge, D. Chen, and H. Li, "Mutual mean-teaching: Pseudo label refinery for unsupervised domain adaptation on person re-identification," *Proc. Int. Conf. Learn. Representations*, 2020.
- [38] Z. Feng, Q. Zhou, Q. Gu, X. Tan, G. Cheng, X. Lu, J. Shi, and L. Ma, "Dmt: Dynamic mutual training for semi-supervised learning," *Pattern Recognition*, p. 108777, 2022.
- [39] Q. Xie, Z. Dai, E. Hovy, T. Luong, and Q. Le, "Unsupervised data augmentation for consistency training," *Proc. Advances in Neural Inf. Process. Syst.*, vol. 33, pp. 6256–6268, 2020.
- [40] B. Hariharan, P. Arbeláez, L. Bourdev, S. Maji, and J. Malik, "Semantic contours from inverse detectors," in *Proc. IEEE Int. Conf. Comp. Vis.*, pp. 991–998, IEEE, 2011.
- [41] M. Contributors, "MMSegmentation: Openmmlab semantic segmentation toolbox and benchmark." https://github.com/open-mmlab/mmsegmentation, 2020.
- [42] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, pp. 770–778, 2016.
- [43] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," Proc. Advances in Neural Inf. Process. Syst., vol. 25, 2012.
- [44] Y. Liu, Y. Tian, Y. Chen, F. Liu, V. Belagiannis, and G. Carneiro, "Perturbed and strict mean teachers for semi-supervised semantic segmentation," arXiv preprint arXiv:2111.12903, 2021.

- [45] T. Miyato, S.-i. Maeda, M. Koyama, and S. Ishii, "Virtual adversarial training: a regularization method for supervised and semi-supervised learning," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 8, pp. 1979–1993, 2018.
- [46] Y. Zhong, B. Yuan, H. Wu, Z. Yuan, J. Peng, and Y.-X. Wang, "Pixel contrastive-consistent semi-supervised semantic segmentation," in *Proc. IEEE Int. Conf. Comp. Vis.*, pp. 7273–7282, 2021.