This CVPR paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

Perturbed and Strict Mean Teachers for Semi-supervised Semantic Segmentation

Yuyuan Liu[1] Yu Tian[1] Yuanhong Chen[1] Fengbei Liu[1]
Vasileios Belagiannis[2] Gustavo Carneiro[1]
[1] Australian Institute for Machine Learning, University of Adelaide
[2] Universität Ulm, Germany

Abstract

Consistency learning using input image, feature, or network perturbations has shown remarkable results in semi-supervised semantic segmentation, but this approach can be seriously affected by inaccurate predictions of unlabelled training images. There are two consequences of these inaccurate predictions: 1) the training based on the "strict" cross-entropy (CE) loss can easily overfit prediction mistakes, leading to confirmation bias; and 2) the perturbations applied to these inaccurate predictions will use potentially erroneous predictions as training signals, degrading consistency learning. In this paper, we address the prediction accuracy problem of consistency learning methods with novel extensions of the mean-teacher (MT) model, which include a new auxiliary teacher, and the replacement of MT's mean square error (MSE) by a stricter confidence-weighted cross-entropy (Conf-CE) loss. The accurate prediction by this model allows us to use a challenging combination of network, input data and feature perturbations to improve the consistency learning generalisation, where the feature perturbations consist of a new adversarial perturbation. Results on public benchmarks show that our approach achieves remarkable improvements over the previous SOTA methods in the field.[1] *Our code is available at* https://github.com/yyliu01/PS-MT.

1. Introduction

Semantic segmentation is an essential pixel-wise classification task that has reached remarkable success in recent years. However, the training of such a task is known to be data-hungry, where the labelling process is particularly costly and time-consuming [34]. To tackle this limitation, semi-supervised semantic segmentation has become an important research direction that has drawn a growing attention recently [9,21,34]. This problem relies on a small set of pixel-level labelled images and a large set of unlabelled images, where both types of images are drawn from the same data distribution. The challenge is how to extract additional and useful training signal from the unlabelled images to allow the training of the model to generalise beyond the small labelled set.

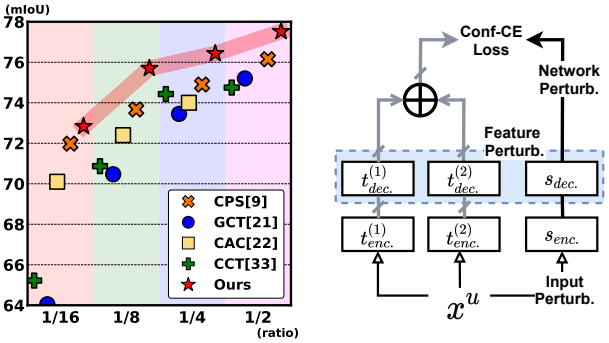


Figure 1. In (a), we compare our method (red star) to the current SOTA on *Pascal VOC 2012* under different partition protocols based on the augmented set [16], and (b) shows a simple diagram of our consistency-based model, which includes two mean teachers (denoted by the encoders $t^{(1)}_{enc.}$ and $t^{(2)}_{enc.}$ and decoders $t^{(1)}_{dec.}$ and $t^{(2)}_{dec.}$) that produce accurate segmentation maps for unlabelled images x^u and the student (encoder $s_{enc.}$ and decoder $s_{dec.}$), with three types of perturbations (network, feature and input image) that are used in the minimisation of our strict Conf-CE loss.

Current state-of-the-art (SOTA) semi-supervised semantic segmentation models are based on consistency learning, which enforces the agreement between the outputs from different views of unlabelled images [9, 14, 34, 44]. These different views can be obtained via perturbations applied to the input image with data augmentation [44] or to the feature space with noise injection [34]. Another way of obtaining different views is with network perturbation, which encourages similar predictions between multiple models trained from different initialization, and has been shown to enable better consistency regularization than input image and feature perturbations [9, 21]. One potential weakness of consistency learning is that it assumes accurate predictions for unlabelled images, such that the perturbation does not push the image feature to the wrong side of the true (hidden) classification decision boundary. Unfortunately, in prac-

[1]Supported by Australian Research Council through grants DP180103232 and FT190100525.

tice this assumption is not always met by SOTA methods, making the training signal of consistency learning methods potentially incorrect. This problem is exacerbated for consistency learning based on network perturbation because incorrect predictions from one model will deteriorate the training for the other model, and vice versa. Another consequence of these inaccurate predictions is that consistency learning methods that rely on a "strict" cross-entropy (CE) loss can easily overfit prediction mistakes, which can lead to confirmation bias.

In this paper, we address the prediction accuracy problem of consistency based methods by extending the mean teacher (MT) model [9, 14, 21, 37] with a new auxiliary teacher, and the replacement of MT's means square error (MSE) loss by a stricter confidence-weighted CE loss (Conf-CE) that has better training convergence. These accurate predictions enable the use of more challenging perturbations, combining input image, feature and network perturbations to improve the generalisation of consistency learning. Furthermore, we propose a new type of adversarial feature perturbation that learns the perturbation to be applied to the student model using virtual adversarial training [33] from the teachers (T-VAT), instead of injecting different types of noise in the image features [34]. To summarise, our contributions are:

- New consistency based semi-supervised semantic segmentation MT model designed to improve the segmentation accuracy of unlabelled training images with a new auxiliary teacher and a replacement of MT's MSE loss by a stricter confidence-weighted CE loss (Conf-CE) that allows stronger convergence and overall better training accuracy;
- A new challenging combination of input data, feature and network perturbations to improve model generalisation; and
- A new type of feature perturbation, called T-VAT, based on an adversarial noise learned from the both teachers of our MT model and applied to the student model, which results in the generation of challenging noise to promote an effective training of the student model.

Our experimental evaluation shows that our approach achieves the best results on *Pascal VOC 2012* [12]. Our approach also shows the best performance on Cityscapes [10].

2. Related Work

Below, we first discuss supervised semantic segmentation, then semi-supervised learning, and then we describe pseudo-labelling and consistency-based SSL methods.

Supervised semantic segmentation consists of classifying all image pixels into a closed set of visual classes. Current models are based on fully convolutional neural networks (FCN) [2, 30, 32] and extensions that explore: 1) multi-scale aspects of the image [11, 29], 2) pyramidal feature maps [15, 17, 43], 3) dilated convolutions [6, 8, 40], and 4) attention mechanisms [7, 27]. SOTA semi-supervised semantic segmentation models rely on the supervised semantic segmentation models DeeplabV3+ [5] and PSPNet [43] as backbone architectures.

Semi-supervised learning (SSL) trains a model using labelled and unlabelled images. Current SSL solutions are formulated based on three assumptions [38]: 1) smoothness: similar images have similar labels; 2) low-density: decision boundary does not pass through high-density areas of the feature space; and 3) manifold: samples on the same low-dimensional manifold embedded in the feature space have the same label. SSL methods can be loosely classified into **pseudo-label based SSL** [1, 3, 36] and **consistency based SSL** [23, 35, 37], with the former generally presenting worse accuracy than the latter. We believe that this is due to the fact that pseudo-label methods disregard part of the unlabelled training set during training, which can reduce their generalisation ability. Below, we focus on consistency based SSL given its superior accuracy on public benchmarks.

Consistency-based SSL methods aims to enforce the agreement between the predictions of perturbed unlabelled images, where perturbations can be applied to the input image, to the feature representation, or to the network. The effectiveness of consistency-based SSL depends on the prediction accuracy for the unlabelled images and the perturbations to challenge the model training. In general, more challenging perturbations target an improved generalisation, but if this perturbation is applied to inaccurate predictions, it can cause the consistency-based method to learn from wrong labels. Prediction accuracy can be improved in many ways, but a simple model ensemble strategies [23,37]. Perturbations can be applied to the input image [44], feature representation [34] or network [9, 21]. Independently of how they are applied, perturbations tend to be more effective when they challenge the classification process by, for example, moving perturbed feature closer to true (but hidden) classification boundaries, such as with virtual adversarial training (VAT) [33].

Consistency-based semi-supervised semantic segmentation methods have shown more competitive results [9, 14, 34] than pseudo-label approaches. Among the SOTA consistency-based semi-supervised semantic segmentation methods, PseudoSeg [44] relies on a new pseudo-labelling strategy and data augmentation consistency training to calibrate pseudo-labels, but its dependence on the generally inaccurate class activation maps can lead to poor training performance. Cross-consistency training (CCT) [34] applies different types of feature perturbations to enforce the

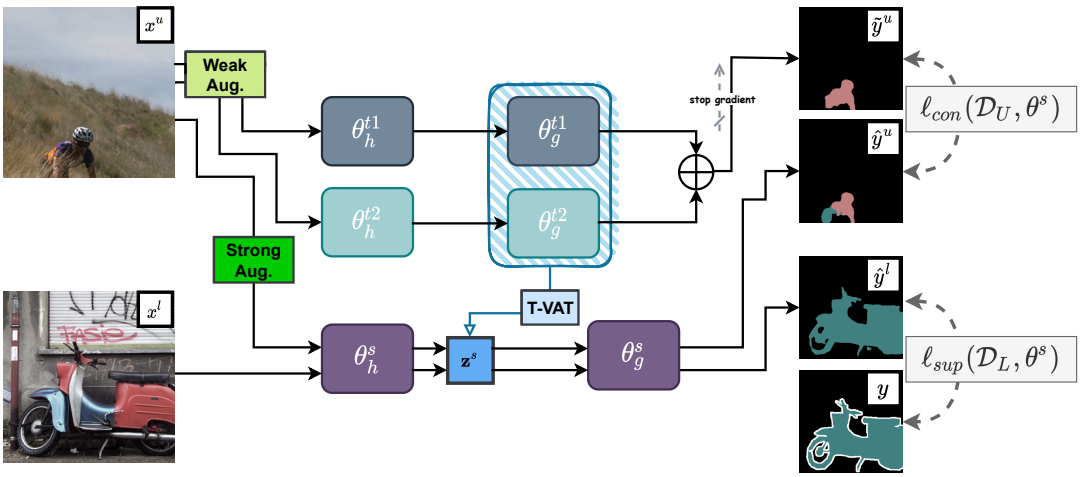


Figure 2. Illustration of our approach. The unlabelled image $\mathbf{x}^u$ is weakly augmented for the mean teachers (encoders parameterised by $\theta_h^{t1,t2}$ and decoders by $\theta_g^{t1,t2}$) that jointly predict the segmentation $\tilde{\mathbf{y}}^u$. The same unlabelled image is strongly augmented for the student (with encoder θ_h^s and decoder θ_g^s) that suffers T-VAT feature perturbation before predicting the segmentation $\hat{\mathbf{y}}^u$. This prediction of the unlabelled image and the prediction $\hat{\mathbf{y}}^l$ of the labelled image $\mathbf{x}^l$ (also perturbed by T-VAT) are used to minimise the consistency loss $\ell_{con}(.)$ and the supervised loss $\ell_{sup}(.)$, respectively, to train the student. The mean teachers are trained with EMA of the student model.

agreement between their semantic segmentation results and the segmentation from the non-perturbed feature. Although the feature perturbations used in CCT are effective, more targeted and accurate adversarial noise can be more helpful for the consistency regularization. Other methods have explored network perturbation [9, 13, 21, 31], where consistency is enforced between the responses of differently initialised models. Perturbation models depend on the ability of the models to produce accurate segmentation results, and as mentioned above, such ability can be improved with the use of model ensembling [37]. French et al. [14] explore model ensembling [37] together with network perturbation and input image perturbation [42]. This is one of the closest methods to our proposal, but we add a more effective model ensembling with multiple mean teachers, and a new adversarial feature perturbation with VAT [33] and challenging input image perturbation with CutMix [42] and Zoom In/Out [7, 28] data augmentation. Also comparing with [14], the more accurate segmentation results produced by our multiple mean teachers allows us to train the model for unlabelled images with the CE loss instead of the MSE used in [14], providing better training convergence and accuracy.

3. Method

Before we describe our model and training process, we introduce our dataset for semi-supervised semantic segmentation. We have a small labelled training set $\mathcal{D}_L = \{(\mathbf{x}_i, \mathbf{y}_i)\}_{i=1}^{|\mathcal{D}_L|}$, where $\mathbf{x}_i \in \mathcal{X} \subset \mathbb{R}^{H\times W\times C}$ is the input image of size $H \times W$ with C colour channels, and $\mathbf{y}_i \in \mathcal{Y} \subset \{0,1\}^{H\times W\times Y}$ is the segmentation map, with the number of visual classes denoted by Y. We also have a large unlabelled training set $\mathcal{D}_U = \{\mathbf{x}_i\}_{i=1}^{|\mathcal{D}_U|}$ is provided, with $|\mathcal{D}_L| << |\mathcal{D}_U|$. These datasets are used to train our proposed MT model with an auxiliary teacher, described in Sec. 3.1. The training of our new MT model, exploring network, feature and input image perturbations, and a strict Conf-CE loss is described in Sec. 3.2.

3.1. Multiple Mean Teachers and Student Models

As explained in Sec. 1, we aim to improve the accuracy of the segmentation of unlabelled training images. To achieve that, we propose the inclusion of an auxiliary teacher, exploring the idea of a double ensembling process to improve segmentation accuracy, namely the ensemble of teacher models, each representing a temporal ensemble of the student model [37]. The teachers and student models have the same network structure denoted by $f_\theta : \mathcal{X} \rightarrow \mathbb{R}^{H\times W\times Y}$, where θ is the model parameter. This model is decomposed into an encoder $h_{\theta_h} : \mathcal{X} \rightarrow \mathcal{Z}$ and a decoder $g_{\theta_g} : \mathcal{Z} \rightarrow \mathcal{Y}$, where $\mathcal{Z} \subset \mathbb{R}^Z$ represents the feature space of dimension Z. Hence, $f_\theta = g_{\theta_g} \circ h_{\theta_h}$, where $\theta = \{\theta_g, \theta_h\}$. The probability output of the network is achieved by applying the pixel-wise softmax function $\sigma(.)$ over the Y classes, as in $p_\theta(\mathbf{x}) = \sigma(f_\theta(\mathbf{x}))$. The multiple mean teacher-student model is represented with the respective parameter superscripts: $\theta^{t1} = \{\theta_g^{t1}, \theta_h^{t1}\}$ and $\theta^{t2} = \{\theta_g^{t2}, \theta_h^{t2}\}$ for the teachers, and $\theta^s = \{\theta_g^s, \theta_h^s\}$ for the student.

3.2. Training with Multiple Perturbations and a Strict Confidence-weighted CE Loss

In this section, we present the training process of our new MT model using a confidence-weighted CE loss, which is

optimised with perturbations to the network, feature representations and input images.

Training. The full training loss for the student model is

$$\ell(\mathcal{D}_L, \mathcal{D}_U, \theta^s) = \ell_{sup}(\mathcal{D}_L, \theta^s) + \beta \ell_{con}(\mathcal{D}_U, \theta^s), \quad (1)$$

where the first loss is the supervised segmentation loss, defined as:

$$\ell_{sup}(\mathcal{D}_L, \theta^s) = \frac{1}{|\mathcal{D}_L||\Omega|} \sum_{(\mathbf{x},\mathbf{y})\in\mathcal{D}_L} \sum_{\omega\in\Omega} \ell(\mathbf{y}(\omega), p_{\theta^s}(\mathbf{x})(\omega)), \quad (2)$$

where Ω is the image lattice of size $H \times W$, and $\ell(.)$ denotes the CE loss between the annotation $\mathbf{y}$ and segmentation prediction from $p_{\theta^s}(\mathbf{x})$ at pixel address $\omega \in \Omega$. The second term in (1) is the consistency loss, denoted by the confidence weighted CE loss (**Conf-CE**), with $\beta \in \mathbb{R}$ weighting its contribution and its definition being as follows:

$$\ell_{con}(\mathcal{D}_U, \theta^s) = \frac{1}{|\mathcal{D}_U||\Omega|} \sum_{\mathbf{x}\in\mathcal{D}_U} \sum_{\omega\in\Omega} c(\omega)\ell(\tilde{\mathbf{y}}(\omega), p_{\theta^s}(\mathbf{x})(\omega)), \quad (3)$$

where $\ell(.)$ represents the CE loss, ω denotes the pixel address of the output lattice Ω of the segmentation map, $\tilde{\mathbf{y}}(\omega) \in \{0,1\}^Y$ is the segmentation prediction from the teacher models at ω, $p_{\theta^s}(\mathbf{x})(\omega) \in [0,1]^Y$ is the student model segmentation prediction at ω, and $c(\omega) \in [0,1]$ represents the segmentation prediction confidence from the teacher models at ω, defined below in (4).

The **network perturbation** is obtained from the predictions of the mean teacher models and the student model. The soft segmentation map produced by the ensemble of the mean teachers is estimated as:

$$\hat{\mathbf{y}} = \sigma(0.5 \times (f_{\theta^{t1}}(\mathbf{x}) + f_{\theta^{t2}}(\mathbf{x}))), \quad (4)$$

where $\sigma(.)$ denotes the softmax function. The hard segmentation prediction by the ensemble of teachers, denoted by $\tilde{\mathbf{y}} \in \mathcal{Y}$ in (3), is obtained from the one-hot representation computed from $\hat{\mathbf{y}} \in [0,1]^{H\times W\times Y}$ in (4). The segmentation prediction confidence $c(\omega)$ in (3) is computed with $c(\omega) = \tilde{\mathbf{y}}(\omega)^\top \hat{\mathbf{y}}(\omega) \times \mathbb{I}(\tilde{\mathbf{y}}(\omega)^\top \hat{\mathbf{y}}(\omega) > \tau)$, where $\mathbb{I}(.)$ denotes an indicator function and τ is a minimum confidence to enable a value larger than zero for $c(\omega)$. Following the MT framework [37], while the student model is trained via stochastic gradient descent (SGD) to minimise the cost function in (3), both teacher models are trained with exponential moving average (EMA) [37] of the student model and batch norm (BN) parameters [4], with:

$$\theta^k = \gamma \times \theta^k + (1-\gamma) \times \theta^s, \quad (5)$$

where $k \in \{t_1, t_2\}$, and $\gamma \in (0,1)$ controls the transfer weight between epochs. For the training of teacher models, we update the parameters of only one of the two teachers at each training epoch.

The **feature perturbation** consists of a challenging adversarial feature perturbation that is designed to violate the cluster, or low-density, assumption [34, 38] by pushing the image features, computed from the model encoder, toward the classification boundaries in the feature space. One effective way to produce such adversarial feature noise is with virtual adversarial training (VAT) [33] that optimises a perturbation vector to maximise the divergence between correct and adversarial classification. Current methods estimate the adversarial noise using the same single network where the consistency loss will be applied [34]. However, the use of VAT to perturb the training of MT in semi-supervised semantic segmentation is new, to the best of our knowledge. In an MT model, the feature perturbation can be applied to the student model, but given that it has less accurate predictions than the teacher model, this approach may not be conducive to effective training. Hence we propose to estimate the adversarial noise using the more accurate teachers, and then apply this estimated noise to the feature of the student model – we call this feature perturbation **T-VAT**. The student output to be used in the loss in (3) is $p_{\theta^s}(\mathbf{x}) = \sigma(g_{\theta^s_g}(h_{\theta^s_h}(\mathbf{x}) + \mathbf{r}_{adv}))$, where the adversarial feature perturbation $\mathbf{r}_{adv} \in \mathcal{Z}$ is estimated from the response of the ensemble of teacher models with:

$$\begin{aligned} &\text{maximise} \;\; d\Big(\sigma\big(0.5 \times (g_{\theta^{t1}_g}(\mathbf{z}^s) + g_{\theta^{t2}_g}(\mathbf{z}^s))\big), \\ &\qquad\qquad \sigma\big(0.5 \times (g_{\theta^{t1}_g}(\mathbf{z}^s + \mathbf{r}_{adv}) + g_{\theta^{t2}_g}(\mathbf{z}^s + \mathbf{r}_{adv}))\big)\Big), \\ &\text{subject to} \;\; ||\mathbf{r}_{adv}||_2 <= \epsilon, \end{aligned} \quad (6)$$

where $\mathbf{z}^s = f_{\theta^s}(\mathbf{x})$, $d(.)$ is the the sum of the pixel-wise Kullback-Leibler (KL) divergence between the original and perturbed pixel predictions.

The **input image perturbation** is based on the weak-strong augmentation pairs [24], where weak augmentations (image flipping, cropping and scaling) are applied to the images to be processed by the teacher models, and in addition to those weak augmentation, strong augmentations [9, 21] (e.g., colour jitter, randomise grayscale and blur) are applied to the images fed to the student model to improve the overall generalization capability.

On top of the strong augmentations, we also apply the CutMix [42] and Zoom In/Out [7, 28] data augmentations to the student model images. As defined in [14], the CutMix augmentation is achieved by applying a binary mask $\mathbf{m} \in \{0,1\}^{H\times W}$ that combines two images using the function $\mu(\mathbf{x}_i, \mathbf{x}_j, \mathbf{m}) = (1-\mathbf{m}) \odot \mathbf{x}_i + \mathbf{m} \odot \mathbf{x}_j$. We can apply CutMix by combining two input images and minimise the consistency loss (3) with the prediction from (4) [9] (referred to as *CutMix before prediction*), or we can minimise the consistency loss using the CutMix combination of the images and their predictions, as in

$$\ell(\mu(\tilde{\mathbf{y}}_i, \tilde{\mathbf{y}}_j, \mathbf{m}), p_{\theta^s}(\mu(\mathbf{x}_i, \mathbf{x}_j, \mathbf{m}))), \quad (7)$$

with $\tilde{\mathbf{y}}$ defined in (4). The perturbation used in (7) is referred to as *CutMix after prediction*, which we argue to

produce a cleaner prediction for the consistency loss than *CutMix before prediction* because its prediction does not contain the artifacts introduced by the prediction from the CutMix images. The Zoom In/Out augmentation [7, 28] is defined by the function $\zeta(\mathbf{x}, s)$ that zooms in or out the image using the parameter $s \in \mathbb{R}_+$. The input image consistency loss in (3) for the zoom in/out augmentation for the ensemble results of teacher models is defined by

$$\ell(\zeta(\tilde{y}, s), p_{\theta^s}(\zeta(\mathbf{x}, s))), \tag{8}$$

and $\tilde{y}$ is defined in (4).

The segmentation loss $\ell(.)$ in (3) for previous consistency-based semi-supervised semantic segmentation models [9, 14] is usually based on the L2 loss. Even though L2 loss is known to be robust, which is advantageous when dealing with the noisy predictions produced by consistency-based methods, it is also known to have poor converge and to possibly lead to vanishing gradients. Given the reliability of the segmentation predictions produced by our extended MT model, we instead use the more effective cross entropy (CE) loss, constrained to be computed at regions of high-confidence segmentation results, represented by $c(\omega)$ in (3), following the strategy applied in self-training approaches [18, 39, 41].

Inference. The semantic segmentation of a test image is obtained from the teachers, as described in (4).

4. Experiments

We firstly introduce the experimental setting that we used to evaluate our approach. In Sec. 4.2 we demonstrate our approach for both datasets under different partition protocols by comparing them with the supervised baselines and other previous SOTA approaches. Moreover, we also carry out detailed results based on the few supervision studies in Sec. 4.3. Lastly, we perform the ablation study in Sec. 4.4 and demonstrate an extension experiment based on the exploring of the image-level data in Sec. 4.6.

4.1. Experimental Setup

Datasets. *Pascal VOC 2012* [12] is the standard dataset used for evaluating the performance of the semi-supervised segmentation models. The dataset contains more than $13,000$ images with 21 classes, providing $1,464$ images with pixel-level labels for training, $1,449$ images for validation and $1,456$ for testing. Following previous papers [9, 34], we adopt the additional labels from [16] and our entire training set contains $10,582$ images. Note that the labels from [16] are of low-quality, and may contain noise. *Cityscapes* [10] is the urban driving scene dataset, which consists of $2,975$ images for training, 500 for validation and $1,525$ for testing. Each image in the dataset has resolution $2,048 \times 1,024$, and there are 19 classes in total.

Following [9, 21], we random split the full labelled set in *Pascal VOC 2012* and *Cityscapes* with different ratios. We also provide the results based on the official labelled set (with $1,464$ images for *Pascal VOC 2012*) [41, 44].

Implementation details. Most results are based on using our method to train the DeeplabV3+ [5] model. We load the ImageNet pre-trained checkpoint, and the segmentation heads are initialized randomly. Following previous papers [9, 18, 34], we utilise the following polynomial learning-rate decay: $(1 - \frac{\text{iter}}{\text{max_iter}})^{0.9}$. We also test our method on PSPNet [22, 34] to show the generalization of our approach.

During training, we apply data augmentation with random scaling in $\{0.5, 0.75, 1.25\}$ and random flipping of both labelled and unlabelled images. On *Pascal VOC 2012*, we crop images to 512×512 pixels for DeeplabV3+, train for 80 epochs with base learning rate set to 0.01, using batch size 32, for both labelled and unlabelled images, following [9]. For PSPNet, we follow [18] and crop images to 321×321 pixels and use batch size 8. On *Cityscapes*, due to hardware limitation, we crop images to 712×712 pixels, train for 450 epochs with base learning rate set to 0.0038 with batch size 8 for both architectures. Because the teacher's predictions are unstable at the early stage of the training, we apply the Gaussian ramp-up to the consistency loss weight β in (1). For both datasets, the supervised loss is the cross-entropy loss.

Evaluation metrics. Following previous papers [9, 21], we report the mean Intersection-over-Union (mIoU) for validation set for both datasets. All the results are based on the single scale inference.

4.2. Results on Different Partition Protocols

In this section, following [9, 21], we evaluate our method based on sub-sampling the datasets with ratio $1/n$ for the labelled set and $(1 - 1/n)$ unlabelled set. Specifically, in the *Pascal VOC 2012* dataset, we split the entire training set (with $10,582$ images) with the ratios of $1/16$, $1/4$, $1/8$, $1/2$ for be labelled set. In the *Cityscapes*, we similarly utilise the ratios $1/8$, $1/4$, $1/2$ for the labelled set. All the partition protocols are from [9].

Improvements over Supervised Baselines. We first compare our results with fully supervised learning (trained with the same ratio of labelled set) using DeepLabV3+ architecture, and illustrate the improvements in Fig. 3. This figure demonstrates that our approach successfully exploits unlabelled data, with a dramatic performance boost. On *Pascal VOC 2012*, Fig. 3-(a) shows that our approach outperforms the supervised baseline by a large gap, especially for small labelled partitions. Specifically, in the $1/16$ ratio (with 662 labelled images), our approach surpasses the fully supervised baseline with 6.01% and 5.97% for the ResNet50 and ResNet101, respectively. On the other set-

Table 1. **Comparison with SOTA on Pascal VOC 2012.** All approaches are based on the DeeplabV3+ architecture, under different partition protocols [9]. The $*$ indicates the approaches re-implemented by [9]. Best results are in bold.

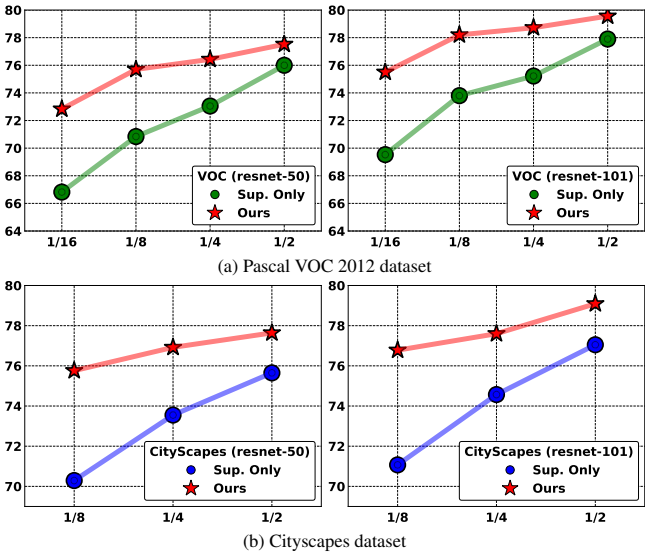
Method	ResNet-50				ResNet-101			
	1/16(662)	1/8(1323)	1/4(2646)	1/2(5291)	1/16(662)	1/8(1323)	1/4(2646)	1/2(5291)
MT* [37]	66.77	70.78	73.22	75.41	70.59	73.20	76.62	77.61
French et al.* [14]	68.90	70.70	72.46	74.49	72.56	72.69	74.25	75.89
CCT* [34]	65.22	70.87	73.43	74.75	67.94	73.00	76.17	77.56
GCT* [21]	64.05	70.47	73.45	75.20	69.77	73.30	75.25	77.14
ECS [31]	-	67.38	70.70	72.89	-	-	-	-
CPS [9]	71.98	73.67	74.90	76.15	74.48	76.44	77.68	78.64
CAC [22]	70.10	72.40	74.00	-	72.40	74.60	76.30	-
Ours	**72.83**	**75.70**	**76.43**	**77.88**	**75.50**	**78.20**	**78.72**	**79.76**

Table 2. **Comparison with SOTA on Cityscapes** under different partition protocols. All the approaches are based on DeeplabV3+. The † indicates the experimental settings (e.g., supervised loss, resolution) follow CPS [9].

Method	Backbone	1/8	1/4	1/2
ECS [31]	ResNet50	67.38	70.70	72.89
CAC [22]	ResNet50	69.70	72.70	-
Ours	ResNet50	74.37	75.15	76.02
Ours (sliding eval.)	ResNet50	75.76	76.92	77.64
	ResNet101	76.89	77.60	79.09
GCT [21]†	ResNet50	71.33	75.30	77.09
CPS [9]†	ResNet50	76.61	77.83	78.77
Ours†	ResNet50	77.12	78.38	79.22

tings, our approach also shows consistent improvements between 2% and 5% for ratios 1/8, 1/4, and 1/2. On *Cityscapes*, we use the **sliding evaluation** to evaluate our final results following [9]. Fig. 3-(b) shows that our approach surpasses the supervised baseline by 2% and 6% for ResNet50 and ResNet101 for all protocols.

Comparison to SOTA. For *Pascal VOC 2012*, Tab. 1 shows that our approach is the best for all partition protocols, using DeepLabV3+ and ResNet50 and ResNet101 backbones. Comparing to French et al. [14], which is considered a strong baseline, our approach improves by 3% to 6% in all cases. Our approach also provides a significant boost for the original MT in all cases. The results also show that our approach is better than the current SOTA CPS [9] by around 1% to 2% for all cases. In some partition protocols, our approach is better than the CPS [9] with fewer labelled samples. For example, our approach trained with $1,323$ labelled images outperforms CPS [9] trained with $2,646$ labelled images using both backbones. This demonstrates that our perturbed and strict mean teachers yield more accurate results than any other method in the field. On *Cityscapes*, we use the settings from [22] and show results that use similar settings (in terms of image resolution, batch size, and supervised loss function) for fair comparison. Our approach outperforms CAC [22] by nearly 4.6% and 2.4% for the 1/8 and 1/4 partition protocols. The sliding evaluation also boosts our performance by approximately 1.30% for all the ratios. This shows that the the sliding process improves the performance of our approach in large resolution images.



(a) Pascal VOC 2012 dataset

(b) Cityscapes dataset

Figure 3. **Improvements over the supervised baseline.** mIoU vs. partition protocols results of our approach and the supervised baseline on *Pascal VOC 2012* (top) and *Cityscapes* (bottom) using ResNet50 and ResNet101 backbones and DeepLabV3+.

4.3. Results on Official Labelled Set of Pascal VOC 2012

In this section, we report the results based on the official $1,464$ high quality labelled data of the *Pascal VOC 2012*. We compare our approach for both PSPNet and DeepLabV3+ architectures. Table 3 shows that our approach outperforms the SOTA methods for each architecture and backbone settings. For PSPNet, our approach outperforms DARS [18] by 0.7% mIoU and CCT [34] by 5.19%. In the experiments, our approach outperforms other SOTA approaches by a large gap. Yuan et. al [41] only

Table 3. **Comparison using the official** $(1,464)$ **labelled images** on the Pascal VOC 2012 under different network settings. Best results are in bold.

Method	Architecture	Backbone	mIoU
CCT [34]	PSPNet	ResNet50	69.40
DARS [18]	PSPNet	ResNet50	73.89
ours	PSPNet	ResNet50	**74.59**
CAC [22]	DeeplabV3+	ResNet50	74.50
ours	DeeplabV3+	ResNet50	**78.08**
PseudoSeg [44]	DeeplabV3+	ResNet101	73.20
Yuan et al. [41]	DeeplabV3+	ResNet101	75.00
ours	DeeplabV3+	ResNet101	**80.01**

Table 4. **Comparison with SOTA approaches with few-supervision using the official** $(1,464)$ **labelled images** on the Pascal VOC 2012. Our approach follows the same protocols as CPS [9] and PseudoSeg [44]. Best results are in bold.

Method	732	366	183	92
AdvSemSeg [20]	65.27	59.97	47.58	39.69
CCT [34]	62.10	58.80	47.60	33.10
MT [37]	69.16	63.01	55.81	48.70
GCT [21]	70.67	64.71	54.98	46.04
VAT	63.34	56.88	49.35	36.92
French et al. [14]	69.84	68.36	63.20	55.58
PseudoSeg [44]	72.41	69.14	65.50	57.60
CPS [9]	75.88	71.71	67.42	64.07
ours	**78.42**	**76.57**	**69.58**	**65.80**

utilize the single network to produce the pseudo labels in a self-training manner. For instance, our approach outperforms Yuan et. al [41] by 5.01% mIoU, which shows the value of our teachers and student network and several perturbation strategies, compared with their self-training single network approach.

Few-supervision study. We subsample with partitions $1/2$, $1/4$, $1/8$ and $1/16$ using the official $1,464$ labelled images. The remaining data are combined with the augmented set [16] (around $9K$ images) to be the unlabelled data in the experiments. Table 4 shows that our approach yields the best mIoU results for all cases. For example, our approach outperforms CPS [9] by 4.86% for 366 labelled images. We argue that our effective perturbations allowed the better generalisation of our model under this limited labelled data conditions.

4.4. Ablation Study

In this section, we study the roles of the confidence weighted CE loss (conf-CE), T-VAT perturbation and auxiliary teacher (AT) of our approach. All the experiments are run on *Pascal VOC 2012* under $1/8$ ratio, and we use DeeplabV3+ to evaluate our results. Table 5 demonstrates the improvements of each component mentioned above,

Table 5. **Ablation study** using the $1/8$ labelled ratio on *Pascal VOC 2012* under DeeplabV3+ architecture.

MT	conf-CE	T-VAT	AT	Backbone	
				ResNet-50	ResNet-101
✓				71.49	73.50
✓	✓			73.79	76.39
✓	✓	✓		74.87	77.36
✓	✓	✓	✓	**75.70**	**78.20**

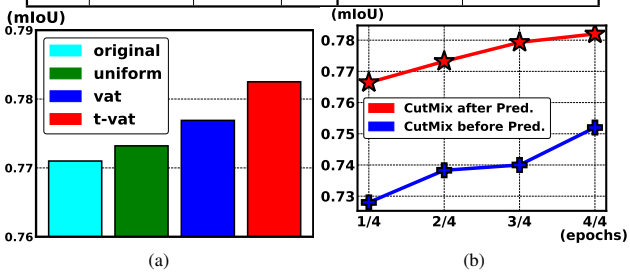


Figure 4. **T-VAT and CutMix Perturbation Effectiveness.** (a) mIoU under different feature perturbations. (b) mIoU of validation set during training for CutMix applied before or after prediction, as described in (7).

where we use MT [37] trained with the input image perturbations from Sec. 3.2 and MSE loss as baseline. We note that by replacing MSE by our conf-CE increases mIoU by 2.30% and 2.81% for the ResNet50 and ResNet101. T-VAT perturbation yields nearly 1% improvements, showing the effectiveness of our proposed feature perturbation. The more accurate predictions by the auxiliary teacher allows a further improvement of 0.83% and 0.84% for the two backbones.

T-VAT perturbation. Fig. 4-(a) shows the performance under different types of feature perturbations, namely: original (no feature perturbation), uniform (feature noise randomly sampled from uniform distribution), vat (VAT noise learned from the student model), t-vat (T-VAT noise learned from the teacher model as in (6)). Our proposed T-VAT outperforms uniform and VAT perturbations by 0.93% and 0.62%, respectively. Additionally, it also surpasses original by 1.10%.

Empirical results of the CutMix before or after prediction, as described in (7). In Fig 4-(b), we show the mIoU results on the validation set during training epochs. Applying CutMix before predictions may introduce extra semantic complexity and yield inaccurate pseudo-labels, leading to ineffective optimisation. In contrast, the result indicates that applying the CutMix after the prediction improves mIoU by around 3%.

Average gradient magnitudes difference between MSE and Conf-CE. Fig. 5 shows the average gradient magnitudes per layer of the student model after being trained with MSE and Conf-CE losses to optimise consistency in (3), at the last stages of training (70th out of 80

Table 6. **Comparison using the official** $(1,464)$ **labelled images with combined pseudo-label and consistency-based losses** on the Pascal VOC 2012 under different network settings. Best results per architecture are in bold.

Method	Architecture	Backbone	mIoU
CCT [34]	PSPNet	ResNet50	73.2
ours	PSPNet	ResNet50	**75.74**
AdvCAM [26]	Deeplabv2	ResNet101	77.8
PseudoSeg [44]	DeeplabV3+	ResNet50	73.8
CAC [22]	DeeplabV3+	ResNet50	76.1
ours	DeeplabV3+	ResNet50	78.71
ours	DeeplabV3+	ResNet101	**81.19**

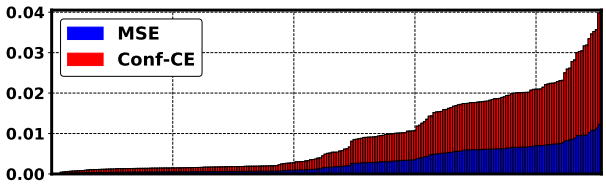


Figure 5. **Different gradient magnitudes**. Average gradient magnitudes per layer of the student model after being trained with **MSE** and **Conf-CE** losses, at the last stages of training.

epochs). Note that our Conf-CE shows larger gradient magnitudes than MSE [9, 14, 21], suggesting that it can lead to stronger convergence than MSE.

4.5. Qualitative Results

Figure 6 shows the supervised, student and mean teachers results on *Pascal VOC 2012* images. The supervised results display the worst accuracy in column (c), caused by the limited labelled training samples. Our final results in column (e) significantly improves the baseline performance, which demonstrates the effectiveness of our approach. Moreover, our final results in column(e) are also more accurate than the student results in column (d).

4.6. Combining Pseudo-label and Consistency-based losses

Current consistency-based methods [19, 22, 25, 34] also include a pseudo-labelling loss, involving the use of class activation maps (CAM) from the model to generate pseudo-labels $\bar{\mathbf{y}}$. We follow a similar strategy as in CCT [34] and add the CAM loss below to the cost function in (1) to train the student model:

$$\ell_{cam}(\mathcal{D}_U, \theta^s) = \frac{1}{|\mathcal{D}_U||\Omega|} \sum_{\mathbf{x}\in\mathcal{D}_U} \sum_{\omega\in\Omega} (1 - c(\omega))\ell(\bar{\mathbf{y}}(\omega), f_{\theta^s}(\mathbf{x})(\omega)), \tag{9}$$

where $c(\omega)$ is the segmentation confidence from the teachers defined in (3), and $\ell(.)$ is the CE loss. Hence, the non-confident predictions from the teachers are then handled by this CAM loss in (9).

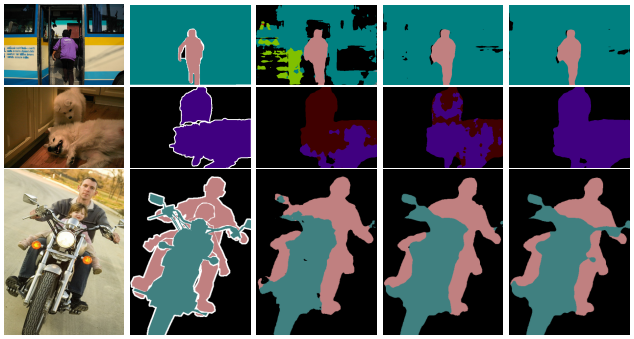


Figure 6. **Qualitative results** from *Pascal VOC 2012*. (a) input images, (b) ground truth, (c) supervised baseline results, (d) student results and (e) results by our approach.

We run experiments based on the official $1,464$ labelled images on *Pascal VOC 2012* with the additional $\approx 9K$ images used to minimise the CAM loss in (9). Results on Tab. 6 show that our method outperforms all previous works that use a similar strategy [22, 34, 44]. Moreover, the addition of this CAM loss in (9) to our cost function boosts the performance by 0.63% and 1.18% for the two backbones under DeeplabV3+ architecture, compared with our previous results in Tab. 3.

5. Conclusion

In this paper, we proposed a new consistency-based semi-supervised semantic segmentation method. Among our contributions, we introduced a new MT model, based on multiple mean teachers and a student network, which shows more accurate predictions for unlabelled images that facilitate consistency learning, allowing us to use a stricter confidence-based CE loss than the original MT's MSE loss. This more accurate predictions also allowed us to use a challenging combination of network, feature and input image perturbations that showed better generalisation. Furthermore, we proposed a new adversarial feature perturbation, called T-VAT, that further improved the generalisation of our approach. Our method outperforms previous methods on Pascal VOC 2012 and Cityscapes, becoming the new SOTA for the semi-supervised semantic segmentation field. Regarding the limitations of our model, it can be argued that the strict Conf-CE loss has the potential to overfit the remaining prediction mistakes, so we will focus on improving the robustness of the Conf-CE loss. Another limitation that we plan to address is to work on an approach that can handle high-resolution images without using the time-consuming sliding evaluation.

References

[1] Eric Arazo, Diego Ortego, Paul Albert, Noel E. O'Connor, and Kevin McGuinness. Pseudo-labeling and confirmation bias in deep semi-supervised learning. In *2020 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8, 2020. 2

[2] Vijay Badrinarayanan, Alex Kendall, and Roberto Cipolla. Segnet: A deep convolutional encoder-decoder architecture for image segmentation. *IEEE transactions on pattern analysis and machine intelligence*, 39(12):2481–2495, 2017. 2

[3] David Berthelot, Nicholas Carlini, Ian Goodfellow, Nicolas Papernot, Avital Oliver, and Colin Raffel. Mixmatch: A holistic approach to semi-supervised learning. *arXiv preprint arXiv:1905.02249*, 2019. 2

[4] Zhaowei Cai, Avinash Ravichandran, Subhransu Maji, Charless Fowlkes, Zhuowen Tu, and Stefano Soatto. Exponential moving average normalization for self-supervised and semi-supervised learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 194–203, 2021. 4

[5] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE transactions on pattern analysis and machine intelligence*, 40(4):834–848, 2017. 2, 5

[6] Liang-Chieh Chen, George Papandreou, Florian Schroff, and Hartwig Adam. Rethinking atrous convolution for semantic image segmentation. *arXiv preprint arXiv:1706.05587*, 2017. 2

[7] Liang-Chieh Chen, Yi Yang, Jiang Wang, Wei Xu, and Alan L Yuille. Attention to scale: Scale-aware semantic image segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3640–3649, 2016. 2, 3, 4, 5

[8] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *Proceedings of the European conference on computer vision (ECCV)*, pages 801–818, 2018. 2

[9] Xiaokang Chen, Yuhui Yuan, Gang Zeng, and Jingdong Wang. Semi-supervised semantic segmentation with cross pseudo supervision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2613–2622, 2021. 1, 2, 3, 4, 5, 6, 7, 8

[10] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3213–3223, 2016. 2, 5

[11] Jifeng Dai, Kaiming He, and Jian Sun. Convolutional feature masking for joint object and stuff segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3992–4000, 2015. 2

[12] Mark Everingham, SM Ali Eslami, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes challenge: A retrospective. *International journal of computer vision*, 111(1):98–136, 2015. 2, 5

[13] Zhengyang Feng, Qianyu Zhou, Guangliang Cheng, Xin Tan, Jianping Shi, and Lizhuang Ma. Semi-supervised semantic segmentation via dynamic self-training and classbalanced curriculum. *arXiv preprint arXiv:2004.08514*, 1(2):5, 2020. 3

[14] Geoff French, Timo Aila, Samuli Laine, Michal Mackiewicz, and Graham Finlayson. Semi-supervised semantic segmentation needs strong, high-dimensional perturbations. 2019. 1, 2, 3, 4, 5, 6, 7, 8

[15] Golnaz Ghiasi and Charless C Fowlkes. Laplacian pyramid reconstruction and refinement for semantic segmentation. In *European conference on computer vision*, pages 519–534. Springer, 2016. 2

[16] Bharath Hariharan, Pablo Arbeláez, Lubomir Bourdev, Subhransu Maji, and Jitendra Malik. Semantic contours from inverse detectors. In *2011 International Conference on Computer Vision*, pages 991–998. IEEE, 2011. 1, 5, 7

[17] Junjun He, Zhongying Deng, Lei Zhou, Yali Wang, and Yu Qiao. Adaptive pyramid context network for semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7519–7528, 2019. 2

[18] Ruifei He, Jihan Yang, and Xiaojuan Qi. Re-distributing biased pseudo labels for semi-supervised semantic segmentation: A baseline investigation. *arXiv preprint arXiv:2107.11279*, 2021. 5, 6, 7

[19] Zilong Huang, Xinggang Wang, Jiasi Wang, Wenyu Liu, and Jingdong Wang. Weakly-supervised semantic segmentation network with deep seeded region growing. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7014–7023, 2018. 8

[20] Wei-Chih Hung, Yi-Hsuan Tsai, Yan-Ting Liou, Yen-Yu Lin, and Ming-Hsuan Yang. Adversarial learning for semi-supervised semantic segmentation. *arXiv preprint arXiv:1802.07934*, 2018. 7

[21] Zhanghan Ke, Di Qiu, Kaican Li, Qiong Yan, and Rynson WH Lau. Guided collaborative training for pixel-wise semi-supervised learning. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XIII 16*, pages 429–445. Springer, 2020. 1, 2, 3, 4, 5, 6, 7, 8

[22] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1205–1214, 2021. 5, 6, 7, 8

[23] Samuli Laine and Timo Aila. Temporal ensembling for semi-supervised learning. *arXiv preprint arXiv:1610.02242*, 2016. 2

[24] Dong-Hyun Lee et al. Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks. In *Workshop on challenges in representation learning, ICML*, volume 3, page 896, 2013. 4

[25] Jungbeom Lee, Eunji Kim, Sungmin Lee, Jangho Lee, and Sungroh Yoon. Ficklenet: Weakly and semi-supervised semantic image segmentation using stochastic inference. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5267–5276, 2019. 8

[26] Jungbeom Lee, Eunji Kim, and Sungroh Yoon. Anti-adversarially manipulated attributions for weakly and semi-supervised semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4071–4080, 2021. 8

[27] Xia Li, Zhisheng Zhong, Jianlong Wu, Yibo Yang, Zhouchen Lin, and Hong Liu. Expectation-maximization attention networks for semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9167–9176, 2019. 2

[28] Di Lin, Yuanfeng Ji, Dani Lischinski, Daniel Cohen-Or, and Hui Huang. Multi-scale context intertwining for semantic segmentation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 603–619, 2018. 3, 4, 5

[29] Guosheng Lin, Chunhua Shen, Anton Van Den Hengel, and Ian Reid. Efficient piecewise training of deep structured models for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3194–3203, 2016. 2

[30] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3431–3440, 2015. 2

[31] Robert Mendel, Luis Antonio De Souza, David Rauber, João Paulo Papa, and Christoph Palm. Semi-supervised segmentation based on error-correcting supervision. In *European Conference on Computer Vision*, pages 141–157. Springer, 2020. 3, 6

[32] Shervin Minaee, Yuri Y Boykov, Fatih Porikli, Antonio J Plaza, Nasser Kehtarnavaz, and Demetri Terzopoulos. Image segmentation using deep learning: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021. 2

[33] 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. 2, 3, 4

[34] Yassine Ouali, Céline Hudelot, and Myriam Tami. Semi-supervised semantic segmentation with cross-consistency training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12674–12684, 2020. 1, 2, 4, 5, 6, 7, 8

[35] Boris T Polyak and Anatoli B Juditsky. Acceleration of stochastic approximation by averaging. *SIAM journal on control and optimization*, 30(4):838–855, 1992. 2

[36] 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. *arXiv preprint arXiv:2001.07685*, 2020. 2

[37] Antti Tarvainen and Harri Valpola. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. *arXiv preprint arXiv:1703.01780*, 2017. 2, 3, 4, 6, 7

[38] Jesper E Van Engelen and Holger H Hoos. A survey on semi-supervised learning. *Machine Learning*, 109(2):373–440, 2020. 2, 4

[39] Lihe Yang, Wei Zhuo, Lei Qi, Yinghuan Shi, and Yang Gao. St++: Make self-training work better for semi-supervised semantic segmentation. *arXiv preprint arXiv:2106.05095*, 2021. 5

[40] Fisher Yu and Vladlen Koltun. Multi-scale context aggregation by dilated convolutions. *arXiv preprint arXiv:1511.07122*, 2015. 2

[41] Jianlong Yuan, Yifan Liu, Chunhua Shen, Zhibin Wang, and Hao Li. A simple baseline for semi-supervised semantic segmentation with strong data augmentation. *arXiv preprint arXiv:2104.07256*, 2021. 5, 6, 7

[42] 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 *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6023–6032, 2019. 3, 4

[43] Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia. Pyramid scene parsing network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2881–2890, 2017. 2

[44] Yuliang Zou, Zizhao Zhang, Han Zhang, Chun-Liang Li, Xiao Bian, Jia-Bin Huang, and Tomas Pfister. Pseudoseg: Designing pseudo labels for semantic segmentation. *arXiv preprint arXiv:2010.09713*, 2020. 1, 2, 5, 7, 8