

A Three-Stage Self-Training Framework for Semi-Supervised Semantic Segmentation

Rihuan Ke^{*1}, Angelica Aviles-Rivero^{*1}, Saurabh Pandey³, Saikumar Reddy⁴, and Carola-Bibiane Schönlieb¹

¹Centre for Mathematical Sciences, University of Cambridge, Cambridge CB3 0WA, UK. [†]

²KritiKal Solutions[‡]

Abstract

Semantic segmentation has been widely investigated in the community, in which the state of the art techniques are based on supervised models. Those models have reported unprecedented performance at the cost of requiring a large set of high quality segmentation masks. To obtain such annotations is highly expensive and time consuming, in particular, in semantic segmentation where pixel-level annotations are required. In this work, we address this problem by proposing a holistic solution framed as a three-stage self-training framework for semi-supervised semantic segmentation. The key idea of our technique is the extraction of the pseudo-masks statistical information to decrease uncertainty in the predicted probability whilst enforcing segmentation consistency in a multi-task fashion. We achieve this through a three-stage solution. Firstly, we train a segmentation network to produce rough pseudo-masks which predicted probability is highly uncertain. Secondly, we then decrease the uncertainty of the pseudo-masks using a multi-task model that enforces consistency whilst exploiting the rich statistical information of the data. We compare our approach with existing methods for semi-supervised semantic segmentation and demonstrate its state-of-the-art performance with extensive experiments.

1. Introduction

Semantic segmentation is a fundamental task in computer vision, it aims to assign a label, from a set of predefined classes, to each pixel in the image. This task has been widely explored in the literature yet not fully solved. The current state-of-the-art models are based on building upon deep nets [30, 5, 34, 53, 54]. Whilst these techniques have reported unprecedented results, they rely on a very high la-

belled data regime. This is a strong assumption as annotations are pixel-level; which is expensive, time consuming and inherent to human bias. To address the lack of a large and well-representative set of labels, one could rely more on unlabelled samples.

An alternative is to use solely unlabelled data – i.e. unsupervised learning. However, this paradigm has not been successful for semantic segmentation as the performance substantially degrades due to the lack of correspondence between the samples and classes. Another option is to use weakly supervised techniques [25, 40], nonetheless, the rich information from unlabelled samples is not fully exploited and performance is still limited. A feasible option is to use semi-supervised learning that leverages on a vast amount of labelled data and a tiny set of annotations. Although semi-supervised learning (SSL) has been widely developed [4] in the community, the progress of deep semi-supervised learning has been only noticeable in the last few years, and mainly for the task of image classification e.g. [24, 32, 44], which principles have been used recently in the context of semantic segmentation e.g. [17, 11, 36, 10].

The existing techniques for SSL can be broadly divided in entropy minimisation [13], generative models [23], graph based techniques [56], proxy-based techniques (building upon pseudo-labels/self-training) [46], consistency regularisation [24, 32] and holistic approaches that take the best of each principle. For semantic segmentation existing techniques use generative models and consistency regularisation techniques. Although the results reported for this task are promising, there is plenty of room for improvement, and in particular, regarding how to improve the confidence score predictions.

With this purpose, we propose a new framework for semantic segmentation under the assumption of a very low labelled data regime. Our holistic solution is framed as a three-stage self-training technique for semi-supervised learning. Every stage of our framework has a purpose: **Stage 1** generates initial pseudo-masks (in the sense of

^{*}equal contribution

[†]Email: {rk621, ai323, cbs31}@cam.ac.uk

[‡]Email: {saurabh, saikumar.reddy}@kritikalvision.ai

pseudo-labelling) by a segmentation network that is trained using the tiny labelled set in a supervised fashion, [Stage 2](#) is cast as a multi-task model that learns the pseudo mask statistics (Task 2) and the segmentation (Task 1) that are used to produce higher quality pseudo-masks, and in [Stage 3](#) the updated pseudo-masks along with the tiny annotation set are used to train the segmentation network. Our contributions are as follows:

- We propose a novel self-training framework for deep semi-supervised semantic segmentation, in which we highlight:
 - An end-to-end optimisation model framed into a three-stages solution. Our model reduces the uncertainty of the pseudo-mask predicted probability using a multi-task model, which enforces segmentation consistency whilst exploiting the statistical information of the data.
 - We introduce a new perspective for semantic segmentation – the holistic principle. We demonstrate that whilst consistency regularisation is important, one needs to account for uncertainty in the predicted pseudo-masks. We show that learning information from both sources allows for increasing the certainty predictions.
- We validate our technique on a range of numerical and visual results and compared it against the current state-of-the-art techniques for deep semi-supervised semantic segmentation. We demonstrate that our well-designed model fulfils its purpose and, at this point in time, the proposed technique outperforms the current SOTA semantic segmentation models that rely on very limited ground truth annotations.

2. Related Work

Since the seminal work of Long [\[30\]](#), where fully convolutional networks demonstrated potentials to transfer the learnt representation to the task of segmentation, substantial progress has been done using deep supervised techniques for segmentation including [\[2, 5, 34, 38\]](#). More recently, sophisticated mechanisms have been combined to overcome the performance limits of existing techniques including architecture search [\[33, 29, 52, 27\]](#), attention mechanism [\[53, 50, 51, 16\]](#) and re-designing the principles of several architectures [\[47, 26, 37, 54, 55\]](#).

Despite the astonishing results reported by these techniques, a major communal drawback is the assumption of a setting under high labelled data regime. This has motivated the fast development of techniques relying less on annotations, such as the weakly supervised segmentation e.g. [\[1, 25, 40, 19, 42, 18\]](#). A more recent focus of attention has been recent developments of deep semi-supervised

learning, which is the focus of this work. In this section, we review the existing techniques in turn.

Deep Semi-Supervised Learning. Semi-supervised learning (SSL) has been widely investigated since early developments in the area [\[4\]](#). However, in the last few years there has been a substantial increase of interest on this paradigm, this is particularly because the underpinning theory of SSL has been combined with the power of deep nets reporting impressive results that readily compete with fully supervised techniques. This performance gain has been mainly reported in the context of image classification, where several techniques have been developed e.g. [\[24, 32, 44\]](#). However, *there is a significant difference between image classification and semantic segmentation as the last one involves more dense and complex predictions.*

Deep SSL for image classification can be achieved using different principles, in which the major success has been achieved through consistency regularisation [\[24, 43, 39, 32, 3\]](#). The core idea of this philosophy is that unlabelled samples ($x_u \in \mathcal{D}_u$ denoting the unlabelled set) under induced perturbations, δ , should not change the performance output such that $f(x_u) = f(x_u + \delta)$. This principle relaxes the clustering assumption of SSL by enforcing an equivalent form of it by pushing the decision boundary to low-density regions. Although the principles applied in image classification can be somehow extrapolated to semantic segmentation *the inherent gap between tasks prevent semantic segmentation techniques to reach similar high performance than the one reported in classification* e.g. [\[11\]](#), and therefore, one needs to rethink the design of deep SSL for semantic segmentation.

Deep Semi-Supervised Semantic Segmentation. The annotations quality plays a fundamental role in the performance of the techniques. In particular, in the task of semantic segmentation labelling is overly expensive. For example, a single image from the segmentation benchmark dataset Cityspaces [\[6\]](#) having a resolution of 1024×2048 , involves more than 1M pixel-wise labels with implicit prone to annotation errors, and one needs to account for the problem of ambiguous pixels. SSL is a perfect fit for the task at hand as the prior relies on a tiny set of labels. Deep SSL for semantic segmentation has been only explored recently in a few works.

Early techniques rely on using GANs [\[12\]](#) principles. The authors of [\[41\]](#) proposed to enlarge the training set by generating GAN-type synthetic images to enrich the feature space and strengthen the relationship between unlabelled and labelled samples. Hung et al. [\[17\]](#) proposed a GAN based technique to differentiate the predicted probability maps from the ground truth segmentations. Similarly, Mittal et al. [\[31\]](#) proposed a two-branch solution composed of: i) a GAN branch that generated per-pixel class labels for the input sample and ii) a multi-label Mean Teacher [\[43\]](#)

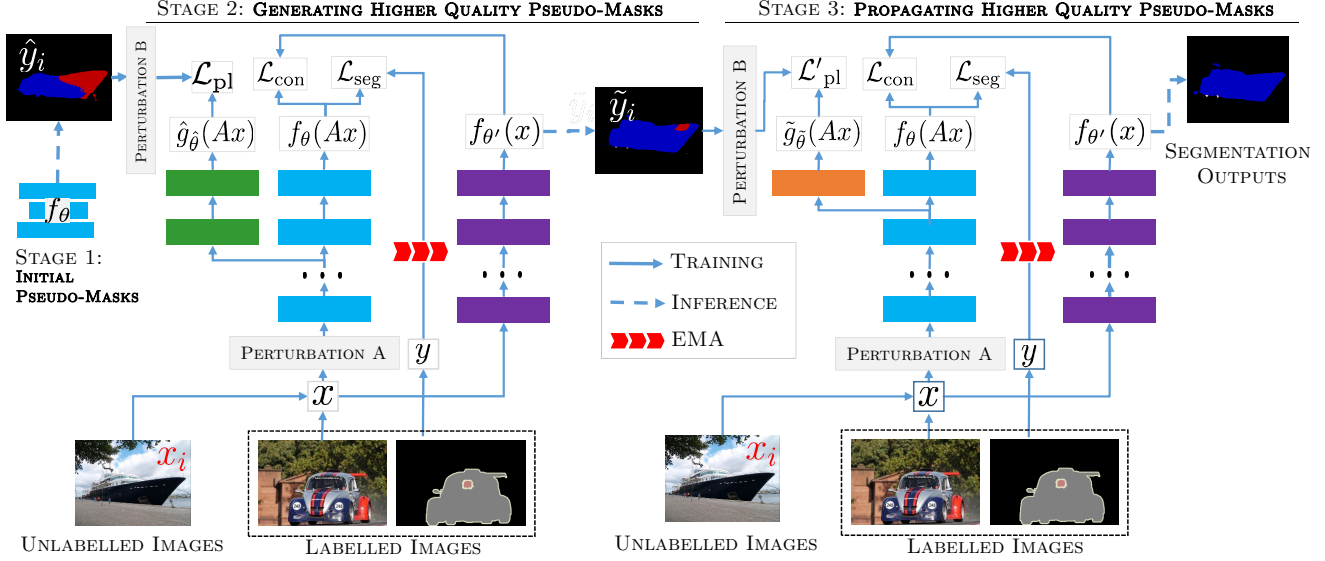


Figure 1. The proposed self-training framework consists of three stages. Stage 1 uses a segmentation network f_θ trained on the tiny labelled set to produce initial pseudo-masks. For Stage 2, the pseudo-masks are used to learn an auxiliary task (with loss \mathcal{L}_{pl}), and consistency regularisation \mathcal{L}_{con} is imposed in the segmentation task which then generates higher quality pseudo mask. The two tasks are modelled using a multi-task network (having two branches \hat{g}_θ and f_θ for the two tasks. The weights $\hat{\theta}$ and θ are shared between the branches except in the last two blocks). The last stage aims to propagate the information of the higher quality pseudo-masks during the optimisation process using another network \tilde{g}_θ (sharing the weights with f_θ except in the last block).

branch to remove false positive predictions.

More recently, the authors of that [11] extended CutMix [48] to the context of semantic segmentation. In that paper, authors applied the principle of strong augmentations, including Cutout, derived from findings in image classification. The authors of [36] enforced consistency between feature-based, prediction-based and random perturbations. Ke et al. [20] used a flaw probability map along with the extension of the dual student [21] to pixel-wise tasks. An offline self-training scheme based pseudo-labels with data augmentation has been proposed in [10] to enforce consistency between the labelled and unlabelled sets.

3. Proposed Technique

This section contains three core parts of our proposed framework: (i) how we generate the initial pseudo-masks, (ii) a multi-task model to improve the quality of the initial pseudo-masks and (iii) how to propagate the high quality pseudo-mask in the final model. The overall workflow is displayed in Figure 1.

Our Deep SSL Setting. Our holistic technique is framed in a three-stage self-training technique for deep semi-supervised learning. In our setting, we work under the assumption that we have a vast amount of unlabelled data and only a tiny set of annotations. Formally, let the set of labelled samples be denoted by $\{(x_i, y_i) \mid i \in L\}$ where x_i is the image and y_i refers to the ground truth segmentation mask corresponding to x_i . The set of unlabelled samples are

expressed as $\{x_i \mid i \in U\}$. We consider the learning problem where only a tiny fraction of the images are labelled, i.e. $|L| \ll |U|$. In what follows, we discuss our technique in details.

3.1. Consistency Regularisation with Strong Augmentations

A key issue in semi-supervised learning is how to compute labels with high certainty for the unlabelled data x_i ($i \in U$). In the context of deep learning, prior knowledge has been investigated to better exploit the rich unlabelled data.

To formulate the consistency regularisation in the setting of SSL segmentation, we define an operator as $A : \mathbb{R}^{n \times m \times c} \rightarrow \mathbb{R}^{n \times m \times c}$ that takes the image x to its randomly perturbed versions Ax . Common examples of such perturbations include rotation, flipping, translation and their combinations. We also define an operator as $B : \mathbb{R}^{n \times n \times c'} \rightarrow \mathbb{R}^{n \times n \times c'}$ which maps the pixels of a segmentation mask to a new mask in the same way as A does. The consistency loss is then defined as:

$$\mathcal{L}_{con} := \sum_{i \in L \cup U} d_c(f_\theta(Ax_i), Bf_{\theta'}(x_i)), \quad (1)$$

where f_θ and $f_{\theta'}$ are deep neural network models parametrised by θ and θ' respectively, and d_c measures the distance between $f_\theta(Ax_i)$ and $Bf_{\theta'}(x_i)$. In this work, A

and B are random operators – that is, they are different for different samples.

One key factor of the consistency regularisation (1) is the perturbation operators A and B . The minimisation of the cost (1) smooths the network’s prediction over a neighbour of the data sample. Therefore, the choice of the operator A reflects the size of the neighbour, which affects the segmentation results. In particular, we follow the principle of using strong augmentations as one enforces the perturbations to be diverse and natural yielding to boost the SSL performance [45]. We then implement the random operators A and B as strong augmenters based on the combination of RandAugment [7] and Cutout [8].

The consistency loss (1) is built on the relationship between two network models. In the literature, the networks f_θ and $f_{\theta'}$ are often termed as the student model and the teacher model, respectively. The parameter θ' of the teacher model can be determined in several ways. In the Γ model [24], θ' is set to θ . In the mean teacher (MT) scheme [43], θ' is computed as an exponential moving average (EMA) of θ during the training process. In the MT, as a desired propriety of the convergence of the student model, θ' gets closer to the weight θ of the student as the learning process keeps going [21]. Consequently, the teacher model $f_{\theta'}$ is lack of effective guidance as the number of training steps gets large.

Motivated by the fact that in consistency regularisation based approaches (e.g. [24, 43]), the consistency of the models’ outputs depends on the coupling of the teacher and the student during the training, we propose a framework that integrates additional guidance signals via multi-task learning. In the following, we present a three-stage self-training approach that incorporates pseudo masks statistics to help in learning better consistent outputs on the unlabelled data, which in return improve the quality of pseudo masks within the different stages.

3.2. A Three-Stage Self-Training Framework

The proposed framework consists of three stages (See Figure 1) as described next.

✧ **Stage 1: Generating Initial Pseudo-Masks.** In this stage, we use the tiny labelled images set to train a segmentation network f_θ and generate initial pseudo-masks that have high uncertainty (i.e. low quality pseudo masks). To do this, we use the loss function defined as:

$$\mathcal{L}_{\text{seg}} := \sum_{i \in L} d_s(f_\theta(x_i), y_i) \quad (2)$$

where d_s is the cross entropy loss. Once the network is trained, pseudo segmentation masks are generated from the output of the model $\hat{y}_i := \arg \max f_\theta(x_i)$ where the $\arg \max$ function is performed pixelwise. The main purpose of this stage is to generate initial pseudo masks, which

the maximum predicted probability is highly uncertain due to the limited amount of ground truth labels.

✧ **Stage 2: Increasing Certainty for Pseudo-Masks.** In this stage, the labelled samples, the unlabelled images, and the pseudo masks \hat{y}_i , generated from the first stage, are used together to produce higher quality pseudo-masks via training a second model. As the masks \hat{y}_i are not accurate, we do not use them to fit the segmentation network, but instead learn to reproduce them in a multi-task model. To do this, we consider two tasks, namely the segmentation task (Task 1) and the auxiliary task (Task 2). Our Task 2 aims to extract the pseudo-mask statistical information from \hat{y}_i .

Our two tasks are framed in a multi-task network as shown in Figure 1. The first network f_θ is the segmentation network composed of N blocks (i.e., the blue blocks). Whilst the second network, \hat{g}_θ for Task 2, is composed of N blocks along with the first $N - 2$ blocks shared with f_θ (see the left side of Figure 1), and $\hat{\theta}$ denotes the shared weights and the own weight of \hat{g} . For Task 1, the segmentation loss (2) is applied. Additionally, to promote the segmentation accuracy, we include the consistency loss (1) to this task. The weights θ' of the teacher network are updated using the exponential moving average (EMA) of θ during training [43]. For Task 2 and using pseudo-masks \hat{y}_i , the loss is defined as:

$$\mathcal{L}_{\text{pl}} := \sum_{i \in L \cup U} d_c(\hat{g}_\theta(Ax_i), B\hat{y}_i). \quad (3)$$

We underline that the output of \hat{g}_θ is not identical to that of f_θ , as the pseudo masks are not seen by Task 1. However, the pseudo masks in Task 2 produce additional signals for Task 1 during the optimisation process. Overall, the loss we use for this stage reads:

$$\mathcal{L} := \mathcal{L}_{\text{seg}} + \lambda_1 \mathcal{L}_{\text{con}} + \lambda_2 \mathcal{L}_{\text{pl}}, \quad (4)$$

where λ_1 and λ_2 are hyper-parameters balancing the tasks. Once the multi-task network is trained, the high quality pseudo-masks are computed by $\tilde{y}_i := \arg \max f_\theta(x_i)$.

✧ **Stage 3: Propagating Higher Quality Masks Information.** Finally, in this stage the high quality pseudo-masks \tilde{y}_i from Stage 2 are used to supervise the final models. The high quality pseudo-masks are integrated in a similar way as in Stage 2, except that we replace the network \hat{g}_θ by \tilde{g}_θ which has N blocks and shares its first $N - 1$ blocks (see the right side of Figure 1) with f_θ (the network \hat{g}_θ , however, has only $N - 2$ common blocks with f_θ). This makes the pseudo-mask information to better propagate to f_θ , and it is because \tilde{y}_i are more accurate than \hat{y}_i being used in Stage 2. The loss used for this stage is given by:

$$\mathcal{L}' := \mathcal{L}_{\text{seg}} + \lambda_1 \mathcal{L}_{\text{con}} + \lambda_2 \mathcal{L}'_{\text{pl}}, \quad (5)$$

where $\mathcal{L}'_{\text{pl}} := \sum_{i \in L \cup U} d_c(\tilde{g}_\theta(Ax_i), B\tilde{y}_i)$.

Algorithm 1 Our Three-stage Self-training Scheme

- 1: **Input:** Labelled samples $\{(x_i, y_i) \mid i \in L\}$ and unlabelled images $\{x_i \mid i \in U\}$. The parameters λ_1 and λ_2 . The random augmentation operators A and B .
 - 2: **# Stage 1:**
 - 3: Initialise θ of the segmentation network
 - 4: Minimise the loss \mathcal{L}_{seg} in (2) for θ
 - 5: Compute the initial pseudo masks $\hat{y}_i = \arg \max f_\theta(x_i)$, $i \in L \cup U$
 - 6: **# Stage 2:**
 - 7: Initialise $\tilde{\theta}$ and reinitialise θ
 - 8: With the pseudo-masks \hat{y}_i , minimise the loss \mathcal{L} in (4) for parameters θ and $\tilde{\theta}$ of the multi-task network. The weight θ' of the teacher network is computed as the EMA of θ .
 - 9: With the new θ , compute $\tilde{y}_i := \arg \max f_\theta(x_i)$ $i \in L \cup U$
 - 10: **# Stage 3:**
 - 11: Initialise $\tilde{\theta}$ and reinitialise θ
 - 12: Minimise the loss \mathcal{L}' in (5) for θ and $\tilde{\theta}$, where \tilde{y}_i are computed from Stage 2, and θ' is the EMA of θ
 - 13: **Output:** Return the parameter θ of the segmentation network
-

The overall process of our technique – in which previous individual stages are combined to solve the semantic segmentation problem – is listed in Algorithm 1.

4. Experimental Results

In this section, we detail the range of experiments that were conducted to evaluate our proposed technique ¹.

4.1. Data Description and Evaluation Protocol

In the experiments, we use two major benchmark datasets for semantic segmentation. Our first dataset is **Cityscapes** [6]: this is an urban scene dataset composed of 2975, 500 and 1525 images for training, validation and test respectively, and has 19 classes. The images are of size 1024×2048 . In our experiments we downsampled all images to the size 512×1024 , following the experimental setting used in [17, 31]. Our second dataset is the augmented **PASCAL VOC 2012** [9]. It is a natural scenes dataset that captures objects from 20 classes + 1 background class. The dataset is composed of 10582 and 1449 images for training and validation respectively (parts of the annotations come from [14]).

To evaluate the performance of the proposed method on the different amount of labels, we vary the fraction of labelled images for both datasets. Specifically, for the Cityscapes dataset, we carry out experiments on 100, 372, 744 and 1448 labelled images taken from the whole training set, respectively. We note that all 2975 training images have ground truth segmentation masks, but some of the segmentation masks are ignored during the training and the corresponding images are treated as unlabelled. Similarly, for the augmented PASCAL dataset, we report results for the

models trained on 1/100, 1/50, 1/20 and 1/8 of the training images (together with the other training images treated as unlabelled), respectively. In both datasets, we follow the same split for the labelled images and unlabelled images as in the work of [31, 17].

We address our evaluation protocol from both quantitative and qualitative point of views. The former is based on the widely-used metric called the mean Intersection-over-Union (mIoU) that is used for all our comparisons. The numerical comparison of our technique was performed against the state-of-the-art methods for deep semantic segmentation: Hung et al. [17], Mittal et al. [31], French et al. [11], VAT [32], ICT [44], Feng et al. [10], Olsson et al. [35] and Ke et al. [20]. The latter is based on a visual inspection of the segmentation outputs.

4.2. Implementation Details

We provide the training details of our techniques for the datasets as well as outlining the reproduced baselines.

Network architecture and pretraining scheme. In our experiments, we use as our segmentation network, f_θ , Deeplabv2 [5] with ResNet-101 backbone [15], which has been also considered in the works of that [17, 31, 11]. For the experiments on Cityscapes [6], the ResNet-101 is pre-trained on ImageNet, whilst for the experiments on PASCAL VOC [9], we use two schemes, the first one of which uses ImageNet pretraining whilst the second one considers additionally MSCOCO [28] pretraining for f_θ .

The networks \hat{g}_θ and \tilde{g}_θ are built on the top of f_θ . Specifically, \hat{g}_θ has the first $N-2$ common blocks with f_θ , by sharing its weights of conv1, conv2x, conv3x, conv4x with f_θ , while having its free parameter on conv5x and ASPP layer [5]. We refer the readers to Table 1 in [15] for the introduction of conv1, ..., conv5x. The network \tilde{g}_θ shares the blocks conv1, ..., conv5x with f_θ while having its own parameter for the ASPP layer. A more detailed description of the structure of f_θ , \hat{g}_θ and \tilde{g}_θ is provided in the supplementary material.

We underline that though \hat{g}_θ (for Stage 2) and \tilde{g}_θ (for Stage 3) are introduced during the optimisation process, the segmentation network architecture f_θ remains unchanged and is therefore comparable to the works of [17, 31, 11]. In fact, in the test phase, \hat{g}_θ and \tilde{g}_θ are not included, and therefore the network being evaluated is identical to the DeeplabV2 in e.g., [11].

Training Scheme. The training details regarding both datasets are described next. For both Cityscape and PASCAL VOC 2012, the overall training procedure of the proposed method follows Algorithm 1, and the losses for the three stages are listed therein. In particular, in the consistency loss \mathcal{L}_{con} (1) and \mathcal{L}_{pl} (3), we use the cross entropy for the measure d_c . The random operators A and B in the definition of these two losses are implemented as the com-

¹The code will be made available at https://github.com/RK621/ThreeStageSelftraining_SemanticSegmentation

CITYSCAPES	# LABELS				
TECHNIQUE	1/30 (100)	1/8 (372)	1/4 (744)	1/2 (1488)	Full(2975)
DeeplabV2 [5]	–	56.2	60.2	–	66.0
Hung et al. [17]	–	57.1	60.5	–	66.2
Mittal et al. [31]	–	59.3	61.9	–	65.8
French (Cutout) [11]	47.21	57.72	61.96	–	67.47
French et al. (CutMix) [11]	51.20	60.34	63.87	–	67.88
Hung et al. [+MSCOCO] [17]	–	58.8	62.3	65.7	–
Olsson et al. [+MSCOCO] [35]	54.07	61.35	63.63	66.29	–
Ours	54.85	62.82	65.80	67.11	–

Table 1. Numerical comparison of our technique vs the state-of-the-art models for semantic segmentation. All results reported are based on the Deeplabv2 network (ResNet-101 backbone). The numerical values are computed as the mIoU metric (in percentage) over different label counts. ‘+MSCOCO’ denotes that the pre-training considers additionally MSCOCO dataset and only ImageNet otherwise.

# WITHOUT COCO PRE-TRAINING					
PASCAL VOC	# LABELS				
TECHNIQUE	1/100	1/50	1/20	1/8	Full(10582)
DeeplabV2 [5]	–	48.3	56.8	62.0	70.7
VAT [32]	38.81	48.55	58.50	62.93	72.18
ICT [44]	35.82	46.28	53.17	59.63	71.50
Hung et al. [17]	–	49.2	59.1	64.3	71.4
Mittal et al. (s4GAN) [31]	–	58.1	60.9	65.4	71.2
Mittal et al. [31]	–	60.4	62.9	67.3	73.2
French et al. (Cutout) [11]	48.73	58.26	64.37	66.76	72.03
French et al. (CutMix) [11]	53.79	64.81	66.48	67.60	72.54
Ours	55.13	62.71	68.23	69.36	–

# WITH MSCOCO PRE-TRAINING					
PASCAL VOC	# LABELS				
TECHNIQUE	1/100	1/50	1/20	1/8	Full(10582)
DeeplabV2 [5]	–	53.2	58.7	65.2	73.6
Hung et al. [17]	–	57.2	64.7	69.5	74.9
Zhai et al. [49]	–	–	–	68.65	75.38
Mittal et al. (s4GAN) [31]	–	60.9	66.4	69.8	73.9
Mittal et al. [31]	–	63.3	67.2	71.4	75.6
French et al. (CutMix)† [11]	60.19	67.30	70.33	71.82	–
Feng et al. [10]	61.6	65.5	69.3	70.7	73.5
Olsson et al. [35]	54.18	66.15	67.77	71.0	–
Ke et al. [20]	–	–	–	72.14	75.73
Ours	63.82	70.77	71.90	72.95	–

Table 2. Numerical comparison between our technique and existing techniques for semi-supervised semantic segmentation. The segmentation network architectures are based on Deeplabv2 with ResNet-101 backbone. The numerical values reflect the mIoU metric (in percentage) over different label counts. The left table displays the results from using only ImageNet pre-training whilst the right one uses MSCOCO pretraining. The best results are shown in bold font. † denotes results reproduced by us.

bination of RandAugment [7] and Cutout [8]. At the beginning of each stage, the network parameters, θ , are initialised using the associated pretrained model (either ImageNet or MSCOCO). In each stage, the network is trained using the Adam optimiser [22] with a learning rate of 3×10^{-5} , and 60,000 optimisation steps. In each step of the stochastic optimisation, the losses are computed on a minibatch. More specifically, the loss \mathcal{L}_{con} and \mathcal{L}_{pl} are averaged over all images in a minibatch, and the segmentation loss \mathcal{L}_{seg} is averaged over the subset of images for which the ground truth masks are available.

For Cityscapes, all images are downsampled into 512×1024 throughout the experiments. During the training phase, all images are cropped into 256×512 before feeding to the network. Random horizontal flipping is applied to augment the dataset. The same setting was used in the works of [17, 31]. To apply our three-stage self-training method, we use a minibatch size of 5. We consider different labelled images counts using $|L| = 100$ (1/30 of the whole set of images), 372, 744 and 1448 respectively. We include 1 labelled image in each of the minibatch for the cases $|L| = 100, 372$ and 744, while increasing this number to 3 when $|L|$ increases to 1448. For all cases, unless specified otherwise, the parameter λ_1 and λ_2 are set to 1/2.

For PASCAL VOCC 2012, the images are cropped into sizes of 321×321 during the training. Augmentations, including random horizontal flipping and random scaling, are applied, which follows the setting of semi-supervised segmentation benchmarks [17, 31]. We trained the network using 1/100, 1/50, 1/20 and 1/8 labelled images respectively. The size of the minibatch is set to 10. In each minibatch, only 1 image has the ground truth masks for the setting with 1/100, 1/50, 1/20 labels, and this number is increased to 2 when the fraction of labelled images increases to 1/8. For this experiment, the parameter λ_1 is set to 1/4, 1/4, 5/4, 5/8 for the 4 different amounts of labels respectively, and $\lambda_2 = \lambda_1$.

4.3. Results and Discussions

In this section, we discuss the findings drawn from our numerical and visual results.

Comparison of our Technique vs SOTA Models. We start by comparing our technique against the existing semi-supervised semantic segmentation models for both datasets. *Performance Comparison using Cityscapes.* The results are reported in Table 1. We report two case studies in this table. The first case is regarding using ImageNet as pre-training

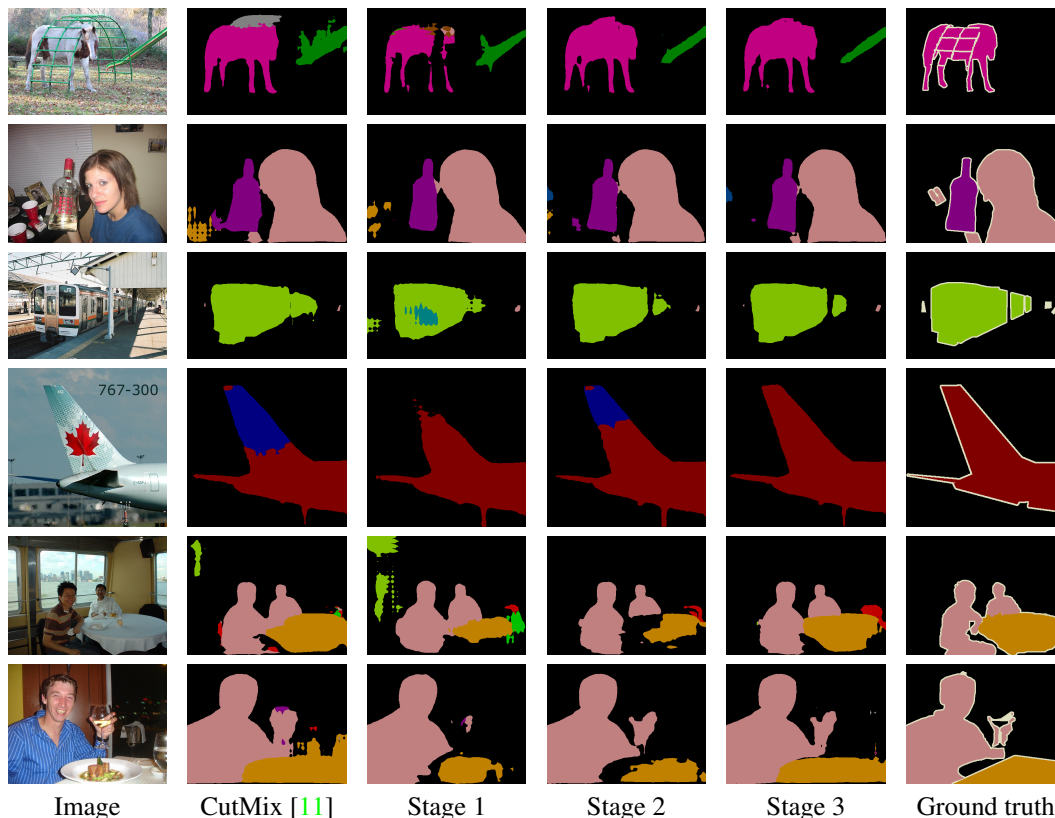


Figure 2. Visual comparison of the different stages of our technique vs a SOTA technique on a selection of images from PASCAL VOC validation sets. The models are trained using 212 ground truth segmentation masks. The first column displays the sample input whilst the others the segmentation outputs for both techniques.

(see the upper part of Table 1), which is used by our approach. From this comparison, we can observe that our approach significantly improves over all compared techniques and for all labels counts. Interestingly, we observe that only with 1/4 and 1/2 of labels, we can reach the performance that only is reachable for the compared techniques using the full labelled set. The second one is against techniques that additionally use MSCOCO (see techniques with the tag '[+MSCOCO]') in the pre-training stage, we observe that our technique without such additional data is able to outperform these methods.

Performance Comparison using PASCAL. We further investigate the performance of our technique by a set of comparisons using PASCAL VOC dataset. The results are displayed in Table 4.2. Similarly than in the previous set of comparisons, we have two cases with and without (i.e. only ImageNet) MSCOCO pre-training. We begin the comparison using the case without MSCOCO and the results are reported at the left part of Table 4.2. In a closer look at the results, we observe that our technique is consistent with the previous ones on Cityscapes as our technique substantially improves over the compared algorithmic techniques for all label counts. The only exception is for [11] in the case of

1/50 labelled data. However, one can see that our technique substantially improves (from 62.71% to 70.77% in mIoU) when using MSCOCO for this particular case. An interesting case is when using VAT and ICT techniques, which drawn directly from image classification. From the numerical results, we observe that these two techniques report low performance which highlights the substantial gap in modelling hypothesis between classification and semantic segmentation, where the second one requires more dense and complex annotations.

The results on PASCAL for the case with MSCOCO pre-training are displayed at the right part of Table 4.2. By inspection, we observe that the performance of our technique is prevalently outperforming all compared techniques for all labels counts. Particularly, our method reports large improvement for this case up to $\sim 33\%$, most notably, when a lower label regime is used i.e. 1/100, 1/50, 1/20. Unlike Cityscapes where the performance gain using MSCOCO pretraining is small, one can observe that for the case of PASCAL VOC the improvement is significant. This is mainly because the samples content from MSCOCO is similar than PASCAL VOC. This allows to further enrich the data generalising better the performance for this case.

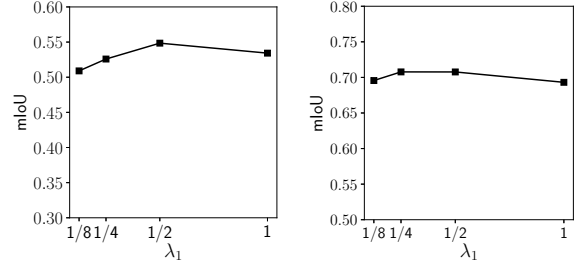
S1	S2	S3	CR	T2	SA	N2	CS (100)	P VOC (212)
✓			NA	NA	NA	NA	46.91	55.20
✓	✓		✓	✓	✓	✓	53.25	68.50
✓	✓	✓		✓	✓	✓	50.2	59.82
✓	✓	✓	✓		✓	✓	52.63	68.21
✓	✓	✓	✓	✓		✓	47.96	64.37
✓	✓	✓	✓	✓		✓	53.80	69.52
✓	✓	✓	✓	✓	✓	✓	54.85	70.77

Table 3. Ablation study on Cityscapes (CS) and PASCAL VOC (P VOC) datasets. We refer to S1, S2 and S3 as each corresponding stage of our technique. Moreover, we indicate consistency regularisation as CR, Task 2 as T2 and strong augmentation as SA. Finally, the option N2 means the use of $\tilde{g}_{\hat{\theta}}$ in Stage 3 (otherwise $\hat{g}_{\hat{\theta}}$ is used in Stage 3 instead).

Figure 2 compares the segmentation outputs of the three stages, along with that of CutMix [11] and the ground truth annotations, for images taken from the PASCAL validation set. As shown in the figure, more details of the foreground objects are captured from Stage 1 to Stage 3, and the final segmentation results are more accurate at the object boundaries than the segmentation by CutMix.

Ablation study on the tasks and the augmentations. Results of ablation study for Cityscapes (with 100 labelled images and ImageNet pretraining) and PASCAL VOC (with 212 labelled images and MSCOCO pretraining) are reported in Table 3. Starting from Stage 1 (S1), the performance is substantially improved by including the second stage (S2). Stage 3 (S3) further improves the mIoU by 1.6% for Cityscapes and 2.2% for PASCAL. A significant drop of the score ($\geq 4\%$) appears when the consistency regularisation (CR) \mathcal{L}_{con} is not included, which suggests that CR helps to improve the segmentation accuracy in addition to the pseudo-masks. A loss of around 2% in the mIoU is observed when Task 2 (T2) is dropped, indicating the importance of Task 2 (using the pseudo-masks). Strong augmentation (SA) is an important component of the framework, as excluding SA decreases the score by around 6.5% for both datasets. Finally, changing the network $\tilde{g}_{\hat{\theta}}$ for Stage 3 into $\hat{g}_{\hat{\theta}}$ (used by Stage 2) results in about 1%’s performance loss, which indicates that $\tilde{g}_{\hat{\theta}}$ (sharing more blocks with f_{θ}) allows better propagation of the pseudo mask information to f_{θ} .

The weights of the losses. We study how the choice of the weights λ_1 and λ_2 affects the performance of the proposed method. To do this, we report the results for both datasets over different values of the weights in Figure 3. In this test, for Cityscapes, 100 ground truth segmentation masks are used for training, and the segmentation network DeeplabV2 is pretrained on ImageNet. For the augmented PASCAL dataset, we use 2% labelled images and MSCOCO pretraining. We set equal weights for the consistency loss and the pseudo-masks loss term, i.e., $\lambda_2 = \lambda_1$.



(a). Cityscapes (b). PASCAL
Figure 3. Behaviour curves of the mIoU vs the values of λ_1 .

From the results in Figure 3 (a), the best performance for the Cityscapes dataset is reached at $\lambda_1 = 1/2$, and the mIoU decreases by around 2% and 1.5% when λ_1 is changed to 1/4 and 1, respectively. For the augmented PASCAL dataset, the highest mIoU (0.708) is archived at $\lambda_1 = 1/4$ and $\lambda_1 = 1/2$ (See Figure 3(b)). We observe a small drop in the mIoU (around 1%) when λ_1 is set to 1/8 or 1.

The multi-task network and extensions. The multi-task networks for Stage 2 and Stage 3 play an important role in balancing the uncertainty in the pseudo masks and the consistency of the prediction. This is crucial especially when the labelled data is small and the initial pseudo masks reflect low accuracy. As described in Section 3, the network $\hat{g}_{\hat{\theta}}$ is constructed by reusing the first $N - 2$ blocks of the segmentation network f_{θ} , while having its free parameters for the last 2 blocks. The results displayed in this work are based on Deeplabv2, but we remark that the multitask network can be constructed in a similar way if other segmentation networks are adopted. Furthermore, it is possible to extend the proposed three stages into 4 or more stages depending on the applications.

5. Conclusion

In this work, we propose a three-stage self-training method for semi-supervised semantic segmentation. In last two stages, we maintain a segmentation task with consistency constraint via a moving teacher network, and introduce an auxiliary task to propagate the information of pseudo masks (which are stationary during the training) to the student network during the optimisation process. We demonstrate that our framework effectively improves the pseudo masks quality using a multi-task model, given only a very tiny amount of ground truth segmentation masks. The proposed method reaches the state-of-the-art semi-supervised segmentation results on two major benchmark datasets. Extension of the proposed method can therefore be implemented by combining it with additional pseudo masks enhancement techniques including uncertainty weighting and warm-up, which further improve the quality of pseudo masks.

References

- [1] Jiwoon Ahn and Suha Kwak. Learning pixel-level semantic affinity with image-level supervision for weakly supervised semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4981–4990, 2018. 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 A Raffel. Mixmatch: A holistic approach to semi-supervised learning. In *Advances in Neural Information Processing Systems*, pages 5049–5059, 2019. 2
- [4] Olivier Chapelle, Bernhard Scholkopf, and Alexander Zien. Semi-supervised learning (chapelle, o. et al., eds.; 2006)[book reviews]. *MIT Press*, 2006. 1, 2
- [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. 1, 2, 5, 6
- [6] 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
- [7] Ekin D Cubuk, Barret Zoph, Jonathon Shlens, and Quoc V Le. Randaugment: Practical automated data augmentation with a reduced search space. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 702–703, 2020. 4, 6
- [8] Terrance DeVries and Graham W Taylor. Improved regularization of convolutional neural networks with cutout. *arXiv preprint arXiv:1708.04552*, 2017. 4, 6
- [9] Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *International journal of computer vision*, 88(2):303–338, 2010. 5
- [10] Zhengyang Feng, Qianyu Zhou, Guangliang Cheng, Xin Tan, Jianping Shi, and Lizhuang Ma. Semi-supervised semantic segmentation via dynamic self-training and class-balanced curriculum. *arXiv preprint arXiv:2004.08514*, 2020. 1, 3, 5, 6
- [11] Geoff French, Samuli Laine, Timo Aila, and Michal Mackiewicz. Semi-supervised semantic segmentation needs strong, varied perturbations. *British Machine Vision Conference (BMVC)*, 2020. 1, 2, 3, 5, 6, 7, 8
- [12] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680, 2014. 2
- [13] Yves Grandvalet and Yoshua Bengio. Semi-supervised learning by entropy minimization. In *Advances in neural information processing systems*, pages 529–536, 2005. 1
- [14] Bharath Hariharan, Pablo Arbelaez, Lubomir Bourdev, Subhransu Maji, and Jitendra Malik. Semantic contours from inverse detectors. In *International Conference on Computer Vision (ICCV)*, 2011. 5
- [15] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 5
- [16] Zilong Huang, Xinggang Wang, Lichao Huang, Chang Huang, Yunchao Wei, and Wenyu Liu. Ccnet: Criss-cross attention for semantic segmentation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 603–612, 2019. 2
- [17] Wei Chih Hung, Yi Hsuan Tsai, Yan Ting Liou, Yen Yu Lin, and Ming Hsuan Yang. Adversarial learning for semi-supervised semantic segmentation. In *British Machine Vision Conference (BMVC)*, 2018. 1, 2, 5, 6
- [18] Rihuan Ke, Aurélie Bugeau, Nicolas Papadakis, Mark Kirkland, Peter Schuetz, and Carola-Bibiane Schönlieb. Multi-task deep learning for image segmentation using recursive approximation tasks. *arXiv preprint arXiv:2005.13053*, 2020. 2
- [19] Rihuan Ke, Aurélie Bugeau, Nicolas Papadakis, Peter Schuetz, and Carola-Bibiane Schönlieb. Learning to segment microscopy images with lazy labels. *arXiv preprint arXiv:1906.12177*, 2019. 2
- [20] Zhanghan Ke, Di Qiu, Kaican Li, Qiong Yan, and Rynson W.H. Lau. Guided collaborative training for pixel-wise semi-supervised learning. In *European Conference on Computer Vision (ECCV)*, August 2020. 3, 5, 6
- [21] Zhanghan Ke, Daoye Wang, Qiong Yan, Jimmy Ren, and Rynson WH Lau. Dual student: Breaking the limits of the teacher in semi-supervised learning. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 6728–6736, 2019. 3, 4
- [22] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 6
- [23] Durk P Kingma, Shakir Mohamed, Danilo Jimenez Rezende, and Max Welling. Semi-supervised learning with deep generative models. *Advances in neural information processing systems*, 27:3581–3589, 2014. 1
- [24] Samuli Laine and Timo Aila. Temporal ensembling for semi-supervised learning. *International Conference of Learning Representation*, 2017. 1, 2, 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 conference on computer vision and pattern recognition*, pages 5267–5276, 2019. 1, 2
- [26] Hanchao Li, Pengfei Xiong, Haoqiang Fan, and Jian Sun. Dfanet: Deep feature aggregation for real-time semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 9522–9531, 2019. 2

- [27] Peiwen Lin, Peng Sun, Guangliang Cheng, Sirui Xie, Xi Li, and Jianping Shi. Graph-guided architecture search for real-time semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4203–4212, 2020. 2
- [28] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European Conference on Computer Vision*, pages 740–755. Springer, 2014. 5
- [29] Chenxi Liu, Liang-Chieh Chen, Florian Schroff, Hartwig Adam, Wei Hua, Alan L Yuille, and Li Fei-Fei. Auto-deeplab: Hierarchical neural architecture search for semantic image segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 82–92, 2019. 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. 1, 2
- [31] Sudhanshu Mittal, Maxim Tatarchenko, and Thomas Brox. Semi-supervised semantic segmentation with high-and low-level consistency. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2019. 2, 5, 6
- [32] 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. 1, 2, 5, 6
- [33] Vladimir Nekrasov, Hao Chen, Chunhua Shen, and Ian Reid. Fast neural architecture search of compact semantic segmentation models via auxiliary cells. In *Proceedings of the IEEE Conference on computer vision and pattern recognition*, pages 9126–9135, 2019. 2
- [34] Hyeonwoo Noh, Seunghoon Hong, and Bohyung Han. Learning deconvolution network for semantic segmentation. In *Proceedings of the IEEE international conference on computer vision*, pages 1520–1528, 2015. 1, 2
- [35] Viktor Olsson, Wilhelm Trane, Julian Pinto, and Lennart Svensson. Classmix: Segmentation-based data augmentation for semi-supervised learning. *arXiv preprint arXiv:2007.07936*, 2020. 5, 6
- [36] 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, 3
- [37] Rudra PK Poudel, Stephan Liwicki, and Roberto Cipolla. Fast-scnn: Fast semantic segmentation network. *arXiv preprint arXiv:1902.04502*, 2019. 2
- [38] Eduardo Romera, José M Alvarez, Luis M Bergasa, and Roberto Arroyo. Erfnet: Efficient residual factorized convnet for real-time semantic segmentation. *IEEE Transactions on Intelligent Transportation Systems*, 19(1):263–272, 2017. 2
- [39] Mehdi Sajjadi, Mehran Javanmardi, and Tolga Tasdizen. Regularization with stochastic transformations and perturbations for deep semi-supervised learning. In *Advances in neural information processing systems*, pages 1163–1171, 2016. 2
- [40] Wataru Shimoda and Keiji Yanai. Self-supervised difference detection for weakly-supervised semantic segmentation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 5208–5217, 2019. 1, 2
- [41] Nasim Souly, Concetto Spampinato, and Mubarak Shah. Semi supervised semantic segmentation using generative adversarial network. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 5688–5696, 2017. 2
- [42] Guolei Sun, Wenguan Wang, Jifeng Dai, and Luc Van Gool. Mining cross-image semantics for weakly supervised semantic segmentation. *European Conference of Computer Vision*, 2020. 2
- [43] Antti Tarvainen and Harri Valpola. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. In *Advances in neural information processing systems*, pages 1195–1204, 2017. 2, 4
- [44] Vikas Verma, Alex Lamb, Juho Kannala, Yoshua Bengio, and David Lopez-Paz. Interpolation consistency training for semi-supervised learning. In *Proceedings of the 28th International Joint Conference on Artificial Intelligence*, pages 3635–3641, 2019. 1, 2, 5, 6
- [45] Qizhe Xie, Zihang Dai, Eduard Hovy, Minh-Thang Luong, and Quoc V Le. Unsupervised data augmentation for consistency training. *arXiv preprint arXiv:1904.12848*, 2019. 4
- [46] Qizhe Xie, Minh-Thang Luong, Eduard Hovy, and Quoc V Le. Self-training with noisy student improves imagenet classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10687–10698, 2020. 1
- [47] Changqian Yu, Jingbo Wang, Chao Peng, Changxin Gao, Gang Yu, and Nong Sang. Bisenet: Bilateral segmentation network for real-time semantic segmentation. In *Proceedings of the European conference on computer vision (ECCV)*, pages 325–341, 2018. 2
- [48] 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 International Conference on Computer Vision*, pages 6023–6032, 2019. 3
- [49] Xiaohua Zhai, Avital Oliver, Alexander Kolesnikov, and Lucas Beyer. S4l: Self-supervised semi-supervised learning. In *Proceedings of the IEEE international conference on computer vision*, pages 1476–1485, 2019. 6
- [50] Hang Zhang, Kristin Dana, Jianping Shi, Zhongyue Zhang, Xiaoqiang Wang, Amrith Tyagi, and Amit Agrawal. Context encoding for semantic segmentation. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 7151–7160, 2018. 2
- [51] Yang Zhang, Philip David, Hassan Foroosh, and Boqing Gong. A curriculum domain adaptation approach to the semantic segmentation of urban scenes. *IEEE transactions on pattern analysis and machine intelligence*, 2019. 2

- [52] Yiheng Zhang, Zhaofan Qiu, Jingen Liu, Ting Yao, Dong Liu, and Tao Mei. Customizable architecture search for semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 11641–11650, 2019. 2
- [53] Hengshuang Zhao, Yi Zhang, Shu Liu, Jianping Shi, Chen Change Loy, Dahua Lin, and Jiaya Jia. Pscanet: Point-wise spatial attention network for scene parsing. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 267–283, 2018. 1, 2
- [54] Mingmin Zhen, Jinglu Wang, Lei Zhou, Shiwei Li, Tianwei Shen, Jiaxiang Shang, Tian Fang, and Long Quan. Joint semantic segmentation and boundary detection using iterative pyramid contexts. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020. 1, 2
- [55] Zilong Zhong, Zhong Qiu Lin, Rene Bidart, Xiaodan Hu, Ibrahim Ben Daya, Zhifeng Li, Wei-Shi Zheng, Jonathan Li, and Alexander Wong. Squeeze-and-attention networks for semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13065–13074, 2020. 2
- [56] Dengyong Zhou, Olivier Bousquet, Thomas Lal, Jason Weston, and Bernhard Schölkopf. Learning with local and global consistency. *Advances in neural information processing systems*, 16:321–328, 2003. 1