Pseudo-Label Noise Suppression Techniques for Semi-Supervised Semantic Segmentation

Sebastian Scherer
sebastian.scherer1@uni-a.de

Robin Schön
robin.schoen@uni-a.de

Rainer Lienhart
rainer.lienhart@uni-a.de

University of Augsburg
Machine Learning and Computer Vision Lab
Augsburg, Germany

Abstract

Semi-supervised learning (SSL) can reduce the need for large labelled datasets by incorporating unlabelled data into the training. This is particularly interesting for semantic segmentation, where labelling data is very costly and time-consuming. Current SSL approaches use an initially supervised trained model to generate predictions for unlabelled images, called pseudo-labels, which are subsequently used for training a new model from scratch. Since the predictions usually do not come from an error-free neural network, they are naturally full of errors. However, training with partially incorrect labels often reduce the final model performance. Thus, it is crucial to manage errors/noise of pseudo-labels wisely. In this work, we use three mechanisms to control pseudo-label noise and errors: (1) We construct a solid base framework by mixing images with cow-patterns on unlabelled images to reduce the negative impact of wrong pseudo-labels. Nevertheless, wrong pseudo-labels still have a negative impact on the performance. Therefore, (2) we propose a simple and effective loss weighting scheme for pseudo-labels defined by the feedback of the model trained on these pseudo-labels. This allows us to soft-weight the pseudo-label training examples based on their determined confidence score during training. (3) We also study the common practice to ignore pseudo-labels with low confidence and empirically analyse the influence and effect of pseudo-labels with different confidence ranges on SSL and the contribution of pseudo-label filtering to the achievable performance gains. We show that our method performs superior to state of-the-art alternatives on various datasets. Furthermore, we show that our findings also transfer to other tasks such as human pose estimation. Our code is available at https://github.com/ChristmasFan/SSL_Denoising_Segmentation.

1 Introduction

Semantic segmentation has accomplished amazing performance on annotated data and has become an important task in computer vision. However, labelling data for semantic segmentation requires assigning a class label to each pixel in an image, which is an extremely tedious and time-consuming task. As an example, annotating a single image of the Cityscapes

dataset [7], which consists of images of urban scenes, takes up to 90 minutes [7]. However, simply obtaining images is by far not as time-consuming. The goal of semi-supervised learning (SSL) is to incorporate unlabelled data into standard supervised training, improving overall robustness and performance while reducing the amount of labelled data needed.

In SSL, most existing approaches treat model predictions as ground truth for unlabelled images during training, while human-generated annotations are used for labelled images. In this context, the predictions from a model that will be used as targets in the loss calculation are called *pseudo-labels*. Due to the fact that the predictions and thus the pseudo-labels are inevitably error-prone, the network is partly trained on wrong annotations. Such pseudo-labels are sometimes called noisy labels, and DNNs are susceptible to them. There is usually a large gap between networks trained on clean and noisy datasets [6, 22].

In this paper, three mechanisms to handle error-prone pseudo-labels from unlabelled data will be analysed to improve the performance for semantic segmentation. Among them, we study the common pseudo-label filtering technique as well as a novel loss function that weights pseudo-labels according to the feedback of the trained model to reduce the influence of wrong pseudo-labels during training. This novel weighted loss function boosts SSL performance substantially and is a core contribution of this paper. We analyse this approach on several datasets to build a general and better understanding. We then integrate it into a naive pseudo-label refinement method and combine it with a cow-pattern based image mixing for data augmentation. Furthermore, we demonstrate the effectiveness of our approach on several datasets. Specially, we provide results at a very low data regime and show that we can significantly improve the performance when training e.g. only with 15 labelled examples on the Cityscapes dataset [7]. We also demonstrate the effectiveness of our proposed mechanisms on a different task than semantic segmentation, namely human pose estimation.

2   Related Work

Semi-Supervised Learning (SSL) aims to train a model using labelled and unlabelled images. Current SSL approaches utilise the prediction of a model as supervision for the training on unlabelled data and can be weakly separated into consistency regularization [1, 2, 11, 21, 30] and self-training approaches [26, 28]. Both have the usage of pseudo-labels in common. Typically, the parameters of the model that generates pseudo-labels is fixed at a training iteration in self-training, while it changes at consistency regularization. Both approaches utilise strong data augmentation such as CutOut [9], CutMix [11], ClassMix [19] and colour space augmentations [21, 28, 34] when training with pseudo-labels. In this context, data augmentation is called perturbation because it makes it difficult for the model to reproduce the pseudo-label prediction. It is known that utilising perturbations has a strong effect on the final performance.

Some methods further apply confidence-based filtering of pseudo labels [20, 21, 27, 30, 31, 33]. FixMatch [21] uses a fixed confidence threshold for all classes to remove uncertain labels for image classification. FlexMatch [30] further extends this by filtering low-confidence samples with class-wise thresholds. Other approaches use different methods than model confidence to measure uncertainty. Zhou *et al.* [33] use multiple forward passes of different augmented version of the same image to estimate uncertainty and filter pseudo-labels. Yang *et al.* [27] uses multiple model training snapshots to determine a measure of uncertainty. However, none of them apply a comprehensive study on the effect of confidence-based filtering for semantic segmentation.

Our work is mostly related to those of [28]. In comparison to their work, we use simple image mixing with a diverse and dynamic mask instead of data augmentation strategies on pixel-level. We further apply pseudo-label filtering and dynamic weighting as denoising strategies to handle error-prone pseudo-labels. All approaches are examined in more detail on several datasets.

3 Methodology

In this section we first define import notation and the problem to solve in Sec. 3.1. Then we introduce the general SSL framework in Sec. 3.2. Finally, we describe pseudo-labels filtering in Sec. 3.3. Our loss function with a dynamical weighting scheme is described at Sec. 3.4.

3.1 Problem Definition

Let $\mathcal{D}$ be a dataset consisting of a set $\mathcal{D}_{\mathcal{L}}$ of labelled and a set $\mathcal{D}_{\mathcal{U}}$ of unlabelled images. Their respective sizes are N_l and $N_{\mathcal{U}}$. $\mathbf{x}_l \in \mathcal{D}_{\mathcal{L}}$ and $\mathbf{x}_u \in \mathcal{D}_{\mathcal{U}}$ denote any labelled and unlabelled image from the corresponding set of images. The labelled dataset has access to the ground truth segmentation masks $\mathbf{y}_l$, i.e., $\mathcal{D}_{\mathcal{L}} = \{(\mathbf{x}_l^i, \mathbf{y}_l^i)\}_{i=1}^{N_l}$. The unlabelled dataset $\mathcal{D}_{\mathcal{U}}$ has no segmentation masks, i.e., $\mathcal{D}_{\mathcal{U}} = \{(\mathbf{x}_u^i)\}_{i=1}^{N_u}$. To receive training feedback for unlabelled images, we use the predictions of a segmentation model as pseudo-labels $\hat{\mathbf{y}}$. $\mathbf{y}$, in contrast, denotes ground truth annotations from the dataset. The aim is to improve the performance of the segmentation model by exploiting unlabelled images with generated pseudo-labels.

3.2 SSL Base Approach

We follow the simple idea of iterative pseudo-label refinement [4, 26] and first train an initial segmentation model on the labelled dataset. We then use the model as teacher F_T to generate pseudo-labels for the unlabelled images by taking the argmax over the class dimension of the prediction $F_T(\mathbf{x}_u)$ for all input images $\mathbf{x}_u$, i.e., $\hat{\mathbf{y}} = argmax(F_T(\mathbf{x}_u))$. Then we train a student model F_S with identical architecture as the teacher model F_T on the labelled and unlabelled dataset $D = \{(\mathbf{x}_l^i, \mathbf{y}_l^i)\}_{i=1}^{N_l} \cup \{(\mathbf{x}_u^i, \hat{\mathbf{y}}_u^i)\}_{i=1}^{N_u}$ from scratch. After training, the pseudo-label dataset is updated by the predictions of the student model and a new model is trained from scratch. This process can be repeated several times.

Previous work based on consistency regularization has shown that adding perturbations at the image space when training with pseudo-labels helps to improve the performance [11, 19, 21]. Such perturbations are usually data augmentation strategies. They can prevent the model from easily memorizing the noisy pseudo-labels. Another motivation for the perturbations is that it may avoid ending up in the same local minima as the teacher during training. Motivated by those works, we utilise image mixing as perturbation for semantic segmentation when training with pseudo-labels. Here the image is changed by the Cutmix [29] augmentation using a CowMask [12]. With this augmentation, individual parts of the image are replaced by another image. The CowMask enables a greater diversity compared to CutMix [29] while looking similar to the typical black and white skin pattern of a cow. To generate a training sample of unlabelled data, we take two unlabelled images $\mathbf{x}_u^1$ and $\mathbf{x}_u^2$ and calculate their pseudo-labels $\hat{\mathbf{y}}_u^1$ and $\hat{\mathbf{y}}_u^2$ with the teacher model. Based on a binary CowMask

(a) first image   (b) second image   (c) cow mask   (d) resulting image

Figure 1: Example of the image mixing perturbation. The teacher network produces pseudo-labels on the original images. The pseudo-labels are mixed identical as the images. The student network is then trained to produce the mixed pseudo-labels on the mixed image.

$M \in \{0,1\}^{W \times H}$, a new augmented image $\bar{\mathbf{x}}_u$ and its pseudo-label is calculated by:

$$
\begin{aligned}
\bar{\mathbf{x}}_u &= M \odot \mathbf{x}_u^1 + (1-M) \odot \mathbf{x}_u^2, \\
\hat{\mathbf{y}}_u &= M \odot \hat{\mathbf{y}}_u^1 + (1-M) \odot \hat{\mathbf{y}}_u^1,
\end{aligned}
\tag{1}
$$

where $\odot$ is the element-wise multiplication. An illustration for this image mixing method is shown in Figure 1. For the calculation of a random CowMask, please refer to [12]. This perturbation simulates an occlusion of objects, as individual object parts are not visible within the image. It also simulates an additional segmentation task defined by the cow mask M.

3.3 Pseudo-Label Filtering

As stated in [22], wrong/noisy pseudo-labels in the training dataset generally have a negative impact on the training process, in the sense that the final performance is worse than if the false pseudo-labels were removed. This is due to that neural networks can overfit to wrong labels owing to their high capacity in totally memorising training samples [6, 22]. We made the same observation in an iteration of our method on assumed unlabelled data using pseudo labels. If all wrong pseudo-labels were not used as feedback at training, the subsequent performance increases enormously, which shows that wrong pseudo-labels do indeed lead to a poorer performing model. The critical question is: *How can we identify wrong pseudo-labels?* The common idea is to remove pseudo labels with low confidence from the training and assume that a low confidence correlates with wrong pseudo-labels. The confidence is defined by the output probability after the softmax function of the predicted class of the model. A higher confidence is equated with a higher certainty that the pseudo-label is correct. Another group of methods [20, 31, 33] measures uncertainty based on several different predictions per image. In doing so, we assume that the more the different predictions deviate from each other, the higher the uncertainty is. The different predictions can be generated, for example, by gathering the predictions of a model using different augmented versions of the same image, by using different dropout configurations within the model while forwarding the same image several times or by using other independently trained models where each model is trained with a different data-stream. Given a set of predictions in the form of probability distributions for the same original image, we can average the distributions and take their probability at the pseudo-label index as a measure of confidence.

We have also found that pixels at the class boundaries of the segmentation mask in particular, usually object edges, have higher uncertainty within an image than distant pixels. Therefore, handling all pixel-wise pseudo-labels identical may be misleading. Motivated by this, we propose to include the information of the segmentation borders of the image space at the pseudo-label filtering process. We do this by dividing the pixel-wise pseudo-labels into two groups, depending on the distance to the nearest segmentation boundary.

3.4 Pseudo-Label Weighting

Recent studies on the memorisation effect of deep neural networks show that they first memorise training data of clean labels and then those of noisy labels [6, 13]. This leads to our observation that the confidences of correct label predictions grow faster than the confidences of wrong label predictions over the course of model training. Motivated by this observation, we propose a weighting scheme with the aim to assign smaller weights to possibly wrong pseudo-labels and therefore reduce the contribution of noisy pseudo-labels at training. We calculate the weight of a pseudo-label from the current confidence value of the student model for the given pseudo-label. Similar to previous work [28] we further adapt the Symmetric Cross Entropy loss [23] (SCE), since it has been shown that it improve the performance when training on noisy datasets. The resulting loss $l^i \in \mathbb{R}^{H \times W}$ for an image of shape $H \times W$ is defined as:

$$l^i = \mathbf{w}^i \odot (\alpha l_{ce}(\mathbf{p}^i, \hat{\mathbf{y}}_u^i) + \beta l_{ce}(\hat{\mathbf{y}}_u^i, \mathbf{p}^i)) \tag{2}$$

where α and β are balancing coefficients and set to 2.0 and 1.0 respectively. $\mathbf{p}^i$ is the prediction of the student model on the perturbed image $\bar{\mathbf{x}}_u^i$, i.e. $\mathbf{p}^i = F_S(\bar{\mathbf{x}}_u^i)$. $\hat{\mathbf{y}}_u^i$ is the pseudo label from the current teacher model, i.e. $\hat{\mathbf{y}}_u^i = argmax(F_T(\mathbf{x}_u^i))$. The dynamic weights $\mathbf{w}^i \in \mathbb{R}^{H \times W}$ is defined by the confidence of the softmax output of the student model F_S and calculated as follows:

$$\mathbf{w}^i = P(\hat{\mathbf{y}}_u^i) \quad \text{with} \quad P = F_S(\mathbf{x}_u^i). \tag{3}$$

Compared to pseudo-label filtering, the proposed weighting scheme does not have the disadvantage that the resulting pseudo-label dataset is smaller than before, since no labels are removed. Furthermore, compared to the pseudo-label filtering approach, we use the confidence from the trained model as feedback instead of the confidence from the teacher model.

4 Experiments

4.1 Implementation Details

Dataset. We evaluate our framework on three different datasets, Cityscapes [7], PASCAL VOC 2012 [10] and Mapillary [18]. The Cityscapes dataset contains of images with 19 individual classes, where 2975 and 500 are used for training and validation respectively. We use an image size of 1024×2048 for our experiments. We further provide results for the PASCAL VOC 2012 [10] dataset, which consists of 1464 training and 1449 validation images. Following [17, 24, 28], we use SBD [14] as the augmented set with 9118 additional training images. We follow common practise and use center-crops of size 513×513. We further incorporate the Mapillary [18] dataset for our ablation study, since it contains 18000 diverse training images and 2000 images for validation. We additionally chose Mapillary because it contains a larger amount of diverse images. We resize the images to 512×1024 and incorporate the same 19 classes as the Cityscapes dataset. For each dataset, we randomly sample different partitions from the training set and set them as labelled images, while using the rest of the images as unlabelled. We also analyse our solution for the case that only very few labelled data are available, for example 15 labelled images at Cityscapes. Those images are taken such that all 19 classes are highly present.

Training. Following previous work [17, 24, 28], we use the DeepLabv3+ model [3] with a ResNet101 [15] backbone pretrained on Imagenet [8], if not stated otherwise. For the Cityscapes experiments with 15 labelled images, we initialize the backbone with COCO

Model	CM	ST	ST_{CM}	PLF	PLW	PLW_{SCE}	Iter.	$C_{(15)}$	$M_{(100)}$	$P_{(183)}$
DL3+								53.0	50.7	57.0
DL3+	✓							52.1	50.2	56.2
DL3+		✓						56.2	53.9	66.3
DL3+			✓					60.3	56.1	68.1
DL3+			✓	✓				60.9	56.8	67.8
DL3+			✓		✓			61.6	57.0	68.9
DL3+			✓			✓		62.4	57.8	69.5
DL3+			✓	✓		✓		63.0	58.0	69.3
DL3+			✓	✓		✓	✓	66.5	60.1	-

Table 1: Ablation study on the proposed SSL framework. "C", "M", "P" stands for the Cityscapes, Mapillary and PASCAL dataset respectively. The number in brackets stands for the number of labelled images used. 'CM' denotes CowMask augmentation. 'ST' denotes self-training with pseudo-labels. 'ST_{CM}' uses image mixing augmentation for training on pseudo-labels. 'PLF' and 'PLW' refers to pseudo-label filtering and weighting. 'SCE' refers to the symmetric cross-entropy loss.

weights. The network is trained using stochastic gradient descent. We use identical training details as in [24, 28]. We use the poly scheduling to decay the learning rate during the training. The learning rates are set to 0.001 and 0.01 for PASCAL VOC 2012 and Cityscapes dataset respectively. For the Cityscapes dataset, we use a crop-size of 769×769 during training. Batch size has been set to 8, where 2 samples are laballed and 6 unlabelled. We perform up to three iterations where we reinitialise the student network. We adopt the mean intersection over union (mIoU) as evaluation metric. Following [28], we apply random rescale and horizontal flip as augmentation. When using pseudo-labels for supervision, we additionally use the image mixing method. To filter pseudo-labels according to their confidence, we first calculate a confidence histogram for each class from the pseudo-label dataset. From that we can calculate class-wise threshold to remove certain percentage of pseudo-labels.

4.2   Ablation Study

To better investigate the influence of the individual methods we perform our ablation study on all three mentioned datasets. We also specifically assume a scenario with only a few labelled samples, as this presents the greatest difficulty and has a very large difference in performance compared to the fully supervised model. There is also a higher number of wrong pseudo-labels, which makes pseudo-label filtering more interesting.

4.2.1   Impact of Image Mixing Augmentation

In this section, we evaluate the contribution of the image mixing method. Table 1 shows on the one hand the resulting performance after a single iteration with additional training on unlabelled data with pseudo-labels and on the other hand the performance when we further apply the image mixing method when training on pseudo-labels. At the Cityscapes dataset, using pseudo-labels improves the performance from 53.0% to 56.2%. Further applying the image mixing approach, the performance increases to 60.2%. From this it can be seen that the use of an image mixing perturbation in addition to pseudo-labels leads to a similar per-
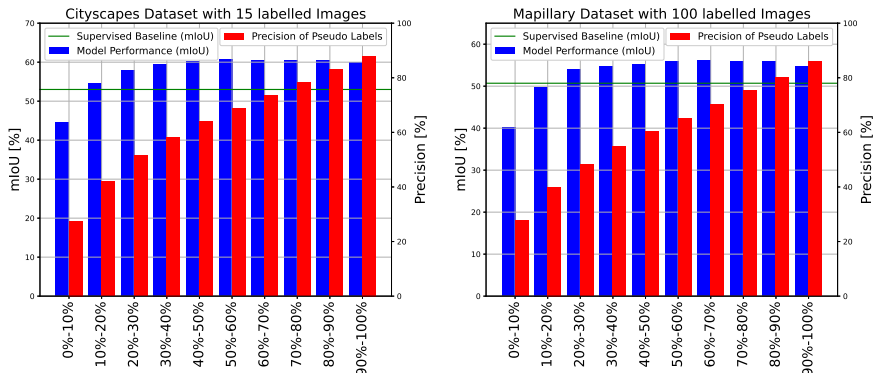
Figure 2: Analysis of different pseudo-label datasets ordered in 10% steps according to confidence. We report the class-wise precision of those pseudo-labels as well as the model performance when incorporating a single pseudo-label subset into our SSL training. The green horizontal line shows the supervised only model.

formance boost as the use of pseudo-labels in general. To verify that the improvement does not come from the data augmentation strategy, we also trained the fully supervised baseline model with the image mixing augmentation. Here we can observe a small deterioration in performance, showing that this augmentation strategy only improves the result when applied to images with pseudo-labels, not to images where ground-truth annotation is available. The same findings can be seen at the PASCAL VOC 2012 and Mapillary dataset. Table 4 further compares the usage of a CowMask to CutMix [29]. It can be seen, that CowMask is more effective and yields better results.

4.2.2 Impact of Pseudo-Label Filtering

First we analysed whether the pseudo-label filtering by the softmax confidence of the model performs different as filtering based on several predictions as introduced in section 3.3. We did so by adding the most confident examples to a dataset and analyse this dataset for correctness. Here we observed that with all methods we obtain a dataset with equal quality. The multiple model approach also did not lead to an improvement, which probably comes from the fact that only a small amount of labelled data was available. Due to the small differences between the methods and the simplicity of the approach using the model confidence, we keep it for the rest of this work.

Next, we analysed pseudo-labels with different uncertainty in more detail. Given the complete pseudo-label dataset, we split it into ten subsets according to the confidence, taking into account different classes. For example, one dataset contains pseudo-labels of each class with the lowest confidence, i.e. $0\% - 10\%$. Then we performed a semi-supervised training iteration using one of the ten pseudo-label datasets. Figure 2 shows the quality of the pseudo-label subsets and the resulting performance of the trained model when additionally using a single pseudo-label subset for training. As expected, we can observe that the confidence of the model prediction indeed correlates with the correctness of the pseudo-labels. For the Mapillary dataset, from the top 10% confident samples per class 86% are correct on average, while only 28% of the worst 10% are correct. A similar correlation can be seen at the Cityscapes experiment. Comparing the final performance when incorporating

the pseudo-label subset into a semi-supervised training, we can observe a weaker correlation. For both datasets, taking the pseudo-labels with the worst 10% confidence decreases the final performance. At the Mapillary dataset we drop in performance from 50.7% to $\sim 40\%$ when we additionally use those pseudo-labels for training. This shows that those pseudo-labels have a harmful influence at the training. The same can be observed at the Cityscapes experiment. However, this effect no longer occurs in samples with a slightly higher confidence, most probably because strong perturbations already suppress the negative effect of wrong pseudo-labels. At the Cityscapes experiment, using the weakest $10\% - 20\%$ samples additionally for training already improves the performance slightly. Even more interestingly, we can observe that the worst $30\% - 40\%$ have an almost equal effect in performance boost as the top 10% confident samples $(90\% - 100\%)$ on both datasets. This is probably because the correct pseudo-labels from the lower confident subsets are samples that the model struggles with. In other words, the correct pseudo-labels from lower confident subsets improve the performance more than those from the high confident samples. High confident samples are probably more similar to examples from the labelled training set and correct pseudo-labels with low confidence have a higher impact on the resulting decision boundary.

Based on this findings, we removed the most uncertain 20% of each class from the complete pseudo-label dataset. The result is shown in Table 1. The performance increases slightly for the Cityscapes and Mapillary experiment, respectively. At the PASCAL VOC experiment we can observe a small deterioration. Here we also observed that additionally using the worst 10% of the pseudo-labels lead to a similar performance as the supervised baseline. From the experiments, it can be concluded that a confidence-based filtering does not lead to the hoped-for significant improvement. The Mapillary experiment further verified, that the marginal improvement does not come from the smaller dataset. From these results, we conclude that removing incorrect pseudo-labels lead to an improvement, but removing correct pseudo-labels leads to a deterioration, and that both effects ultimately overlap. Thus, confidence-based filtering is not good enough at detecting wrong pseudo-labels because there are too many correct pseudo-labels with low confidence.

4.2.3   Impact of the Robust Loss Function

To improve the performance of the model and reduce the impact of wrong pseudo-labels, we employ a weighting scheme of pseudo-labels defined by the confidence of the trained model for the pseudo-labels of the teacher. From Table 1 we can see that the weighting scheme improves the performance on all three experimental settings. We further verified whether wrong pseudo-labels indeed have a lower weight during training than correct pseudo-labels. Here we found that the weights of the correct pseudo-labels were on average twice as high as those of wrong pseudo-labels. This shows that the influence of wrong pseudo-labels during training is reduced and therefore an improvement in performance occurs. Applying the SCE loss instead of the standard cross-entropy loss further improves the performance slightly. We provide sensitivity analysis of the alpha and beta parameters of the SCE loss in the Appendix.

4.3   Comparison with State-of-the-art Methods

We compare our method with the most recent semi-supervised semantic segmentation methods that are mostly related to our work: Simple Baseline [28], U2PL [24], ST++ [27] and PS-MT [17]. We further compare our solution to CPS [5] which is slightly different to our work, since here two independent networks are trained simultaneously. We report results

Method	Model	$\frac{1}{200}$ (15)	$\frac{1}{20}$ (148)	$\frac{1}{16}$ (186)	$\frac{1}{8}$ (372)	$\frac{1}{4}$ (744)	Full
SupOnly (ours)	DL3+	53.0	64.1	64.8	69.0	73.3	78.7
CutMix⋆ [11]	DL3+	-	-	67.1	71.8	76.4	-
Yuan et al. [28]	DL3+	-	-	-	74.1	77.8	78.7
U2PL [24]	DL3+	-	-	74.9	76.5	78.5	-
PS-MT [17]	DL3+	-	-	-	77.1	78.4	-
CPS [5]	DL3+	-	-	74.7	77.6	**79.1**	80.4
Ours	DL3+	**66.5**	**74.4**	**75.7**	**78.0**	78.7	-

Table 2: Comparison of state-of-the-art methods for different portions of labelled data on the Cityscapes validation set. ⋆ means approach has been reproduced by [24].

Method	Model	$\frac{1}{16}$ (92)	$\frac{1}{8}$ (183)	$\frac{1}{4}$ (366)	$\frac{1}{2}$ (732)	Full (1464)
SupOnly (ours)	DL3+	51.0	57.0	66.8	72.1	73.4
CutMix⋆ [11]	DL3+	52.2	63.5	69.5	73.7	76.5
PseudoSeg [34]	DL3+	57.6	65.5	69.1	72.4	73.2
PC^{2}Seg [32]	DL3+	57.0	66.3	69.8	73.1	74.2
CPS [5]	DL3+	64.1	67.4	71.7	75.9	-
U2PL [24]	DL3+	68.0	69.2	73.7	76.4	**79.5**
ST++ [27]	DL3+	65.2	71.0	**74.6**	77.3	79.1
Ours	DL3+	**68.6**	**71.4**	74.4	**77.5**	**79.5**

Table 3: Comparison of state-of-the-art methods for different portions of labelled data on the PASCAL VOC 2012 validation set. The labelled images are selected from the original VOC training dataset. The rest plus the images from the SBD [14] dataset are regarded as unlabelled data. ⋆ means approach has been reproduced by [24].

without using pseudo-label filtering with the exception of 15 labelled image at Cityscapes. The results for the Cityscapes dataset are shown in Table 2. Our approach improves the performance from our supervised baseline by +13.5%, +10.3%, +10.9%, +9% and +5.4% under $^1/_{200}$, $^1/_{20}$, $^1/_{16}$, $^1/_8$ and $^1/_4$ partition protocols respectively. The proposed approach performs favourably against state-of-the art approaches that are mostly related to our work when assuming $\frac{1}{16}$, $\frac{1}{8}$ and $\frac{1}{4}$ to be labelled. We can notice that using $\frac{1}{4}$ of the labelled data in our SSL framework achieves a similar performance as using the whole dataset for supervision. Comparing the results to those of [28], the improvement comes from the image mixing method using a CowMask and the pseudo-label weighting strategy, as shown in our ablation study. We further provide results on the classic PASCAL VOC 2012 dataset in Table 3. Here our approach improves the performance from our supervised baseline by +17.6%, +14.4%, +7.6% and +5.4% under $^1/_{16}$, $^1/_8$, $^1/_4$ and $^1/_2$ partition protocols respectively.

4.4 Human Pose Estimation

To verify if our findings are not specific to semantic segmentation, we also apply our method on the human pose estimation (HPE) task. The experiments are conducted on the LSP dataset [16]. It contains of 2000 images with 14 annotated joints. For our ablations we

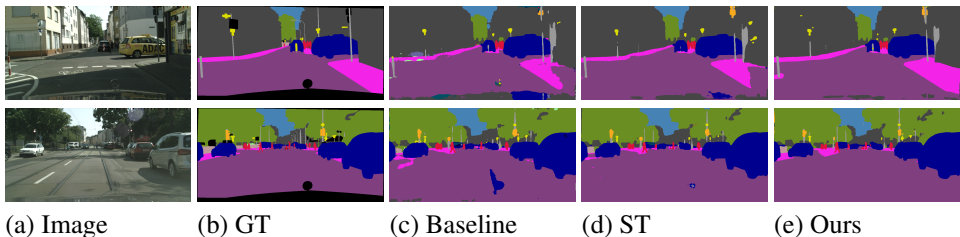| (a) Image | (b) GT | (c) Baseline | (d) ST | (e) Ours |

Figure 3: Qualitative examples on the Cityscapes val set. All models are trained using the annotation of 15 images. The Baseline is trained supervised only. "ST" represents the naive self-training approach without our proposed extensions. Our proposed solution produces improved results compared to the baseline and the naive self-training approach.

Method	$C_{(15)}$	$M_{(100)}$	$P_{(183)}$
PL	56.2	53.9	66.3
PL + CutMix	58.3	55.0	66.9
PL + CowMask	60.3	56.1	68.1

Table 4: Ablation study comparing CowMask and CutMix augmentation on pseudo-labels at the first iteration. "C", "M", "P" refer to the different datasets.

Method	$\frac{1}{16}(100)$	$\frac{1}{4}(400)$	Full$_{(1600)}$
Sup. only	62.3	73.5	82.6
PL	65.9	76.4	-
+ CowMask	68.2	78.2	-
+ PLW	69.6	78.8	-
+ PLF	69.4	78.5	-
+ iter	71.3	79.4	-

Table 5: Ablation study of our proposed methods on a common HPE task.

use 1600 images as training set and the rest for evaluation. We use *SimpleBaseline* [25] to estimate heatmaps and ResNet18 as its backbone for simplicity. Note that our approach can be applied to other pose estimators as well. We use the standard Percentage of Correct Key-points (PCK) measurement that quantifies the fraction of correct predictions within an error threshold λ normalised by the size of the torso. We set λ to 0.2, i.e. PCK@0.2. We perform ablations with image mixing using a CowMask, pseudo label weighting as well as filtering. The results are summarised at Table 5. Similar to semantic segmentation, we can observe an improvement when using image mixing as well as our proposed pseudo-label weighting scheme, while pseudo-label filtering is less effective. Since the findings for HPE are similar to those of semantic segmentation, we can assume that they also apply to other domains and tasks.

5 Conclusion

We have proposed a simple and efficient semi-supervised semantic segmentation framework. It employs efficient image mixing with a diverse mask and pseudo-label denoising strategies. We have analysed the effect of confident-based pseudo label filtering in more detail and shown that it leads to little, if any, improvement. In addition to the pseudo-label filtering we have also analysed our novel weighting of pseudo-labels given the current feedback of the model. This approach has shown to be effective and yield to an overall improvement on three different datasets. Quantitative and qualitative comparisons have shown that the proposed method performs favourably against state-of-the art approaches and also apply to other tasks such as human pose estimation.

References

[1] David Berthelot, Nicholas Carlini, Ekin D Cubuk, Alex Kurakin, Kihyuk Sohn, Han Zhang, and Colin Raffel. Remixmatch: Semi-supervised learning with distribution alignment and augmentation anchoring. *arXiv preprint arXiv:1911.09785*, 2019.

[2] David Berthelot, Nicholas Carlini, Ian Goodfellow, Nicolas Papernot, Avital Oliver, and Colin A Raffel. Mixmatch: A holistic approach to semi-supervised learning. *Advances in neural information processing systems*, 32, 2019.

[3] 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.

[4] Liang-Chieh Chen, Raphael Gontijo Lopes, Bowen Cheng, Maxwell D Collins, Ekin D Cubuk, Barret Zoph, Hartwig Adam, and Jonathon Shlens. Naive-student: Leveraging semi-supervised learning in video sequences for urban scene segmentation. In *European Conference on Computer Vision*, pages 695–714. Springer, 2020.

[5] 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.

[6] Filipe R Cordeiro and Gustavo Carneiro. A survey on deep learning with noisy labels: How to train your model when you cannot trust on the annotations? In *2020 33rd SIBGRAPI conference on graphics, patterns and images (SIBGRAPI)*, pages 9–16. IEEE, 2020.

[7] 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 *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[8] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

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

[10] 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.

[11] Geoff French, Samuli Laine, Timo Aila, Michal Mackiewicz, and Graham Finlayson. Semi-supervised semantic segmentation needs strong, varied perturbations. *arXiv preprint arXiv:1906.01916*, 2019.

[12] Geoff French, Avital Oliver, and Tim Salimans. Milking cowmask for semi-supervised image classification. *arXiv preprint arXiv:2003.12022*, 2020.

[13] Bo Han, Quanming Yao, Xingrui Yu, Gang Niu, Miao Xu, Weihua Hu, Ivor Tsang, and Masashi Sugiyama. Co-teaching: Robust training of deep neural networks with extremely noisy labels. *Advances in neural information processing systems*, 31, 2018.

[14] 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.

[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.

[16] Sam Johnson and Mark Everingham. Clustered pose and nonlinear appearance models for human pose estimation. In *Proceedings of the British Machine Vision Conference*, 2010. doi:10.5244/C.24.12.

[17] Yuyuan Liu, Yu Tian, Yuanhong Chen, Fengbei Liu, Vasileios Belagiannis, and Gustavo Carneiro. Perturbed and strict mean teachers for semi-supervised semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4258–4267, 2022.

[18] Gerhard Neuhold, Tobias Ollmann, Samuel Rota Bulo, and Peter Kontschieder. The mapillary vistas dataset for semantic understanding of street scenes. In *Proceedings of the IEEE international conference on computer vision*, pages 4990–4999, 2017.

[19] Viktor Olsson, Wilhelm Tranheden, Juliano Pinto, and Lennart Svensson. Classmix: Segmentation-based data augmentation for semi-supervised learning. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1369–1378, 2021.

[20] Mamshad Nayeem Rizve, Kevin Duarte, Yogesh S Rawat, and Mubarak Shah. In defense of pseudo-labeling: An uncertainty-aware pseudo-label selection framework for semi-supervised learning. *arXiv preprint arXiv:2101.06329*, 2021.

[21] Kihyuk Sohn, David Berthelot, Nicholas Carlini, Zizhao Zhang, Han Zhang, Colin A Raffel, Ekin Dogus Cubuk, Alexey Kurakin, and Chun-Liang Li. Fixmatch: Simplifying semi-supervised learning with consistency and confidence. *Advances in Neural Information Processing Systems*, 33:596–608, 2020.

[22] Hwanjun Song, Minseok Kim, Dongmin Park, Yooju Shin, and Jae-Gil Lee. Learning from noisy labels with deep neural networks: A survey. *arXiv preprint arXiv:2007.08199*, 2020.

[23] Yisen Wang, Xingjun Ma, Zaiyi Chen, Yuan Luo, Jinfeng Yi, and James Bailey. Symmetric cross entropy for robust learning with noisy labels. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 322–330, 2019.

[24] Yuchao Wang, Haochen Wang, Yujun Shen, Jingjing Fei, Wei Li, Guoqiang Jin, Liwei Wu, Rui Zhao, and Xinyi Le. Semi-supervised semantic segmentation using unreliable pseudo-labels. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4248–4257, 2022.

[25] Bin Xiao, Haiping Wu, and Yichen Wei. Simple baselines for human pose estimation and tracking. In *Proceedings of the European conference on computer vision (ECCV)*, pages 466–481, 2018.

[26] Qizhe Xie, Eduard H. Hovy, Minh-Thang Luong, and Quoc V. Le. Self-training with noisy student improves imagenet classification. *CoRR*, abs/1911.04252, 2019. URL http://arxiv.org/abs/1911.04252.

[27] Lihe Yang, Wei Zhuo, Lei Qi, Yinghuan Shi, and Yang Gao. St++: Make self-training work better for semi-supervised semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4268–4277, 2022.

[28] Jianlong Yuan, Yifan Liu, Chunhua Shen, Zhibin Wang, and Hao Li. A simple baseline for semi-supervised semantic segmentation with strong data augmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8229–8238, 2021.

[29] 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.

[30] Bowen Zhang, Yidong Wang, Wenxin Hou, Hao Wu, Jindong Wang, Manabu Okumura, and Takahiro Shinozaki. Flexmatch: Boosting semi-supervised learning with curriculum pseudo labeling. *Advances in Neural Information Processing Systems*, 34: 18408–18419, 2021.

[31] Zhedong Zheng and Yi Yang. Rectifying pseudo label learning via uncertainty estimation for domain adaptive semantic segmentation. *International Journal of Computer Vision*, 129(4):1106–1120, 2021.

[32] Yuanyi Zhong, Bodi Yuan, Hong Wu, Zhiqiang Yuan, Jian Peng, and Yu-Xiong Wang. Pixel contrastive-consistent semi-supervised semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7273–7282, 2021.

[33] Qianyu Zhou, Zhengyang Feng, Qiqi Gu, Guangliang Cheng, Xuequan Lu, Jianping Shi, and Lizhuang Ma. Uncertainty-aware consistency regularization for cross-domain semantic segmentation. *arXiv preprint arXiv:2004.08878*, 2020.

[34] 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.