Discovering Latent Classes for Semi-Supervised Semantic Segmentation

Olga Zatsarynna[1⋆], Johann Sawatzky[1,2⋆], and Juergen Gall[1]

[1] University of Bonn
[2] EyewareTech
{s6olzats, jsawatzk, jgall} @ uni-bonn.de

Abstract. High annotation costs are a major bottleneck for the training of semantic segmentation approaches. Therefore, methods working with less annotation effort are of special interest. This paper studies the problem of semi-supervised semantic segmentation, that is only a small subset of the training images is annotated. In order to leverage the information present in the unlabeled images, we propose to learn a second task that is related to semantic segmentation but that is easier to learn and requires less annotated images. For the second task, we learn latent classes that are on one hand easy enough to be learned from the small set of labeled data and are on the other hand as consistent as possible with the semantic classes. While the latent classes are learned on the labeled data, the branch for inferring latent classes provides on the unlabeled data an additional supervision signal for the branch for semantic segmentation. In our experiments, we show that the latent classes boost the accuracy for semi-supervised semantic segmentation and that the proposed method achieves state-of-the-art results on the Pascal VOC 2012 and Cityscapes datasets.

Keywords: Semantic Segmentation; Semi-Supervised Learning; Generative Adversarial Networks.

1 Introduction

In recent years, deep convolutional neural networks (DCNNs) have achieved astonishing performance for the task of semantic segmentation. However, to achieve good results, DCNN-based methods require an enormous amount of high-quality annotated training data and acquiring it takes a lot of effort and time. This problem is especially acute for the task of semantic segmentation, due to the need for per-pixel labels for every training image. To mitigate the annotation expenses, Hung et al. [14] proposed a semi-supervised algorithm that employs images without annotation during training. On labeled data, the authors train a discriminator network that distinguishes segmentation predictions and ground-truth annotations. On unlabeled data, they use the discriminator to obtain two

⋆ contributed equally

(a) Image (b) Latent Classes (c) Semantic Classes

Fig. 1. Our network learns not only semantic but also latent classes that are easier to predict. The figure shows an example of latent and semantic class segmentation for an image that is not part of the training data. As it can be seen, the learned latent classes are very intuitive since the vehicles are grouped into one latent class and objects that are difficult to segment like pedestrians, bicycles, and signs are grouped into another latent class.

kinds of supervision signals. First, they use an adversarial loss to enforce realism in the predictions. Second, they use the discriminator to locate regions of sufficient realism in the prediction. These regions are then annotated by the semantic class with the highest probability. Finally, the network for semantic segmentation is trained on the labeled images and the estimated regions of the unlabeled images. Recently, Mittal et al. [28] introduced an extension to [14] by improving the adversarial training and adding a semi-supervised classification module. The latter is used for refining the predictions at the inference time. Although these approaches report impressive results for semi-supervised semantic segmentation, they do not leverage the entire information which is present in the unlabeled images since they discard large parts of the images.

In this work, we propose an approach for semi-supervised semantic segmentation that does not discard any information. Our key observation is that the difficulty of the semantic segmentation task depends on the definition of the semantic classes. This means that the task can be simplified if some classes are grouped together or if the classes are defined in a different way, which is more consistent with the similarity of the instances in the feature space. If the segmentation task becomes easier, less labeled data will be required to train the network. This approach is in contrast to [14, 28] that focus on regions in the unlabeled images that are easy to segment, whereas we focus to learn a simpler segmentation task with latent classes on the labeled data that is then used as additional guidance to learn the original task on the labeled and unlabeled data. Figure 1 shows an example of inferred latent classes and semantic classes.

Our network consists of two branches and is trained on labeled and unlabeled images jointly in an end-to-end fashion as illustrated in Figure 2. While the semantic branch learns to infer the given semantic classes, the latent branch learns latent classes and infers the learned latent classes. In contrast to the semantic branch, the loss for the latent branch takes only the labeled images into account. The purpose of the latent branch is to discover latent classes that are simple enough such that they can be learned on the small set of labeled data. Without any constraints this would result in a single latent class. We there-

fore introduce a conditional entropy loss that minimizes the variety of semantic classes that are assigned to a particular latent class. In other words, the latent classes should be on one hand easy enough to be learned from the small set of labeled data and on the other hand they should be as consistent as possible with the semantic classes. Since the latent branch solves a simpler semantic segmentation task, we use it as additional supervision for the semantic branch on the unlabeled images. After training, the latent branch is discarded and only the semantic branch is used for inference.

We demonstrate that our model achieves state-of-the-art results on PASCAL VOC 2012 [8] and Cityscapes [6]. Additionally, we show that the learned latent classes are superior to manually defined supercategories.

2 Related Work

The expensive acquisition of pixel-wise annotated images has been recognized as a major bottleneck for the training of deep semantic segmentation models. Consequently, the community sought ways to reduce the amount of annotated images while loosing as little performance as possible.

Weakly-supervised semantic segmentation methods learn to segment images from cheaper image annotations, i.e. pixel-wise labels are exchanged for cheaper annotations for all the images in the training set. The proposed types of annotations include bounding boxes [16, 23, 31, 41], scribbles [25, 42, 43] or human annotated keypoints [2]. Image level class tags have attracted special attention. A minority of works in this area first detect potential object regions and then identify the object class using the class tags [9, 32, 34]. The majority of approaches use class activation maps (CAMs) [49] to initially locate the classes of interest. Pinheiro et al. [33, 40] pioneered in this area and several methods have improved this approach [1, 3, 4, 10, 12, 13, 17, 30, 37, 43–47]. A few works leverage additional data available on the Internet. For example, [11, 15, 20] use videos. While the works mentioned above mainly focus on refining the localization cues obtained from the CAM, recently the task of improving the CAM itself received attention [19, 20, 22].

Some of the works mentioned above consider a setup where some images have pixel-wise annotations and the other images are weakly labeled. They combine fully supervised learning with weakly supervised learning. Papandreou et al. [31] proposed an expectation maximization based approach, modelling the pixel-wise labels as hidden variables and the image labels or bounding boxes as the observed ones. Lee et al. [19] introduce a sophisticated dropout method to obtain better class activation maps on unlabeled images. Earlier, Li et al. [22] improved the CAMs by automatically erasing the most discriminative parts of an object. Wei et al. [47] examine what improvement in CAMs can be achieved by dilated convolutions. Different from previous approaches, Zilong et al. [13] do not improve the CAM but focus on refining high confidence regions obtained from the CAM by deep seeded region growing. The semi-supervised setting without any additional weak supervision has been so far only addressed by [14, 28].

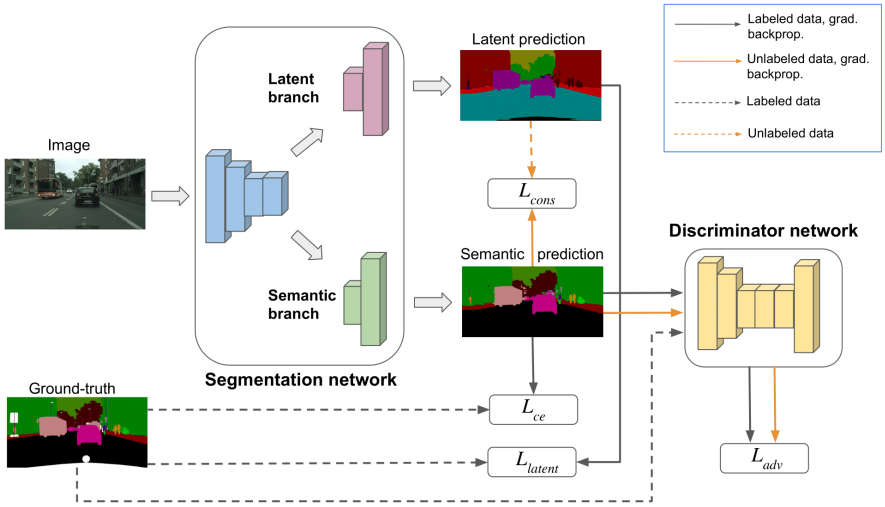


Fig. 2. Overview of the proposed method. While the semantic branch infers pixel-wise class labels, the latent branch learns latent classes and infers the learned latent classes. The latent classes are learned only on the labeled images using the latent loss L_{latent} that ensures that the latent classes are as consistent as possible with the semantic classes. The semantic branch is trained on labeled images with the cross-entropy loss L_{ce} and on unlabeled images the predictions of the latent branch are used as supervision (L_{cons}). Additionally, the semantic branch receives adversarial feedback (L_{adv}) from a discriminator network distinguishing predicted and ground truth segmentations.

While learning an easier auxiliary task as an intermediate step has been investigated in the area of domain adaptation [7, 18, 24, 39, 48], it has not been studied for semi-supervised semantic segmentation. Moreover, using latent classes to facilitate learning has been investigated for object detection [35, 50], joint object detection and pose estimation [21], and weakly-supervised video segmentation [36]. However, apart from addressing a different task, these approaches focus on discovering subcategories of classes while we aim to group the classes.

3 Method

An overview of our method is given in Figure 2. Our proposed model is a two-branch network. While the semantic branch serves to solve the final task, the purpose of the latent branch is to learn to group the semantic classes into latent classes in a data driven way as fine-grained as possible. While the fraction of annotated data is not sufficient to produce good results for the task of semantic segmentation, it is enough to learn the prediction of latent classes reasonably well, since this task is easier. Thus, the predictions of the latent branch can then serve as a supervision signal for the semantic branch on unlabeled data.

3.1 Semantic Branch

The task of the semantic branch S_c is to solve the final task of semantic segmentation, that is to predict the semantic classes for the input image. This branch is trained both on labeled and unlabeled data.

On labeled data, we optimize the semantic branch with respect to two loss terms. The first term is the cross-entropy loss:

$$L_{ce} = -\sum_{h,w,n}\sum_{c\in C} Y_n^{(h,w,c)} \log(S_c(X_n)^{(h,w,c)}) \tag{1}$$

where $X_n \in \mathbb{R}^{H\times W\times 3}$ is the image, $Y_n \in \mathbb{R}^{H\times W\times|\mathcal{C}|}$ is the one-hot encoded ground truth for semantic classes, and S_c is the predicted probability of the semantic classes. To enforce realism in the semantic predictions, we additionally apply an adversarial loss:

$$L_{adv} = -\sum_{n,h,w} \log(D(S_c(X_n))^{(h,w)}) \tag{2}$$

Details of the discriminator network D are given in Section 3.4

On unlabeled data, the loss function for the semantic branch also consists of two terms. The first one is the adversarial term (2) and the second term is the consistency loss that is described in Section 3.3.

3.2 Latent Branch

In order to provide additional supervision for the semantic branch on the unlabeled data, we introduce a latent branch S_l that is trained only on the labeled data. The purpose of the latent branch is to learn latent classes that are easier to distinguish than the semantic classes and that can be better learned on a small set of labeled images. Figure 1 shows an example of latent classes where for instance semantic similar classes like vehicles are grouped together. One of the latent classes often corresponds to a stuff class that includes all difficult classes. This is desirable since having several latent classes that are easy to recognize and one latent class that contains the rest results in a simple segmentation task that can be learned from a small set of labeled images. However, we have to prevent a trivial solution where a single latent class contains all semantic classes. We therefore propose a loss that ensures that the latent classes $l \in \mathcal{L}$ have to provide as much information about semantic classes $c \in \mathcal{C}$ as possible.

To this end, we use the conditional entropy as loss:

$$L_{latent} = -\sum_{l\in\mathcal{L}}\sum_{c\in\mathcal{C}} P_b(c,l)\log(P_b(c|l)). \tag{3}$$

The loss is minimized if the variety of possible semantic classes for each latent class l is as low as possible. In the best case, there is a one-to-one mapping

between the latent and semantic classes. The index b denotes that the probability is calculated batchwise. We first estimate the joint probability

$$P_b(c,l) = \frac{1}{NHW} \sum_{h,w,n} S_l(X_n)^{(h,w,l)} Y_n^{(h,w,c)} \tag{4}$$

where H and W are the image height and width, N is the number of images in the batch, S_l is the predicted probability of the latent classes, and $Y_n \in \mathbb{R}^{H \times W \times |\mathcal{C}|}$ is the one-hot encoded ground truth for the semantic classes. From this, we obtain

$$P_b(c|l) = \frac{P_b(c,l)}{\sum_c P_b(c,l)}. \tag{5}$$

Obtaining the conditional entropy from multiple batches is in principle desirable, but it requires the storage of feature maps from multiple batches. Therefore we compute it per batch.

3.3 Consistency Loss

While the latent branch is trained only on the labeled data, the purpose of the latent branch is to provide additional supervision for the semantic branch on the unlabeled data. Given that the latent branch solves a simpler task than the semantic branch, we can expect that the latent classes are more accurately predicted than the semantic classes. We therefore propose a loss that measures the consistency of the prediction of the semantic branch with the prediction of the latent branch. Since the number of latent classes is less or equal than the number of semantic classes, we map the prediction of the semantic branch S_c to a probability distribution of latent classes $S_{\hat{l}_c}$:

$$S_{\hat{l}_c}(X_n)^{(h,w,l)} = \sum_{c \in \mathcal{C}} P(l|c) S_c(X_n)^{(h,w,c)}. \tag{6}$$

We estimate $P(l|c)$ from the predictions of the latent branch on the labeled data. We keep track of how often semantic and latent classes co-occur with an exponentially moving average:

$$M_{c,l}^{(i)} = (1-\alpha) M_{c,l}^{(i-1)} + \alpha \sum_{h,w,n} Y_n^{(h,w,c)} S_l(X_n)^{(h,w,l)} \tag{7}$$

where i denotes the number of the batch. The initialization is $M_{c,l}^0 = 0$. The parameter $0 < \alpha < 1$ controls how fast we update the average. We set α to the batch size divided by the number of images in the data set. Using the acquired co-occurence matrix M, $P(l|c)$ is estimated as:

$$P(l|c) = \frac{M_{c,l}}{\sum_{k \in \mathcal{L}} M_{c,k}}. \tag{8}$$

The consistency loss is then defined by the mean cross entropy between the latent variable maps predicted by the latent branch S_l and the ones constructed based on the prediction of the semantic branch $S_{\hat{l}_c}$:

$$L_{cons} = -\frac{1}{NHW} \sum_{n,h,w} \sum_{l \in \mathcal{L}} S_l(X_n)^{(h,w,l)} \log(S_{\hat{l}_c}(X_n)^{(h,w,l)}). \tag{9}$$

The minimization of this loss forces the semantic branch to predict classes which are assigned to highly probable latent classes.

3.4 Discriminator Network

Our discriminator network D is a fully-convolutional network [27] with 5 layers and Leaky-ReLu as nonlinearity. It takes label probability maps from the segmentation network or ground-truth maps as input and predicts spatial confidence maps. Each pixel represents the confidence of the discriminator about whether the corresponding pixel in a semantic label map was sampled from the ground-truth map or the segmentation prediction. We train the discriminator network with the help of the spatial cross-entropy loss using both labeled and unlabeled data:

$$L_D = -\sum_{h,w} (1 - y_n) \log(1 - D(S_c(X_n))^{h,w}) + y_n \log(D(Y_n)^{h,w}) \tag{10}$$

where $y_n = 0$ if a sample is drawn from the segmentation network, and $y_n = 1$ if it is a ground-truth map. By minimizing such a loss, the discriminator learns to distinguish between the generated and ground-truth probability maps.

4 Experiments

4.1 Implementation Details

For a fair comparison with Hung et al. [14] and Mittal et al. [28], we choose the same backbone architecture and keep the same hyper-parameters where appropriate. For the segmentation network, we use a single scale ResNet-based DeepLab-v2 [5] architecture that is pre-trained on ImageNet [38] and MSCOCO [26]. We branch the proposed network at the last layer by applying Atrous Spatial Pyramid Pooling (ASPP) [5] two times for the semantic and latent branch. Finally, we use bilinear upsampling to make the predictions match the initial image size.

For the discriminator network, we use a fully convolutional network, which contains 5 convolutional layers with kernels of the sizes 4×4 and 64, 128, 256, 512 and 1 channels, applied with a stride equal to 2. Each convolutional layer, except for the last one, is followed by a Leaky-ReLU with the leakage coefficient equal to 0.2.

Table 1. Comparison to the state-of-the-art on Pascal VOC 2012 using mIoU (%).

	Fraction of annotated images					
Method	1/50	1/20	1/8	1/4	1/2	Full
Hung et al. [14]	55.6	64.6	69.5	72.1	73.8	74.9
Mittal et al. [28]	**63.3**	67.2	71.4	-	-	75.6
Proposed	59.6	68.2	71.3	**72.4**	**73.9**	75.0
Proposed + Classifier	61.8	**69.3**	**72.2**	-	-	75.3

We train the segmentation network on labeled and unlabeled data jointly with $L = L_{labeled} + 0.1 \cdot L_{unlabeled}$ where the weight factor is the same as in [14]. The loss for the labeled and unlabeled data are given by

$$L_{labeled} = L_{ce} + L_{latent} + 0.01 \cdot L_{adv}, \tag{11}$$

$$L_{unlabeled} = L_{cons} + 0.01 \cdot L_{adv}. \tag{12}$$

The weight for the adversarial loss is also the same as in [14]. By default, we limit the number of latent classes to 20. Additional details are provided as part of the supplementary material.

We conducted our experiments on three datasets for semantic segmentation: Pascal VOC 2012 [8], Cityscapes [6] and IIT Affordances [29]. We report the results for the IIT Affordances dataset [29] in the supplementary material. The Pascal VOC 2012 dataset contains images with objects from 20 foreground classes and one background class. There are 10528 training and 1449 validation images in total. The testing of the resulting model is carried out on the validation set. The Cityscapes dataset comprises images extracted from 50 driving videos. It contains 2975, 500 and 1525 images in the training, validation and test set, respectively, with annotated objects from 19 categories. We report the results of testing the resulting model on the validation set. As an evaluation metric, we use mean-intersection-over-union (mIoU).

4.2 Comparison with the State-of-the-Art

PASCAL VOC 2012. On the PASCAL VOC 2012 dataset, we conducted our experiments on five fractions of annotated images, as shown in Table 1, where the rest of the images are used as unlabeled data. Since [14] report the results only for the latest three fractions, we evaluate the performance of their method for the unreported fractions based on the publicly available code. The improvement is especially pronounced, if we look at the sparsely labeled data fractions, such as 1/50, 1/20 and 1/8. Our method performs on par with [28] and the leading method varies from data fraction to data fraction. However, our approach of learning latent variables is complementary to [28] and we can also add a classifier for refinement as in [28]. We show some qualitative results of our method in the supplementary material.

Table 2. Comparison to the state-of-the-art on Cityscapes using mIoU (%).

		Fraction of annotated images			
Method	Pre-training	1/8	1/4	1/2	Full
Mittal et al. [28]		59.3	61.9	-	65.8
Proposed		61.0	63.1	-	64.9
Hung et al. [14]	COCO	58.8	62.3	65.7	67.7
Proposed	COCO	**63.3**	**65.4**	**66.1**	66.3

Table 3. Impact of the loss terms. The evaluation is performed on Pascal VOC 2012 where 1/8 of the data is labeled. $L_{adv}^{labeled}$ denotes that the adversarial loss is only used for the labeled images.

Loss	mIoU (%)
L_{ce}	64.1
$L_{ce} + L_{latent}$	64.6
$L_{ce} + L_{latent} + L_{cons}$	67.3
$L_{ce} + L_{adv}^{labeled}$	68.7
$L_{ce} + L_{adv}$	69.4
$L_{ce} + L_{latent} + L_{cons} + L_{adv}$	71.3

Cityscapes. For the Cityscapes dataset, we follow the semi-supervised learning protocol that was proposed in [14]. This means that 1/8, 1/4 or 1/2 of the training images are annotated and the other images are used without any annotations. We report the results in Table 2. Since [28] does not pre-train the segmentation network on COCO, we evaluated our method also without COCO pre-training. We outperform both [14] and [28] on all annotated data fractions. We show some qualitative results of our method in the supplementary material.

4.3 Ablation Experiments

In our ablation experiments, we evaluate the impact of each loss term. Then we examine the impact of the number of latent classes and show that they form meaningful supercategories of the semantic classes. Finally, we show that the learned latent classes outperform supercategories that are defined by humans.

Impact of the loss terms. For analyzing the impact of the loss terms L_{ce} (1), L_{adv} (2), L_{latent} (3), and L_{cons} (9), we use the Pascal VOC 2012 dataset where 1/8 of the data is labeled. The results for different combinations of loss terms are reported in Table 3.

We start using only the entropy loss L_{ce} since this loss is always required. In this setting only the semantic branch is used and trained only on the labeled data. This setting achieves 64.1% mIoU. Adding the latent loss L_{latent} improves the performance by 0.5%. In this setting, the semantic and latent branch are used, but they are both only trained on the labeled data. Adding the consistency loss

L_{cons} boosts the accuracy by 2.7%. This shows that the latent branch provides additional supervision for the semantic branch on the unlabeled data.

So far, we did not use the adversarial loss L_{adv}. When we add the adversarial loss only for the labeled data $L_{adv}^{labeled}$ to the entropy loss L_{ce}, the performance grows by 4.6%. In this setting, only the labeled data is used for training. If we use the adversarial loss also for the unlabeled data, the accuracy increases by 0.7%. This shows that the adversarial loss improves semi-supervised learning, but the gain is not as high compared to additionally using the latent branch to supervise the semantic branch on the unlabeled data. In this setting, all loss terms are used and the accuracy increases further by 1.9%. Compared to the entropy loss L_{ce}, the proposed loss terms increase the accuracy by 7.2%.

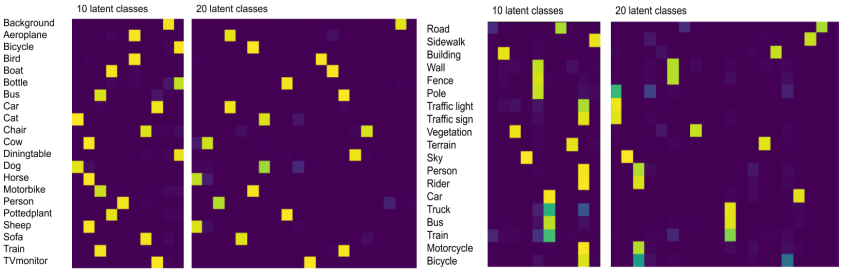
Impact of number of latent classes. For our approach, we need to specify the maximum number of latent classes. While we used by default 20 in our previous experiments, we now evaluate it for 2, 4, 6, 10, and 20 latent classes on Pascal VOC 2012 with 1/8 of the data being labeled. The results are reported in Table 4. The performance grows monotonically with the number of latent classes reaching its peak for 20.

In the same table, we also report the number of effective latent classes. We consider a latent class l to be effectively used at threshold t, if $P(l|c) > t$ for at least one semantic class c. We report this number for $t = 0.1$ and $t = 0.9$. The number of effective latent classes differs only slightly for these two thresholds. This shows that a latent class typically either constitutes a supercategory of at least one semantic class or it is not used at all. We observe that until 10, all latent classes are used. If we allow up to 20 latent classes, only 14 latent classes are effectively used. In practice, we recommend to set the number of maximum latent classes to the number of semantic classes. The approach will then select as many latent classes as needed. Although we assume that the number of latent classes is less or equal to the number of semantic classes, we also evaluated the approach for 50 latent classes. As expected, the accuracy drops but the approach remains stable. The number of effectively used latent classes also remains at 14. In practice, this setting should not be used since it violates the assumptions of the approach and can lead to unexpected behavior in some cases.

To see if a semantic class is typically mapped to a single latent class, we plot $P(l|c)$ for inference on Pascal VOC 2012 as well as on Cityscapes and show the results in Figure 3(a) and Figure 3(b), respectively. Indeed, the mapping from semantic classes to latent classes is very sparse. Typically, for each semantic class c, there is one dominant latent class l, i.e., $P(l|c) > 0.9$. If the number of latent classes increases to 20, some of the latent classes are not used. On Pascal VOC 2012, similar categories like cat and dog or cow, horse, and sheep are grouped. Some groupings are based on the common background like aeroplane and bird. The grouping bicycle, bottle, and dining table combines the most difficult classes of the dataset. However, we observed that there are small variations of the groupings for different runs when the number of latent classes is very small. On Cityscapes with 20 latent classes, the semantic classes pole, traffic light, and

Table 4. Impact of the number of latent classes. The evaluation is performed on Pascal VOC 2012 where 1/8 of the data is labeled. A latent class l is considered effective, if there exists a semantic class c so that $P(l|c) > t$. The third column shows this number for $t = 0.1$ and the fourth for $t = 0.9$.

		Effective latent classes	
Max. latent classes	mIoU (%)	$t = 0.1$	$t = 0.9$
2	69.7	2	2
4	70.2	4	4
6	70.3	6	6
10	70.7	10	10
20	71.3	16	14
50	70.8	18	14



(a) $P(l|c)$ on Pascal VOC 2012 for 10 and 20 latent classes

(b) $P(l|c)$ on Cityscapes for 10 and 20 latent classes.

Fig. 3. The distribution of latent classes for both datasets is pretty sparse, essentially the latent classes form supercategories of semantic classes that are similar in appearance. The grouping bicycle, bottle, and dining table for 10 latent classes seems to be unexpected, but due to the low number of latent classes the network is forced to group additional semantic classes. In this case, the network tends to group the most difficult classes of the dataset. In case of 20 latent classes, the merged classes are very intuitive, but not all latent classes are effectively used.

traffic sign; person, rider, motorcycle, and bicycle; wall and fence; truck, bus, and train are grouped together. These groupings are very intuitive.

Comparison of learned latent classes with manually defined latent classes. Since the latent classes typically learn supercategories of the semantic classes, the question arises if the same effect can be achieved with manually defined supercategories. In this experiment, the latent classes are replaced with 10 manually defined supercategories. More details regarding these supercategories are provided in the supplementary material. In this setting, the latent branch is trained to predict these supercategories on the labeled data using the cross-entropy loss. For unlabeled data, everything remains the same as for the proposed method. We report the results in Table 5. The performance using the

Table 5. Comparison of learned latent classes with manually defined latent classes. The evaluation is performed on Pascal VOC 2012 where 1/8 of the data is labeled. In case of learned latent classes, the second column reports the maximum number of latent classes. In case of manually defined latent classes, the exact number of classes is reported.

Method	Classes	mIoU (%)
Manual	10	69.0
Learned	10	70.7
Semantic classes	21	68.5
Semantic classes (KL)	21	69.1
Learned	20	71.3

supercategories is only 69.0%, which is significantly below the proposed method for 10 latent variables.

Another approach would be to learn all semantic classes instead of the latent classes in the latent branch. In this case, both branches learn the same semantic classes. This gives 68.5%, which is also worse than the learned latent classes. If both branches predict the same semantic classes, we can also train them symmetrically. Being more specific, on labeled data they are both trained with the cross-entropy loss as well as the adversarial loss. On unlabeled data, we apply the adversarial loss to both of them and use the symmetric KullbackLeibler divergence (KL) as a consistency loss. This approach performs better, giving 69.1%, but it is still inferior to our proposed method. Overall, this shows the necessity to learn the latent classes in a data-driven way.

5 Conclusion

In this work, we addressed the task of semi-supervised semantic segmentation, where a small fraction of the data set is labeled in a pixel-wise manner, while most images do not have any types of labeling. Our key contribution is a two-branch segmentation architecture, which uses latent classes learned in a data-driven way on labeled data to supervise the semantic segmentation branch on unlabeled data. We evaluated our approach on the Pascal VOC 2012 and the Cityscapes dataset where the proposed method achieves state-of-the-art results.

Acknowledgement

This work was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) GA 1927/5-1 and under Germanys Excellence Strategy EXC 2070 390732324.

References

1. Ahn, J., Kwak, S.: Learning pixel-level semantic affinity with image-level supervision for weakly supervised semantic segmentation. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 4981–4990 (2018) 3
2. Bearman, A., Russakovsky, O., Ferrari, V., Fei-Fei, L.: What's the point: Semantic segmentation with point supervision. In: European Conference on Computer Vision (ECCV). pp. 549–565 (2016) 3
3. Briq, R., Moeller, M., Gall, J.: Convolutional simplex projection network for weakly supervised semantic segmentation (2018) 3
4. Chaudhry, A., Dokania, P.K., Torr, P.H.: Discovering Class-Specific Pixels for Weakly-Supervised Semantic Segmentation. In: British Machine Vision Conference (BMVC) (2017) 3
5. Chen, L.C., Papandreou, G., Kokkinos, I., Murphy, K., Yuille, A.L.: 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 (2018) 7
6. Cordts, M., Omran, M., Ramos, S., Rehfeld, T., Enzweiler, M., Benenson, R., Franke, U., Roth, S., Schiele, B.: The cityscapes dataset for semantic urban scene understanding. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2016) 3, 8
7. Dai, D., Sakaridis, C., Hecker, S., Van Gool, L.: Curriculum model adaptation with synthetic and real data for semantic foggy scene understanding. International Journal of Computer Vision **128**, 1182–1204 (2020) 4
8. Everingham, M., Eslami, S.M., Van Gool, L., Williams, C.K.I., Winn, J., Zisserman, A.: The Pascal Visual Object Classes Challenge: A Retrospective. International Journal of Computer Vision (IJCV) **111**(1), 98–136 (2014) 3, 8
9. Fan, R., Hou, Q., Cheng, M.M., Yu, G., Martin, R.R., Hu, S.M.: Associating inter-image salient instances for weakly supervised semantic segmentation. In: European Conference on Computer Vision (ECCV). pp. 371–388 (2018) 3
10. Ge, W., Yang, S., Yu, Y.: Multi-evidence filtering and fusion for multi-label classification, object detection and semantic segmentation based on weakly supervised learning. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 1277–1286 (2018) 3
11. Hong, S., Yeo, D., Kwak, S., Lee, H., Han, B.: Weakly supervised semantic segmentation using web-crawled videos. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 2224–2232 (2017) 3
12. Hou, Q., Massiceti, D., Dokania, P.K., Wei, Y., Cheng, M.M., Torr, P.H.S.: Bottom-up top-down cues for weakly-supervised semantic segmentation. In: Energy Minimization Methods in Computer Vision and Pattern Recognition. pp. 263–277 (2018) 3
13. Huang, Z., Wang, X., Wang, J., Liu, W., Wang, J.: Weakly-supervised semantic segmentation network with deep seeded region growing. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 7014–7023 (2018) 3
14. Hung, W.C., Tsai, Y.H., Liou, Y.T., Lin, Y.Y., Yang, M.H.: Adversarial learning for semi-supervised semantic segmentation. In: Proceedings of the British Machine Vision Conference (BMVC) (2018) 1, 2, 3, 7, 8, 9
15. Jin, B., Segovia, M.V.O., Ssstrunk, S.: Webly supervised semantic segmentation. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 1705–1714 (2017) 3

16. Khoreva, A., Benenson, R., Hosang, J., Hein, M., Schiele, B.: Simple does it: Weakly supervised instance and semantic segmentation. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 1665–1674 (2017) 3
17. Kolesnikov, A., Lampert, C.H.: Seed, expand and constrain: Three principles for weakly-supervised image segmentation. In: European Conference on Computer Vision (ECCV). pp. 695–711 (2016) 3
18. Kurmi, V.K., Bajaj, V., Venkatesh, K.S., Namboodiri, V.P.: Curriculum based dropout discriminator for domain adaptation. In: British Machine Vision Conference (BMVC) (2019) 4
19. Lee, J., Kim, E., Lee, S., Lee, J., Yoon, S.: Ficklenet: Weakly and semi-supervised semantic image segmentation using stochastic inference. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2019) 3
20. Lee, J., Kim, E., Lee, S., Lee, J., Yoon, S.: Frame-to-frame aggregation of active regions in web videos for weakly supervised semantic segmentation. In: IEEE International Conference on Computer Vision (ICCV) (2019) 3
21. Li, H., He, X., Barnes, N., Mingwen, W.: Learning Hough transform with latent structures for joint object detection and pose estimation. In: International Conference on Multimedia Modeling. pp. 116–129 (2016) 4
22. Li, K., Wu, Z., Peng, K., Ernst, J., Fu, Y.: Guided attention inference network. IEEE Transactions on Pattern Analysis and Machine Intelligence (2019) 3
23. Li, Q., Arnab, A., Torr, P.H.: Weakly- and semi-supervised panoptic segmentation. In: European Conference on Computer Vision (ECCV). pp. 106–124 (2018) 3
24. Lian, Q., Lv, F., Duan, L., Gong, B.: Constructing self-motivated pyramid curriculums for cross-domain semantic segmentation: A non-adversarial approach. In: IEEE International Conference on Computer Vision (ICCV) (2019) 4
25. Lin, D., Dai, J., Jia, J., He, K., Sun, J.: Scribblesup: Scribble-supervised convolutional networks for semantic segmentation. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 3159–3167 (2016) 3
26. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft COCO: Common objects in context. In: European Conference on Computer Vision (ECCV), pp. 740–755 (2014) 7
27. Long, J., Shelhamer, E., Darrell, T.: Fully convolutional networks for semantic segmentation. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 3431–3440 (2015) 7
28. Mittal, S., Tatarchenko, M., Brox, T.: Semi-supervised semantic segmentation with high- and low-level consistency. IEEE Transactions on Pattern Analysis and Machine Intelligence (2019) 2, 3, 7, 8, 9
29. Nguyen, A., Kanoulas, D., Caldwell, D.G., Tsagarakis, N.: Object-based affordances detection with convolutional neural networks and dense conditional random fields. In: IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS) (2017) 8
30. Oh, S.J., Benenson, R., Khoreva, A., Akata, Z., Fritz, M., Schiele, B.: Exploiting saliency for object segmentation from image level labels. In: IEEE International Conference on Computer Vision and Pattern Recognition (CVPR). pp. 5038–5047 (2017) 3
31. Papandreou, G., Chen, L.C., Murphy, K.P., Yuille, A.L.: Weakly- and semi-supervised learning of a deep convolutional network for semantic image segmentation. In: International Conference on Computer Vision (ICCV). pp. 1742–1750 (2015) 3

32. Pathak, D., Krähenbühl, P., Darrell, T.: Constrained convolutional neural networks for weakly supervised segmentation. In: International Conference on Computer Vision (ICCV). pp. 1796–1804 (2015) 3
33. Pinheiro, P.H.O., Collobert, R.: From image-level to pixel-level labeling with convolutional networks. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 1713–1721 (2015) 3
34. Qi, X., Liu, Z., Shi, J., Zhao, H., Jia, J.: Augmented feedback in semantic segmentation under image level supervision. In: European Conference on Computer Vision (ECCV). pp. 90–105 (2016) 3
35. Razavi, N., Gall, J., Kohli, P., Van Gool, L.: Latent Hough transform for object detection. In: European Conference on Computer Vision (ECCV). pp. 312–325 (10 2012) 4
36. Richard, A., Kuehne, H., Gall, J.: Weakly supervised action learning with RNN based fine-to-coarse modeling. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 1273–1282 (2017) 4
37. Roy, A., Todorovic, S.: Combining bottom-up, top-down, and smoothness cues for weakly supervised image segmentation. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 7282–7291 (2017) 3
38. Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A.C., Fei-Fei, L.: ImageNet Large Scale Visual Recognition Challenge. International Journal of Computer Vision (IJCV) **115**(3), 211–252 (2015) 7
39. Sakaridis, C., Dai, D., Van Gool, L.: Guided curriculum model adaptation and uncertainty-aware evaluation for semantic nighttime image segmentation. In: IEEE International Conference on Computer Vision (ICCV) (2019) 4
40. Shimoda, W., Yanai, K.: Distinct class-specific saliency maps for weakly supervised semantic segmentation. In: European Conference on Computer Vision (ECCV). pp. 218–234 (2016) 3
41. Song, C., Huang, Y., Ouyang, W., Wang, L.: Box-driven class-wise region masking and filling rate guided loss for weakly supervised semantic segmentation. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2019) 3
42. Tang, M., Djelouah, A., Perazzi, F., Boykov, Y., Schroers, C.: Normalized cut loss for weakly-supervised cnn segmentation. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 1818–1827 (2018) 3
43. Tang, M., Perazzi, F., Djelouah, A., Ben Ayed, I., Schroers, C., Boykov, Y.: On regularized losses for weakly-supervised CNN segmentation. In: European Conference on Computer Vision (ECCV). pp. 524–540 (2018) 3
44. Wang, X., You, S., Li, X., Ma, H.: Weakly-supervised semantic segmentation by iteratively mining common object features. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 1354–1362 (2018) 3
45. Wei, Y., Feng, J., Liang, X., Cheng, M.M., Zhao, Y., Yan, S.: Object region mining with adversarial erasing: A simple classification to semantic segmentation approach. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 6488–6496 (2017) 3
46. Wei, Y., Liang, X., Chen, Y., Shen, X., Cheng, M.M., Feng, J., Zhao, Y., Yan, S.: Stc: A simple to complex framework for weakly-supervised semantic segmentation. IEEE Transactions on Pattern Analysis and Machine Intelligence **39**(11), 2314–2320 (2017) 3
47. Wei, Y., Xiao, H., Shi, H., Jie, Z., Feng, J., Huang, T.S.: Revisiting dilated convolution: A simple approach for weakly- and semi-supervised semantic segmentation.

In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 7268–7277 (2018) 3
48. Zhang, Y., David, P., Gong, B.: Curriculum domain adaptation for semantic segmentation of urban scenes. In: IEEE International Conference on Computer Vision (ICCV). pp. 2039–2049 (2017) 4
49. Zhou, B., Khosla, A., Lapedriza, A., Oliva, A., Torralba, A.: Learning deep features for discriminative localization. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 2921–2929 (2016) 3
50. Zhu, X., Anguelov, D., Ramanan, D.: Capturing long-tail distributions of object subcategories pp. 915–922 (2014) 4