MODELING UNKNOWN SEMANTIC LABELS AS UNCERTAINTY IN THE PREDICTION: EVIDENTIAL DEEP LEARNING FOR CLASS-INCREMENTAL SEMANTIC SEGMENTATION

Anonymous authors
Paper under double-blind review

ABSTRACT

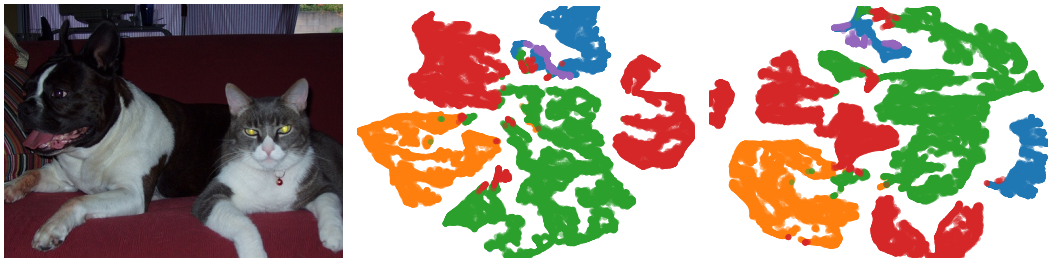
Class-Incremental Learning is an essential component for expanding the knowledge of previously trained neural networks. This is especially useful if the system needs to be able to handle new objects but the original training data is unavailable. While the semantic segmentation problem has received less attention than classification, it is faced with its own set of challenges in terms of unlabeled classes in the images. In this paper we address the problem of how to model unlabeled classes to avoid unnecessary feature clustering of uncorrelated classes. We propose to use Evidential Deep Learning to model the evidence of the classes with a Dirichlet distribution. Our method factorizes the problem into a separate foreground class probability, calculated by the expected value of the Dirichlet distribution, and an unknown class probability corresponding to the uncertainty of the estimate. In our novel formulation the background probability is implicitly modelled, avoiding the feature space clustering that comes from forcing the model to output a high background score for these pixels. Experiments on the incremental Pascal VOC and ADE20k show that our method is superior to state-of-the-art methods, especially when repeatedly learning new classes.

1 INTRODUCTION

Current state-of-the art machine learning methods for semantic segmentation achieve impressive results but are limited to the set of classes specified during training. To expand the capabilities of the model to new classes, retraining or fine-tuning on both the old and new classes is commonly necessary. Class-Incremental Learning (CIL) addresses this limitation by studying how to add new class knowledge, e.g. in the case where the labeled data of old classes is unavailable or when time-constraints prohibits training on the full dataset.

The main objective of incremental learning is to be able to incrementally add new class knowledge to the model while maintaining the ability to correctly classify all known classes $k \in \mathbb{K}$. Compared to incremental learning for classification, each input sample will contain multiple labels as well as unlabeled regions. The choice of which labels that may be present, but unlabeled, in the background give rise to a few different scenarios, two of which are the 'overlap' and the 'disjoint' setups. In the 'disjoint' setting, no future classes are allowed when learning the first few classes, but, previously learned classes may still be present. In contrast to this, the 'overlap' settings is likely more common in practical scenarios, putting no constraint on the classes that may appear unlabeled. This is especially true for scenarios were we cannot know before-hand that we would like to learn certain classes and, as such, cannot make sure that they are not present in the data.

Class-incremental learning for semantic segmentation differs to classification in how different labels interact and most classification approaches are not easily adapted to the semantic segmentation problem since they do not model the unlabeled regions. However, Cermelli et al. (2020) proposed to model the unlabeled background as a mixture of a general background class and the currently unlabeled classes. This lead to a clear improvement compared to ignoring the unlabeled classes.



(a) RGB image (b) MiB (Cermelli et al., 2020) (c) Our

Figure 1: Image showing the t-SNE visualization of the features for each class after training on a subset of the classes (e.g. cats and dogs). Both methods successfully learns the labeled classes (cats-orange, dogs-green). But our proposed method gives a larger freedom of the embedding features to the unknown regions, facilitating learning new classes (sofa-red, monitor-purple).

In this paper we expand on this idea by implicitly modeling the background through the Evidential Deep Learning framework (Sensoy et al., 2018), based on Subjective Logic (SL) and Dempster–Shafer theory (DST). We learn a belief mass, b_i, for each of the foreground classes which lets us estimate an uncertainty mass, u. Together, we let these masses correspond to the final class probability, P_i.

Our main contributions To summarize, our main contributions are:

- We propose a novel modeling of the background using Deep Evidential Learning
- In contrast to previous methods, we use the model uncertainty to identify unlabeled regions.
- We recognize that the current metrics can be biased towards increments containing a larger amount of classes and propose a new incremental metric that better illustrate the incremental nature of the task
- We perform exhaustive large-scale experiments and show that our method is superior to state-of-the-art, especially for larger number of increments and in the overlap scenario.

2 Related work

Class-Incremental Learning has been studied during a long time but has recently begun to receive a large amount of attention, leading to studies of the CIL scenario previously unstudied fields, e.g. semantic segmentation. Even though more attention has turned on the task of semantic segmentation (Cermelli et al., 2020; Michieli & Zanuttigh, 2021), a lot of the attention has previously been on the classification task. Most related work can be separated into a few categories, 'Weight regularization methods', 'Memory-based methods' and 'Distillation-based methods', either employed on its own or together with components from another category.

2.1 Weight regularization methods

The plasticity of a neural network affects both the ability to learn new things as well as forget old knowledge. The weight regularization methods attempts to control the plasticity of the model to minimize the loss of old knowledge while allowing the efficient learning of new knowledge. Elastic Weight Consolidation (EWC) (Kirkpatrick et al., 2017) estimates the importance of each weight of the model offline using the Fisher information matrix, while Memory Aware Synapses (MAS) (Aljundi et al., 2018) estimates it from the gradients with respect to the output given a small amount of noise in the parameter space. Both methods relatively computationally heavy. Similarly there are methods that focus on preserving the topology of the feature space, either directly (Tao et al., 2020a;b) or by compensating for the drift caused by learning new classes (Yu et al., 2020).

2.2 MEMORY-BASED METHODS

Complementary to the weight regularization methods, the memory-based methods address the issue where the data distribution may vary greatly between different different increments. To account for this, they keep a set of exemplars from previous increments and use during training the new classes. One of the first works utilizing the concept of exemplars was iCaRL by Rebuffi et al. (2017) which combined a limited memory bank of previous data samples with a distillation loss. Another method is ReMIX (Mi et al., 2020) which combines the idea of exemplar replay (ER) with Mixup (Zhang et al., 2017) to augment the dataset and learn more robust representations. While an explicit memory of previous samples has been shown to be beneficial for maintaining the previous knowledge it comes with an increased memory footprint and requires that an efficent sampling process to select the samples to use as exemplars.

2.3 DISTILLATION-BASED METHODS

Both of the previous categories of methods are commonly combined with a distillation loss, which is also commonly used for model compression (Hinton et al., 2015). The use of distillation for incremental learning was presented by Li & Hoiem (2017), who proposed to use the distillation loss to maintain the model predictions close to that of the model from a previous time-step. The role of distillation has been studied by Zhao et al. (2020) and different variations on distillation has been proposed which apply it at a feature-level instead of an output-level and also to combine two models trained on different increments (Zhang et al., 2020).

2.4 CLASS INCREMENTAL SEMANTIC SEGMENTATION

Semantic segmentation introduces a set of new difficulties in how each data sample contain multiple distinct classes and also may contain unlabeled examples of previous or future classes. Cermelli et al. (2020) showed the importance of modeling the background for semantic segmentation when they formulated the background as a mixture of different unlabeled classes. However, while this formulation facilitates maintaining old class knowledge, we found it to force unlabeled classes of future increments to be clustered closely together. Another recent method by Michieli & Zanuttigh (2021), takes inspiration from representation learning and show the importance of learning a good feature space. They do this by clustering the features of the same class together and forcing the clusters apart and they also proposed to facilitate the addition of new knowledge by encouraging the feature representations to be sparse. Our work is taking one step further in the direction of how to model the problem to facilitate learning and we show that SDR (Michieli & Zanuttigh, 2021) also benefit our method.

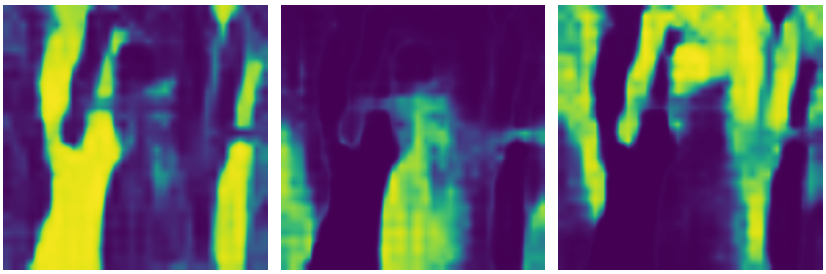
2.5 EVIDENTIAL DEEP LEARNING

The framework of Evidential Deep Learning as proposed by Sensoy et al. (2018) has been used for open-set action recognition (Wentao Bao & Kong, 2021) and for regression (Amini et al., 2019). The ideas is based on the subjective logic were the problem is formulated as K-disjoint frames or for our case classes, each associated with a belief-mass or evidence as well as a uncertainty-mass corresponding to the epistemic uncertainty. For the classification task, the Dirichlet distribution is commonly used allowing formulating the class-probabilities and the uncertainty in a systematic way.

Evidential Deep Learning has previously been used both for regression (Amini et al., 2019) and for classification tasks (Wentao Bao & Kong, 2021). Primarily, with the goal of a more efficient uncertainty estimation than full Bayesian networks or ensemble networks.

Evidential Deep Learning has primarily been used as a more efficient method to uncertainty estimation that full Bayesian networks or ensemble networks. This property has been leveraged for detecting out-of-sample samples that the network is uncertainty how to classify. Classical methods that only uses a softmax layer for classifying will always predict something even when uncertain which can be catastrophic if used for decision-making.

(a) RGB image (b) GT labels (c) Prediction



(d) $P_{\text{Bkg}} = u$ (e) P_{Bicycle} (f) P_{Person}

Figure 2: Illustration of the learned uncertainty and foreground probabilities, including the final label predictions.

3 METHOD

Contrary to the most common approaches which models both labeled and unlabeled classes as one of the possible categories of a categorical distribution, we do not model the unlabeled pixels as a unique class. Instead, we model the labeled classes independently of the unlabeled ones and use the epistemic uncertainty, u, of the model to differentiate labeled from unlabeled pixels.

The task is to predict a set of belief masses, b_i, for the labeled classes while maintaining a high uncertainty mass, u, were there are no labels. Evidential Deep Learning as proposed by Sensoy et al. (2018), relates these belief masses to the learned class evidences, $b_i = \frac{e_i}{\sum_j (e_j+1)}$ and the concentration parameters of a Dirichlet distribution, $\alpha_i = e_i + 1$. Furthermore, from subjective logics, they note that the sum of all of the belief masses and the uncertainty mass is one. Which leads to $u = \frac{K}{\sum_j (e_j+1)}$

3.1 EVIDENTIAL DEEP LEARNING FOR SEMANTIC SEGMENTATION

Next, we show how the network learns the class evidence necessary for estimating the class probabilities and the uncertainty.

The network is trained to output K-class scores for the known classes. These output scores are rectified to be strictly larger than zero by a scaled exponential term. Originally a ReLU was proposed for rectification but as noted by Wentao Bao & Kong (2021), using an exponential term facilitates the output of high confidence scores.

One issue with the standard formulation for calculating the expected value of the distribution based on the evidence, e, is that the required evidence needed scales with the number of classes, $|\mathbb{K}|$. In a single increment, this is not an issue but for the incremental setup were the number of classes may vary it can introduce an unwanted bias towards classes in the larger increments. To correct for this, we propose rescaling the output scores based on the fraction of total number of classes that we currently are learning.

3.2 LEARNING NEW CLASSES

Sensoy et al. (2018) discusses multiple choices of loss functions, one of which is the Type II Maximum Likelihood which corresponds to minimizing the negative log-likelihood of a multinomial distribution with the learned Dirichlet distribution as a prior. This equals the following loss function per-pixel,

$$\mathcal{L}_{ML} = \sum_k y_k (\log S - \log \alpha_k), \tag{1}$$

where $S = \sum_k \alpha_k$. However, this formulation does not take into account unlabeled samples. As such, we propose to expand the loss function with these background classes as well and give them a fix evidence of one which leads to the following term for unlabeled pixels.

$$\mathcal{L}_{ML} = \sum_{k \notin \mathbb{K}^t} y_k \log S, \tag{2}$$

To balance the foreground and the unlabeled losses and to direct the learning towards harder examples. We use the focal loss to weight the $\mathcal{L}_{ML}$-loss per pixel in addition to weighting the labeled and unlabaled terms based on the inverse proportion of labeled samples in the current batch.

For regularization purposes we use an uninformed prior on the predictive distribution in which all concentration parameters are equal to 1. This further makes sure that the evidence is kept minimal in the absence of foreground classes. Commonly the KL-divergence between the prior distribution and the predictive distribution is used with the correct label masked as 1. However, in this work we instead chose the Rényi-Dirichlet Divergence with $a = \frac{1}{2}$ as a divergence measure, which correspond to the following regularization loss,

$$\mathcal{L}^{0.5}_{RDh} = \log \frac{\prod_{k=1}^{K} \Gamma(\alpha_k)}{\Gamma(S)\Gamma(K)} - 2 \log \frac{\prod_k \Gamma(\tilde{\alpha}_k)}{\Gamma(\tilde{S})}. \tag{3}$$

Where $\tilde{\alpha}_k = \frac{\alpha_k + 1}{2}$. As is custom in related work, we add an annealing factor that slowly introduces the regularization term to the loss, $\lambda_t = \min[1, \frac{\text{current epoch}}{20}]$.

3.3 MAINTAINING PREVIOUS CLASS KNOWLEDGE

We use the standard output knowledge distillation loss in which we want to maintain the output predictive scores as close to the predictions from the old model. For this we use a cross-entropy term for the foreground class to maintain discriminativeness between them the foreground classes. Additionally we use a binary cross-entropy term for the uncertainty, making sure that we don't inflate or deflate our certainty of the old classes. The use of a KL-divergence based loss was also evaluated but because of the less robust behaviour of this term during learning, the simpler cross-entropy loss was used for all experiments.

$$\mathcal{L}_{KD} = -\sum_k \left(\hat{p}^{t-1} \log \hat{p}^t\right) - \left(u^{t-1} \log u^t + u^t \log u^{t-1}\right) \tag{4}$$

To summerize, the complete loss can be written as,

$$\mathcal{L}_{\text{Total}} = \lambda_{ML}\mathcal{L}_{ML+FL} + \lambda_t \lambda_{RDh}\mathcal{L}_{RDh} + \lambda_{KD}\mathcal{L}_{KD}. \tag{5}$$

The knowledge distillation loss is not calculated for the first step since no previous model exist at this time step.

3.4 WEIGHT INITIALIZATION

The weights of the classification layer is commonly initialized with Xavier initialization (Glorot & Bengio, 2010) in Pytorch. However, this choice of initalization corresponds to a highly biased starting prior for the background class since the weights are initialized with mean zero. This correspond to an initial prediction of $\approx 91\%$ for background. Instead we initialize all the biases to be equal and so that the uncertainty is equal to our prior background probability p_{bkg},

$$b_i = \log\left(10\frac{1-p_{\text{bkg}}}{p_{\text{bkg}}}\right) + \log\left(\frac{|\mathbb{K}^{\{1,\dots,T\}}|}{|\mathbb{K}^t|}\right) \tag{6}$$

The last term follows as a correction term from the evidence scaling.

4 EXPERIMENTS

If we define the total number of increments as T and a specific increment as t. We can denote the current set of classes to be learned as $\mathbb{K}^t$. Similarly, we define the sets of input samples $\mathbb{X}^t$ as all samples which contain any pixels of the classes in $\mathbb{K}^t$. This allows us to describe the different scenarios as follows.

- Overlapped: $\mathbb{X}^t_{\text{ov}} = \mathbb{X}^t$
- Disjoint: $\mathbb{X}^t_{\text{dis}} = \mathbb{X}^t \backslash \mathbb{X}^{\{t+1,\dots,T\}}$
- Joint: $\mathbb{K}_{\text{joint}} = \mathbb{K}$

Note that only the classes in the current class set, $\mathbb{K}^t$, is labeled and that the joint scenario is training all classes jointly in a single increment.

We use the evaluation protocol proposed by Cermelli et al. (2020). The first experiments was done for the 15-5 and the 15-1 scenarios for Pascal VOC (Everingham et al.). Both of these start with a set of 15 base-classes and adds 5 missing classes to these. The 15-5 setup, adds all 5 classes in a single step while the 15-1 setup adds the classes one at a time.

4.1 METRICS

The metrics used for evaluation follows the standards of semantic segmentation where the main metric is the class-averaged Intersection-over-Union, denoted mIoU. Cermelli et al. (2020) reported three different mIoU measures per task.

- Old: mIoU of the base classes trained in the first increment (Pascal: 1-15), omitting the background class.
- New: mIoU of the incrementally added classes (Pascal: 16-20), also omitting the background class. The loss weight used similar to previous work, $\lambda_{ML} = 1$, $\lambda_{KD} = 10$ and $\lambda_{RDh} = 0.1$. For the experiments were we also use add the Sparsity, Cluster separation loss and the clustering loss by Michieli & Zanuttigh (2021), we used their suggested parameter values, $\lambda_{Sparsity} = 10^{-4}$, $\lambda_{Separation} = 10^{-3}$ and $\lambda_{Cluster} = 10^{-2}$.
- All: mIoU of all classes including background (Pascal: 0-20)

However, we find that these metrics can be skewed if there is a large class imbalance between the increments. As such we propose a new metric, Increment Averaged mIoU (abbreviated Inc. mIoU).

This metric is defined as:

$$\text{Inc. mIoU} = \frac{1}{T}\sum_{t=1}^{T}\frac{1}{\mathbb{K}^t}\sum_{k\in\mathbb{K}^t} IoU_k = \frac{1}{T}\sum_{t=1}^{T}\text{mIoU}(\mathbb{K}^t) \tag{7}$$

4.2 EXPERIMENTAL SETUP

We use the same network architecture as done by Cermelli et al. (2020) and Michieli & Zanuttigh (2021) which is a the DeepLab v3 (Chen et al., 2017) with a ResNet101 backbone. In order to facilitate the comparison we are using the same learning rates and the same optimizer and learning rate scheduler as Cermelli et al. (2020), that is, a polynomial learning rate scheduler with decay rate 0.9, a learning rate of $\lambda_0 = 0.01$ and a learning rate of the following steps of $\lambda_{t>0} = 0.001$. The optimizer is SGD with a nesterov momentum of 0.9.

The pretrained weights of the backbone are the same as used by Cermelli et al. (2020), which however, differs from the ones used by Michieli & Zanuttigh (2021). The output stride during test is kept the same as during training, e.g. 16, which means that the final class-scores are upsampled from 32×32 to 512×512. We used a batchsize of 20 and 8 respectively and used 30 epochs for Pascal and 60 epochs for training each increment on ADE20k (Zhou et al., 2019). No early stopping was employed.

The loss weight used similar to previous work, $\lambda_{ML} = 1$, $\lambda_{KD} = 10$ and $\lambda_{RDh} = 0.1$. For the experiments were we also use add the Sparsity, Cluster separation loss and the clustering loss by Michieli & Zanuttigh (2021), we used their suggested parameter values, $\lambda_{Sparsity} = 10^{-4}$, $\lambda_{Separation} = 10^{-3}$ and $\lambda_{Cluster} = 10^{-2}$.

Our choice of background prior, $p_{bkg} = 0.7$ for the Pascal datasets and $p_{bkg} = 0.3$ for the ADE20k dataset. The reason of the large difference in initial bias is that the ADE20k dataset does not contain any pixels labeled as background since it contains both things (e.g. chairs, cars) and stuff classes (e.g. sky, grass). Compared to Pascal that contain classes of things but no stuff, leading to close to 70% background pixels.

5 RESULTS

We present our results, evaluated according to the experimental setup proposed by Cermelli et al. (2020). And also include our proposed incremental averaged mIoU (inc. mIoU) metric to highlight the incremental nature of the task.

5.1 PASCAL VOC RESULTS

The results for different scenarios on the Pascal VOC dataset are shown in table 1 and 2. As can be seen, our proposed method performs better than the state-of-the-art in all settings. Further, the difference becomes larger the more increments we learn and is also larger in the overlapped setting compared to in the disjoint setting.

This is likely caused by our proposed way of modeling the unlabeled probabilities which facilitates the learning of multiple increments by avoiding unnecessary clustering of unlabeled pixels as indicated by the feature space in figure 1. This also explains why we notice a larger improvement in the overlapped setting where we have examples of unlabeled classes that are learned in later increments compared to the disjoint setting were no future classes has been seen.

We train our models with three different random seeds and report their per-metric average and variance. The MiB-results are as reported by Cermelli et al. (2020) and the rest are based on the reported performance by Michieli & Zanuttigh (2021). We follow Cermelli et al. (2020) in that we exclude the background from the 'old' class set since the background is also present in the following increments. However, this differs compared to Michieli & Zanuttigh (2021) and as such, we add a parenthesis to note that the 'Old' class set is not completely comparable.

5.2 ADE20K RESULTS

The ADE20k dataset (Zhou et al., 2019) differs compared to Pascal in that there are no actual background class. To account for this we do not predict background during testing on ADE20k and instead only make our prediction based on the foreground probabilities.

Another improtant factor is that ADE20k is a highly class imbalanced dataset were the classes are ordered according to decreasing frequency. As such we follow Cermelli et al. (2020) in evaluating

	15-5							
	disjoint				overlapped			
	Old	New	All	Inc.	Old	New	All	Inc.
MiB	71.8	43.3	64.7	-	75.5	49.4	69.0	-
MiB (from SDR)	45.5	34.1	44.3	39.8	(73.1)	44.5	66.3	-
SDR	*72.5*	47.3	67.2	59.9	(*75.4*)	52.6	69.9	-
SDR+MiB	**73.6**	44.1	67.3	58.9	(**76.3**)	50.2	70.1	-
Our	71.3	*50.9*	*67.4*	*61.1*	74.8	**55.0**	**70.9**	**64.9**
	±1.8	±1.4	±1.5	±1.5	±1.7	±4.0	±1.6	±2.2
Our+SDR	72.4	**51.1**	**68.2**	**61.7**	74.8	*53.6*	*70.5*	*64.2*
	±0.01	±0.1	±1.5	±0.2	±0.4	±0.8	±0.4	±0.6
Our (joint)	77.1	69.2	75.9	73.1	77.1	69.2	75.9	73.1
	±0.2	±1.0	±0.3	±0.5	±0.2	±1.0	±0.3	±0.5

Table 1: Results on the Pascal VOC 15-5 dataset, best in **bold** and runner-up in *italic*.

	15-1							
	disjoint				overlapped			
	Old	New	All	Inc.	Old	New	All	Inc.
MiB	46.2	12.9	37.9	-	35.1	13.5	29.7	-
MiB (from SDR)	36.9	15.0	33.3	18.7	(44.5)	11.7	36.7	-
SDR	57.4	12.9	48.1	20.3	(44.7)	**21.8**	39.2	-
SDR+MiB	57.6	14.3	48.7	21.6	(47.3)	14.7	39.5	-
Our	*58.6*	*16.0*	*49.2*	*23.1*	*63.6*	18.4	*53.3*	*25.9*
	±0.6	±0.6	±0.4	±0.5	±1.1	±2.3	±0.7	±1.7
Our+SDR	**59.9**	**17.0**	**50.4**	**24.2**	**64.1**	*18.7*	**53.74**	**26.3**
	±0.02	±0.9	±0.01	±0.6	±1.0	±0.01	±1.0	±0.03

Table 2: Results on the Pascal VOC 15-5 dataset, best in **bold** and runner-up in *italic*.

two different class sets and presenting the mean value of these in the table. As can be seen in table 3, our method struggles when encountering highly class imbalanced data but still keeps a similar performance on old and new classes.

	100-50				50-50				100-10			
	Old	New	All	Inc.	Old	New	All	Inc.	Old	New	All	Inc.
MiB	37.9	27.9	34.6	32.9	35.5	22.9	27.0	27.1	31.8	14.1	25.9	17.0
MiB (from SDR)	37.6	24.7	33.3	31.2	39.1	22.6	28.1	20.9	21.0	5.3	15.8	13.2
SDR	37.4	24.8	33.2	-	40.9	23.8	19.5	-	28.9	7.4	21.7	-
SDR+MiB	37.5	25.5	33.5	31.5	42.9	25.4	31.3	34.2	28.9	11.7	23.2	20.3
Our+SDR	18.7	17.9	18.4	18.3	16.1	17.5	17.0	17.0	17.1	12.7	13.4	15.6
Our (joint)	40.8	13.2	31.6	27.0	40.8	13.2	31.6	-	40.8	13.2	31.6	17.8

Table 3: Results on the ADE20k dataset. The MIB results are based on the ones reported by Cermelli et al. (2020) while the MIB(from SDR), SDR and SDR+MIB results are based on the results reported by Michieli & Zanuttigh (2021). The Old set correspond to classes 1-100 and 1-50 respectively, the New set correspond to the incrementally added classes 51-150 and 101-150 respectively. The background class is only included in the All set.

6 Conclusion

In this paper we proposed to implicitly model the unlabeled pixels in class-incremental semantic segmentation. We have shown that this approach facilitates learning and outperforms current state-of-the-art methods, especially when no limitations are imposed on which classes that may exist in the background. This paper represents an important step forward in the modeling necessary for how to incrementally add new class knowledge to a semantic segmentation model.

REFERENCES

Rahaf Aljundi, Francesca Babiloni, Mohamed Elhoseiny, Marcus Rohrbach, and Tinne Tuytelaars. Memory aware synapses: Learning what (not) to forget. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 139–154, 2018.

Alexander Amini, Wilko Schwarting, Ava Soleimany, and Daniela Rus. Deep evidential regression. *arXiv preprint arXiv:1910.02600*, 2019.

Fabio Cermelli, Massimiliano Mancini, Samuel Rota Bulo, Elisa Ricci, and Barbara Caputo. Modeling the background for incremental learning in semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

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

M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The PASCAL Visual Object Classes Challenge 2012 (VOC2012) Results. http://www.pascal-network.org/challenges/VOC/voc2012/workshop/index.html.

Xavier Glorot and Yoshua Bengio. Understanding the difficulty of training deep feedforward neural networks. In *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, pp. 249–256. JMLR Workshop and Conference Proceedings, 2010.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.

James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, et al. Overcoming catastrophic forgetting in neural networks. *Proceedings of the national academy of sciences*, 114(13):3521–3526, 2017.

Zhizhong Li and Derek Hoiem. Learning without forgetting. *IEEE transactions on pattern analysis and machine intelligence*, 40(12):2935–2947, 2017.

Fei Mi, Lingjing Kong, Tao Lin, Kaicheng Yu, and Boi Faltings. Generalized class incremental learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2020.

Umberto Michieli and Pietro Zanuttigh. Continual semantic segmentation via repulsion-attraction of sparse and disentangled latent representations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1114–1124, June 2021.

Sylvestre-Alvise Rebuffi, Alexander Kolesnikov, Georg Sperl, and Christoph H Lampert. icarl: Incremental classifier and representation learning. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pp. 2001–2010, 2017.

Murat Sensoy, Lance Kaplan, and Melih Kandemir. Evidential deep learning to quantify classification uncertainty. *arXiv preprint arXiv:1806.01768*, 2018.

Xiaoyu Tao, Xinyuan Chang, Xiaopeng Hong, Xing Wei, and Yihong Gong. Topology-preserving class-incremental learning. In *European Conference on Computer Vision*, pp. 254–270. Springer, 2020a.

Xiaoyu Tao, Xiaopeng Hong, Xinyuan Chang, Songlin Dong, Xing Wei, and Yihong Gong. Few-shot class-incremental learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12183–12192, 2020b.

Qi Yu Wentao Bao and Yu Kong. Evidential deep learning for open set action recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021.

Lu Yu, Bartlomiej Twardowski, Xialei Liu, Luis Herranz, Kai Wang, Yongmei Cheng, Shangling Jui, and Joost van de Weijer. Semantic drift compensation for class-incremental learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

Hongyi Zhang, Moustapha Cisse, Yann N Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. *arXiv preprint arXiv:1710.09412*, 2017.

Junting Zhang, Jie Zhang, Shalini Ghosh, Dawei Li, Serafettin Tasci, Larry Heck, Heming Zhang, and C.-C. Jay Kuo. Class-incremental learning via deep model consolidation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, March 2020.

Bowen Zhao, Xi Xiao, Guojun Gan, Bin Zhang, and Shu-Tao Xia. Maintaining discrimination and fairness in class incremental learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

Bolei Zhou, Hang Zhao, Xavier Puig, Tete Xiao, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Semantic understanding of scenes through the ade20k dataset. *International Journal of Computer Vision*, 127(3):302–321, 2019.

A Appendix

B Additional results on Pascal-VOC dataset

In addition to the results presented in the paper we also present some additional scenarios here. The $10-5$ scenario was not reported by neither Cermelli et al. (2020) nor Michieli & Zanuttigh (2021), however, future work might be interested in this scenario as well and as such we report our results here.

	10-5									
	disjoint					overlapped				
	1-10	11-15	16-20	All	Inc.	1-10	11-15	16-20	All	Inc.
Our	*62.6*	*50.8*	*48.0*	*57.7*	*53.8*	*72.0*	67.1	*52.2*	*67.0*	*63.8*
	±0.9	±16.8	±0.3	±2.4	±2.43	±0.3	±0.8	±0.1	±0.2	±0.2
Our+SDR	**62.2**	**49.6**	**46.9**	**56.9**	*52.9*	**71.7**	*66.8*	**51.2**	**66.5**	*63.2*
	±1.6	±15.5	±0.5.	±2.2	±2.7	±0.3	±7.1	±3.2	±1.1	±1.5

Table 4: Results on the Pascal VOC dataset, the highest metric is marked by **bold** and the second highest by *italic*. We evaluated our method with three different random seeds and report mean and variance of each metric.