Incremental Learning in Semantic Segmentation from Image Labels

Fabio Cermelli*[1,2], Dario Fontanel*[1], Antonio Tavera*[1], Marco Ciccone[1], Barbara Caputo[1]
[1]Politecnico di Torino, [2]Italian Institute of Technology
{first.last}@polito.it

Abstract

Although existing semantic segmentation approaches achieve impressive results, they still struggle to update their models incrementally as new categories are uncovered. Furthermore, pixel-by-pixel annotations are expensive and time-consuming. This paper proposes a novel framework for Weakly Incremental Learning for Semantic Segmentation, that aims at learning to segment new classes from cheap and largely available image-level labels. As opposed to existing approaches, that need to generate pseudo-labels offline, we use a localizer, trained with image-level labels and regularized by the segmentation model, to obtain pseudo-supervision online and update the model incrementally. We cope with the inherent noise in the process by using soft-labels generated by the localizer. We demonstrate the effectiveness of our approach on the Pascal VOC and COCO datasets, outperforming offline weakly-supervised methods and obtaining results comparable with incremental learning methods with full supervision. [1]

1. Introduction

Semantic segmentation is a fundamental problem in computer vision where significant progress has been made thanks to the surge of deep learning [13–15] and the availability of large-scale human-annotated or synthetic datasets [4, 17, 23, 40, 55]. Despite the fact that many pre-trained models using public datasets are available online, one of their key disadvantages is that they are not meant to be incrementally updated over time and their knowledge is often limited to the predefined set of classes.

A naïve solution to this problem would be to extend existing datasets with new annotated samples and train new models from scratch. However, this approach is impractical in case of frequent updates because training on the entire augmented dataset would take too long, increasing the energy consumption and carbon footprint of machine learning models [51, 58, 61]. Moreover, retraining or fine-tuning becomes infeasible when the original data is no longer available, *e.g.*, due to privacy concerns or intellectual property.

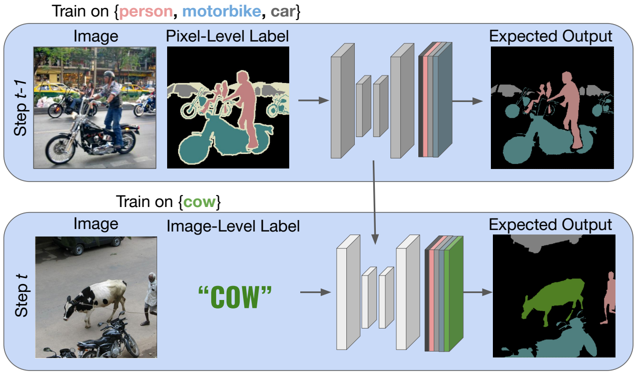

Figure 1. Illustration of WILSS. A model is first pre-trained on a set of classes (*e.g.*, *person, motorbike, car*), using expensive pixel-wise annotations. Then, in the following incremental learning steps, the model is updated to segment new classes (*e.g.*, *cow*) being provided image-level labels and without access to old data.

A better solution is to incrementally add new classes to the pre-existing model, as done in some recent works [8, 21, 43, 45, 46]. Incremental learning approaches update the model's parameters by training only on new data and employing ad-hoc techniques to avoid catastrophic forgetting on old classes [44]. While they reduce the cost of training, they rely on pixel-wise supervision on novel classes, which is expensive and time-consuming to collect, and usually requires expert human annotators [6, 40].

To reduce the annotation cost, different types of weak supervision have been proposed: bounding boxes [18, 32], scribbles [39, 62], points [16], and image-level labels [34, 50, 52]. Image labels can be easily retrieved from image classification benchmarks [19] or the web, dramatically lowering the annotation cost. Nevertheless, their use has never been investigated in an incremental learning setting.

In light of these considerations, we argue that it is crucial to jointly address the problems of incrementally updating the model and reducing the annotation cost of new data for semantic segmentation. To this end, we propose to incrementally train a segmentation model using only image-level labels for the new classes. We call this task *Weakly-*

*Equal contribution

[1] Code can be found at https://github.com/fcdl94/WILSON.

Supervised Incremental Learning for Semantic Segmentation (WILSS). This novel setting combines the advantages of incremental learning (training only on new class data) and weak supervision (cheap and largely available annotations). An illustration of WILSS is reported in Fig. 1.

Directly applying existing weakly-supervised methods to incremental segmentation would require to (i) extract pixel-wise pseudo-supervision offline using a weakly supervised approach [3, 5, 36, 59, 63] and (ii) update the segmentation network resorting to an incremental learning technique [8,21,43]. However, we argue that generating pseudo-labels offline in incremental settings is sub-optimal, as it involves two separate training stages and ignores the model's knowledge on previous classes that can be exploited to learn new classes more efficiently.

Hence, we propose a **W**eakly **I**ncremental **L**earning framework for semantic **S**egmentation that incrementally trains a segmentation model generating **ON**line pseudo-supervision from image-level annotations (WILSON) and exploits previous knowledge to learn new classes. We extend a standard encoder-decoder segmentation architecture [13–15] by introducing a *localizer* on the encoder, from which we extract pseudo-supervision for the segmentation backbone. To improve the pseudo-supervision, we train the localizer with a pixel-wise loss guided by the predictions of the segmentation model. This regularization serves two purposes: i) it acts as a strong prior for the previous class distribution, informing the model on where old classes are located in the image, and (ii) it provides a saliency prior for extracting better object boundaries. To address the noise present in the pseudo-supervision, instead of using hard pseudo-labels as in previous works [5, 36, 63], we obtain soft-labels from the localizer, which provides information on the probability assigned to a pixel to belong to a certain class.

To summarize, the contributions are as follows:

- We propose the Weakly supervised Incremental Learning for Semantic Segmentation (WILSS) task to extend pre-trained segmentation models with new classes using image-level supervision only.
- We propose WILSON, a novel framework that generates pseudo-supervision online using a simple localizer trained with an image-level classification loss and a pixel-wise localization loss that relies on old class knowledge. To model the noise in the pseudo-supervision, we use a convex combination of soft and hard labels that improves the segmentation performance over hard labels only.
- We evaluate our method on the Pascal VOC [23] and COCO [40] datasets, showing that our approach outperforms offline weakly-supervised methods, and that it is comparable or slightly inferior w.r.t. fully supervised incremental learning methods.

2. Related work

Incremental learning semantic segmentation. Incremental learning (IL) aims at addressing the phenomenon known as *catastrophic forgetting* [25, 44]: a model, expanding its knowledge with new classes over time, gradually forgets previously learned ones. Even if in image classification it has been exhaustively studied [1, 11, 20, 22, 24, 30, 38, 41, 54, 57, 68], in semantic segmentation it is still in its early stages [8, 9, 21, 33, 43, 45–47]. As first shown by [8], catastrophic forgetting in segmentation is exacerbated by the background shift problem; hence, they proposed a modified version of the traditional cross-entropy to propagate only the probability of old classes through the incremental steps and a distillation term to preserve previous knowledge. Later, [21] proposed to preserve long and short-range spatial relationships at feature level, while [46] regularized the latent space to improve class-conditional features separations. Alternatively, [43] used samples of old classes with replay methods to mitigate forgetting. Finally, [9] proposed the incremental few-shot segmentation setting, where only a few images to learn new classes are provided.

Differently from these works, we focus on a more challenging scenario where the supervision on new classes is provided as cheap image-level labels.

Weakly supervised semantic segmentation. Collecting accurate pixel-wise annotations for supervising semantic segmentation models is generally costly and time-consuming. To address this issue, Weakly Supervised Semantic Segmentation (WSSS) methods aim to obtain effective segmentation models using cheaper supervisions such as bounding boxes [18, 32, 49], scribbles [39, 60], points [6, 53], and image-level labels [31, 34, 35, 59]. Because of low prices and large availability on the web, image-level supervision gained the most attention over other types of weak supervision. Most image-based weakly supervised approaches [2, 3, 10, 31, 34, 35, 48, 59] use a two-stage procedure: (i) they generate pixel-wise pseudo-labels and then (ii) use them for training a segmentation backbone. The pseudo-labels are often extracted from an image-level classifier exploiting its Class Activation Maps (CAMs) [67]. An exception is [5], which proposes to learn a segmentation model in a single stage. Previous works focused on improving the pseudo-labels through multiple refinements steps [2, 3], additional losses [5, 10, 31, 34, 59, 63], or erasing techniques that force the CAM to expand and focus on non-discriminative parts of the image [12, 29, 64]. Finally, a recent trend uses external information, such as saliency, to improve the object boundaries [36, 66].

Despite the rapid development of pseudo-labels generation techniques from image-level supervision, these works operate in a static scenario where the model learns from a fixed set of classes. Instead, we focus on the more challeng-

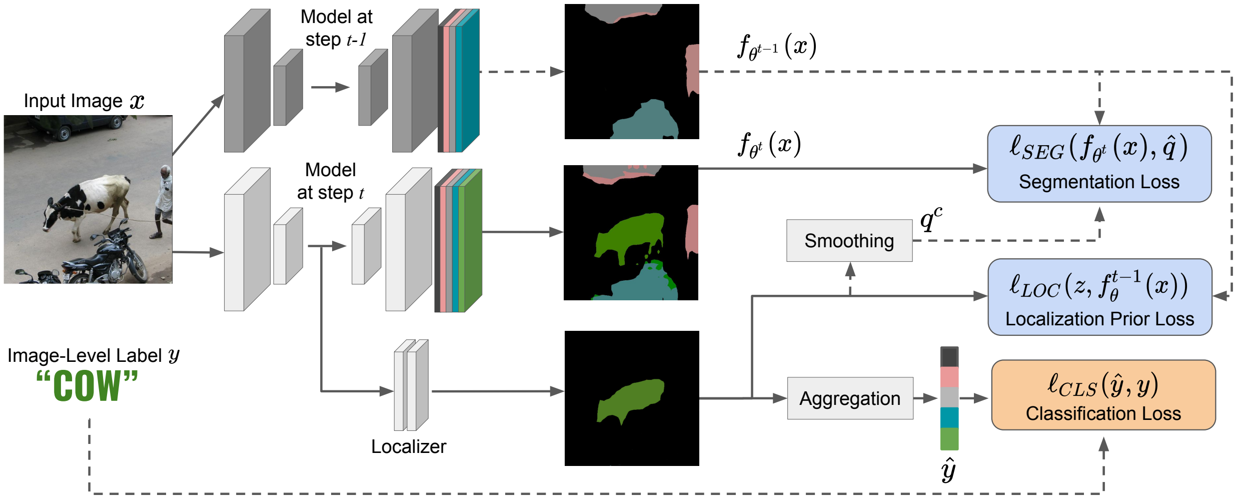

Figure 2. Illustration of the end-to-end training of WILSON. The localizer is directly trained using a classification loss ℓ_{CLS} and the Localization Prior loss ℓ_{LOC}, which exploits the prior information of the old model at step $t-1$. The segmentation model is supervised using CAM and old model output. The gradient is not backpropagated on dotted lines.

ing incremental learning setting where we learn new classes over time, extending a pre-trained segmentation model using only image-level labels.

3. WILSON Framework

Adapting current WSSS methods [3,34,36,37,63] for incremental learning requires generating pseudo-labels offline for the new classes and then training a segmentation model separately. Instead, we propose an end-to-end framework for WILSS that can learn incrementally from pseudo-labels generated online by a localizer attached to the model. In the following, we first define the problem and the notation (Sec. 3.1). Then we illustrate how the classification module can be trained to obtain pseudo-supervision (Sec. 3.2). Finally, in Sec. 3.3 we describe how to train the segmentation model to learn new classes without forgetting old ones. The framework is depicted in Fig. 2.

3.1. Problem Definition and Notation

We consider an input space $\mathcal{X}$ (*i.e.* the image space) and assume, without loss of generality, that each image is composed by a set of pixels $\mathcal{I}$ with constant cardinality $|\mathcal{I}| = H \times W = N$. The output space $\mathcal{Y}^N$ is defined as the product set of N-tuples with elements in a label space $\mathcal{Y}$. In the standard semantic segmentation setting, given an image $x \in \mathcal{X}$, we want to learn a mapping to assign each pixel x_i a label $y_i \in \mathcal{Y}$, representing its semantic class. The mapping is realized by a model $f_\theta = d_{\theta^d} \circ e_{\theta^e} : \mathcal{X} \mapsto \mathbb{R}^{N \times |\mathcal{Y}|}$ from the image space $\mathcal{X}$ to a pixel-wise class probability vector. e and d denote the encoder and decoder of the segmentation network, respectively.

The output segmentation mask is obtained as $y^* = \{\arg\max_{c \in \mathcal{Y}} p_i^c\}_{i=1}^N$, where p_i^c is the model prediction of pixel i for class c.

In the incremental segmentation setting [8], training is realized over multiple *learning steps*. At each learning step t, the previous label set $\mathcal{Y}^{t-1}$ is augmented with novel classes $\mathcal{C}^t$, yielding a new label set $\mathcal{Y}^t = \mathcal{Y}^{t-1} \cup \mathcal{C}^t$. Differently from the original incremental setting, in WILSS we are provided with dense annotations only for the initial step ($t = 0$). That is, the model is pre-trained on a densely-annotated dataset $\mathcal{T}^0 \subset \mathcal{X} \times (\mathcal{C}^0)^N$ only for the initial classes. Then, we learn new classes only from cheap image-level labels for all the following steps. Namely, for ($t > 0$), we have access to multiple training sets with only image-level annotations for novel classes $\mathcal{T}^t \subset \mathcal{X} \times (\mathcal{C}^t)$. As in [8], we assume that data from previous training steps is not accessible anymore, and we want to update the model to perform segmentation on new classes preserving its performance on old classes *i.e.* $f_{\theta^t} : \mathcal{X} \mapsto \mathbb{R}^{N \times |\mathcal{Y}^t|}$.

3.2. Training the Localizer

Inspired by the WSSS literature [3, 5, 34, 36, 37, 63], we introduce a *localizer* g, trained with image-level labels, to produce the pseudo-supervision for the segmentation model. The localizer uses the features from the segmentation encoder e to predict a score for all classes (background, old and new ones) *i.e.* $z = g(e(x)) \in \mathbb{R}^{|\mathcal{Y}^t| \times H \times W}$.

Learning from image-level labels. To learn from image-level labels first we need to aggregate the pixel-level classification scores z. The common solution is to use a Global Average Pooling (GAP) [3, 63].

However, simply averaging the scores produces coarse pseudo-labels [5], as all pixels in the feature map are encouraged to be less discriminative for the target class. For this reason, we use the *normalized Global Weighted Pooling* (nGWP) [5], that weights every pixel based on its relevance for the target class. In particular, the weight of each pixel is computed normalizing the classification scores with the `softmax` operation ψ, *i.e.* $m = \psi(z)$. The aggregated scores are computed as:

$$\hat{y}^{nGWP} = \frac{\sum_{i\in\mathcal{I}} m_i z_i}{\epsilon + \sum_{i\in\mathcal{I}} m_i}, \tag{1}$$

where ϵ is a small constant. Moreover, to encourage the scores to identify all the visible parts of the object, we use the *focal penalty* term introduced by [5], that is obtained as:

$$\hat{y}^{FOC} = (1 - \frac{\sum_{i\in\mathcal{I}} m_i}{|\mathcal{I}|})^{\gamma} log(\lambda + \frac{\sum_{i\in\mathcal{I}} m_i}{|\mathcal{I}|}), \tag{2}$$

where λ and γ are hyper-parameters. We refer the readers to [5] for more details on the nGWP and the focal penalty.

Since WILSS is an incremental learning scenario, we assume to have access only to image-level annotations y for the *new classes* $\mathcal{C}^t$. The localizer is then trained minimizing the *multi-label soft-margin loss*:

$$\ell_{CLS}(\hat{y}, y) = -\frac{1}{|\mathcal{K}|} \sum_{c\in\mathcal{K}} y^c log(\hat{y}^c) + \\ + (1 - y^c) log(1 - \hat{y}^c), \tag{3}$$

where $\mathcal{K} = \mathcal{C}^t$, $\hat{y} = \sigma(\hat{y}^{nGWP} + \hat{y}^{FOC})$, and σ is the logistic function. We note that, while the loss is computed only on new classes, it implicitly depends on the old classes scores due to the softmax-based aggregation in Eq. (1). However, since image-level annotations are cheap, and new images can be easily annotated, we may also consider a relaxed setting in which weak annotations are provide for both old and new classes. In this scenario the classification loss in Eq. (3) is computed on all classes and $\mathcal{K} = \mathcal{Y}^t$.

Localization Prior. The image-level labels provide supervision only on the presence of new classes in the image. However, they do not provide any cue on their boundaries or any information about the location of old classes. We argue that these insights can be freely obtained from the segmentation model learned in previous learning steps. In particular, the background score can be used as a saliency prior to extracting better object boundaries. Moreover, the scores of the old classes guide the localizer in detecting whether and where an old class is present in the image, directing its attention to alternative regions.

Hence, we introduce a direct supervision on the localizer coming from the segmentation model trained on step $t-1$, *i.e.* f_θ^{t-1}. The supervision acts as a *Localization Prior* (LOC) and can be provided as a pixel-wise loss between the segmentation model outputs $\omega = \sigma(f_\theta^{t-1}(x))$ and the classification scores z. Formally, we minimize the following objective function:

$$\ell_{LOC}(z, \omega) = -\frac{1}{|\mathcal{Y}^{t-1}||\mathcal{I}|} \sum_{i\in\mathcal{I}} \sum_{c\in\mathcal{Y}^{t-1}} \omega_i^c log(\sigma(z_i^c)) + \\ + (1 - \omega_i^c) log(1 - \sigma(z_i^c))), \tag{4}$$

where $\sigma(\cdot)$ is the logistic function.

In Eq. (4), the segmentation model provides a dense target on old classes. Unlike the `softmax` operator, which enforces competition among classes, the `logistic` function makes the class probabilities independent which is beneficial for a correct localization prior; in the case of a novel class, both old classes and the background will have a low score, implicitly informing the localizer that the pixel belongs to a new class.

3.3. Learning to Segment from Pseudo-Supervision

A solution often adopted by WSSS methods to train the semantic segmentation network is to extract hard-pseudo labels from an image-level classifier. In particular, these are obtained generating a one-hot distribution $q^{\text{Hard},c}$ for each pixel, attributing value one to the class with the maximum score for each pixel and zero to the others, *i.e.*

$$q_i^{\text{Hard},c} = \begin{cases} 1 & \text{if } c = \arg\max_{k\in\mathcal{Y}^t} m_i^c, \\ 0 & \text{otherwise,} \end{cases} \tag{5}$$

where m is the softmax normalized score extracted from the localizer.

However, it is well-known that pseudo-supervision generated from an image-level classifier is noisy [5, 36, 37, 63], and using $q^{H,c}$ to supervise the segmentation network may be detrimental for learning, causing the model to fit the wrong targets. For this reason, we propose to smooth the pseudo-labels to reduce the noise [42]. Formally, given a class c, the pseudo-supervision q^c is computed as:

$$q^c = \alpha q^{\text{Hard},c} + (1 - \alpha) m^c, \tag{6}$$

where α is a hyper-parameter that controls the smoothness.

Although the localizer produces scores for both new and old classes, the output distribution might be biased towards new classes due to the incremental training step. Thus, using q as a target for the segmentation model would lead to catastrophic forgetting [44]. Inspired by the knowledge distillation framework [28], we replace the pseudo-supervision extracted from the localizer on old classes with the output of the segmentation model trained in the previous learning step. The final pixel-level pseudo-supervision $\hat{q}$ is thus composed as follows:

$$\hat{q}^c = \begin{cases} \min(\sigma(f_{\theta^{t-1}}(x))^c, q^c) & \text{if } c = \mathtt{b}, \\ q^c & \text{if } c \in \mathcal{C}^t, \\ \sigma(f_{\theta^{t-1}}(x))^c & \text{otherwise,} \end{cases} \tag{7}$$

where $\mathtt{b}$ is the background class and $\sigma(\cdot)$ is the logistic function. We note that we utilize the minimum value of the two distributions for the background class, which contributes in modeling the background shift [8].

Since the pseudo-supervision $\hat{q}^c$ is not a probability distribution that sums to one, as required by the standard softmax-based cross-entropy loss, we propose to use a training loss based on the multi-label soft-margin loss:

$$\ell_{SEG}(p, \hat{q}) = -\frac{1}{|\mathcal{I}|} \sum_{i \in \mathcal{I}} \sum_{c \in \mathcal{Y}^t} \hat{q}_i^c log(p_i^c) + \\ + (1 - \hat{q}_i^c) log(1 - \sigma(p_i^c)), \quad (8)$$

where $\mathcal{Y}^t$ is the set of all seen classes and $p = f_{\theta^t}(x)$ is the segmentation model output.

In conclusion, we remark that the localizer is not employed during the testing phase, thus our method does not increase the time required for the inference.

4. Experiments

4.1. Datasets and Settings

We provide an extensive evaluation of WILSON on the two standard benchmarks Pascal VOC 2012 [23] and COCO [40]. Following the standard methodology [3, 34], we augment the Pascal VOC dataset with images from [26] for a total of 10582 images for training and 1449 for validation annotated on 20 object categories. COCO is a large-scale dataset providing 164K images and 80 object classes. We follow the training split and the annotation of [7] that solves the overlapping annotation problem present in [40].

Following prior works [8, 43], we adopt two incremental learning settings on the Pascal VOC dataset: the **15-5 VOC**, where 15 classes are learned in the first learning phase and 5 new classes added in a second step, and the **10-10 VOC**, where two steps of 10 classes are performed. Following [8, 43], we report results using two experimental protocols: (i) the *disjoint* scenario, in which each training step includes images containing only new or previously seen classes; (ii) the *overlap* scenario, in which each training step includes all the images containing at least one pixel from a novel class. In addition, we propose a novel incremental learning scenario, the **COCO-to-VOC**, composed of two training steps. First, we learn the 60 COCO classes not present in the Pascal VOC dataset, removing all the images containing at least one pixel of the latter. Then, in the second step, we learn 20 Pascal VOC classes. Following previous protocols [8, 43], we report the results on the dataset validation sets since the test set labels have not been publicly released. We adopt the standard mean Intersection over Union metric (mIoU) [23] to evaluate the performance of the segmentation model.

We recall that, differently from [8, 43], in the proposed WILSS setting the incremental steps provide only image-level labels for the new classes.

4.2. Baselines

Given that WILSS is a new setting, we compare WILSON with both incremental learning and weakly supervised semantic segmentation approaches. We report eight methods that represent the current state-of-the-art for incremental learning using pixel-wise supervision: LWF [38], LWF-MC [54], ILT [45], MiB [8], PLOP [21], CIL [33], SDR [46], and RECALL [43]. We note that RECALL [43], differently from other methods, uses additional images taken from the Web. For Pascal VOC, we use the results published in [21, 43], while we run the experiments on the COCO-to-VOC setting using the code provided by [8].

Furthermore, we report the performance of several state-of-the-art WSSS methods adapted to operate in the incremental learning scenario. In particular, we first train a classification model using the images available in the incremental learning steps. Then, we generate the hard pseudo-labels offline and train the segmentation model minimizing Eq. (8). We report the results with the pseudo-labels generated from: the class activation maps obtained from a standard image classifier (CAM), SEAM [63], SS [5], and EPS [36]. As for WILSON we followed the same experimental protocols provided by [8], training each method using only the images belonging to disjoint and overlap scenarios. For each method, we used the implementation released by the authors to produce the results. For CAM, we used the implementation of EPS to generate the pseudo-labels. It is important to remark that, while CAM, SS, and SEAM rely only on image-level labels, EPS also makes use of an off-the-shelf saliency detector trained on external data.

4.3. Implementation Details

We employ Deeplab V3 [13] architecture for all the experiments, with a ResNet-101 [27] backbone with output stride equal to 16 for Pascal VOC and a Wide-ResNet-38 [65] with output stride 8 for COCO, both pre-trained on ImageNet. As in [8], we use in-place activated batch normalization [56] to reduce the memory footprint required by the experiments. The localizer used to generate the CAMs is composed of 3 convolutional layers followed by batch normalization and Leaky ReLU, where the first two have kernel size 3×3 while the last 1×1, with channel numbers $\{256, 256, \text{ number of classes}\}$, and stride 1. The model is trained for 40 epochs using batch size 24 and SGD with an initial learning rate of 0.001 (0.01 for the Deeplab head and the localizer), momentum 0.9, and weight decay 10^{-4}. We train only the localizer for the first 5 epochs. Then, we train the whole network by adding the pseudo-supervision from the localizer and decay the learning rate using a poly-

Method	Sup	Disjoint			Overlap		
		1-15	**16-20**	**All**	**1-15**	**16-20**	**All**
Joint ⋆	Pixel	75.5	73.5	75.4	75.5	73.5	75.4
FT ⋆	Pixel	8.4	33.5	14.4	12.5	36.9	18.3
LWF ⋆ [38]	Pixel	39.7	33.3	38.2	67.0	41.8	61.0
LWF-MC ⋆ [54]	Pixel	41.5	25.4	37.6	59.8	22.6	51.0
ILT ⋆ [45]	Pixel	31.5	25.1	30.0	69.0	46.4	63.6
CIL ⋆ [33]	Pixel	42.6	35.0	40.8	14.9	37.3	20.2
MIB ⋆ [8]	Pixel	71.8	43.3	64.7	75.5	49.4	69.0
PLOP ⋄ [21]	Pixel	71.0	42.8	64.3	75.7	51.7	70.1
SDR ⋆ [46]	Pixel	73.5	47.3	67.2	75.4	52.6	69.9
RECALL ⋆ [43]	Pixel	69.2	52.9	66.3	67.7	54.3	65.6
CAM	Image	69.3	26.1	59.4	69.9	25.6	59.7
SEAM [63]	Image	71.0	33.1	62.7	68.3	31.8	60.4
SS [5]	Image	71.6	26.0	61.5	72.2	27.5	62.1
EPS [36]	Image	72.4	38.5	65.2	69.4	34.5	62.1
WILSON (ours)	Image	**73.6**	**43.8**	**67.3**	**74.2**	**41.7**	**67.2**

Table 1. Results on the 15-5 setting of Pascal VOC expressed in mIoU%. The best method using Image-level supervision is bold. The best method using Pixel supervision is underlined. ⋆: results from [43]. ⋄: results from [21].

Method	Sup	Disjoint			Overlap		
		1-10	**11-20**	**All**	**1-10**	**11-20**	**All**
Joint ⋆	Pixel	76.6	74.0	75.4	76.6	74.0	*75.4*
FT ⋆	Pixel	7.7	60.8	33.0	7.8	58.9	32.1
LWF ⋆ [38]	Pixel	63.1	61.1	62.2	70.7	63.4	67.2
LWF-MC ⋆ [54]	Pixel	52.4	42.5	47.7	53.9	43.0	48.7
ILT ⋆ [45]	Pixel	67.7	61.3	64.7	70.3	61.9	66.3
CIL ⋆ [33]	Pixel	37.4	60.6	48.8	38.4	60.0	48.7
MIB ⋆ [8]	Pixel	66.9	57.5	62.4	70.4	63.7	67.2
PLOP [21]	Pixel	63.7	60.2	63.4	69.6	62.2	67.1
SDR ⋆ [46]	Pixel	67.5	57.9	62.9	70.5	63.9	67.4
RECALL ⋆ [43]	Pixel	64.1	56.9	61.9	66.0	58.8	63.7
CAM	Image	**65.4**	41.3	54.5	**70.8**	44.2	58.5
SEAM [63]	Image	65.1	53.5	60.6	67.5	55.4	62.7
SS [5]	Image	60.7	25.7	45.0	69.6	32.8	52.5
EPS [36]	Image	64.2	54.1	60.6	69.0	57.0	64.3
WILSON (ours)	Image	64.5	**54.3**	**60.8**	70.4	**57.1**	**65.0**

Table 2. Results on the 10-10 setting of Pascal VOC expressed in mIoU%. The best method using Image-level supervision is bold. The best method using Pixel supervision is underlined. ⋆:results from [43].

nomial schedule with a power of 0.9. Following [5], we set $\lambda = 0.01$, $\gamma = 3$ of Eq. (2), and after the fifth epoch, we use the self-supervised segmentation loss on the localizer. Finally, we set $\alpha = 0.5$ in Eq. (6) for all the experiments.

4.4. Results

Single step addition of five classes (15-5). In this setting, after the initial learning stage, the following 5 classes of the VOC dataset are added: *plant, sheep, sofa, train, tv-monitor*. We report results in Tab. 1. Despite being trained only with image-level labels, WILSON achieves competitive results in all settings (disjoint and overlap) against approaches trained with pixel-wise supervision. Considering all the classes, in the disjoint scenario, we are able to outperform RECALL [43] by 1.0% and SDR [46] by 0.1%, demonstrating the resilience of WILSON to forgetting without the need for a replay buffer while maintaining enough plasticity for learning new classes. Moreover, in the disjoint scenario, we surpass PLOP [21] by 1.0% and MIB [8] by 0.5% on new classes.

Method	Sup	COCO			VOC
		1-60	**61-80**	**All**	**61-80**
FT	Pixel	1.9	41.7	12.7	75.0
LWF [38]	Pixel	36.7	49.0	40.3	73.6
ILT [45]	Pixel	37.0	43.9	39.3	68.7
MIB [8]	Pixel	34.9	47.8	38.7	73.2
PLOP [21]	Pixel	35.1	39.4	36.8	64.7
CAM	Image	30.7	20.3	28.1	39.1
SEAM [63]	Image	31.2	28.2	30.5	48.0
SS [5]	Image	35.1	36.9	35.5	52.4
EPS [36]	Image	34.9	38.4	35.8	55.3
WILSON (ours)	Image	**39.8**	**41.0**	**40.6**	**55.7**

Table 3. Results on the COCO-to-VOC setting expressed in mIoU%. The best method using Image-level supervision is bold. The best method using Pixel supervision is underlined.

Considering WSSS methods adapted to the WILSS scenario, the results are a demonstration of the strengths of WILSON: the ability to retain the knowledge of past classes and, most importantly, the capability of learning new semantic classes given only image-level annotations. Indeed, when considering new classes, we outperform EPS [36] by +5.3% mIoU in the disjoint scenario, although it uses saliency maps generated from an external off-the-shelf model. Moreover, SEAM [63] is outperformed by 11.7% and SS [5] by 17.8%. These achievements are even more pronounced in the overlap scenario, where WILSON not only preserves all the prior knowledge but also achieves a +7.2% boost when learning new classes w.r.t. EPS. In this situation, the overall improvement is +5.1% when compared to the best methods (SS, EPS).

Single step addition of ten classes (10-10). In this setting, we introduce 10 classes in the incremental step: *dining-table, dog, horse, motorbike, person, plant, sheep, sofa, train, tv-monitor*. Tab. 2 shows consistent results with the 15-5 setting. The differences between WILSON and IL (pixel-wise supervision) methods are quite small and the results are nearly comparable. In terms of accuracy, the gap using the most accurate incremental learning method, ILT, is 3.9% in the disjoint scenario and shrinks to 2.4% in the overlap one when compared to SDR. The efficacy of WILSON is confirmed when compared to the WSSS (image-level supervision) method as well. Indeed, while learning novel semantic classes, our online technique outperforms all offline competitors in the overlap protocols by more than +0.7% overall mIoU, while achieving a comparable result (+0.2%) in the disjoint scenario. In Fig. 3 we report qualitative results demonstrating the superiority of WILSON on both new and old classes.

COCO-to-VOC. This set of experiments can be considered the most challenging. Initially, the network is trained on 60 classes from the COCO dataset (which are not shared with VOC), while additional 20 classes from the VOC dataset are added in the second step. Tab. 3 shows evaluations on both COCO and VOC validation sets. Despite the fact that WILSON performance drops 8% when learning new classes compared to LwF, this experiment better showcases our ability to retain prior information while learning new classes under image-level supervision, surpassing ILT performance on old classes (+2.8%), which is the top competitor trained with pixel-wise supervision. When comparing against WSSS methods, WILSON is the best method, marking 4.8% improvements in terms of mIoU from the best WSSS method (EPS) on COCO. Similar results hold also for the VOC validation set. WILSON outperforms all the previous weakly supervised methods on both the old and new classes, both on COCO and VOC.

4.5. Ablation studies

Localization Prior. To validate the robustness of the pseudo-supervision generation, we perform an ablation study considering different choices for training the localizer. Results are reported in Tab. 4 on the VOC 10-10 disjoint and overlap scenarios. In particular, we compare different strategies for training the localizer: (i) we use a constant value for the old classes, as in [5], (ii) we use a fixed prior, directly concatenating the segmentation output of the old model to the class scores when computing m, (iii) we provide a localization supervision to the localizer with the softmax cross-entropy loss and (iv) with the loss in Eq. (4). Using a constant value and disregarding past knowledge from the old segmentation network results in lower performance when compared to the overall mIoU obtained if using a localization prior, particularly on new classes (-4.4% on disjoint and -5.1 on overlap). This demonstrates that teaching the localizer the location of previous classes might be an effective way to prevent forgetting and improve performance while learning new classes. Thereby, using aggressive priors, such as directly using the segmentation output of the old model, does not allow the network to learn effectively the new classes, thus resulting in a gap of -4.0% on disjoint and -4.3% on overlapped scenario w.r.t. ℓ_{LOC}. Moreover, using the softmax cross-entropy loss to match the segmentation output is detrimental for the performance, achieving poor results on both new and old classes (-6.3% on disjoint and -5.8% on overlapped with respect to ℓ_{LOC}). The reason for this result is that, due to the softmax normalization the cross-entropy loss does not consider each class independently, and forces the localizer to produce high scores for old classes even when they have low segmentation scores.

Smoothing effect on pseudo-supervision. We tune the hyper-parameter α of Eq. (4), which regulates the smoothness of the pseudo-labels supervising the segmentation model. In Fig. 4 we show the final mIoU in the VOC 10-10 disjoint and overlap scenarios, for five distinct α values ranging from 0 to 1. As expected, in the case of $\alpha = 1$, which corresponds to using hard labels, the model fits the noise in the supervision, leading to worst results, forgetting the prior knowledge, and being incapable of learning novel classes. We chose $\alpha = 0.5$ for our experiment since it is a reasonable trade-off in accuracy between learning and remembering. It is crucial to note that changing the values from 0 to 0.7 affects the results by less than 0.5% on average between the disjoint and overlap case, indicating the robustness of WILSON to different α values.

		Disjoint			Overlap		
Prior	**Loss**	**1-10**	**11-20**	**All**	**1-10**	**11-20**	**All**
-	-	64.8	49.9	58.8	69.4	52.0	62.0
Fixed	-	**66.1**	50.3	59.7	**71.4**	52.8	63.4
Learned	CE	61.1	46.0	54.5	67.6	49.5	59.2
Learned	ℓ_{LOC}	64.5	**54.3**	**60.8**	70.4	**57.1**	**65.0**

Table 4. Ablation study to validate the robustness of pseudo-supervision considering different types of localization priors for training the localizer.

	VOC 15-5					
	Disjoint			**Overlap**		
Method	**1-15**	**16-20**	**All**	**1-15**	**16-20**	**All**
CAM	70.5	34.7	62.6	71.6	36.0	63.7
SEAM [63]	71.9	26.9	61.7	70.8	28.1	61.0
SS [5]	71.8	26.3	61.7	72.1	27.6	62.1
EPS [36]	73.5	45.7	67.7	75.3	**47.6**	69.4
WILSON (ours)	**75.0**	**46.0**	**68.9**	**76.1**	45.6	**69.5**
	VOC 10-10					
	Disjoint			**Overlap**		
	1-10	**11-20**	**All**	**1-10**	**11-20**	**All**
CAM	63.1	42.2	53.9	66.6	45.0	56.8
SEAM [63]	66.0	50.4	59.7	70.9	54.6	64.0
SS [5]	60.8	26.0	45.2	69.6	33.0	52.6
EPS [36]	69.1	53.0	62.4	72.9	55.7	65.4
WILSON (ours)	**69.5**	**56.4**	**64.2**	**73.6**	**57.6**	**66.7**

Table 5. Performance evaluation of weakly supervised segmentation methods trained with direct supervision on both old and new classes in the incremental step.

Using supervision for all the classes. Since image-level supervision is cheap, we evaluate the performance of weakly-supervised methods when the supervision is provided for both old and new classes in the incremental steps. Tab. 5 reports the results on VOC. Comparing the results with Tab. 1 and Tab. 2, we note a performance improvement. In particular, all the methods improved, with WILSON achieving, on average, 2% on both old and new classes on the 15-5 and 10-10. This result demonstrates that introducing knowledge about old classes in the pseudo-supervision generation is crucial to both learning new classes and avoiding forgetting. Moreover, we show that also in this scenario

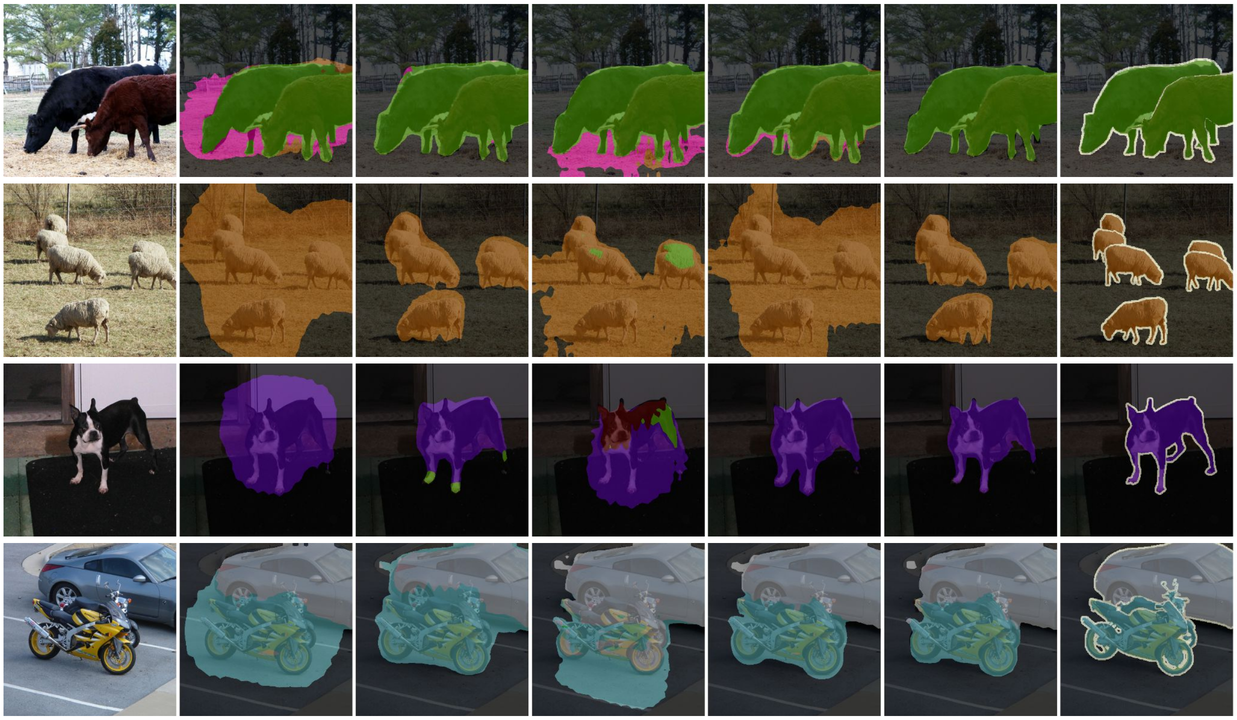

Figure 3. Qualitative results on the 10-10 VOC setting comparing different weakly supervised semantic segmentation methods. The image emphasized the efficiency of WILSON in both learning new classes (e.g. sheep, dog, motorbike) and preserving knowledge of old ones (e.g. cow, car). From left to right: image, CAM, SEAM [63], SS [5], EPS [36], WILSON and the ground-truth. Best viewed in color.

WILSON outperforms the offline WSSS methods. In particular, WILSON achieves better performance on every setting, outperforming EPS by 1.2% and 0.1% in the VOC 15-5 and by 1.8% and 1.3% in the VOC 10-10, respectively for the disjoint and overlapped scenario.

4.6. Limitations

Despite the remarkable results achieved by WILSON, it still has some drawbacks. To begin with, it is unable to perform single-class incremental learning steps, since Eq. 3 requires negative examples to properly guide the training. Moreover, we still need a considerable amount of images to train the model. Investigating learning from a few images could be an interesting future direction.

5. Conclusions

In this paper, we proposed WILSS, a novel setting that aims to extend the knowledge of semantic segmentation models through cheap image-level annotations. Applying current weakly-supervised learning approaches would require to generate the pseudo-supervision offline and then train the segmentation model. Differently, we propose WILSON, that couples the semantic segmentation model with a localizer and use image-level annotations on the new classes to generate online the pseudo-supervision for the segmentation backbone. We show that adding a localization prior from the old model to the localizer improves the generation of the pseudo-labels. We prove the effectiveness of our approach in three incremental settings. We outperform the WSSS baselines that generate pseudo-labels offline and we get results close to fully supervised incremental learning methods.

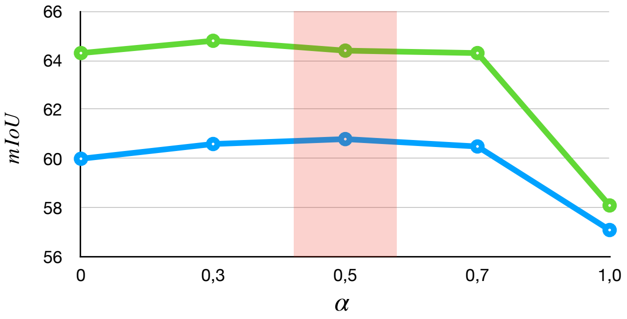

Figure 4. Ablation study about the effect of α to smooth the one-hot pseudo-labels used to supervise the ℓ_{SEG}. Test reporting the mIoU for both the Disjoint and Overlap VOC 10-10 protocols.

Acknowledgments. We thank Francesco Visin, Carlo Masone, and Marco Cannici for their valuable comments on the early draft of the manuscript. We gratefully acknowledge the HPC infrastructure and Team at IIT.

References

[1] Hongjoon Ahn, Jihwan Kwak, Subin Lim, Hyeonsu Bang, Hyojun Kim, and Taesup Moon. Ss-il: Separated softmax for incremental learning. In *ICCV*, pages 844–853, 2021. 2

[2] Jiwoon Ahn, Sunghyun Cho, and Suha Kwak. Weakly supervised learning of instance segmentation with inter-pixel relations. In *CVPR*, pages 2209–2218, 2019. 2

[3] Jiwoon Ahn and Suha Kwak. Learning pixel-level semantic affinity with image-level supervision for weakly supervised semantic segmentation. In *CVPR*, pages 4981–4990, 2018. 2, 3, 5

[4] Emanuele Alberti, Antonio Tavera, Carlo Masone, and Barbara Caputo. Idda: a large-scale multi-domain dataset for autonomous driving. *IEEE Robotics and Automation Letters*, 5(4):5526–5533, 2020. 1

[5] Nikita Araslanov and Stefan Roth. Single-stage semantic segmentation from image labels. In *CVPR*, pages 4253–4262, 2020. 2, 3, 4, 5, 6, 7, 8

[6] Amy Bearman, Olga Russakovsky, Vittorio Ferrari, and Li Fei-Fei. What's the point: Semantic segmentation with point supervision. In *ECCV*, pages 549–565. Springer, 2016. 1, 2

[7] Holger Caesar, Jasper Uijlings, and Vittorio Ferrari. Cocostuff: Thing and stuff classes in context. In *CVPR*, pages 1209–1218, 2018. 5

[8] Fabio Cermelli, Massimiliano Mancini, Samuel Rota Bulo, Elisa Ricci, and Barbara Caputo. Modeling the background for incremental learning in semantic segmentation. In *CVPR*, pages 9233–9242, 2020. 1, 2, 3, 5, 6

[9] Fabio Cermelli, Massimiliano Mancini, Yongqin Xian, Zeynep Akata, and Barbara Caputo. Prototype-based incremental few-shot segmentation. In *BMVC*, 2021. 2

[10] Yu-Ting Chang, Qiaosong Wang, Wei-Chih Hung, Robinson Piramuthu, Yi-Hsuan Tsai, and Ming-Hsuan Yang. Weakly-supervised semantic segmentation via sub-category exploration. In *CVPR*, pages 8991–9000, 2020. 2

[11] Arslan Chaudhry, Puneet K Dokania, Thalaiyasingam Ajanthan, and Philip HS Torr. Riemannian walk for incremental learning: Understanding forgetting and intransigence. In *ECCV*, pages 532–547, 2018. 2

[12] Arslan Chaudhry, Puneet K Dokania, and Philip HS Torr. Discovering class-specific pixels for weakly-supervised semantic segmentation. *arXiv preprint arXiv:1707.05821*, 2017. 2

[13] 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 TPAMI*, 40(4):834–848, 2017. 1, 2, 5

[14] Liang-Chieh Chen, George Papandreou, Florian Schroff, and Hartwig Adam. Rethinking atrous convolution for semantic image segmentation. 2017. 1, 2

[15] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *ECCV*, 2018. 1, 2

[16] Bowen Cheng, Omkar Parkhi, and Alexander Kirillov. Pointly-supervised instance segmentation. *arXiv preprint arXiv:2104.06404*, 2021. 1

[17] 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 *CVPR*, 2016. 1

[18] Jifeng Dai, Kaiming He, and Jian Sun. Boxsup: Exploiting bounding boxes to supervise convolutional networks for semantic segmentation. In *ICCV*, pages 1635–1643, 2015. 1, 2

[19] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, pages 248–255, 2009. 1

[20] Prithviraj Dhar, Rajat Vikram Singh, Kuan-Chuan Peng, Ziyan Wu, and Rama Chellappa. Learning without memorizing. In *CVPR*, pages 5138–5146, 2019. 2

[21] Arthur Douillard, Yifu Chen, Arnaud Dapogny, and Matthieu Cord. Plop: Learning without forgetting for continual semantic segmentation. In *CVPR*, pages 4040–4050, 2021. 1, 2, 5, 6

[22] Arthur Douillard, Matthieu Cord, Charles Ollion, Thomas Robert, and Eduardo Valle. Podnet: Pooled outputs distillation for small-tasks incremental learning. In *ECCV*, pages 86–102. Springer, 2020. 2

[23] Mark Everingham, Luc Van Gool, Christopher K. I. Williams, John M. Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *IJCV*, 88:303–338, 2009. 1, 2, 5

[24] Enrico Fini, Stéphane Lathuiliere, Enver Sangineto, Moin Nabi, and Elisa Ricci. Online continual learning under extreme memory constraints. In *ECCV*, pages 720–735. Springer, 2020. 2

[25] Robert M French. Catastrophic forgetting in connectionist networks. *Trends in cognitive sciences*, 3(4), 1999. 2

[26] Bharath Hariharan, Pablo Arbeláez, Lubomir Bourdev, Subhransu Maji, and Jitendra Malik. Semantic contours from inverse detectors. In *ICCV*, pages 991–998. IEEE, 2011. 5

[27] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016. 5

[28] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. 2015. 4

[29] Qibin Hou, Peng-Tao Jiang, Yunchao Wei, and Ming-Ming Cheng. Self-erasing network for integral object attention. *arXiv preprint arXiv:1810.09821*, 2018. 2

[30] Xinting Hu, Kaihua Tang, Chunyan Miao, Xian-Sheng Hua, and Hanwang Zhang. Distilling causal effect of data in class-incremental learning. In *CVPR*, pages 3957–3966, 2021. 2

[31] Zilong Huang, Xinggang Wang, Jiasi Wang, Wenyu Liu, and Jingdong Wang. Weakly-supervised semantic segmentation network with deep seeded region growing. In *CVPR*, pages 7014–7023, 2018. 2

[32] Anna Khoreva, Rodrigo Benenson, Jan Hosang, Matthias Hein, and Bernt Schiele. Simple does it: Weakly supervised instance and semantic segmentation. In *CVPR*, pages 876–885, 2017. 1, 2

[33] Marvin Klingner, Andreas Bär, Philipp Donn, and Tim Fingscheidt. Class-incremental learning for semantic segmentation re-using neither old data nor old labels. In *2020 IEEE 23rd International Conference on Intelligent Transportation Systems (ITSC)*, pages 1–8. IEEE, 2020. 2, 5, 6

[34] Alexander Kolesnikov and Christoph H Lampert. Seed, expand and constrain: Three principles for weakly-supervised image segmentation. In *ECCV*, pages 695–711. Springer, 2016. 1, 2, 3, 5

[35] Jungbeom Lee, Eunji Kim, Sungmin Lee, Jangho Lee, and Sungroh Yoon. Ficklenet: Weakly and semi-supervised semantic image segmentation using stochastic inference. In *CVPR*, pages 5267–5276, 2019. 2

[36] Seungho Lee, Minhyun Lee, Jongwuk Lee, and Hyunjung Shim. Railroad is not a train: Saliency as pseudo-pixel supervision for weakly supervised semantic segmentation. In *CVPR*, pages 5495–5505, 2021. 2, 3, 4, 5, 6, 7, 8

[37] Yi Li, Zhanghui Kuang, Liyang Liu, Yimin Chen, and Wayne Zhang. Pseudo-mask matters in weakly-supervised semantic segmentation. In *ICCV*, pages 6964–6973, 2021. 3, 4

[38] Zhizhong Li and Derek Hoiem. Learning without forgetting. *IEEE TPAMI*, 40(12):2935–2947, 2017. 2, 5, 6

[39] Di Lin, Jifeng Dai, Jiaya Jia, Kaiming He, and Jian Sun. Scribblesup: Scribble-supervised convolutional networks for semantic segmentation. In *CVPR*, pages 3159–3167, 2016. 1, 2

[40] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *ECCV*, pages 740–755. Springer, 2014. 1, 2, 5

[41] Yaoyao Liu, Yuting Su, An-An Liu, Bernt Schiele, and Qianru Sun. Mnemonics training: Multi-class incremental learning without forgetting. In *CVPR*, pages 12245–12254, 2020. 2

[42] Michal Lukasik, Srinadh Bhojanapalli, Aditya Menon, and Sanjiv Kumar. Does label smoothing mitigate label noise? In *ICML*, pages 6448–6458. PMLR, 2020. 4

[43] Andrea Maracani, Umberto Michieli, Marco Toldo, and Pietro Zanuttigh. Recall: Replay-based continual learning in semantic segmentation. In *ICCV*, pages 7026–7035, 2021. 1, 2, 5, 6

[44] Michael McCloskey and Neal J Cohen. Catastrophic interference in connectionist networks: The sequential learning problem. In *Psychology of learning and motivation*, volume 24, pages 109–165. Elsevier, 1989. 1, 2, 4

[45] Umberto Michieli and Pietro Zanuttigh. Incremental learning techniques for semantic segmentation. In *ICCV-WS*, pages 0–0, 2019. 1, 2, 5, 6

[46] Umberto Michieli and Pietro Zanuttigh. Continual semantic segmentation via repulsion-attraction of sparse and disentangled latent representations. In *CVPR*, pages 1114–1124, 2021. 1, 2, 5, 6

[47] Umberto Michieli and Pietro Zanuttigh. Knowledge distillation for incremental learning in semantic segmentation. *Computer Vision and Image Understanding*, 205:103167, 2021. 2

[48] Seong Joon Oh, Rodrigo Benenson, Anna Khoreva, Zeynep Akata, Mario Fritz, and Bernt Schiele. Exploiting saliency for object segmentation from image level labels. In *CVPR*, pages 5038–5047. IEEE, 2017. 2

[49] George Papandreou, Liang-Chieh Chen, Kevin P Murphy, and Alan L Yuille. Weakly-and semi-supervised learning of a deep convolutional network for semantic image segmentation. In *ICCV*, pages 1742–1750, 2015. 2

[50] Deepak Pathak, Philipp Krahenbuhl, and Trevor Darrell. Constrained convolutional neural networks for weakly supervised segmentation. In *ICCV*, pages 1796–1804, 2015. 1

[51] David Patterson, Joseph Gonzalez, Quoc Le, Chen Liang, Lluis-Miquel Munguia, Daniel Rothchild, David So, Maud Texier, and Jeff Dean. Carbon emissions and large neural network training. *arXiv preprint arXiv:2104.10350*, 2021. 1

[52] Pedro O Pinheiro and Ronan Collobert. From image-level to pixel-level labeling with convolutional networks. In *CVPR*, pages 1713–1721, 2015. 1

[53] Rui Qian, Yunchao Wei, Honghui Shi, Jiachen Li, Jiaying Liu, and Thomas Huang. Weakly supervised scene parsing with point-based distance metric learning. In *AAAI*, volume 33, pages 8843–8850, 2019. 2

[54] Sylvestre-Alvise Rebuffi, Alexander Kolesnikov, Georg Sperl, and Christoph H Lampert. icarl: Incremental classifier and representation learning. In *CVPR*, pages 2001–2010, 2017. 2, 5, 6

[55] Stephan R. Richter, Vibhav Vineet, Stefan Roth, and Vladlen Koltun. Playing for data: Ground truth from computer games. In Bastian Leibe, Jiri Matas, Nicu Sebe, and Max Welling, editors, *European Conference on Computer Vision (ECCV)*, volume 9906 of *LNCS*, pages 102–118. Springer International Publishing, 2016. 1

[56] Samuel Rota Bulò, Lorenzo Porzi, and Peter Kontschieder. In-place activated batchnorm for memory-optimized training of dnns. In *CVPR*, 2018. 5

[57] Hanul Shin, Jung Kwon Lee, Jaehong Kim, and Jiwon Kim. Continual learning with deep generative replay. *arXiv preprint arXiv:1705.08690*, 2017. 2

[58] Emma Strubell, Ananya Ganesh, and Andrew McCallum. Energy and policy considerations for deep learning in nlp. *arXiv preprint arXiv:1906.02243*, 2019. 1

[59] Guolei Sun, Wenguan Wang, Jifeng Dai, and Luc Van Gool. Mining cross-image semantics for weakly supervised semantic segmentation. In *ECCV*. Springer, 2020. 2

[60] Meng Tang, Abdelaziz Djelouah, Federico Perazzi, Yuri Boykov, and Christopher Schroers. Normalized cut loss for weakly-supervised cnn segmentation. In *CVPR*, pages 1818–1827, 2018. 2

[61] Neil C Thompson, Kristjan Greenewald, Keeheon Lee, and Gabriel F Manso. The computational limits of deep learning. *arXiv preprint arXiv:2007.05558*, 2020. 1

[62] Paul Vernaza and Manmohan Chandraker. Learning random-walk label propagation for weakly-supervised semantic segmentation. In *CVPR*, pages 7158–7166, 2017. 1

[63] Yude Wang, Jie Zhang, Meina Kan, Shiguang Shan, and Xilin Chen. Self-supervised equivariant attention mech-

anism for weakly supervised semantic segmentation. In *CVPR*, pages 12275–12284, 2020. 2, 3, 4, 5, 6, 7, 8

[64] Yunchao Wei, Jiashi Feng, Xiaodan Liang, Ming-Ming Cheng, Yao Zhao, and Shuicheng Yan. Object region mining with adversarial erasing: A simple classification to semantic segmentation approach. In *CVPR*, pages 1568–1576, 2017. 2

[65] Zifeng Wu, Chunhua Shen, and Anton Van Den Hengel. Wider or deeper: Revisiting the resnet model for visual recognition. *PR*, 90:119–133, 2019. 5

[66] Yazhou Yao, Tao Chen, Guo-Sen Xie, Chuanyi Zhang, Fumin Shen, Qi Wu, Zhenmin Tang, and Jian Zhang. Non-salient region object mining for weakly supervised semantic segmentation. In *CVPR*, pages 2623–2632, 2021. 2

[67] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization. In *CVPR*, pages 2921–2929, 2016. 2

[68] Fei Zhu, Xu-Yao Zhang, Chuang Wang, Fei Yin, and Cheng-Lin Liu. Prototype augmentation and self-supervision for incremental learning. In *CVPR*, pages 5871–5880, June 2021. 2